IN THE UNITED STATES DISTRICT COURT
FOR THE WESTERN DISTRICT OF WISCONSIN

WILLIAM WHITFORD, ROGER ANCLAM, EMILY BUNTING, MARY LYNNE DONOHUE, HELEN HARRIS, WAYNE JENSEN, WENDY SUE JOHNSON, JANET MITCHELL, ALLISON SEATON, JAMES SEATON, JEROME WALLACE, and DONALD WINTER, Plaintiffs, v. GERALD C. NICHOL, THOMAS BARLAND, JOHN FRANKE, HAROLD V. FROEHLICH, KEVIN J. KENNEDY, ELSA LAMELAS, and TIMOTHY VOCKE, Defendants.	)	No. 15-cv-421-bbc

---

DECLARATION OF ANNABELLE ELIZABETH HARLESS

I, Annabelle Elizabeth Harless, pursuant to 28 U.S.C. § 1746, hereby declare as follows:

1. I am one the attorneys representing Plaintiffs in the above captioned action. I make this declaration based upon my personal knowledge and in support of the Plaintiffs' Motion in Limine to Exclude the Testimony of Sean P. Trende.

2. Attached as Exhibit A is a true and correct copy of the article D.M. Smith & W.N. Venables, *An Introduction to R* (2015).

3. Attached as Exhibit B is a true and correct copy of a computation file relied on by defense expert Sean P. Trende to produce his declaration, entitled "Wisconsin_clustering_computation.R"

4. Attached as Exhibit C is a true and correct copy of the article Andrew Gelman & Gary King, *A Unified Method of Evaluating Electoral Systems and Redistricting Plans,* 38 Am. J. Pol. Sci. 514 (1994).

5. Attached as Exhibit D is a true and correct copy of the article John N. Friedman & Richard T. Holden, *Optimal Gerrymandering: Sometimes Pack, but Never Crack*, 98 Am. Econ. Rev. 113 (2008).

6. Attached as Exhibit E is a true and correct copy of the article Luc Anselin, *Local Indicators of Spatial Association – LISA*, 27 Geographical Analysis 93 (1995).

7. Attached as Exhibit F is a true and correct copy of the article Wendy K. Tam Cho, *Contagion Effects and Ethnic Contribution Networks*, 47 Am. J. Pol. Sci. 368 (2003).

8. Attached as Exhibit G is a true and correct copy of the article Sean F. Reardon & David O'Sullivan, *Measures of Spatial Segregation*, 34 Soc. Methodology 121 (2004).

9. Attached as Exhibit H is a true and correct copy of the article Nancy A. Denton & Douglas S. Massey, *Hypersegregation in U.S. Metropolitan Areas: Black and Hispanic Segregation Along Five Dimensions*, 26 Demography 373 (1989).

I declare under penalty of perjury that the foregoing is true and correct.

Dated this 26th day of January, 2016.

/s/ Annabelle Harless

ANNABELLE E. HARLESS

An Introduction to R

Notes on R: A Programming Environment for Data Analysis and Graphics
Version 3.2.3 (2015-12-10)

W. N. Venables, D. M. Smith
and the R Core Team

This manual is for R, version 3.2.3 (2015-12-10).

Table of Contents

Preface

This introduction to R is derived from an original set of notes describing the S and S-PLUS environments written in 1990–2 by Bill Venables and David M. Smith when at the University of Adelaide. We have made a number of small changes to reflect differences between the R and S programs, and expanded some of the material.

We would like to extend warm thanks to Bill Venables (and David Smith) for granting permission to distribute this modified version of the notes in this way, and for being a supporter of R from way back.

Comments and corrections are always welcome. Please address email correspondence to `R-core@R-project.org`.

Suggestions to the reader

Most R novices will start with the introductory session in Appendix A. This should give some familiarity with the style of R sessions and more importantly some instant feedback on what actually happens.

Many users will come to R mainly for its graphical facilities. See Chapter 12 [Graphics], page 63, which can be read at almost any time and need not wait until all the preceding sections have been digested.

1 Introduction and preliminaries

1.1 The R environment

R is an integrated suite of software facilities for data manipulation, calculation and graphical display. Among other things it has

- an effective data handling and storage facility,
- a suite of operators for calculations on arrays, in particular matrices,
- a large, coherent, integrated collection of intermediate tools for data analysis,
- graphical facilities for data analysis and display either directly at the computer or on hard-copy, and
- a well developed, simple and effective programming language (called 'S') which includes conditionals, loops, user defined recursive functions and input and output facilities. (Indeed most of the system supplied functions are themselves written in the S language.)

The term "environment" is intended to characterize it as a fully planned and coherent system, rather than an incremental accretion of very specific and inflexible tools, as is frequently the case with other data analysis software.

R is very much a vehicle for newly developing methods of interactive data analysis. It has developed rapidly, and has been extended by a large collection of *packages*. However, most programs written in R are essentially ephemeral, written for a single piece of data analysis.

1.2 Related software and documentation

R can be regarded as an implementation of the S language which was developed at Bell Laboratories by Rick Becker, John Chambers and Allan Wilks, and also forms the basis of the S-PLUS systems.

The evolution of the S language is characterized by four books by John Chambers and coauthors. For R, the basic reference is *The New S Language: A Programming Environment for Data Analysis and Graphics* by Richard A. Becker, John M. Chambers and Allan R. Wilks. The new features of the 1991 release of S are covered in *Statistical Models in S* edited by John M. Chambers and Trevor J. Hastie. The formal methods and classes of the **methods** package are based on those described in *Programming with Data* by John M. Chambers. See Appendix F [References], page 99, for precise references.

There are now a number of books which describe how to use R for data analysis and statistics, and documentation for S/S-PLUS can typically be used with R, keeping the differences between the S implementations in mind. See Section "What documentation exists for R?" in *The R statistical system FAQ*.

1.3 R and statistics

Our introduction to the R environment did not mention *statistics*, yet many people use R as a statistics system. We prefer to think of it of an environment within which many classical and modern statistical techniques have been implemented. A few of these are built into the base R environment, but many are supplied as *packages*. There are about 25 packages supplied with R (called "standard" and "recommended" packages) and many more are available through the CRAN family of Internet sites (via `https://CRAN.R-project.org`) and elsewhere. More details on packages are given later (see Chapter 13 [Packages], page 77).

Most classical statistics and much of the latest methodology is available for use with R, but users may need to be prepared to do a little work to find it.

There is an important difference in philosophy between S (and hence R) and the other main statistical systems. In S a statistical analysis is normally done as a series of steps, with intermediate results being stored in objects. Thus whereas SAS and SPSS will give copious output from a regression or discriminant analysis, R will give minimal output and store the results in a fit object for subsequent interrogation by further R functions.

1.4 R and the window system

The most convenient way to use R is at a graphics workstation running a windowing system. This guide is aimed at users who have this facility. In particular we will occasionally refer to the use of R on an X window system although the vast bulk of what is said applies generally to any implementation of the R environment.

Most users will find it necessary to interact directly with the operating system on their computer from time to time. In this guide, we mainly discuss interaction with the operating system on UNIX machines. If you are running R under Windows or OS X you will need to make some small adjustments.

Setting up a workstation to take full advantage of the customizable features of R is a straightforward if somewhat tedious procedure, and will not be considered further here. Users in difficulty should seek local expert help.

1.5 Using R interactively

When you use the R program it issues a prompt when it expects input commands. The default prompt is '`>`', which on UNIX might be the same as the shell prompt, and so it may appear that nothing is happening. However, as we shall see, it is easy to change to a different R prompt if you wish. We will assume that the UNIX shell prompt is '`$`'.

In using R under UNIX the suggested procedure for the first occasion is as follows:

1. Create a separate sub-directory, say `work`, to hold data files on which you will use R for this problem. This will be the working directory whenever you use R for this particular problem.

   ```
   $ mkdir work
   $ cd work
   ```

2. Start the R program with the command

   ```
   $ R
   ```

3. At this point R commands may be issued (see later).
4. To quit the R program the command is

   ```
   > q()
   ```

   At this point you will be asked whether you want to save the data from your R session. On some systems this will bring up a dialog box, and on others you will receive a text prompt to which you can respond *`yes`*, *`no`* or *`cancel`* (a single letter abbreviation will do) to save the data before quitting, quit without saving, or return to the R session. Data which is saved will be available in future R sessions.

Further R sessions are simple.

1. Make `work` the working directory and start the program as before:

   ```
   $ cd work
   $ R
   ```

2. Use the R program, terminating with the `q()` command at the end of the session.

To use R under Windows the procedure to follow is basically the same. Create a folder as the working directory, and set that in the `Start In` field in your R shortcut. Then launch R by double clicking on the icon.

1.6 An introductory session

Readers wishing to get a feel for R at a computer before proceeding are strongly advised to work through the introductory session given in Appendix A [A sample session], page 82.

1.7 Getting help with functions and features

R has an inbuilt help facility similar to the `man` facility of UNIX. To get more information on any specific named function, for example `solve`, the command is

```
> help(solve)
```

An alternative is

```
> ?solve
```

For a feature specified by special characters, the argument must be enclosed in double or single quotes, making it a "character string": This is also necessary for a few words with syntactic meaning including `if`, `for` and `function`.

```
> help("[[")
```

Either form of quote mark may be used to escape the other, as in the string `"It's important"`. Our convention is to use double quote marks for preference.

On most R installations help is available in HTML format by running

```
> help.start()
```

which will launch a Web browser that allows the help pages to be browsed with hyperlinks. On UNIX, subsequent help requests are sent to the HTML-based help system. The 'Search Engine and Keywords' link in the page loaded by `help.start()` is particularly useful as it is contains a high-level concept list which searches though available functions. It can be a great way to get your bearings quickly and to understand the breadth of what R has to offer.

The `help.search` command (alternatively `??`) allows searching for help in various ways. For example,

```
> ??solve
```

Try `?help.search` for details and more examples.

The examples on a help topic can normally be run by

```
> example(topic)
```

Windows versions of R have other optional help systems: use

```
> ?help
```

for further details.

1.8 R commands, case sensitivity, etc.

Technically R is an *expression language* with a very simple syntax. It is *case sensitive* as are most UNIX based packages, so `A` and `a` are different symbols and would refer to different variables. The set of symbols which can be used in R names depends on the operating system and country within which R is being run (technically on the *locale* in use). Normally all alphanumeric symbols are allowed[1] (and in some countries this includes accented letters) plus '`.`' and '`_`', with the restriction that a name must start with '`.`' or a letter, and if it starts with '`.`' the second character must not be a digit. Names are effectively unlimited in length.

Elementary commands consist of either *expressions* or *assignments*. If an expression is given as a command, it is evaluated, printed (unless specifically made invisible), and the value is lost. An assignment also evaluates an expression and passes the value to a variable but the result is not automatically printed.

[1] For portable R code (including that to be used in R packages) only A–Za–z0–9 should be used.

Commands are separated either by a semi-colon (‘`;`’), or by a newline. Elementary commands can be grouped together into one compound expression by braces (‘`{`’ and ‘`}`’). *Comments* can be put almost[2] anywhere, starting with a hashmark (‘`#`’), everything to the end of the line is a comment.

If a command is not complete at the end of a line, R will give a different prompt, by default

```
+
```

on second and subsequent lines and continue to read input until the command is syntactically complete. This prompt may be changed by the user. We will generally omit the continuation prompt and indicate continuation by simple indenting.

Command lines entered at the console are limited[3] to about 4095 bytes (not characters).

1.9 Recall and correction of previous commands

Under many versions of UNIX and on Windows, R provides a mechanism for recalling and re-executing previous commands. The vertical arrow keys on the keyboard can be used to scroll forward and backward through a *command history*. Once a command is located in this way, the cursor can be moved within the command using the horizontal arrow keys, and characters can be removed with the `DEL` key or added with the other keys. More details are provided later: see Appendix C [The command-line editor], page 92.

The recall and editing capabilities under UNIX are highly customizable. You can find out how to do this by reading the manual entry for the **readline** library.

Alternatively, the Emacs text editor provides more general support mechanisms (via ESS, *Emacs Speaks Statistics*) for working interactively with R. See Section “R and Emacs” in *The R statistical system FAQ*.

1.10 Executing commands from or diverting output to a file

If commands[4] are stored in an external file, say `commands.R` in the working directory `work`, they may be executed at any time in an R session with the command

```
> source("commands.R")
```

For Windows **Source** is also available on the **File** menu. The function `sink`,

```
> sink("record.lis")
```

will divert all subsequent output from the console to an external file, `record.lis`. The command

```
> sink()
```

restores it to the console once again.

1.11 Data permanency and removing objects

The entities that R creates and manipulates are known as *objects*. These may be variables, arrays of numbers, character strings, functions, or more general structures built from such components.

During an R session, objects are created and stored by name (we discuss this process in the next session). The R command

```
> objects()
```

(alternatively, `ls()`) can be used to display the names of (most of) the objects which are currently stored within R. The collection of objects currently stored is called the *workspace*.

To remove objects the function `rm` is available:

---

[2] **not** inside strings, nor within the argument list of a function definition

[3] some of the consoles will not allow you to enter more, and amongst those which do some will silently discard the excess and some will use it as the start of the next line.

[4] of unlimited length.

```
> rm(x, y, z, ink, junk, temp, foo, bar)
```

All objects created during an R session can be stored permanently in a file for use in future R sessions. At the end of each R session you are given the opportunity to save all the currently available objects. If you indicate that you want to do this, the objects are written to a file called `.RData`[5] in the current directory, and the command lines used in the session are saved to a file called `.Rhistory`.

When R is started at later time from the same directory it reloads the workspace from this file. At the same time the associated commands history is reloaded.

It is recommended that you should use separate working directories for analyses conducted with R. It is quite common for objects with names `x` and `y` to be created during an analysis. Names like this are often meaningful in the context of a single analysis, but it can be quite hard to decide what they might be when the several analyses have been conducted in the same directory.

[5] The leading "dot" in this file name makes it *invisible* in normal file listings in UNIX, and in default GUI file listings on OS X and Windows.

2 Simple manipulations; numbers and vectors

2.1 Vectors and assignment

R operates on named *data structures*. The simplest such structure is the numeric *vector*, which is a single entity consisting of an ordered collection of numbers. To set up a vector named `x`, say, consisting of five numbers, namely 10.4, 5.6, 3.1, 6.4 and 21.7, use the R command

```
> x <- c(10.4, 5.6, 3.1, 6.4, 21.7)
```

This is an *assignment* statement using the *function* `c()` which in this context can take an arbitrary number of vector *arguments* and whose value is a vector got by concatenating its arguments end to end.[1]

A number occurring by itself in an expression is taken as a vector of length one.

Notice that the assignment operator (‘`<-`’), which consists of the two characters ‘`<`’ (“less than”) and ‘`-`’ (“minus”) occurring strictly side-by-side and it ‘points’ to the object receiving the value of the expression. In most contexts the ‘`=`’ operator can be used as an alternative.

Assignment can also be made using the function `assign()`. An equivalent way of making the same assignment as above is with:

```
> assign("x", c(10.4, 5.6, 3.1, 6.4, 21.7))
```

The usual operator, `<-`, can be thought of as a syntactic short-cut to this.

Assignments can also be made in the other direction, using the obvious change in the assignment operator. So the same assignment could be made using

```
> c(10.4, 5.6, 3.1, 6.4, 21.7) -> x
```

If an expression is used as a complete command, the value is printed *and lost*[2]. So now if we were to use the command

```
> 1/x
```

the reciprocals of the five values would be printed at the terminal (and the value of `x`, of course, unchanged).

The further assignment

```
> y <- c(x, 0, x)
```

would create a vector `y` with 11 entries consisting of two copies of `x` with a zero in the middle place.

2.2 Vector arithmetic

Vectors can be used in arithmetic expressions, in which case the operations are performed element by element. Vectors occurring in the same expression need not all be of the same length. If they are not, the value of the expression is a vector with the same length as the longest vector which occurs in the expression. Shorter vectors in the expression are *recycled* as often as need be (perhaps fractionally) until they match the length of the longest vector. In particular a constant is simply repeated. So with the above assignments the command

```
> v <- 2*x + y + 1
```

generates a new vector `v` of length 11 constructed by adding together, element by element, `2*x` repeated 2.2 times, `y` repeated just once, and `1` repeated 11 times.

---

[1] With other than vector types of argument, such as `list` mode arguments, the action of `c()` is rather different. See Section 6.2.1 [Concatenating lists], page 27.

[2] Actually, it is still available as `.Last.value` before any other statements are executed.

The elementary arithmetic operators are the usual `+`, `-`, `*`, `/` and `^` for raising to a power. In addition all of the common arithmetic functions are available. `log`, `exp`, `sin`, `cos`, `tan`, `sqrt`, and so on, all have their usual meaning. `max` and `min` select the largest and smallest elements of a vector respectively. `range` is a function whose value is a vector of length two, namely `c(min(x), max(x))`. `length(x)` is the number of elements in `x`, `sum(x)` gives the total of the elements in `x`, and `prod(x)` their product.

Two statistical functions are `mean(x)` which calculates the sample mean, which is the same as `sum(x)/length(x)`, and `var(x)` which gives

```
sum((x-mean(x))^2)/(length(x)-1)
```

or sample variance. If the argument to `var()` is an n-by-p matrix the value is a p-by-p sample covariance matrix got by regarding the rows as independent p-variate sample vectors.

`sort(x)` returns a vector of the same size as `x` with the elements arranged in increasing order; however there are other more flexible sorting facilities available (see `order()` or `sort.list()` which produce a permutation to do the sorting).

Note that `max` and `min` select the largest and smallest values in their arguments, even if they are given several vectors. The *parallel* maximum and minimum functions `pmax` and `pmin` return a vector (of length equal to their longest argument) that contains in each element the largest (smallest) element in that position in any of the input vectors.

For most purposes the user will not be concerned if the "numbers" in a numeric vector are integers, reals or even complex. Internally calculations are done as double precision real numbers, or double precision complex numbers if the input data are complex.

To work with complex numbers, supply an explicit complex part. Thus

```
sqrt(-17)
```

will give `NaN` and a warning, but

```
sqrt(-17+0i)
```

will do the computations as complex numbers.

2.3 Generating regular sequences

R has a number of facilities for generating commonly used sequences of numbers. For example `1:30` is the vector `c(1, 2, ..., 29, 30)`. The colon operator has high priority within an expression, so, for example `2*1:15` is the vector `c(2, 4, ..., 28, 30)`. Put `n <- 10` and compare the sequences `1:n-1` and `1:(n-1)`.

The construction `30:1` may be used to generate a sequence backwards.

The function `seq()` is a more general facility for generating sequences. It has five arguments, only some of which may be specified in any one call. The first two arguments, if given, specify the beginning and end of the sequence, and if these are the only two arguments given the result is the same as the colon operator. That is `seq(2,10)` is the same vector as `2:10`.

Arguments to `seq()`, and to many other R functions, can also be given in named form, in which case the order in which they appear is irrelevant. The first two arguments may be named `from=`*value* and `to=`*value*; thus `seq(1,30)`, `seq(from=1, to=30)` and `seq(to=30, from=1)` are all the same as `1:30`. The next two arguments to `seq()` may be named `by=`*value* and `length=`*value*, which specify a step size and a length for the sequence respectively. If neither of these is given, the default `by=1` is assumed.

For example

```
> seq(-5, 5, by=.2) -> s3
```

generates in `s3` the vector `c(-5.0, -4.8, -4.6, ..., 4.6, 4.8, 5.0)`. Similarly

```
> s4 <- seq(length=51, from=-5, by=.2)
```

generates the same vector in `s4`.

The fifth argument may be named `along=`*vector*, which is normally used as the only argument to create the sequence `1, 2, ..., length(`*vector*`)`, or the empty sequence if the vector is empty (as it can be).

A related function is `rep()` which can be used for replicating an object in various complicated ways. The simplest form is

```
> s5 <- rep(x, times=5)
```

which will put five copies of `x` end-to-end in `s5`. Another useful version is

```
> s6 <- rep(x, each=5)
```

which repeats each element of `x` five times before moving on to the next.

2.4 Logical vectors

As well as numerical vectors, R allows manipulation of logical quantities. The elements of a logical vector can have the values `TRUE`, `FALSE`, and `NA` (for "not available", see below). The first two are often abbreviated as `T` and `F`, respectively. Note however that `T` and `F` are just variables which are set to `TRUE` and `FALSE` by default, but are not reserved words and hence can be overwritten by the user. Hence, you should always use `TRUE` and `FALSE`.

Logical vectors are generated by *conditions*. For example

```
> temp <- x > 13
```

sets `temp` as a vector of the same length as `x` with values `FALSE` corresponding to elements of `x` where the condition is *not* met and `TRUE` where it is.

The logical operators are `<`, `<=`, `>`, `>=`, `==` for exact equality and `!=` for inequality. In addition if `c1` and `c2` are logical expressions, then `c1 & c2` is their intersection ("*and*"), `c1 | c2` is their union ("*or*"), and `!c1` is the negation of `c1`.

Logical vectors may be used in ordinary arithmetic, in which case they are *coerced* into numeric vectors, `FALSE` becoming `0` and `TRUE` becoming `1`. However there are situations where logical vectors and their coerced numeric counterparts are not equivalent, for example see the next subsection.

2.5 Missing values

In some cases the components of a vector may not be completely known. When an element or value is "not available" or a "missing value" in the statistical sense, a place within a vector may be reserved for it by assigning it the special value `NA`. In general any operation on an `NA` becomes an `NA`. The motivation for this rule is simply that if the specification of an operation is incomplete, the result cannot be known and hence is not available.

The function `is.na(x)` gives a logical vector of the same size as `x` with value `TRUE` if and only if the corresponding element in `x` is `NA`.

```
> z <- c(1:3,NA);  ind <- is.na(z)
```

Notice that the logical expression `x == NA` is quite different from `is.na(x)` since `NA` is not really a value but a marker for a quantity that is not available. Thus `x == NA` is a vector of the same length as `x` *all* of whose values are `NA` as the logical expression itself is incomplete and hence undecidable.

Note that there is a second kind of "missing" values which are produced by numerical computation, the so-called *Not a Number*, `NaN`, values. Examples are

```
> 0/0
```

or

```
> Inf - Inf
```

which both give NaN since the result cannot be defined sensibly.

In summary, `is.na(xx)` is `TRUE` *both* for `NA` and `NaN` values. To differentiate these, `is.nan(xx)` is only `TRUE` for `NaN`s.

Missing values are sometimes printed as `<NA>` when character vectors are printed without quotes.

2.6 Character vectors

Character quantities and character vectors are used frequently in R, for example as plot labels. Where needed they are denoted by a sequence of characters delimited by the double quote character, e.g., `"x-values"`, `"New iteration results"`.

Character strings are entered using either matching double (") or single (') quotes, but are printed using double quotes (or sometimes without quotes). They use C-style escape sequences, using \ as the escape character, so \\ is entered and printed as \\, and inside double quotes " is entered as \". Other useful escape sequences are \n, newline, \t, tab and \b, backspace—see `?Quotes` for a full list.

Character vectors may be concatenated into a vector by the `c()` function; examples of their use will emerge frequently.

The `paste()` function takes an arbitrary number of arguments and concatenates them one by one into character strings. Any numbers given among the arguments are coerced into character strings in the evident way, that is, in the same way they would be if they were printed. The arguments are by default separated in the result by a single blank character, but this can be changed by the named argument, `sep=`*`string`*, which changes it to *`string`*, possibly empty.

For example

```
> labs <- paste(c("X","Y"), 1:10, sep="")
```

makes `labs` into the character vector

```
c("X1", "Y2", "X3", "Y4", "X5", "Y6", "X7", "Y8", "X9", "Y10")
```

Note particularly that recycling of short lists takes place here too; thus `c("X", "Y")` is repeated 5 times to match the sequence `1:10`.[3]

2.7 Index vectors; selecting and modifying subsets of a data set

Subsets of the elements of a vector may be selected by appending to the name of the vector an *index vector* in square brackets. More generally any expression that evaluates to a vector may have subsets of its elements similarly selected by appending an index vector in square brackets immediately after the expression.

Such index vectors can be any of four distinct types.

1. **A logical vector**. In this case the index vector is recycled to the same length as the vector from which elements are to be selected. Values corresponding to `TRUE` in the index vector are selected and those corresponding to `FALSE` are omitted. For example

   ```
   > y <- x[!is.na(x)]
   ```

   creates (or re-creates) an object `y` which will contain the non-missing values of `x`, in the same order. Note that if `x` has missing values, `y` will be shorter than `x`. Also

   ```
   > (x+1)[(!is.na(x)) & x>0] -> z
   ```

   creates an object `z` and places in it the values of the vector `x+1` for which the corresponding value in `x` was both non-missing and positive.

[3] `paste(..., collapse=`*`ss`*`)` joins the arguments into a single character string putting *ss* in between, e.g., `ss <- "|"`. There are more tools for character manipulation, see the help for `sub` and `substring`.

2. **A vector of positive integral quantities**. In this case the values in the index vector must lie in the set {1, 2, . . . , `length(x)`}. The corresponding elements of the vector are selected and concatenated, *in that order*, in the result. The index vector can be of any length and the result is of the same length as the index vector. For example `x[6]` is the sixth component of `x` and

   ```
   > x[1:10]
   ```

   selects the first 10 elements of `x` (assuming `length(x)` is not less than 10). Also

   ```
   > c("x","y")[rep(c(1,2,2,1), times=4)]
   ```

   (an admittedly unlikely thing to do) produces a character vector of length 16 consisting of `"x", "y", "y", "x"` repeated four times.
3. **A vector of negative integral quantities**. Such an index vector specifies the values to be *excluded* rather than included. Thus

   ```
   > y <- x[-(1:5)]
   ```

   gives `y` all but the first five elements of `x`.
4. **A vector of character strings**. This possibility only applies where an object has a `names` attribute to identify its components. In this case a sub-vector of the names vector may be used in the same way as the positive integral labels in item 2 further above.

   ```
   > fruit <- c(5, 10, 1, 20)
   > names(fruit) <- c("orange", "banana", "apple", "peach")
   > lunch <- fruit[c("apple","orange")]
   ```

   The advantage is that alphanumeric *names* are often easier to remember than *numeric indices*. This option is particularly useful in connection with data frames, as we shall see later.

An indexed expression can also appear on the receiving end of an assignment, in which case the assignment operation is performed *only on those elements of the vector*. The expression must be of the form `vector[`*`index_vector`*`]` as having an arbitrary expression in place of the vector name does not make much sense here.

For example

```
> x[is.na(x)] <- 0
```

replaces any missing values in `x` by zeros and

```
> y[y < 0] <- -y[y < 0]
```

has the same effect as

```
> y <- abs(y)
```

2.8 Other types of objects

Vectors are the most important type of object in R, but there are several others which we will meet more formally in later sections.

- *matrices* or more generally *arrays* are multi-dimensional generalizations of vectors. In fact, they *are* vectors that can be indexed by two or more indices and will be printed in special ways. See Chapter 5 [Arrays and matrices], page 18.
- *factors* provide compact ways to handle categorical data. See Chapter 4 [Factors], page 16.
- *lists* are a general form of vector in which the various elements need not be of the same type, and are often themselves vectors or lists. Lists provide a convenient way to return the results of a statistical computation. See Section 6.1 [Lists], page 26.
- *data frames* are matrix-like structures, in which the columns can be of different types. Think of data frames as 'data matrices' with one row per observational unit but with (possibly)

both numerical and categorical variables. Many experiments are best described by data frames: the treatments are categorical but the response is numeric. See Section 6.3 [Data frames], page 27.

- *functions* are themselves objects in R which can be stored in the project's workspace. This provides a simple and convenient way to extend R. See Chapter 10 [Writing your own functions], page 42.

3 Objects, their modes and attributes

3.1 Intrinsic attributes: mode and length

The entities R operates on are technically known as *objects*. Examples are vectors of numeric (real) or complex values, vectors of logical values and vectors of character strings. These are known as "atomic" structures since their components are all of the same type, or *mode*, namely *numeric*[1], *complex*, *logical*, *character* and *raw*.

Vectors must have their values *all of the same mode*. Thus any given vector must be unambiguously either *logical*, *numeric*, *complex*, *character* or *raw*. (The only apparent exception to this rule is the special "value" listed as `NA` for quantities not available, but in fact there are several types of `NA`). Note that a vector can be empty and still have a mode. For example the empty character string vector is listed as `character(0)` and the empty numeric vector as `numeric(0)`.

R also operates on objects called *lists*, which are of mode *list*. These are ordered sequences of objects which individually can be of any mode. *lists* are known as "recursive" rather than atomic structures since their components can themselves be lists in their own right.

The other recursive structures are those of mode *function* and *expression*. Functions are the objects that form part of the R system along with similar user written functions, which we discuss in some detail later. Expressions as objects form an advanced part of R which will not be discussed in this guide, except indirectly when we discuss *formulae* used with modeling in R.

By the *mode* of an object we mean the basic type of its fundamental constituents. This is a special case of a "property" of an object. Another property of every object is its *length*. The functions `mode(`*`object`*`)` and `length(`*`object`*`)` can be used to find out the mode and length of any defined structure[2].

Further properties of an object are usually provided by `attributes(`*`object`*`)`, see Section 3.3 [Getting and setting attributes], page 14. Because of this, *mode* and *length* are also called "intrinsic attributes" of an object.

For example, if `z` is a complex vector of length 100, then in an expression `mode(z)` is the character string `"complex"` and `length(z)` is `100`.

R caters for changes of mode almost anywhere it could be considered sensible to do so, (and a few where it might not be). For example with

```
> z <- 0:9
```

we could put

```
> digits <- as.character(z)
```

after which `digits` is the character vector `c("0", "1", "2", ..., "9")`. A further *coercion*, or change of mode, reconstructs the numerical vector again:

```
> d <- as.integer(digits)
```

Now `d` and `z` are the same.[3] There is a large collection of functions of the form `as.`*`something`*`()` for either coercion from one mode to another, or for investing an object with some other attribute it may not already possess. The reader should consult the different help files to become familiar with them.

---

[1] *numeric* mode is actually an amalgam of two distinct modes, namely *integer* and *double* precision, as explained in the manual.

[2] Note however that `length(`*`object`*`)` does not always contain intrinsic useful information, e.g., when *`object`* is a function.

[3] In general, coercion from numeric to character and back again will not be exactly reversible, because of roundoff errors in the character representation.

3.2 Changing the length of an object

An "empty" object may still have a mode. For example

```
> e <- numeric()
```

makes `e` an empty vector structure of mode numeric. Similarly `character()` is a empty character vector, and so on. Once an object of any size has been created, new components may be added to it simply by giving it an index value outside its previous range. Thus

```
> e[3] <- 17
```

now makes `e` a vector of length 3, (the first two components of which are at this point both `NA`). This applies to any structure at all, provided the mode of the additional component(s) agrees with the mode of the object in the first place.

This automatic adjustment of lengths of an object is used often, for example in the `scan()` function for input. (see Section 7.2 [The scan() function], page 31.)

Conversely to truncate the size of an object requires only an assignment to do so. Hence if `alpha` is an object of length 10, then

```
> alpha <- alpha[2 * 1:5]
```

makes it an object of length 5 consisting of just the former components with even index. (The old indices are not retained, of course.) We can then retain just the first three values by

```
> length(alpha) <- 3
```

and vectors can be extended (by missing values) in the same way.

3.3 Getting and setting attributes

The function `attributes(`*`object`*`)` returns a list of all the non-intrinsic attributes currently defined for that object. The function `attr(`*`object`*`, `*`name`*`)` can be used to select a specific attribute. These functions are rarely used, except in rather special circumstances when some new attribute is being created for some particular purpose, for example to associate a creation date or an operator with an R object. The concept, however, is very important.

Some care should be exercised when assigning or deleting attributes since they are an integral part of the object system used in R.

When it is used on the left hand side of an assignment it can be used either to associate a new attribute with *`object`* or to change an existing one. For example

```
> attr(z, "dim") <- c(10,10)
```

allows R to treat `z` as if it were a 10-by-10 matrix.

3.4 The class of an object

All objects in R have a *class*, reported by the function `class`. For simple vectors this is just the mode, for example `"numeric"`, `"logical"`, `"character"` or `"list"`, but `"matrix"`, `"array"`, `"factor"` and `"data.frame"` are other possible values.

A special attribute known as the *class* of the object is used to allow for an object-oriented style[4] of programming in R. For example if an object has class `"data.frame"`, it will be printed in a certain way, the `plot()` function will display it graphically in a certain way, and other so-called generic functions such as `summary()` will react to it as an argument in a way sensitive to its class.

To remove temporarily the effects of class, use the function `unclass()`. For example if `winter` has the class `"data.frame"` then

[4] A different style using 'formal' or 'S4' classes is provided in package `methods`.

```
> winter
```

will print it in data frame form, which is rather like a matrix, whereas

```
> unclass(winter)
```

will print it as an ordinary list. Only in rather special situations do you need to use this facility, but one is when you are learning to come to terms with the idea of class and generic functions.

Generic functions and classes will be discussed further in Section 10.9 [Object orientation], page 49, but only briefly.

4 Ordered and unordered factors

A *factor* is a vector object used to specify a discrete classification (grouping) of the components of other vectors of the same length. R provides both *ordered* and *unordered* factors. While the "real" application of factors is with model formulae (see Section 11.1.1 [Contrasts], page 53), we here look at a specific example.

4.1 A specific example

Suppose, for example, we have a sample of 30 tax accountants from all the states and territories of Australia[1] and their individual state of origin is specified by a character vector of state mnemonics as

```
> state <- c("tas", "sa",  "qld", "nsw", "nsw", "nt",  "wa",  "wa",
             "qld", "vic", "nsw", "vic", "qld", "qld", "sa",  "tas",
             "sa",  "nt",  "wa",  "vic", "qld", "nsw", "nsw", "wa",
             "sa",  "act", "nsw", "vic", "vic", "act")
```

Notice that in the case of a character vector, "sorted" means sorted in alphabetical order.

A *factor* is similarly created using the `factor()` function:

```
> statef <- factor(state)
```

The `print()` function handles factors slightly differently from other objects:

```
> statef
 [1] tas sa  qld nsw nsw nt  wa  wa  qld vic nsw vic qld qld sa
[16] tas sa  nt  wa  vic qld nsw nsw wa  sa  act nsw vic vic act
Levels:  act nsw nt qld sa tas vic wa
```

To find out the levels of a factor the function `levels()` can be used.

```
> levels(statef)
[1] "act" "nsw" "nt"  "qld" "sa"  "tas" "vic" "wa"
```

4.2 The function `tapply()` and ragged arrays

To continue the previous example, suppose we have the incomes of the same tax accountants in another vector (in suitably large units of money)

```
> incomes <- c(60, 49, 40, 61, 64, 60, 59, 54, 62, 69, 70, 42, 56,
               61, 61, 61, 58, 51, 48, 65, 49, 49, 41, 48, 52, 46,
               59, 46, 58, 43)
```

To calculate the sample mean income for each state we can now use the special function `tapply()`:

```
> incmeans <- tapply(incomes, statef, mean)
```

giving a means vector with the components labelled by the levels

```
   act    nsw     nt    qld     sa    tas    vic     wa
44.500 57.333 55.500 53.600 55.000 60.500 56.000 52.250
```

The function `tapply()` is used to apply a function, here `mean()`, to each group of components of the first argument, here `incomes`, defined by the levels of the second component, here `statef`[2],

[1] Readers should note that there are eight states and territories in Australia, namely the Australian Capital Territory, New South Wales, the Northern Territory, Queensland, South Australia, Tasmania, Victoria and Western Australia.

[2] Note that `tapply()` also works in this case when its second argument is not a factor, e.g., '`tapply(incomes, state)`', and this is true for quite a few other functions, since arguments are *coerced* to factors when necessary (using `as.factor()`).

as if they were separate vector structures. The result is a structure of the same length as the levels attribute of the factor containing the results. The reader should consult the help document for more details.

Suppose further we needed to calculate the standard errors of the state income means. To do this we need to write an R function to calculate the standard error for any given vector. Since there is an builtin function `var()` to calculate the sample variance, such a function is a very simple one liner, specified by the assignment:

```
> stderr <- function(x) sqrt(var(x)/length(x))
```

(Writing functions will be considered later in Chapter 10 [Writing your own functions], page 42, and in this case was unnecessary as R also has a builtin function `sd()`.) After this assignment, the standard errors are calculated by

```
> incster <- tapply(incomes, statef, stderr)
```

and the values calculated are then

```
> incster
act    nsw  nt    qld     sa tas   vic     wa
1.5 4.3102 4.5 4.1061 2.7386 0.5 5.244 2.6575
```

As an exercise you may care to find the usual 95% confidence limits for the state mean incomes. To do this you could use `tapply()` once more with the `length()` function to find the sample sizes, and the `qt()` function to find the percentage points of the appropriate t-distributions. (You could also investigate R's facilities for t-tests.)

The function `tapply()` can also be used to handle more complicated indexing of a vector by multiple categories. For example, we might wish to split the tax accountants by both state and sex. However in this simple instance (just one factor) what happens can be thought of as follows. The values in the vector are collected into groups corresponding to the distinct entries in the factor. The function is then applied to each of these groups individually. The value is a vector of function results, labelled by the `levels` attribute of the factor.

The combination of a vector and a labelling factor is an example of what is sometimes called a *ragged array*, since the subclass sizes are possibly irregular. When the subclass sizes are all the same the indexing may be done implicitly and much more efficiently, as we see in the next section.

4.3 Ordered factors

The levels of factors are stored in alphabetical order, or in the order they were specified to `factor` if they were specified explicitly.

Sometimes the levels will have a natural ordering that we want to record and want our statistical analysis to make use of. The `ordered()` function creates such ordered factors but is otherwise identical to `factor`. For most purposes the only difference between ordered and unordered factors is that the former are printed showing the ordering of the levels, but the contrasts generated for them in fitting linear models are different.

5 Arrays and matrices

5.1 Arrays

An array can be considered as a multiply subscripted collection of data entries, for example numeric. R allows simple facilities for creating and handling arrays, and in particular the special case of matrices.

A dimension vector is a vector of non-negative integers. If its length is k then the array is k-dimensional, e.g. a matrix is a 2-dimensional array. The dimensions are indexed from one up to the values given in the dimension vector.

A vector can be used by R as an array only if it has a dimension vector as its *dim* attribute. Suppose, for example, `z` is a vector of 1500 elements. The assignment

```
> dim(z) <- c(3,5,100)
```

gives it the *dim* attribute that allows it to be treated as a 3 by 5 by 100 array.

Other functions such as `matrix()` and `array()` are available for simpler and more natural looking assignments, as we shall see in Section 5.4 [The array() function], page 20.

The values in the data vector give the values in the array in the same order as they would occur in FORTRAN, that is "column major order," with the first subscript moving fastest and the last subscript slowest.

For example if the dimension vector for an array, say `a`, is `c(3,4,2)` then there are $3 \times 4 \times 2 = 24$ entries in `a` and the data vector holds them in the order `a[1,1,1], a[2,1,1], ..., a[2,4,2], a[3,4,2]`.

Arrays can be one-dimensional: such arrays are usually treated in the same way as vectors (including when printing), but the exceptions can cause confusion.

5.2 Array indexing. Subsections of an array

Individual elements of an array may be referenced by giving the name of the array followed by the subscripts in square brackets, separated by commas.

More generally, subsections of an array may be specified by giving a sequence of *index vectors* in place of subscripts; however *if any index position is given an empty index vector, then the full range of that subscript is taken.*

Continuing the previous example, `a[2,,]` is a 4×2 array with dimension vector `c(4,2)` and data vector containing the values

```
c(a[2,1,1], a[2,2,1], a[2,3,1], a[2,4,1],
  a[2,1,2], a[2,2,2], a[2,3,2], a[2,4,2])
```

in that order. `a[,,]` stands for the entire array, which is the same as omitting the subscripts entirely and using `a` alone.

For any array, say `Z`, the dimension vector may be referenced explicitly as `dim(Z)` (on either side of an assignment).

Also, if an array name is given with just *one subscript or index vector*, then the corresponding values of the data vector only are used; in this case the dimension vector is ignored. This is not the case, however, if the single index is not a vector but itself an array, as we next discuss.

5.3 Index matrices

As well as an index vector in any subscript position, a matrix may be used with a single *index matrix* in order either to assign a vector of quantities to an irregular collection of elements in the array, or to extract an irregular collection as a vector.

A matrix example makes the process clear. In the case of a doubly indexed array, an index matrix may be given consisting of two columns and as many rows as desired. The entries in the index matrix are the row and column indices for the doubly indexed array. Suppose for example we have a 4 by 5 array `X` and we wish to do the following:

- Extract elements `X[1,3]`, `X[2,2]` and `X[3,1]` as a vector structure, and
- Replace these entries in the array `X` by zeroes.

In this case we need a 3 by 2 subscript array, as in the following example.

```
> x <- array(1:20, dim=c(4,5))    # Generate a 4 by 5 array.
> x
     [,1] [,2] [,3] [,4] [,5]
[1,]    1    5    9   13   17
[2,]    2    6   10   14   18
[3,]    3    7   11   15   19
[4,]    4    8   12   16   20
> i <- array(c(1:3,3:1), dim=c(3,2))
> i                               # i is a 3 by 2 index array.
     [,1] [,2]
[1,]    1    3
[2,]    2    2
[3,]    3    1
> x[i]                            # Extract those elements
[1] 9 6 3
> x[i] <- 0                       # Replace those elements by zeros.
> x
     [,1] [,2] [,3] [,4] [,5]
[1,]    1    5    0   13   17
[2,]    2    0   10   14   18
[3,]    0    7   11   15   19
[4,]    4    8   12   16   20
>
```

Negative indices are not allowed in index matrices. `NA` and zero values are allowed: rows in the index matrix containing a zero are ignored, and rows containing an `NA` produce an `NA` in the result.

As a less trivial example, suppose we wish to generate an (unreduced) design matrix for a block design defined by factors `blocks` (`b` levels) and `varieties` (`v` levels). Further suppose there are `n` plots in the experiment. We could proceed as follows:

```
> Xb <- matrix(0, n, b)
> Xv <- matrix(0, n, v)
> ib <- cbind(1:n, blocks)
> iv <- cbind(1:n, varieties)
> Xb[ib] <- 1
> Xv[iv] <- 1
> X <- cbind(Xb, Xv)
```

To construct the incidence matrix, `N` say, we could use

```
> N <- crossprod(Xb, Xv)
```

However a simpler direct way of producing this matrix is to use `table()`:

```
> N <- table(blocks, varieties)
```

Index matrices must be numerical: any other form of matrix (e.g. a logical or character matrix) supplied as a matrix is treated as an indexing vector.

5.4 The `array()` function

As well as giving a vector structure a `dim` attribute, arrays can be constructed from vectors by the `array` function, which has the form

```
> Z <- array(data_vector, dim_vector)
```

For example, if the vector `h` contains 24 or fewer, numbers then the command

```
> Z <- array(h, dim=c(3,4,2))
```

would use `h` to set up 3 by 4 by 2 array in `Z`. If the size of `h` is exactly 24 the result is the same as

```
> Z <- h ; dim(Z) <- c(3,4,2)
```

However if `h` is shorter than 24, its values are recycled from the beginning again to make it up to size 24 (see Section 5.4.1 [The recycling rule], page 20) but `dim(h) <- c(3,4,2)` would signal an error about mismatching length. As an extreme but common example

```
> Z <- array(0, c(3,4,2))
```

makes `Z` an array of all zeros.

At this point `dim(Z)` stands for the dimension vector `c(3,4,2)`, and `Z[1:24]` stands for the data vector as it was in `h`, and `Z[]` with an empty subscript or `Z` with no subscript stands for the entire array as an array.

Arrays may be used in arithmetic expressions and the result is an array formed by element-by-element operations on the data vector. The `dim` attributes of operands generally need to be the same, and this becomes the dimension vector of the result. So if `A`, `B` and `C` are all similar arrays, then

```
> D <- 2*A*B + C + 1
```

makes `D` a similar array with its data vector being the result of the given element-by-element operations. However the precise rule concerning mixed array and vector calculations has to be considered a little more carefully.

5.4.1 Mixed vector and array arithmetic. The recycling rule

The precise rule affecting element by element mixed calculations with vectors and arrays is somewhat quirky and hard to find in the references. From experience we have found the following to be a reliable guide.

- The expression is scanned from left to right.
- Any short vector operands are extended by recycling their values until they match the size of any other operands.
- As long as short vectors and arrays *only* are encountered, the arrays must all have the same `dim` attribute or an error results.
- Any vector operand longer than a matrix or array operand generates an error.
- If array structures are present and no error or coercion to vector has been precipitated, the result is an array structure with the common `dim` attribute of its array operands.

5.5 The outer product of two arrays

An important operation on arrays is the *outer product*. If `a` and `b` are two numeric arrays, their outer product is an array whose dimension vector is obtained by concatenating their two dimension vectors (order is important), and whose data vector is got by forming all possible products of elements of the data vector of `a` with those of `b`. The outer product is formed by the special operator `%o%`:

```
> ab <- a %o% b
```

An alternative is

```
> ab <- outer(a, b, "*")
```

The multiplication function can be replaced by an arbitrary function of two variables. For example if we wished to evaluate the function $f(x; y) = \cos(y)/(1 + x^2)$ over a regular grid of values with x- and y-coordinates defined by the R vectors `x` and `y` respectively, we could proceed as follows:

```
> f <- function(x, y) cos(y)/(1 + x^2)
> z <- outer(x, y, f)
```

In particular the outer product of two ordinary vectors is a doubly subscripted array (that is a matrix, of rank at most 1). Notice that the outer product operator is of course non-commutative. Defining your own R functions will be considered further in Chapter 10 [Writing your own functions], page 42.

An example: Determinants of 2 by 2 single-digit matrices

As an artificial but cute example, consider the determinants of 2 by 2 matrices $[a, b; c, d]$ where each entry is a non-negative integer in the range $0, 1, \ldots, 9$, that is a digit.

The problem is to find the determinants, $ad - bc$, of all possible matrices of this form and represent the frequency with which each value occurs as a *high density* plot. This amounts to finding the probability distribution of the determinant if each digit is chosen independently and uniformly at random.

A neat way of doing this uses the `outer()` function twice:

```
> d <- outer(0:9, 0:9)
> fr <- table(outer(d, d, "-"))
> plot(as.numeric(names(fr)), fr, type="h",
       xlab="Determinant", ylab="Frequency")
```

Notice the coercion of the `names` attribute of the frequency table to numeric in order to recover the range of the determinant values. The "obvious" way of doing this problem with `for` loops, to be discussed in Chapter 9 [Loops and conditional execution], page 40, is so inefficient as to be impractical.

It is also perhaps surprising that about 1 in 20 such matrices is singular.

5.6 Generalized transpose of an array

The function `aperm(a, perm)` may be used to permute an array, `a`. The argument `perm` must be a permutation of the integers $\{1, \ldots, k\}$, where k is the number of subscripts in `a`. The result of the function is an array of the same size as `a` but with old dimension given by `perm[j]` becoming the new `j`-th dimension. The easiest way to think of this operation is as a generalization of transposition for matrices. Indeed if `A` is a matrix, (that is, a doubly subscripted array) then `B` given by

```
> B <- aperm(A, c(2,1))
```

is just the transpose of `A`. For this special case a simpler function `t()` is available, so we could have used `B <- t(A)`.

5.7 Matrix facilities

As noted above, a matrix is just an array with two subscripts. However it is such an important special case it needs a separate discussion. R contains many operators and functions that are available only for matrices. For example `t(X)` is the matrix transpose function, as noted above. The functions `nrow(A)` and `ncol(A)` give the number of rows and columns in the matrix `A` respectively.

5.7.1 Matrix multiplication

The operator `%*%` is used for matrix multiplication. An n by 1 or 1 by n matrix may of course be used as an n-vector if in the context such is appropriate. Conversely, vectors which occur in matrix multiplication expressions are automatically promoted either to row or column vectors, whichever is multiplicatively coherent, if possible, (although this is not always unambiguously possible, as we see later).

If, for example, `A` and `B` are square matrices of the same size, then

```
> A * B
```

is the matrix of element by element products and

```
> A %*% B
```

is the matrix product. If `x` is a vector, then

```
> x %*% A %*% x
```

is a quadratic form.[1]

The function `crossprod()` forms "crossproducts", meaning that `crossprod(X, y)` is the same as `t(X) %*% y` but the operation is more efficient. If the second argument to `crossprod()` is omitted it is taken to be the same as the first.

The meaning of `diag()` depends on its argument. `diag(v)`, where `v` is a vector, gives a diagonal matrix with elements of the vector as the diagonal entries. On the other hand `diag(M)`, where `M` is a matrix, gives the vector of main diagonal entries of `M`. This is the same convention as that used for `diag()` in MATLAB. Also, somewhat confusingly, if `k` is a single numeric value then `diag(k)` is the `k` by `k` identity matrix!

5.7.2 Linear equations and inversion

Solving linear equations is the inverse of matrix multiplication. When after

```
> b <- A %*% x
```

only `A` and `b` are given, the vector `x` is the solution of that linear equation system. In R,

```
> solve(A,b)
```

solves the system, returning `x` (up to some accuracy loss). Note that in linear algebra, formally $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$ where $\mathbf{A}^{-1}$ denotes the *inverse* of $\mathbf{A}$, which can be computed by

```
solve(A)
```

but rarely is needed. Numerically, it is both inefficient and potentially unstable to compute `x <- solve(A) %*% b` instead of `solve(A,b)`.

The quadratic form $\mathbf{x^T A^{-1} x}$ which is used in multivariate computations, should be computed by something like[2] `x %*% solve(A,x)`, rather than computing the inverse of `A`.

---

[1] Note that `x %*% x` is ambiguous, as it could mean either $\mathbf{x}^T\mathbf{x}$ or $\mathbf{x}\mathbf{x}^T$, where `x` is the column form. In such cases the smaller matrix seems implicitly to be the interpretation adopted, so the scalar $\mathbf{x}^T\mathbf{x}$ is in this case the result. The matrix $\mathbf{x}\mathbf{x}^T$ may be calculated either by `cbind(x) %*% x` or `x %*% rbind(x)` since the result of `rbind()` or `cbind()` is always a matrix. However, the best way to compute $\mathbf{x}^T\mathbf{x}$ or $\mathbf{x}\mathbf{x}^T$ is `crossprod(x)` or `x %o% x` respectively.

[2] Even better would be to form a matrix square root B with $A = BB^T$ and find the squared length of the solution of $By = x$, perhaps using the Cholesky or eigen decomposition of A.

5.7.3 Eigenvalues and eigenvectors

The function `eigen(Sm)` calculates the eigenvalues and eigenvectors of a symmetric matrix `Sm`. The result of this function is a list of two components named `values` and `vectors`. The assignment

```
> ev <- eigen(Sm)
```

will assign this list to `ev`. Then `ev$val` is the vector of eigenvalues of `Sm` and `ev$vec` is the matrix of corresponding eigenvectors. Had we only needed the eigenvalues we could have used the assignment:

```
> evals <- eigen(Sm)$values
```

`evals` now holds the vector of eigenvalues and the second component is discarded. If the expression

```
> eigen(Sm)
```

is used by itself as a command the two components are printed, with their names. For large matrices it is better to avoid computing the eigenvectors if they are not needed by using the expression

```
> evals <- eigen(Sm, only.values = TRUE)$values
```

5.7.4 Singular value decomposition and determinants

The function `svd(M)` takes an arbitrary matrix argument, `M`, and calculates the singular value decomposition of `M`. This consists of a matrix of orthonormal columns `U` with the same column space as `M`, a second matrix of orthonormal columns `V` whose column space is the row space of `M` and a diagonal matrix of positive entries `D` such that `M = U %*% D %*% t(V)`. `D` is actually returned as a vector of the diagonal elements. The result of `svd(M)` is actually a list of three components named `d`, `u` and `v`, with evident meanings.

If `M` is in fact square, then, it is not hard to see that

```
> absdetM <- prod(svd(M)$d)
```

calculates the absolute value of the determinant of `M`. If this calculation were needed often with a variety of matrices it could be defined as an R function

```
> absdet <- function(M) prod(svd(M)$d)
```

after which we could use `absdet()` as just another R function. As a further trivial but potentially useful example, you might like to consider writing a function, say `tr()`, to calculate the trace of a square matrix. [Hint: You will not need to use an explicit loop. Look again at the `diag()` function.]

R has a builtin function `det` to calculate a determinant, including the sign, and another, `determinant`, to give the sign and modulus (optionally on log scale),

5.7.5 Least squares fitting and the QR decomposition

The function `lsfit()` returns a list giving results of a least squares fitting procedure. An assignment such as

```
> ans <- lsfit(X, y)
```

gives the results of a least squares fit where `y` is the vector of observations and `X` is the design matrix. See the help facility for more details, and also for the follow-up function `ls.diag()` for, among other things, regression diagnostics. Note that a grand mean term is automatically included and need not be included explicitly as a column of `X`. Further note that you almost always will prefer using `lm(.)` (see Section 11.2 [Linear models], page 54) to `lsfit()` for regression modelling.

Another closely related function is `qr()` and its allies. Consider the following assignments

```
> Xplus <- qr(X)
> b <- qr.coef(Xplus, y)
> fit <- qr.fitted(Xplus, y)
> res <- qr.resid(Xplus, y)
```

These compute the orthogonal projection of `y` onto the range of `X` in `fit`, the projection onto the orthogonal complement in `res` and the coefficient vector for the projection in `b`, that is, `b` is essentially the result of the MATLAB 'backslash' operator.

It is not assumed that `X` has full column rank. Redundancies will be discovered and removed as they are found.

This alternative is the older, low-level way to perform least squares calculations. Although still useful in some contexts, it would now generally be replaced by the statistical models features, as will be discussed in Chapter 11 [Statistical models in R], page 51.

5.8 Forming partitioned matrices, `cbind()` and `rbind()`

As we have already seen informally, matrices can be built up from other vectors and matrices by the functions `cbind()` and `rbind()`. Roughly `cbind()` forms matrices by binding together matrices horizontally, or column-wise, and `rbind()` vertically, or row-wise.

In the assignment

```
> X <- cbind(arg_1, arg_2, arg_3, ...)
```

the arguments to `cbind()` must be either vectors of any length, or matrices with the same column size, that is the same number of rows. The result is a matrix with the concatenated arguments *arg_1*, *arg_2*, ... forming the columns.

If some of the arguments to `cbind()` are vectors they may be shorter than the column size of any matrices present, in which case they are cyclically extended to match the matrix column size (or the length of the longest vector if no matrices are given).

The function `rbind()` does the corresponding operation for rows. In this case any vector argument, possibly cyclically extended, are of course taken as row vectors.

Suppose `X1` and `X2` have the same number of rows. To combine these by columns into a matrix `X`, together with an initial column of `1`s we can use

```
> X <- cbind(1, X1, X2)
```

The result of `rbind()` or `cbind()` always has matrix status. Hence `cbind(x)` and `rbind(x)` are possibly the simplest ways explicitly to allow the vector `x` to be treated as a column or row matrix respectively.

5.9 The concatenation function, `c()`, with arrays

It should be noted that whereas `cbind()` and `rbind()` are concatenation functions that respect `dim` attributes, the basic `c()` function does not, but rather clears numeric objects of all `dim` and `dimnames` attributes. This is occasionally useful in its own right.

The official way to coerce an array back to a simple vector object is to use `as.vector()`

```
> vec <- as.vector(X)
```

However a similar result can be achieved by using `c()` with just one argument, simply for this side-effect:

```
> vec <- c(X)
```

There are slight differences between the two, but ultimately the choice between them is largely a matter of style (with the former being preferable).

5.10 Frequency tables from factors

Recall that a factor defines a partition into groups. Similarly a pair of factors defines a two way cross classification, and so on. The function `table()` allows frequency tables to be calculated from equal length factors. If there are k factor arguments, the result is a k-way array of frequencies.

Suppose, for example, that `statef` is a factor giving the state code for each entry in a data vector. The assignment

```
> statefr <- table(statef)
```

gives in `statefr` a table of frequencies of each state in the sample. The frequencies are ordered and labelled by the `levels` attribute of the factor. This simple case is equivalent to, but more convenient than,

```
> statefr <- tapply(statef, statef, length)
```

Further suppose that `incomef` is a factor giving a suitably defined "income class" for each entry in the data vector, for example with the `cut()` function:

```
> factor(cut(incomes, breaks = 35+10*(0:7))) -> incomef
```

Then to calculate a two-way table of frequencies:

```
> table(incomef,statef)
          statef
incomef    act nsw nt qld sa tas vic wa
  (35,45]    1   1  0   1  0   0   1  0
  (45,55]    1   1  1   1  2   0   1  3
  (55,65]    0   3  1   3  2   2   2  1
  (65,75]    0   1  0   0  0   0   1  0
```

Extension to higher-way frequency tables is immediate.

6 Lists and data frames

6.1 Lists

An R *list* is an object consisting of an ordered collection of objects known as its *components*.

There is no particular need for the components to be of the same mode or type, and, for example, a list could consist of a numeric vector, a logical value, a matrix, a complex vector, a character array, a function, and so on. Here is a simple example of how to make a list:

```
> Lst <- list(name="Fred", wife="Mary", no.children=3,
              child.ages=c(4,7,9))
```

Components are always *numbered* and may always be referred to as such. Thus if `Lst` is the name of a list with four components, these may be individually referred to as `Lst[[1]]`, `Lst[[2]]`, `Lst[[3]]` and `Lst[[4]]`. If, further, `Lst[[4]]` is a vector subscripted array then `Lst[[4]][1]` is its first entry.

If `Lst` is a list, then the function `length(Lst)` gives the number of (top level) components it has.

Components of lists may also be *named*, and in this case the component may be referred to either by giving the component name as a character string in place of the number in double square brackets, or, more conveniently, by giving an expression of the form

```
> name$component_name
```

for the same thing.

This is a very useful convention as it makes it easier to get the right component if you forget the number.

So in the simple example given above:

`Lst$name` is the same as `Lst[[1]]` and is the string `"Fred"`,

`Lst$wife` is the same as `Lst[[2]]` and is the string `"Mary"`,

`Lst$child.ages[1]` is the same as `Lst[[4]][1]` and is the number `4`.

Additionally, one can also use the names of the list components in double square brackets, i.e., `Lst[["name"]]` is the same as `Lst$name`. This is especially useful, when the name of the component to be extracted is stored in another variable as in

```
> x <- "name"; Lst[[x]]
```

It is very important to distinguish `Lst[[1]]` from `Lst[1]`. '`[[...]]`' is the operator used to select a single element, whereas '`[...]`' is a general subscripting operator. Thus the former is the *first object in the list* `Lst`, and if it is a named list the name is *not* included. The latter is a *sublist of the list* `Lst` *consisting of the first entry only. If it is a named list, the names are transferred to the sublist.*

The names of components may be abbreviated down to the minimum number of letters needed to identify them uniquely. Thus `Lst$coefficients` may be minimally specified as `Lst$coe` and `Lst$covariance` as `Lst$cov`.

The vector of names is in fact simply an attribute of the list like any other and may be handled as such. Other structures besides lists may, of course, similarly be given a *names* attribute also.

6.2 Constructing and modifying lists

New lists may be formed from existing objects by the function `list()`. An assignment of the form

```
> Lst <- list(name_1=object_1, ..., name_m=object_m)
```

sets up a list `Lst` of m components using *object_1*, ..., *object_m* for the components and giving them names as specified by the argument names, (which can be freely chosen). If these names are omitted, the components are numbered only. The components used to form the list are *copied* when forming the new list and the originals are not affected.

Lists, like any subscripted object, can be extended by specifying additional components. For example

```
> Lst[5] <- list(matrix=Mat)
```

6.2.1 Concatenating lists

When the concatenation function `c()` is given list arguments, the result is an object of mode list also, whose components are those of the argument lists joined together in sequence.

```
> list.ABC <- c(list.A, list.B, list.C)
```

Recall that with vector objects as arguments the concatenation function similarly joined together all arguments into a single vector structure. In this case all other attributes, such as `dim` attributes, are discarded.

6.3 Data frames

A *data frame* is a list with class `"data.frame"`. There are restrictions on lists that may be made into data frames, namely

- The components must be vectors (numeric, character, or logical), factors, numeric matrices, lists, or other data frames.
- Matrices, lists, and data frames provide as many variables to the new data frame as they have columns, elements, or variables, respectively.
- Numeric vectors, logicals and factors are included as is, and by default[1] character vectors are coerced to be factors, whose levels are the unique values appearing in the vector.
- Vector structures appearing as variables of the data frame must all have the *same length*, and matrix structures must all have the same *row size*.

A data frame may for many purposes be regarded as a matrix with columns possibly of differing modes and attributes. It may be displayed in matrix form, and its rows and columns extracted using matrix indexing conventions.

6.3.1 Making data frames

Objects satisfying the restrictions placed on the columns (components) of a data frame may be used to form one using the function `data.frame`:

```
> accountants <- data.frame(home=statef, loot=incomes, shot=incomef)
```

A list whose components conform to the restrictions of a data frame may be *coerced* into a data frame using the function `as.data.frame()`

The simplest way to construct a data frame from scratch is to use the `read.table()` function to read an entire data frame from an external file. This is discussed further in Chapter 7 [Reading data from files], page 30.

[1] Conversion of character columns to factors is overridden using the `stringsAsFactors` argument to the `data.frame()` function.

6.3.2 `attach()` and `detach()`

The `$` notation, such as `accountants$home`, for list components is not always very convenient. A useful facility would be somehow to make the components of a list or data frame temporarily visible as variables under their component name, without the need to quote the list name explicitly each time.

The `attach()` function takes a 'database' such as a list or data frame as its argument. Thus suppose `lentils` is a data frame with three variables `lentils$u`, `lentils$v`, `lentils$w`. The attach

```
> attach(lentils)
```

places the data frame in the search path at position 2, and provided there are no variables `u`, `v` or `w` in position 1, `u`, `v` and `w` are available as variables from the data frame in their own right. At this point an assignment such as

```
> u <- v+w
```

does not replace the component `u` of the data frame, but rather masks it with another variable `u` in the working directory at position 1 on the search path. To make a permanent change to the data frame itself, the simplest way is to resort once again to the `$` notation:

```
> lentils$u <- v+w
```

However the new value of component `u` is not visible until the data frame is detached and attached again.

To detach a data frame, use the function

```
> detach()
```

More precisely, this statement detaches from the search path the entity currently at position 2. Thus in the present context the variables `u`, `v` and `w` would be no longer visible, except under the list notation as `lentils$u` and so on. Entities at positions greater than 2 on the search path can be detached by giving their number to `detach`, but it is much safer to always use a name, for example by `detach(lentils)` or `detach("lentils")`

> **Note:** In R lists and data frames can only be attached at position 2 or above, and what is attached is a *copy* of the original object. You can alter the attached values *via* `assign`, but the original list or data frame is unchanged.

6.3.3 Working with data frames

A useful convention that allows you to work with many different problems comfortably together in the same working directory is

- gather together all variables for any well defined and separate problem in a data frame under a suitably informative name;
- when working with a problem attach the appropriate data frame at position 2, and use the working directory at level 1 for operational quantities and temporary variables;
- before leaving a problem, add any variables you wish to keep for future reference to the data frame using the `$` form of assignment, and then `detach()`;
- finally remove all unwanted variables from the working directory and keep it as clean of left-over temporary variables as possible.

In this way it is quite simple to work with many problems in the same directory, all of which have variables named `x`, `y` and `z`, for example.

6.3.4 Attaching arbitrary lists

`attach()` is a generic function that allows not only directories and data frames to be attached to the search path, but other classes of object as well. In particular any object of mode `"list"` may be attached in the same way:

```
> attach(any.old.list)
```

Anything that has been attached can be detached by `detach`, by position number or, preferably, by name.

6.3.5 Managing the search path

The function `search` shows the current search path and so is a very useful way to keep track of which data frames and lists (and packages) have been attached and detached. Initially it gives

```
> search()
[1] ".GlobalEnv"   "Autoloads"    "package:base"
```

where `.GlobalEnv` is the workspace.[2]

After `lentils` is attached we have

```
> search()
[1] ".GlobalEnv"   "lentils"      "Autoloads"    "package:base"
> ls(2)
[1] "u" "v" "w"
```

and as we see `ls` (or `objects`) can be used to examine the contents of any position on the search path.

Finally, we detach the data frame and confirm it has been removed from the search path.

```
> detach("lentils")
> search()
[1] ".GlobalEnv"   "Autoloads"    "package:base"
```

[2] See the on-line help for `autoload` for the meaning of the second term.

7 Reading data from files

Large data objects will usually be read as values from external files rather than entered during an R session at the keyboard. R input facilities are simple and their requirements are fairly strict and even rather inflexible. There is a clear presumption by the designers of R that you will be able to modify your input files using other tools, such as file editors or Perl[1] to fit in with the requirements of R. Generally this is very simple.

If variables are to be held mainly in data frames, as we strongly suggest they should be, an entire data frame can be read directly with the `read.table()` function. There is also a more primitive input function, `scan()`, that can be called directly.

For more details on importing data into R and also exporting data, see the *R Data Import/Export* manual.

7.1 The `read.table()` function

To read an entire data frame directly, the external file will normally have a special form.

- The first line of the file should have a *name* for each variable in the data frame.
- Each additional line of the file has as its first item a *row label* and the values for each variable.

If the file has one fewer item in its first line than in its second, this arrangement is presumed to be in force. So the first few lines of a file to be read as a data frame might look as follows.

Input file form with names and row labels:

```
     Price    Floor     Area   Rooms     Age  Cent.heat
01   52.00    111.0      830     5       6.2      no
02   54.75    128.0      710     5       7.5      no
03   57.50    101.0     1000     5       4.2      no
04   57.50    131.0      690     6       8.8      no
05   59.75     93.0      900     5       1.9     yes
...
```

By default numeric items (except row labels) are read as numeric variables and non-numeric variables, such as `Cent.heat` in the example, as factors. This can be changed if necessary.

The function `read.table()` can then be used to read the data frame directly

```
> HousePrice <- read.table("houses.data")
```

Often you will want to omit including the row labels directly and use the default labels. In this case the file may omit the row label column as in the following.

Input file form without row labels:

```
Price    Floor     Area   Rooms     Age  Cent.heat
52.00    111.0      830     5       6.2      no
54.75    128.0      710     5       7.5      no
57.50    101.0     1000     5       4.2      no
57.50    131.0      690     6       8.8      no
59.75     93.0      900     5       1.9     yes
...
```

[1] Under UNIX, the utilities `sed` or `awk` can be used.

The data frame may then be read as

```
> HousePrice <- read.table("houses.data", header=TRUE)
```

where the `header=TRUE` option specifies that the first line is a line of headings, and hence, by implication from the form of the file, that no explicit row labels are given.

7.2 The `scan()` function

Suppose the data vectors are of equal length and are to be read in parallel. Further suppose that there are three vectors, the first of mode character and the remaining two of mode numeric, and the file is `input.dat`. The first step is to use `scan()` to read in the three vectors as a list, as follows

```
> inp <- scan("input.dat", list("",0,0))
```

The second argument is a dummy list structure that establishes the mode of the three vectors to be read. The result, held in `inp`, is a list whose components are the three vectors read in. To separate the data items into three separate vectors, use assignments like

```
> label <- inp[[1]]; x <- inp[[2]]; y <- inp[[3]]
```

More conveniently, the dummy list can have named components, in which case the names can be used to access the vectors read in. For example

```
> inp <- scan("input.dat", list(id="", x=0, y=0))
```

If you wish to access the variables separately they may either be re-assigned to variables in the working frame:

```
> label <- inp$id; x <- inp$x; y <- inp$y
```

or the list may be attached at position 2 of the search path (see Section 6.3.4 [Attaching arbitrary lists], page 28).

If the second argument is a single value and not a list, a single vector is read in, all components of which must be of the same mode as the dummy value.

```
> X <- matrix(scan("light.dat", 0), ncol=5, byrow=TRUE)
```

There are more elaborate input facilities available and these are detailed in the manuals.

7.3 Accessing builtin datasets

Around 100 datasets are supplied with R (in package **datasets**), and others are available in packages (including the recommended packages supplied with R). To see the list of datasets currently available use

```
data()
```

All the datasets supplied with R are available directly by name. However, many packages still use the obsolete convention in which `data` was also used to load datasets into R, for example

```
data(infert)
```

and this can still be used with the standard packages (as in this example). In most cases this will load an R object of the same name. However, in a few cases it loads several objects, so see the on-line help for the object to see what to expect.

7.3.1 Loading data from other R packages

To access data from a particular package, use the `package` argument, for example

```
data(package="rpart")
data(Puromycin, package="datasets")
```

If a package has been attached by `library`, its datasets are automatically included in the search.

User-contributed packages can be a rich source of datasets.

7.4 Editing data

When invoked on a data frame or matrix, `edit` brings up a separate spreadsheet-like environment for editing. This is useful for making small changes once a data set has been read. The command

```
> xnew <- edit(xold)
```

will allow you to edit your data set `xold`, and on completion the changed object is assigned to `xnew`. If you want to alter the original dataset `xold`, the simplest way is to use `fix(xold)`, which is equivalent to `xold <- edit(xold)`.

Use

```
> xnew <- edit(data.frame())
```

to enter new data via the spreadsheet interface.

8 Probability distributions

8.1 R as a set of statistical tables

One convenient use of R is to provide a comprehensive set of statistical tables. Functions are provided to evaluate the cumulative distribution function $P(X \leq x)$, the probability density function and the quantile function (given q, the smallest x such that $P(X \leq x) > q$), and to simulate from the distribution.

Distribution	R name	additional arguments
beta	`beta`	`shape1, shape2, ncp`
binomial	`binom`	`size, prob`
Cauchy	`cauchy`	`location, scale`
chi-squared	`chisq`	`df, ncp`
exponential	`exp`	`rate`
F	`f`	`df1, df2, ncp`
gamma	`gamma`	`shape, scale`
geometric	`geom`	`prob`
hypergeometric	`hyper`	`m, n, k`
log-normal	`lnorm`	`meanlog, sdlog`
logistic	`logis`	`location, scale`
negative binomial	`nbinom`	`size, prob`
normal	`norm`	`mean, sd`
Poisson	`pois`	`lambda`
signed rank	`signrank`	`n`
Student's t	`t`	`df, ncp`
uniform	`unif`	`min, max`
Weibull	`weibull`	`shape, scale`
Wilcoxon	`wilcox`	`m, n`

Prefix the name given here by '`d`' for the density, '`p`' for the CDF, '`q`' for the quantile function and '`r`' for simulation (*r*andom deviates). The first argument is `x` for `dxxx`, `q` for `pxxx`, `p` for `qxxx` and `n` for `rxxx` (except for `rhyper`, `rsignrank` and `rwilcox`, for which it is `nn`). In not quite all cases is the non-centrality parameter `ncp` currently available: see the on-line help for details.

The `pxxx` and `qxxx` functions all have logical arguments `lower.tail` and `log.p` and the `dxxx` ones have `log`. This allows, e.g., getting the cumulative (or "integrated") *hazard* function, $H(t) = -\log(1 - F(t))$, by

```
- pxxx(t, ..., lower.tail = FALSE, log.p = TRUE)
```

or more accurate log-likelihoods (by `dxxx(..., log = TRUE)`), directly.

In addition there are functions `ptukey` and `qtukey` for the distribution of the studentized range of samples from a normal distribution, and `dmultinom` and `rmultinom` for the multinomial distribution. Further distributions are available in contributed packages, notably **SuppDists** (`https://CRAN.R-project.org/package=SuppDists`).

Here are some examples

```
> ## 2-tailed p-value for t distribution
> 2*pt(-2.43, df = 13)
> ## upper 1% point for an F(2, 7) distribution
> qf(0.01, 2, 7, lower.tail = FALSE)
```

See the on-line help on `RNG` for how random-number generation is done in R.

8.2 Examining the distribution of a set of data

Given a (univariate) set of data we can examine its distribution in a large number of ways. The simplest is to examine the numbers. Two slightly different summaries are given by `summary` and `fivenum` and a display of the numbers by `stem` (a "stem and leaf" plot).

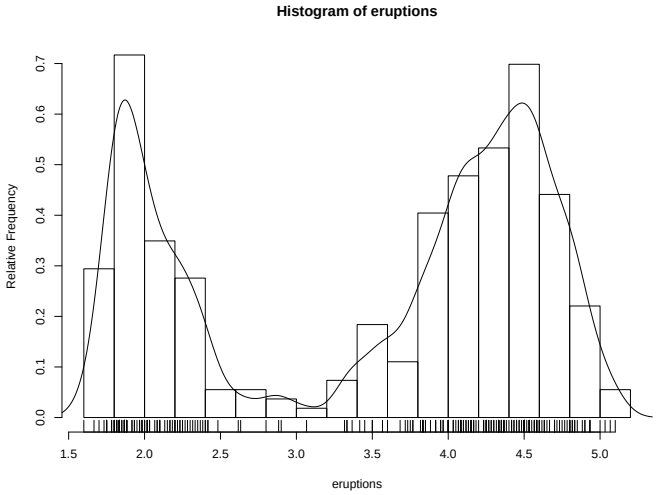
```
> attach(faithful)
> summary(eruptions)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.600   2.163   4.000   3.488   4.454   5.100
> fivenum(eruptions)
[1] 1.6000 2.1585 4.0000 4.4585 5.1000
> stem(eruptions)

  The decimal point is 1 digit(s) to the left of the |

  16 | 070355555588
  18 | 000022233333335577777777888822335777888
  20 | 00002223378800035778
  22 | 0002335578023578
  24 | 00228
  26 | 23
  28 | 080
  30 | 7
  32 | 2337
  34 | 250077
  36 | 0000823577
  38 | 2333335582225577
  40 | 000000335778888800223355557778
  42 | 033355557788002333335555777 78
  44 | 022223355577800000000233333577 78888
  46 | 00002333577000000 23578
  48 | 00000022335800333
  50 | 0370
```

A stem-and-leaf plot is like a histogram, and R has a function `hist` to plot histograms.

```
> hist(eruptions)
## make the bins smaller, make a plot of density
> hist(eruptions, seq(1.6, 5.2, 0.2), prob=TRUE)
> lines(density(eruptions, bw=0.1))
> rug(eruptions) # show the actual data points
```

More elegant density plots can be made by `density`, and we added a line produced by `density` in this example. The bandwidth `bw` was chosen by trial-and-error as the default gives

too much smoothing (it usually does for "interesting" densities). (Better automated methods of bandwidth choice are available, and in this example `bw = "SJ"` gives a good result.)

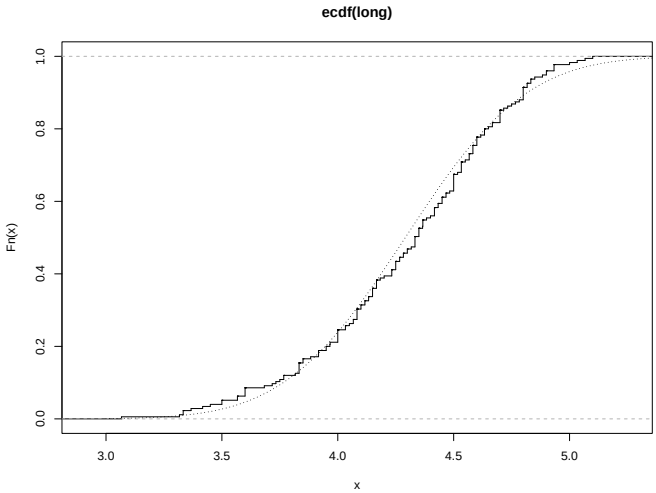


We can plot the empirical cumulative distribution function by using the function `ecdf`.

```
> plot(ecdf(eruptions), do.points=FALSE, verticals=TRUE)
```

This distribution is obviously far from any standard distribution. How about the right-hand mode, say eruptions of longer than 3 minutes? Let us fit a normal distribution and overlay the fitted CDF.

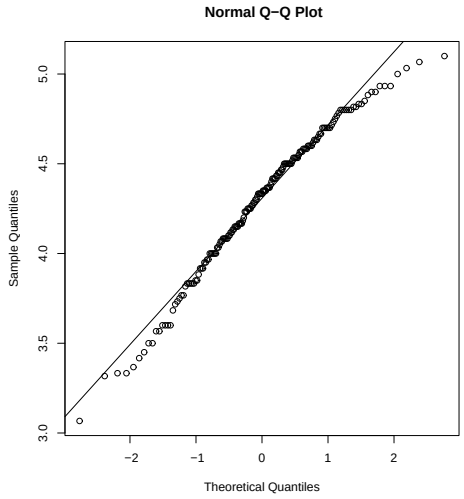
```
> long <- eruptions[eruptions > 3]
> plot(ecdf(long), do.points=FALSE, verticals=TRUE)
> x <- seq(3, 5.4, 0.01)
> lines(x, pnorm(x, mean=mean(long), sd=sqrt(var(long))), lty=3)
```



Quantile-quantile (Q-Q) plots can help us examine this more carefully.

```
par(pty="s")       # arrange for a square figure region
qqnorm(long); qqline(long)
```

which shows a reasonable fit but a shorter right tail than one would expect from a normal distribution. Let us compare this with some simulated data from a t distribution




```
x <- rt(250, df = 5)
qqnorm(x); qqline(x)
```

which will usually (if it is a random sample) show longer tails than expected for a normal. We can make a Q-Q plot against the generating distribution by

```
qqplot(qt(ppoints(250), df = 5), x, xlab = "Q-Q plot for t dsn")
qqline(x)
```

Finally, we might want a more formal test of agreement with normality (or not). R provides the Shapiro-Wilk test

```
> shapiro.test(long)

        Shapiro-Wilk normality test

data:  long
W = 0.9793, p-value = 0.01052
```

and the Kolmogorov-Smirnov test

```
> ks.test(long, "pnorm", mean = mean(long), sd = sqrt(var(long)))

        One-sample Kolmogorov-Smirnov test

data:  long
D = 0.0661, p-value = 0.4284
alternative hypothesis: two.sided
```

(Note that the distribution theory is not valid here as we have estimated the parameters of the normal distribution from the same sample.)

8.3 One- and two-sample tests

So far we have compared a single sample to a normal distribution. A much more common operation is to compare aspects of two samples. Note that in R, all "classical" tests including the ones used below are in package **stats** which is normally loaded.

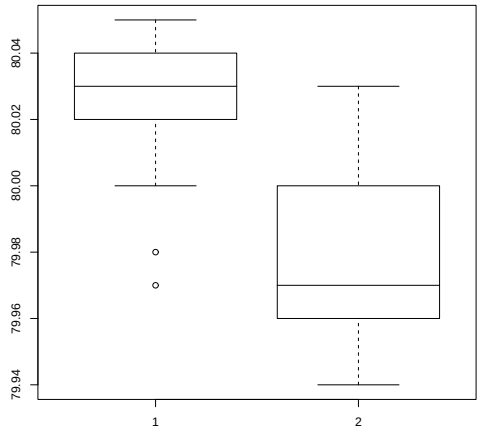
Consider the following sets of data on the latent heat of the fusion of ice (*cal/gm*) from Rice (1995, p.490)

```
Method A: 79.98 80.04 80.02 80.04 80.03 80.03 80.04 79.97
          80.05 80.03 80.02 80.00 80.02
Method B: 80.02 79.94 79.98 79.97 79.97 80.03 79.95 79.97
```

Boxplots provide a simple graphical comparison of the two samples.

```
A <- scan()
79.98 80.04 80.02 80.04 80.03 80.03 80.04 79.97
80.05 80.03 80.02 80.00 80.02

B <- scan()
80.02 79.94 79.98 79.97 79.97 80.03 79.95 79.97

boxplot(A, B)
```

which indicates that the first group tends to give higher results than the second.



To test for the equality of the means of the two examples, we can use an *unpaired* t-test by

```
> t.test(A, B)

        Welch Two Sample t-test

data:  A and B
t = 3.2499, df = 12.027, p-value = 0.00694
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.01385526 0.07018320
sample estimates:
mean of x mean of y
 80.02077  79.97875
```

which does indicate a significant difference, assuming normality. By default the R function does not assume equality of variances in the two samples (in contrast to the similar S-PLUS `t.test` function). We can use the F test to test for equality in the variances, provided that the two samples are from normal populations.

```
> var.test(A, B)

        F test to compare two variances
```

```
data:  A and B
F = 0.5837, num df = 12, denom df =  7, p-value = 0.3938
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1251097 2.1052687
sample estimates:
ratio of variances
         0.5837405
```

which shows no evidence of a significant difference, and so we can use the classical t-test that assumes equality of the variances.

```
> t.test(A, B, var.equal=TRUE)

        Two Sample t-test

data:  A and B
t = 3.4722, df = 19, p-value = 0.002551
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.01669058 0.06734788
sample estimates:
mean of x mean of y
 80.02077  79.97875
```

All these tests assume normality of the two samples. The two-sample Wilcoxon (or Mann-Whitney) test only assumes a common continuous distribution under the null hypothesis.

```
> wilcox.test(A, B)

        Wilcoxon rank sum test with continuity correction

data:  A and B
W = 89, p-value = 0.007497
alternative hypothesis: true location shift is not equal to 0

Warning message:
Cannot compute exact p-value with ties in: wilcox.test(A, B)
```

Note the warning: there are several ties in each sample, which suggests strongly that these data are from a discrete distribution (probably due to rounding).

There are several ways to compare graphically the two samples. We have already seen a pair of boxplots. The following

```
> plot(ecdf(A), do.points=FALSE, verticals=TRUE, xlim=range(A, B))
> plot(ecdf(B), do.points=FALSE, verticals=TRUE, add=TRUE)
```

will show the two empirical CDFs, and `qqplot` will perform a Q-Q plot of the two samples. The Kolmogorov-Smirnov test is of the maximal vertical distance between the two ecdf's, assuming a common continuous distribution:

```
> ks.test(A, B)

        Two-sample Kolmogorov-Smirnov test

data:  A and B
D = 0.5962, p-value = 0.05919
alternative hypothesis: two-sided
```

```
Warning message:
cannot compute correct p-values with ties in: ks.test(A, B)
```

9 Grouping, loops and conditional execution

9.1 Grouped expressions

R is an expression language in the sense that its only command type is a function or expression which returns a result. Even an assignment is an expression whose result is the value assigned, and it may be used wherever any expression may be used; in particular multiple assignments are possible.

Commands may be grouped together in braces, {*expr_1*; ...; *expr_m*}, in which case the value of the group is the result of the last expression in the group evaluated. Since such a group is also an expression it may, for example, be itself included in parentheses and used a part of an even larger expression, and so on.

9.2 Control statements

9.2.1 Conditional execution: `if` statements

The language has available a conditional construction of the form

```
> if (expr_1) expr_2 else expr_3
```

where *expr_1* must evaluate to a single logical value and the result of the entire expression is then evident.

The "short-circuit" operators `&&` and `||` are often used as part of the condition in an `if` statement. Whereas `&` and `|` apply element-wise to vectors, `&&` and `||` apply to vectors of length one, and only evaluate their second argument if necessary.

There is a vectorized version of the `if`/`else` construct, the `ifelse` function. This has the form `ifelse(condition, a, b)` and returns a vector of the length of its longest argument, with elements `a[i]` if `condition[i]` is true, otherwise `b[i]`.

9.2.2 Repetitive execution: `for` loops, `repeat` and `while`

There is also a `for` loop construction which has the form

```
> for (name in expr_1) expr_2
```

where *name* is the loop variable. *expr_1* is a vector expression, (often a sequence like `1:20`), and *expr_2* is often a grouped expression with its sub-expressions written in terms of the dummy *name*. *expr_2* is repeatedly evaluated as *name* ranges through the values in the vector result of *expr_1*.

As an example, suppose `ind` is a vector of class indicators and we wish to produce separate plots of `y` versus `x` within classes. One possibility here is to use `coplot()`,[1] which will produce an array of plots corresponding to each level of the factor. Another way to do this, now putting all plots on the one display, is as follows:

```
> xc <- split(x, ind)
> yc <- split(y, ind)
> for (i in 1:length(yc)) {
    plot(xc[[i]], yc[[i]])
    abline(lsfit(xc[[i]], yc[[i]]))
  }
```

(Note the function `split()` which produces a list of vectors obtained by splitting a larger vector according to the classes specified by a factor. This is a useful function, mostly used in connection with boxplots. See the `help` facility for further details.)

[1] to be discussed later, or use `xyplot` from package **lattice** (`https://CRAN.R-project.org/package=lattice`).

Warning: `for()` loops are used in R code much less often than in compiled languages. Code that takes a 'whole object' view is likely to be both clearer and faster in R.

Other looping facilities include the

```
> repeat expr
```

statement and the

```
> while (condition) expr
```

statement.

The `break` statement can be used to terminate any loop, possibly abnormally. This is the only way to terminate `repeat` loops.

The `next` statement can be used to discontinue one particular cycle and skip to the "next".

Control statements are most often used in connection with *functions* which are discussed in Chapter 10 [Writing your own functions], page 42, and where more examples will emerge.

10 Writing your own functions

As we have seen informally along the way, the R language allows the user to create objects of mode *function*. These are true R functions that are stored in a special internal form and may be used in further expressions and so on. In the process, the language gains enormously in power, convenience and elegance, and learning to write useful functions is one of the main ways to make your use of R comfortable and productive.

It should be emphasized that most of the functions supplied as part of the R system, such as `mean()`, `var()`, `postscript()` and so on, are themselves written in R and thus do not differ materially from user written functions.

A function is defined by an assignment of the form

```
> name <- function(arg_1, arg_2, ...) expression
```

The *expression* is an R expression, (usually a grouped expression), that uses the arguments, *arg_i*, to calculate a value. The value of the expression is the value returned for the function.

A call to the function then usually takes the form `name(expr_1, expr_2, ...)` and may occur anywhere a function call is legitimate.

10.1 Simple examples

As a first example, consider a function to calculate the two sample t-statistic, showing "all the steps". This is an artificial example, of course, since there are other, simpler ways of achieving the same end.

The function is defined as follows:

```
> twosam <- function(y1, y2) {
    n1  <- length(y1); n2  <- length(y2)
    yb1 <- mean(y1);   yb2 <- mean(y2)
    s1  <- var(y1);    s2  <- var(y2)
    s <- ((n1-1)*s1 + (n2-1)*s2)/(n1+n2-2)
    tst <- (yb1 - yb2)/sqrt(s*(1/n1 + 1/n2))
    tst
  }
```

With this function defined, you could perform two sample t-tests using a call such as

```
> tstat <- twosam(data$male, data$female); tstat
```

As a second example, consider a function to emulate directly the MATLAB backslash command, which returns the coefficients of the orthogonal projection of the vector y onto the column space of the matrix, X. (This is ordinarily called the least squares estimate of the regression coefficients.) This would ordinarily be done with the `qr()` function; however this is sometimes a bit tricky to use directly and it pays to have a simple function such as the following to use it safely.

Thus given a n by 1 vector y and an n by p matrix X then $X\ y$ is defined as $(X^TX)^-X^Ty$, where $(X^TX)^-$ is a generalized inverse of $X'X$.

```
> bslash <- function(X, y) {
  X <- qr(X)
  qr.coef(X, y)
}
```

After this object is created it may be used in statements such as

```
> regcoeff <- bslash(Xmat, yvar)
```

and so on.

The classical R function `lsfit()` does this job quite well, and more[1]. It in turn uses the functions `qr()` and `qr.coef()` in the slightly counterintuitive way above to do this part of the calculation. Hence there is probably some value in having just this part isolated in a simple to use function if it is going to be in frequent use. If so, we may wish to make it a matrix binary operator for even more convenient use.

10.2 Defining new binary operators

Had we given the `bslash()` function a different name, namely one of the form

%anything%

it could have been used as a *binary operator* in expressions rather than in function form. Suppose, for example, we choose `!` for the internal character. The function definition would then start as

```
> "%!%" <- function(X, y) { ... }
```

(Note the use of quote marks.) The function could then be used as `X %!% y`. (The backslash symbol itself is not a convenient choice as it presents special problems in this context.)

The matrix multiplication operator, `%*%`, and the outer product matrix operator `%o%` are other examples of binary operators defined in this way.

10.3 Named arguments and defaults

As first noted in Section 2.3 [Generating regular sequences], page 8, if arguments to called functions are given in the "*name=object*" form, they may be given in any order. Furthermore the argument sequence may begin in the unnamed, positional form, and specify named arguments after the positional arguments.

Thus if there is a function `fun1` defined by

```
> fun1 <- function(data, data.frame, graph, limit) {
    [function body omitted]
  }
```

then the function may be invoked in several ways, for example

```
> ans <- fun1(d, df, TRUE, 20)
> ans <- fun1(d, df, graph=TRUE, limit=20)
> ans <- fun1(data=d, limit=20, graph=TRUE, data.frame=df)
```

are all equivalent.

In many cases arguments can be given commonly appropriate default values, in which case they may be omitted altogether from the call when the defaults are appropriate. For example, if `fun1` were defined as

```
> fun1 <- function(data, data.frame, graph=TRUE, limit=20) { ... }
```

it could be called as

```
> ans <- fun1(d, df)
```

which is now equivalent to the three cases above, or as

```
> ans <- fun1(d, df, limit=10)
```

which changes one of the defaults.

It is important to note that defaults may be arbitrary expressions, even involving other arguments to the same function; they are not restricted to be constants as in our simple example here.

[1] See also the methods described in Chapter 11 [Statistical models in R], page 51

10.4 The '...' argument

Another frequent requirement is to allow one function to pass on argument settings to another. For example many graphics functions use the function `par()` and functions like `plot()` allow the user to pass on graphical parameters to `par()` to control the graphical output. (See Section 12.4.1 [The par() function], page 68, for more details on the `par()` function.) This can be done by including an extra argument, literally '`...`', of the function, which may then be passed on. An outline example is given below.

```
fun1 <- function(data, data.frame, graph=TRUE, limit=20, ...) {
  [omitted statements]
  if (graph)
    par(pch="*", ...)
  [more omissions]
}
```

Less frequently, a function will need to refer to components of '`...`'. The expression `list(...)` evaluates all such arguments and returns them in a named list, while `..1`, `..2`, etc. evaluate them one at a time, with '`..n`' returning the n'th unmatched argument.

10.5 Assignments within functions

Note that *any ordinary assignments done within the function are local and temporary and are lost after exit from the function.* Thus the assignment `X <- qr(X)` does not affect the value of the argument in the calling program.

To understand completely the rules governing the scope of R assignments the reader needs to be familiar with the notion of an evaluation *frame*. This is a somewhat advanced, though hardly difficult, topic and is not covered further here.

If global and permanent assignments are intended within a function, then either the "superassignment" operator, `<<-` or the function `assign()` can be used. See the `help` document for details. S-PLUS users should be aware that `<<-` has different semantics in R. These are discussed further in Section 10.7 [Scope], page 46.

10.6 More advanced examples

10.6.1 Efficiency factors in block designs

As a more complete, if a little pedestrian, example of a function, consider finding the efficiency factors for a block design. (Some aspects of this problem have already been discussed in Section 5.3 [Index matrices], page 19.)

A block design is defined by two factors, say `blocks` (`b` levels) and `varieties` (`v` levels). If R and K are the v by v and b by b *replications* and *block size* matrices, respectively, and N is the b by v incidence matrix, then the efficiency factors are defined as the eigenvalues of the matrix

$$E = I_v - R^{-1/2} N^T K^{-1} N R^{-1/2} = I_v - A^T A,$$

where $A = K^{-1/2} N R^{-1/2}$. One way to write the function is given below.

```
> bdeff <- function(blocks, varieties) {
    blocks <- as.factor(blocks)              # minor safety move
    b <- length(levels(blocks))
    varieties <- as.factor(varieties)        # minor safety move
    v <- length(levels(varieties))
    K <- as.vector(table(blocks))            # remove dim attr
    R <- as.vector(table(varieties))         # remove dim attr
```

```
    N <- table(blocks, varieties)
    A <- 1/sqrt(K) * N * rep(1/sqrt(R), rep(b, v))
    sv <- svd(A)
    list(eff=1 - sv$d^2, blockcv=sv$u, varietycv=sv$v)
}
```

It is numerically slightly better to work with the singular value decomposition on this occasion rather than the eigenvalue routines.

The result of the function is a list giving not only the efficiency factors as the first component, but also the block and variety canonical contrasts, since sometimes these give additional useful qualitative information.

10.6.2 Dropping all names in a printed array

For printing purposes with large matrices or arrays, it is often useful to print them in close block form without the array names or numbers. Removing the `dimnames` attribute will not achieve this effect, but rather the array must be given a `dimnames` attribute consisting of empty strings. For example to print a matrix, `X`

```
> temp <- X
> dimnames(temp) <- list(rep("", nrow(X)), rep("", ncol(X)))
> temp; rm(temp)
```

This can be much more conveniently done using a function, `no.dimnames()`, shown below, as a "wrap around" to achieve the same result. It also illustrates how some effective and useful user functions can be quite short.

```
no.dimnames <- function(a) {
  ## Remove all dimension names from an array for compact printing.
  d <- list()
  l <- 0
  for(i in dim(a)) {
    d[[l <- l + 1]] <- rep("", i)
  }
  dimnames(a) <- d
  a
}
```

With this function defined, an array may be printed in close format using

```
> no.dimnames(X)
```

This is particularly useful for large integer arrays, where patterns are the real interest rather than the values.

10.6.3 Recursive numerical integration

Functions may be recursive, and may themselves define functions within themselves. Note, however, that such functions, or indeed variables, are not inherited by called functions in higher evaluation frames as they would be if they were on the search path.

The example below shows a naive way of performing one-dimensional numerical integration. The integrand is evaluated at the end points of the range and in the middle. If the one-panel trapezium rule answer is close enough to the two panel, then the latter is returned as the value. Otherwise the same process is recursively applied to each panel. The result is an adaptive integration process that concentrates function evaluations in regions where the integrand is farthest from linear. There is, however, a heavy overhead, and the function is only competitive with other algorithms when the integrand is both smooth and very difficult to evaluate.

The example is also given partly as a little puzzle in R programming.

```
area <- function(f, a, b, eps = 1.0e-06, lim = 10) {
  fun1 <- function(f, a, b, fa, fb, a0, eps, lim, fun) {
    ## function 'fun1' is only visible inside 'area'
    d <- (a + b)/2
    h <- (b - a)/4
    fd <- f(d)
    a1 <- h * (fa + fd)
    a2 <- h * (fd + fb)
    if(abs(a0 - a1 - a2) < eps || lim == 0)
      return(a1 + a2)
    else {
      return(fun(f, a, d, fa, fd, a1, eps, lim - 1, fun) +
             fun(f, d, b, fd, fb, a2, eps, lim - 1, fun))
    }
  }
  fa <- f(a)
  fb <- f(b)
  a0 <- ((fa + fb) * (b - a))/2
  fun1(f, a, b, fa, fb, a0, eps, lim, fun1)
}
```

10.7 Scope

The discussion in this section is somewhat more technical than in other parts of this document. However, it details one of the major differences between S-PLUS and R.

The symbols which occur in the body of a function can be divided into three classes; formal parameters, local variables and free variables. The formal parameters of a function are those occurring in the argument list of the function. Their values are determined by the process of *binding* the actual function arguments to the formal parameters. Local variables are those whose values are determined by the evaluation of expressions in the body of the functions. Variables which are not formal parameters or local variables are called free variables. Free variables become local variables if they are assigned to. Consider the following function definition.

```
f <- function(x) {
  y <- 2*x
  print(x)
  print(y)
  print(z)
}
```

In this function, `x` is a formal parameter, `y` is a local variable and `z` is a free variable.

In R the free variable bindings are resolved by first looking in the environment in which the function was created. This is called *lexical scope*. First we define a function called `cube`.

```
cube <- function(n) {
  sq <- function() n*n
  n*sq()
}
```

The variable `n` in the function `sq` is not an argument to that function. Therefore it is a free variable and the scoping rules must be used to ascertain the value that is to be associated with it. Under static scope (S-PLUS) the value is that associated with a global variable named `n`. Under lexical scope (R) it is the parameter to the function `cube` since that is the active binding for the variable `n` at the time the function `sq` was defined. The difference between evaluation

in R and evaluation in S-PLUS is that S-PLUS looks for a global variable called `n` while R first looks for a variable called `n` in the environment created when `cube` was invoked.

```
## first evaluation in S
S> cube(2)
Error in sq(): Object "n" not found
Dumped
S> n <- 3
S> cube(2)
[1] 18
## then the same function evaluated in R
R> cube(2)
[1] 8
```

Lexical scope can also be used to give functions *mutable state*. In the following example we show how R can be used to mimic a bank account. A functioning bank account needs to have a balance or total, a function for making withdrawals, a function for making deposits and a function for stating the current balance. We achieve this by creating the three functions within `account` and then returning a list containing them. When `account` is invoked it takes a numerical argument `total` and returns a list containing the three functions. Because these functions are defined in an environment which contains `total`, they will have access to its value.

The special assignment operator, `<<-`, is used to change the value associated with `total`. This operator looks back in enclosing environments for an environment that contains the symbol `total` and when it finds such an environment it replaces the value, in that environment, with the value of right hand side. If the global or top-level environment is reached without finding the symbol `total` then that variable is created and assigned to there. For most users `<<-` creates a global variable and assigns the value of the right hand side to it[2]. Only when `<<-` has been used in a function that was returned as the value of another function will the special behavior described here occur.

```
open.account <- function(total) {
  list(
    deposit = function(amount) {
      if(amount <= 0)
        stop("Deposits must be positive!\n")
      total <<- total + amount
      cat(amount, "deposited.  Your balance is", total, "\n\n")
    },
    withdraw = function(amount) {
      if(amount > total)
        stop("You don't have that much money!\n")
      total <<- total - amount
      cat(amount, "withdrawn.  Your balance is", total, "\n\n")
    },
    balance = function() {
      cat("Your balance is", total, "\n\n")
    }
  )
}

ross <- open.account(100)
```

[2] In some sense this mimics the behavior in S-PLUS since in S-PLUS this operator always creates or assigns to a global variable.

```
robert <- open.account(200)

ross$withdraw(30)
ross$balance()
robert$balance()

ross$deposit(50)
ross$balance()
ross$withdraw(500)
```

10.8 Customizing the environment

Users can customize their environment in several different ways. There is a site initialization file and every directory can have its own special initialization file. Finally, the special functions `.First` and `.Last` can be used.

The location of the site initialization file is taken from the value of the `R_PROFILE` environment variable. If that variable is unset, the file `Rprofile.site` in the R home subdirectory `etc` is used. This file should contain the commands that you want to execute every time R is started under your system. A second, personal, profile file named `.Rprofile`[3] can be placed in any directory. If R is invoked in that directory then that file will be sourced. This file gives individual users control over their workspace and allows for different startup procedures in different working directories. If no `.Rprofile` file is found in the startup directory, then R looks for a `.Rprofile` file in the user's home directory and uses that (if it exists). If the environment variable `R_PROFILE_USER` is set, the file it points to is used instead of the `.Rprofile` files.

Any function named `.First()` in either of the two profile files or in the `.RData` image has a special status. It is automatically performed at the beginning of an R session and may be used to initialize the environment. For example, the definition in the example below alters the prompt to `$` and sets up various other useful things that can then be taken for granted in the rest of the session.

Thus, the sequence in which files are executed is, `Rprofile.site`, the user profile, `.RData` and then `.First()`. A definition in later files will mask definitions in earlier files.

```
> .First <- function() {
  options(prompt="$ ", continue="+\t")  # $ is the prompt
  options(digits=5, length=999)         # custom numbers and printout
  x11()                                 # for graphics
  par(pch = "+")                        # plotting character
  source(file.path(Sys.getenv("HOME"), "R", "mystuff.R"))
                                        # my personal functions
  library(MASS)                         # attach a package
}
```

Similarly a function `.Last()`, if defined, is (normally) executed at the very end of the session. An example is given below.

```
> .Last <- function() {
  graphics.off()                        # a small safety measure.
  cat(paste(date(),"\nAdios\n"))        # Is it time for lunch?
}
```

[3] So it is hidden under UNIX.

10.9 Classes, generic functions and object orientation

The class of an object determines how it will be treated by what are known as *generic* functions. Put the other way round, a generic function performs a task or action on its arguments *specific to the class of the argument itself*. If the argument lacks any `class` attribute, or has a class not catered for specifically by the generic function in question, there is always a *default action* provided.

An example makes things clearer. The class mechanism offers the user the facility of designing and writing generic functions for special purposes. Among the other generic functions are `plot()` for displaying objects graphically, `summary()` for summarizing analyses of various types, and `anova()` for comparing statistical models.

The number of generic functions that can treat a class in a specific way can be quite large. For example, the functions that can accommodate in some fashion objects of class `"data.frame"` include

```
[      [[<-     any     as.matrix
[<-    mean     plot    summary
```

A currently complete list can be got by using the `methods()` function:

```
> methods(class="data.frame")
```

Conversely the number of classes a generic function can handle can also be quite large. For example the `plot()` function has a default method and variants for objects of classes `"data.frame"`, `"density"`, `"factor"`, and more. A complete list can be got again by using the `methods()` function:

```
> methods(plot)
```

For many generic functions the function body is quite short, for example

```
> coef
function (object, ...)
UseMethod("coef")
```

The presence of `UseMethod` indicates this is a generic function. To see what methods are available we can use `methods()`

```
> methods(coef)
[1] coef.aov*          coef.Arima*        coef.default*      coef.listof*
[5] coef.nls*          coef.summary.nls*

   Non-visible functions are asterisked
```

In this example there are six methods, none of which can be seen by typing its name. We can read these by either of

```
> getAnywhere("coef.aov")
A single object matching 'coef.aov' was found
It was found in the following places
  registered S3 method for coef from namespace stats
  namespace:stats
with value

function (object, ...)
{
    z <- object$coef
    z[!is.na(z)]
}
```

```
> getS3method("coef", "aov")
function (object, ...)
{
    z <- object$coef
    z[!is.na(z)]
}
```

A function named *gen.cl* will be invoked by the generic *gen* for class *cl*, so do not name functions in this style unless they are intended to be methods.

The reader is referred to the *R Language Definition* for a more complete discussion of this mechanism.

11 Statistical models in R

This section presumes the reader has some familiarity with statistical methodology, in particular with regression analysis and the analysis of variance. Later we make some rather more ambitious presumptions, namely that something is known about generalized linear models and nonlinear regression.

The requirements for fitting statistical models are sufficiently well defined to make it possible to construct general tools that apply in a broad spectrum of problems.

R provides an interlocking suite of facilities that make fitting statistical models very simple. As we mention in the introduction, the basic output is minimal, and one needs to ask for the details by calling extractor functions.

11.1 Defining statistical models; formulae

The template for a statistical model is a linear regression model with independent, homoscedastic errors

$$y_i = \sum_{j=0}^{p} \beta_j x_{ij} + e_i, \qquad e_i \sim \text{NID}(0, \sigma^2), \qquad i = 1, \ldots, n$$

In matrix terms this would be written

$$y = X\beta + e$$

where the y is the response vector, X is the *model matrix* or *design matrix* and has columns $x_0, x_1, \ldots, x_p$, the determining variables. Very often x_0 will be a column of ones defining an *intercept* term.

Examples

Before giving a formal specification, a few examples may usefully set the picture.

Suppose `y`, `x`, `x0`, `x1`, `x2`, ... are numeric variables, `X` is a matrix and `A`, `B`, `C`, ... are factors. The following formulae on the left side below specify statistical models as described on the right.

`y ~ x`
`y ~ 1 + x` Both imply the same simple linear regression model of y on x. The first has an implicit intercept term, and the second an explicit one.

`y ~ 0 + x`
`y ~ -1 + x`
`y ~ x - 1` Simple linear regression of y on x through the origin (that is, without an intercept term).

`log(y) ~ x1 + x2`
Multiple regression of the transformed variable, $\log(y)$, on $x1$ and $x2$ (with an implicit intercept term).

`y ~ poly(x,2)`
`y ~ 1 + x + I(x^2)`
Polynomial regression of y on x of degree 2. The first form uses orthogonal polynomials, and the second uses explicit powers, as basis.

`y ~ X + poly(x,2)`
Multiple regression y with model matrix consisting of the matrix X as well as polynomial terms in x to degree 2.

`y ~ A` Single classification analysis of variance model of y, with classes determined by A.

`y ~ A + x` Single classification analysis of covariance model of y, with classes determined by A, and with covariate x.

```
y ~ A*B
y ~ A + B + A:B
y ~ B %in% A
y ~ A/B
```
Two factor non-additive model of y on A and B. The first two specify the same crossed classification and the second two specify the same nested classification. In abstract terms all four specify the same model subspace.

```
y ~ (A + B + C)^2
y ~ A*B*C - A:B:C
```
Three factor experiment but with a model containing main effects and two factor interactions only. Both formulae specify the same model.

```
y ~ A * x
y ~ A/x
y ~ A/(1 + x) - 1
```
Separate simple linear regression models of y on x within the levels of A, with different codings. The last form produces explicit estimates of as many different intercepts and slopes as there are levels in A.

```
y ~ A*B + Error(C)
```
An experiment with two treatment factors, A and B, and error strata determined by factor C. For example a split plot experiment, with whole plots (and hence also subplots), determined by factor C.

The operator `~` is used to define a *model formula* in R. The form, for an ordinary linear model, is

response `~` *op_1 term_1 op_2 term_2 op_3 term_3 ...*

where

response is a vector or matrix, (or expression evaluating to a vector or matrix) defining the response variable(s).

op_i is an operator, either `+` or `-`, implying the inclusion or exclusion of a term in the model, (the first is optional).

term_i is either

- a vector or matrix expression, or `1`,
- a factor, or
- a *formula expression* consisting of factors, vectors or matrices connected by *formula operators*.

In all cases each term defines a collection of columns either to be added to or removed from the model matrix. A `1` stands for an intercept column and is by default included in the model matrix unless explicitly removed.

The *formula operators* are similar in effect to the Wilkinson and Rogers notation used by such programs as Glim and Genstat. One inevitable change is that the operator '`.`' becomes '`:`' since the period is a valid name character in R.

The notation is summarized below (based on Chambers & Hastie, 1992, p.29):

Y `~` *M* Y is modeled as M.

M_1 `+` *M_2* Include *M_1* and *M_2*.

`M_1 - M_2` Include *M_1* leaving out terms of *M_2*.

`M_1 : M_2` The tensor product of *M_1* and *M_2*. If both terms are factors, then the "subclasses" factor.

`M_1 %in% M_2`
Similar to `M_1:M_2`, but with a different coding.

`M_1 * M_2` `M_1 + M_2 + M_1:M_2`.

`M_1 / M_2` `M_1 + M_2 %in% M_1`.

`M^n` All terms in *M* together with "interactions" up to order *n*

`I(M)` Insulate *M*. Inside *M* all operators have their normal arithmetic meaning, and that term appears in the model matrix.

Note that inside the parentheses that usually enclose function arguments all operators have their normal arithmetic meaning. The function `I()` is an identity function used to allow terms in model formulae to be defined using arithmetic operators.

Note particularly that the model formulae specify the *columns of the model matrix*, the specification of the parameters being implicit. This is not the case in other contexts, for example in specifying nonlinear models.

11.1.1 Contrasts

We need at least some idea how the model formulae specify the columns of the model matrix. This is easy if we have continuous variables, as each provides one column of the model matrix (and the intercept will provide a column of ones if included in the model).

What about a k-level factor `A`? The answer differs for unordered and ordered factors. For *unordered* factors $k-1$ columns are generated for the indicators of the second, ..., kth levels of the factor. (Thus the implicit parameterization is to contrast the response at each level with that at the first.) For *ordered* factors the $k-1$ columns are the orthogonal polynomials on $1, \ldots, k$, omitting the constant term.

Although the answer is already complicated, it is not the whole story. First, if the intercept is omitted in a model that contains a factor term, the first such term is encoded into k columns giving the indicators for all the levels. Second, the whole behavior can be changed by the `options` setting for `contrasts`. The default setting in R is

```
options(contrasts = c("contr.treatment", "contr.poly"))
```

The main reason for mentioning this is that R and S have different defaults for unordered factors, S using Helmert contrasts. So if you need to compare your results to those of a textbook or paper which used S-PLUS, you will need to set

```
options(contrasts = c("contr.helmert", "contr.poly"))
```

This is a deliberate difference, as treatment contrasts (R's default) are thought easier for newcomers to interpret.

We have still not finished, as the contrast scheme to be used can be set for each term in the model using the functions `contrasts` and `C`.

We have not yet considered interaction terms: these generate the products of the columns introduced for their component terms.

Although the details are complicated, model formulae in R will normally generate the models that an expert statistician would expect, provided that marginality is preserved. Fitting, for example, a model with an interaction but not the corresponding main effects will in general lead to surprising results, and is for experts only.

11.2 Linear models

The basic function for fitting ordinary multiple models is `lm()`, and a streamlined version of the call is as follows:

```
> fitted.model <- lm(formula, data = data.frame)
```

For example

```
> fm2 <- lm(y ~ x1 + x2, data = production)
```

would fit a multiple regression model of y on $x1$ and $x2$ (with implicit intercept term).

The important (but technically optional) parameter `data = production` specifies that any variables needed to construct the model should come first from the `production` *data frame. This is the case regardless of whether data frame* `production` *has been attached on the search path or not.*

11.3 Generic functions for extracting model information

The value of `lm()` is a fitted model object; technically a list of results of class `"lm"`. Information about the fitted model can then be displayed, extracted, plotted and so on by using generic functions that orient themselves to objects of class `"lm"`. These include

```
add1     deviance   formula     predict  step
alias    drop1      kappa       print    summary
anova    effects    labels      proj     vcov
coef     family     plot        residuals
```

A brief description of the most commonly used ones is given below.

`anova(`*object_1*`,` *object_2*`)`
: Compare a submodel with an outer model and produce an analysis of variance table.

`coef(`*object*`)`
: Extract the regression coefficient (matrix).

  Long form: `coefficients(`*object*`)`.

`deviance(`*object*`)`
: Residual sum of squares, weighted if appropriate.

`formula(`*object*`)`
: Extract the model formula.

`plot(`*object*`)`
: Produce four plots, showing residuals, fitted values and some diagnostics.

`predict(`*object*`, newdata=`*data.frame*`)`
: The data frame supplied must have variables specified with the same labels as the original. The value is a vector or matrix of predicted values corresponding to the determining variable values in *data.frame*.

`print(`*object*`)`
: Print a concise version of the object. Most often used implicitly.

`residuals(`*object*`)`
: Extract the (matrix of) residuals, weighted as appropriate.

  Short form: `resid(`*object*`)`.

`step(`*object*`)`
: Select a suitable model by adding or dropping terms and preserving hierarchies. The model with the smallest value of AIC (Akaike's An Information Criterion) discovered in the stepwise search is returned.

`summary(`*object*`)`
: Print a comprehensive summary of the results of the regression analysis.

`vcov(`*object*`)`
: Returns the variance-covariance matrix of the main parameters of a fitted model object.

11.4 Analysis of variance and model comparison

The model fitting function `aov(`*formula*`, data=`*data.frame*`)` operates at the simplest level in a very similar way to the function `lm()`, and most of the generic functions listed in the table in Section 11.3 [Generic functions for extracting model information], page 54 apply.

It should be noted that in addition `aov()` allows an analysis of models with multiple error strata such as split plot experiments, or balanced incomplete block designs with recovery of inter-block information. The model formula

```
response ~ mean.formula + Error(strata.formula)
```

specifies a multi-stratum experiment with error strata defined by the *strata.formula*. In the simplest case, *strata.formula* is simply a factor, when it defines a two strata experiment, namely between and within the levels of the factor.

For example, with all determining variables factors, a model formula such as that in:

```
> fm <- aov(yield ~ v + n*p*k + Error(farms/blocks), data=farm.data)
```

would typically be used to describe an experiment with mean model `v + n*p*k` and three error strata, namely "between farms", "within farms, between blocks" and "within blocks".

11.4.1 ANOVA tables

Note also that the analysis of variance table (or tables) are for a sequence of fitted models. The sums of squares shown are the decrease in the residual sums of squares resulting from an inclusion of *that term* in the model at *that place* in the sequence. Hence only for orthogonal experiments will the order of inclusion be inconsequential.

For multistratum experiments the procedure is first to project the response onto the error strata, again in sequence, and to fit the mean model to each projection. For further details, see Chambers & Hastie (1992).

A more flexible alternative to the default full ANOVA table is to compare two or more models directly using the `anova()` function.

```
> anova(fitted.model.1, fitted.model.2, ...)
```

The display is then an ANOVA table showing the differences between the fitted models when fitted in sequence. The fitted models being compared would usually be an hierarchical sequence, of course. This does not give different information to the default, but rather makes it easier to comprehend and control.

11.5 Updating fitted models

The `update()` function is largely a convenience function that allows a model to be fitted that differs from one previously fitted usually by just a few additional or removed terms. Its form is

```
> new.model <- update(old.model, new.formula)
```

In the *new.formula* the special name consisting of a period, '`.`', only, can be used to stand for "the corresponding part of the old model formula". For example,

```
> fm05 <- lm(y ~ x1 + x2 + x3 + x4 + x5, data = production)
> fm6  <- update(fm05, . ~ . + x6)
> smf6 <- update(fm6, sqrt(.) ~ .)
```

would fit a five variate multiple regression with variables (presumably) from the data frame `production`, fit an additional model including a sixth regressor variable, and fit a variant on the model where the response had a square root transform applied.

Note especially that if the `data=` argument is specified on the original call to the model fitting function, this information is passed on through the fitted model object to `update()` and its allies.

The name '`.`' can also be used in other contexts, but with slightly different meaning. For example

```
> fmfull <- lm(y ~ . , data = production)
```

would fit a model with response `y` and regressor variables *all other variables in the data frame* `production`.

Other functions for exploring incremental sequences of models are `add1()`, `drop1()` and `step()`. The names of these give a good clue to their purpose, but for full details see the on-line help.

11.6 Generalized linear models

Generalized linear modeling is a development of linear models to accommodate both non-normal response distributions and transformations to linearity in a clean and straightforward way. A generalized linear model may be described in terms of the following sequence of assumptions:

- There is a response, y, of interest and stimulus variables x_1, x_2, ..., whose values influence the distribution of the response.
- The stimulus variables influence the distribution of y through *a single linear function, only.* This linear function is called the *linear predictor*, and is usually written

$$\eta = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p,$$

  hence x_i has no influence on the distribution of y if and only if $\beta_i = 0$.
- The distribution of y is of the form

$$f_Y(y; \mu, \varphi) = \exp\left[\frac{A}{\varphi}\left\{y\lambda(\mu) - \gamma\left(\lambda(\mu)\right)\right\} + \tau(y, \varphi)\right]$$

  where φ is a *scale parameter* (possibly known), and is constant for all observations, A represents a prior weight, assumed known but possibly varying with the observations, and μ is the mean of y. So it is assumed that the distribution of y is determined by its mean and possibly a scale parameter as well.
- The mean, μ, is a smooth invertible function of the linear predictor:

$$\mu = m(\eta), \qquad \eta = m^{-1}(\mu) = \ell(\mu)$$

  and this inverse function, $\ell()$, is called the *link function.*

These assumptions are loose enough to encompass a wide class of models useful in statistical practice, but tight enough to allow the development of a unified methodology of estimation and inference, at least approximately. The reader is referred to any of the current reference works on the subject for full details, such as McCullagh & Nelder (1989) or Dobson (1990).

11.6.1 Families

The class of generalized linear models handled by facilities supplied in R includes *gaussian*, *binomial*, *poisson*, *inverse gaussian* and *gamma* response distributions and also *quasi-likelihood* models where the response distribution is not explicitly specified. In the latter case the *variance function* must be specified as a function of the mean, but in other cases this function is implied by the response distribution.

Each response distribution admits a variety of link functions to connect the mean with the linear predictor. Those automatically available are shown in the following table:

Family name	**Link functions**
`binomial`	`logit`, `probit`, `log`, `cloglog`
`gaussian`	`identity`, `log`, `inverse`
`Gamma`	`identity`, `inverse`, `log`
`inverse.gaussian`	`1/mu^2`, `identity`, `inverse`, `log`
`poisson`	`identity`, `log`, `sqrt`
`quasi`	`logit`, `probit`, `cloglog`, `identity`, `inverse`, `log`, `1/mu^2`, `sqrt`

The combination of a response distribution, a link function and various other pieces of information that are needed to carry out the modeling exercise is called the *family* of the generalized linear model.

11.6.2 The `glm()` function

Since the distribution of the response depends on the stimulus variables through a single linear function *only*, the same mechanism as was used for linear models can still be used to specify the linear part of a generalized model. The family has to be specified in a different way.

The R function to fit a generalized linear model is `glm()` which uses the form

```
> fitted.model <- glm(formula, family=family.generator, data=data.frame)
```

The only new feature is the *family.generator*, which is the instrument by which the family is described. It is the name of a function that generates a list of functions and expressions that together define and control the model and estimation process. Although this may seem a little complicated at first sight, its use is quite simple.

The names of the standard, supplied family generators are given under "Family Name" in the table in Section 11.6.1 [Families], page 57. Where there is a choice of links, the name of the link may also be supplied with the family name, in parentheses as a parameter. In the case of the `quasi` family, the variance function may also be specified in this way.

Some examples make the process clear.

The `gaussian` family

A call such as

```
> fm <- glm(y ~ x1 + x2, family = gaussian, data = sales)
```

achieves the same result as

```
> fm <- lm(y ~ x1+x2, data=sales)
```

but much less efficiently. Note how the gaussian family is not automatically provided with a choice of links, so no parameter is allowed. If a problem requires a gaussian family with a nonstandard link, this can usually be achieved through the `quasi` family, as we shall see later.

The `binomial` family

Consider a small, artificial example, from Silvey (1970).

On the Aegean island of Kalythos the male inhabitants suffer from a congenital eye disease, the effects of which become more marked with increasing age. Samples of islander males of various ages were tested for blindness and the results recorded. The data is shown below:

Age:	20	35	45	55	70
No. tested:	50	50	50	50	50
No. blind:	6	17	26	37	44

The problem we consider is to fit both logistic and probit models to this data, and to estimate for each model the LD50, that is the age at which the chance of blindness for a male inhabitant is 50%.

If y is the number of blind at age x and n the number tested, both models have the form

$$y \sim \mathrm{B}(n, F(\beta_0 + \beta_1 x))$$

where for the probit case, $F(z) = \Phi(z)$ is the standard normal distribution function, and in the logit case (the default), $F(z) = e^z/(1 + e^z)$. In both cases the LD50 is

$$\mathrm{LD50} = -\beta_0/\beta_1$$

that is, the point at which the argument of the distribution function is zero.

The first step is to set the data up as a data frame

```
> kalythos <- data.frame(x = c(20,35,45,55,70), n = rep(50,5),
                         y = c(6,17,26,37,44))
```

To fit a binomial model using `glm()` there are three possibilities for the response:

- If the response is a *vector* it is assumed to hold *binary* data, and so must be a 0/1 vector.
- If the response is a *two-column matrix* it is assumed that the first column holds the number of successes for the trial and the second holds the number of failures.
- If the response is a *factor*, its first level is taken as failure (0) and all other levels as 'success' (1).

Here we need the second of these conventions, so we add a matrix to our data frame:

```
> kalythos$Ymat <- cbind(kalythos$y, kalythos$n - kalythos$y)
```

To fit the models we use

```
> fmp <- glm(Ymat ~ x, family = binomial(link=probit), data = kalythos)
> fml <- glm(Ymat ~ x, family = binomial, data = kalythos)
```

Since the logit link is the default the parameter may be omitted on the second call. To see the results of each fit we could use

```
> summary(fmp)
> summary(fml)
```

Both models fit (all too) well. To find the LD50 estimate we can use a simple function:

```
> ld50 <- function(b) -b[1]/b[2]
> ldp <- ld50(coef(fmp)); ldl <- ld50(coef(fml)); c(ldp, ldl)
```

The actual estimates from this data are 43.663 years and 43.601 years respectively.

Poisson models

With the Poisson family the default link is the `log`, and in practice the major use of this family is to fit surrogate Poisson log-linear models to frequency data, whose actual distribution is often multinomial. This is a large and important subject we will not discuss further here. It even forms a major part of the use of non-gaussian generalized models overall.

Occasionally genuinely Poisson data arises in practice and in the past it was often analyzed as gaussian data after either a log or a square-root transformation. As a graceful alternative to the latter, a Poisson generalized linear model may be fitted as in the following example:

```
> fmod <- glm(y ~ A + B + x, family = poisson(link=sqrt),
              data = worm.counts)
```

Quasi-likelihood models

For all families the variance of the response will depend on the mean and will have the scale parameter as a multiplier. The form of dependence of the variance on the mean is a characteristic of the response distribution; for example for the poisson distribution $\mathrm{Var}[y] = \mu$.

For quasi-likelihood estimation and inference the precise response distribution is not specified, but rather only a link function and the form of the variance function as it depends on the mean. Since quasi-likelihood estimation uses formally identical techniques to those for the gaussian distribution, this family provides a way of fitting gaussian models with non-standard link functions or variance functions, incidentally.

For example, consider fitting the non-linear regression

$$y = \frac{\theta_1 z_1}{z_2 - \theta_2} + e$$

which may be written alternatively as

$$y = \frac{1}{\beta_1 x_1 + \beta_2 x_2} + e$$

where $x_1 = z_2/z_1$, $x_2 = -1/z_1$, $\beta_1 = 1/\theta_1$ and $\beta_2 = \theta_2/\theta_1$. Supposing a suitable data frame to be set up we could fit this non-linear regression as

```
> nlfit <- glm(y ~ x1 + x2 - 1,
               family = quasi(link=inverse, variance=constant),
               data = biochem)
```

The reader is referred to the manual and the help document for further information, as needed.

11.7 Nonlinear least squares and maximum likelihood models

Certain forms of nonlinear model can be fitted by Generalized Linear Models (`glm()`). But in the majority of cases we have to approach the nonlinear curve fitting problem as one of nonlinear optimization. R's nonlinear optimization routines are `optim()`, `nlm()` and `nlminb()`, which provide the functionality (and more) of S-PLUS's `ms()` and `nlminb()`. We seek the parameter values that minimize some index of lack-of-fit, and they do this by trying out various parameter values iteratively. Unlike linear regression for example, there is no guarantee that the procedure will converge on satisfactory estimates. All the methods require initial guesses about what parameter values to try, and convergence may depend critically upon the quality of the starting values.

11.7.1 Least squares

One way to fit a nonlinear model is by minimizing the sum of the squared errors (SSE) or residuals. This method makes sense if the observed errors could have plausibly arisen from a normal distribution.

Here is an example from Bates & Watts (1988), page 51. The data are:

```
> x <- c(0.02, 0.02, 0.06, 0.06, 0.11, 0.11, 0.22, 0.22, 0.56, 0.56,
         1.10, 1.10)
> y <- c(76, 47, 97, 107, 123, 139, 159, 152, 191, 201, 207, 200)
```

The fit criterion to be minimized is:

```
> fn <- function(p) sum((y - (p[1] * x)/(p[2] + x))^2)
```

In order to do the fit we need initial estimates of the parameters. One way to find sensible starting values is to plot the data, guess some parameter values, and superimpose the model curve using those values.

```
> plot(x, y)
> xfit <- seq(.02, 1.1, .05)
> yfit <- 200 * xfit/(0.1 + xfit)
> lines(spline(xfit, yfit))
```

We could do better, but these starting values of 200 and 0.1 seem adequate. Now do the fit:

```
> out <- nlm(fn, p = c(200, 0.1), hessian = TRUE)
```

After the fitting, `out$minimum` is the SSE, and `out$estimate` are the least squares estimates of the parameters. To obtain the approximate standard errors (SE) of the estimates we do:

```
> sqrt(diag(2*out$minimum/(length(y) - 2) * solve(out$hessian)))
```

The 2 which is subtracted in the line above represents the number of parameters. A 95% confidence interval would be the parameter estimate $\pm$ 1.96 SE. We can superimpose the least squares fit on a new plot:

```
> plot(x, y)
> xfit <- seq(.02, 1.1, .05)
> yfit <- 212.68384222 * xfit/(0.06412146 + xfit)
> lines(spline(xfit, yfit))
```

The standard package **stats** provides much more extensive facilities for fitting non-linear models by least squares. The model we have just fitted is the Michaelis-Menten model, so we can use

```
> df <- data.frame(x=x, y=y)
> fit <- nls(y ~ SSmicmen(x, Vm, K), df)
> fit
Nonlinear regression model
  model:  y ~ SSmicmen(x, Vm, K)
   data:  df
         Vm            K
212.68370711   0.06412123
 residual sum-of-squares:  1195.449
> summary(fit)

Formula: y ~ SSmicmen(x, Vm, K)

Parameters:
    Estimate Std. Error t value Pr(>|t|)
Vm 2.127e+02  6.947e+00  30.615 3.24e-11
K  6.412e-02  8.281e-03   7.743 1.57e-05

Residual standard error: 10.93 on 10 degrees of freedom

Correlation of Parameter Estimates:
      Vm
K 0.7651
```

11.7.2 Maximum likelihood

Maximum likelihood is a method of nonlinear model fitting that applies even if the errors are not normal. The method finds the parameter values which maximize the log likelihood, or

equivalently which minimize the negative log-likelihood. Here is an example from Dobson (1990), pp. 108–111. This example fits a logistic model to dose-response data, which clearly could also be fit by `glm()`. The data are:

```
> x <- c(1.6907, 1.7242, 1.7552, 1.7842, 1.8113,
         1.8369, 1.8610, 1.8839)
> y <- c( 6, 13, 18, 28, 52, 53, 61, 60)
> n <- c(59, 60, 62, 56, 63, 59, 62, 60)
```

The negative log-likelihood to minimize is:

```
> fn <- function(p)
   sum( - (y*(p[1]+p[2]*x) - n*log(1+exp(p[1]+p[2]*x))
           + log(choose(n, y)) ))
```

We pick sensible starting values and do the fit:

```
> out <- nlm(fn, p = c(-50,20), hessian = TRUE)
```

After the fitting, `out$minimum` is the negative log-likelihood, and `out$estimate` are the maximum likelihood estimates of the parameters. To obtain the approximate SEs of the estimates we do:

```
> sqrt(diag(solve(out$hessian)))
```

A 95% confidence interval would be the parameter estimate ± 1.96 SE.

11.8 Some non-standard models

We conclude this chapter with just a brief mention of some of the other facilities available in R for special regression and data analysis problems.

- **Mixed models.** The recommended **nlme** (`https://CRAN.R-project.org/package=nlme`) package provides functions `lme()` and `nlme()` for linear and non-linear mixed-effects models, that is linear and non-linear regressions in which some of the coefficients correspond to random effects. These functions make heavy use of formulae to specify the models.
- **Local approximating regressions.** The `loess()` function fits a nonparametric regression by using a locally weighted regression. Such regressions are useful for highlighting a trend in messy data or for data reduction to give some insight into a large data set.

  Function `loess` is in the standard package **stats**, together with code for projection pursuit regression.
- **Robust regression.** There are several functions available for fitting regression models in a way resistant to the influence of extreme outliers in the data. Function `lqs` in the recommended package **MASS** (`https://CRAN.R-project.org/package=MASS`) provides state-of-art algorithms for highly-resistant fits. Less resistant but statistically more efficient methods are available in packages, for example function `rlm` in package **MASS** (`https://CRAN.R-project.org/package=MASS`).
- **Additive models.** This technique aims to construct a regression function from smooth additive functions of the determining variables, usually one for each determining variable. Functions `avas` and `ace` in package **acepack** (`https://CRAN.R-project.org/package=acepack`) and functions `bruto` and `mars` in package **mda** (`https://CRAN.R-project.org/package=mda`) provide some examples of these techniques in user-contributed packages to R. An extension is **Generalized Additive Models**, implemented in user-contributed packages **gam** (`https://CRAN.R-project.org/package=gam`) and **mgcv** (`https://CRAN.R-project.org/package=mgcv`).
- **Tree-based models.** Rather than seek an explicit global linear model for prediction or interpretation, tree-based models seek to bifurcate the data, recursively, at critical points of the determining variables in order to partition the data ultimately into groups that are

as homogeneous as possible within, and as heterogeneous as possible between. The results often lead to insights that other data analysis methods tend not to yield.

Models are again specified in the ordinary linear model form. The model fitting function is `tree()`, but many other generic functions such as `plot()` and `text()` are well adapted to displaying the results of a tree-based model fit in a graphical way.

Tree models are available in R *via* the user-contributed packages **rpart** (`https://CRAN.R-project.org/package=rpart`) and **tree** (`https://CRAN.R-project.org/package=tree`).

12 Graphical procedures

Graphical facilities are an important and extremely versatile component of the R environment. It is possible to use the facilities to display a wide variety of statistical graphs and also to build entirely new types of graph.

The graphics facilities can be used in both interactive and batch modes, but in most cases, interactive use is more productive. Interactive use is also easy because at startup time R initiates a graphics *device driver* which opens a special *graphics window* for the display of interactive graphics. Although this is done automatically, it may useful to know that the command used is `X11()` under UNIX, `windows()` under Windows and `quartz()` under OS X. A new device can always be opened by `dev.new()`.

Once the device driver is running, R plotting commands can be used to produce a variety of graphical displays and to create entirely new kinds of display.

Plotting commands are divided into three basic groups:

- **High-level** plotting functions create a new plot on the graphics device, possibly with axes, labels, titles and so on.
- **Low-level** plotting functions add more information to an existing plot, such as extra points, lines and labels.
- **Interactive** graphics functions allow you interactively add information to, or extract information from, an existing plot, using a pointing device such as a mouse.

In addition, R maintains a list of *graphical parameters* which can be manipulated to customize your plots.

This manual only describes what are known as 'base' graphics. A separate graphics subsystem in package **grid** coexists with base – it is more powerful but harder to use. There is a recommended package **lattice** (`https://CRAN.R-project.org/package=lattice`) which builds on **grid** and provides ways to produce multi-panel plots akin to those in the *Trellis* system in S.

12.1 High-level plotting commands

High-level plotting functions are designed to generate a complete plot of the data passed as arguments to the function. Where appropriate, axes, labels and titles are automatically generated (unless you request otherwise.) High-level plotting commands always start a new plot, erasing the current plot if necessary.

12.1.1 The `plot()` function

One of the most frequently used plotting functions in R is the `plot()` function. This is a *generic* function: the type of plot produced is dependent on the type or *class* of the first argument.

`plot(x, y)`
`plot(xy)` If x and y are vectors, `plot(x, y)` produces a scatterplot of y against x. The same effect can be produced by supplying one argument (second form) as either a list containing two elements x and y or a two-column matrix.

`plot(x)` If x is a time series, this produces a time-series plot. If x is a numeric vector, it produces a plot of the values in the vector against their index in the vector. If x is a complex vector, it produces a plot of imaginary versus real parts of the vector elements.

`plot(f)`
`plot(f, y)`
f is a factor object, y is a numeric vector. The first form generates a bar plot of f; the second form produces boxplots of y for each level of f.

`plot(`*df*`)`
`plot(~ `*expr*`)`
`plot(`*y* `~ `*expr*`)`
: *df* is a data frame, *y* is any object, *expr* is a list of object names separated by '`+`' (e.g., `a + b + c`). The first two forms produce distributional plots of the variables in a data frame (first form) or of a number of named objects (second form). The third form plots *y* against every object named in *expr*.

12.1.2 Displaying multivariate data

R provides two very useful functions for representing multivariate data. If `X` is a numeric matrix or data frame, the command

```
> pairs(X)
```

produces a pairwise scatterplot matrix of the variables defined by the columns of `X`, that is, every column of `X` is plotted against every other column of `X` and the resulting $n(n-1)$ plots are arranged in a matrix with plot scales constant over the rows and columns of the matrix.

When three or four variables are involved a *coplot* may be more enlightening. If `a` and `b` are numeric vectors and `c` is a numeric vector or factor object (all of the same length), then the command

```
> coplot(a ~ b | c)
```

produces a number of scatterplots of `a` against `b` for given values of `c`. If `c` is a factor, this simply means that `a` is plotted against `b` for every level of `c`. When `c` is numeric, it is divided into a number of *conditioning intervals* and for each interval `a` is plotted against `b` for values of `c` within the interval. The number and position of intervals can be controlled with `given.values=` argument to `coplot()`—the function `co.intervals()` is useful for selecting intervals. You can also use two *given* variables with a command like

```
> coplot(a ~ b | c + d)
```

which produces scatterplots of `a` against `b` for every joint conditioning interval of `c` and `d`.

The `coplot()` and `pairs()` function both take an argument `panel=` which can be used to customize the type of plot which appears in each panel. The default is `points()` to produce a scatterplot but by supplying some other low-level graphics function of two vectors `x` and `y` as the value of `panel=` you can produce any type of plot you wish. An example panel function useful for coplots is `panel.smooth()`.

12.1.3 Display graphics

Other high-level graphics functions produce different types of plots. Some examples are:

`qqnorm(x)`
`qqline(x)`
`qqplot(x, y)`
: Distribution-comparison plots. The first form plots the numeric vector `x` against the expected Normal order scores (a normal scores plot) and the second adds a straight line to such a plot by drawing a line through the distribution and data quartiles. The third form plots the quantiles of `x` against those of `y` to compare their respective distributions.

`hist(x)`
`hist(x, nclass=`*n*`)`
`hist(x, breaks=`*b*`, ...)`
: Produces a histogram of the numeric vector `x`. A sensible number of classes is usually chosen, but a recommendation can be given with the `nclass=` argument. Alternatively, the breakpoints can be specified exactly with the `breaks=` argument.

If the `probability=TRUE` argument is given, the bars represent relative frequencies divided by bin width instead of counts.

`dotchart(x, ...)`
: Constructs a dotchart of the data in `x`. In a dotchart the y-axis gives a labelling of the data in `x` and the x-axis gives its value. For example it allows easy visual selection of all data entries with values lying in specified ranges.

`image(x, y, z, ...)`
`contour(x, y, z, ...)`
`persp(x, y, z, ...)`
: Plots of three variables. The `image` plot draws a grid of rectangles using different colours to represent the value of `z`, the `contour` plot draws contour lines to represent the value of `z`, and the `persp` plot draws a 3D surface.

12.1.4 Arguments to high-level plotting functions

There are a number of arguments which may be passed to high-level graphics functions, as follows:

`add=TRUE` Forces the function to act as a low-level graphics function, superimposing the plot on the current plot (some functions only).

`axes=FALSE`
: Suppresses generation of axes—useful for adding your own custom axes with the `axis()` function. The default, `axes=TRUE`, means include axes.

`log="x"`
`log="y"`
`log="xy"` Causes the x, y or both axes to be logarithmic. This will work for many, but not all, types of plot.

`type=` The `type=` argument controls the type of plot produced, as follows:

- `type="p"` Plot individual points (the default)
- `type="l"` Plot lines
- `type="b"` Plot points connected by lines (*both*)
- `type="o"` Plot points overlaid by lines
- `type="h"` Plot vertical lines from points to the zero axis (*high-density*)
- `type="s"`
  `type="S"` Step-function plots. In the first form, the top of the vertical defines the point; in the second, the bottom.
- `type="n"` No plotting at all. However axes are still drawn (by default) and the coordinate system is set up according to the data. Ideal for creating plots with subsequent low-level graphics functions.

`xlab=`*`string`*
`ylab=`*`string`*
: Axis labels for the x and y axes. Use these arguments to change the default labels, usually the names of the objects used in the call to the high-level plotting function.

`main=`*`string`*
: Figure title, placed at the top of the plot in a large font.

`sub=`*`string`*
: Sub-title, placed just below the x-axis in a smaller font.

12.2 Low-level plotting commands

Sometimes the high-level plotting functions don't produce exactly the kind of plot you desire. In this case, low-level plotting commands can be used to add extra information (such as points, lines or text) to the current plot.

Some of the more useful low-level plotting functions are:

`points(x, y)`
`lines(x, y)`
: Adds points or connected lines to the current plot. `plot()`'s `type=` argument can also be passed to these functions (and defaults to `"p"` for `points()` and `"l"` for `lines()`.)

`text(x, y, labels, ...)`
: Add text to a plot at points given by `x, y`. Normally `labels` is an integer or character vector in which case `labels[i]` is plotted at point `(x[i], y[i])`. The default is `1:length(x)`.

Note: This function is often used in the sequence

```
> plot(x, y, type="n"); text(x, y, names)
```

The graphics parameter `type="n"` suppresses the points but sets up the axes, and the `text()` function supplies special characters, as specified by the character vector `names` for the points.

`abline(a, b)`
`abline(h=y)`
`abline(v=x)`
`abline(lm.obj)`
: Adds a line of slope `b` and intercept `a` to the current plot. `h=y` may be used to specify y-coordinates for the heights of horizontal lines to go across a plot, and `v=x` similarly for the x-coordinates for vertical lines. Also *lm.obj* may be list with a `coefficients` component of length 2 (such as the result of model-fitting functions,) which are taken as an intercept and slope, in that order.

`polygon(x, y, ...)`
: Draws a polygon defined by the ordered vertices in (`x`, `y`) and (optionally) shade it in with hatch lines, or fill it if the graphics device allows the filling of figures.

`legend(x, y, legend, ...)`
: Adds a legend to the current plot at the specified position. Plotting characters, line styles, colors etc., are identified with the labels in the character vector `legend`. At least one other argument v (a vector the same length as `legend`) with the corresponding values of the plotting unit must also be given, as follows:

  `legend( , fill=v)`
  : Colors for filled boxes

  `legend( , col=v)`
  : Colors in which points or lines will be drawn

  `legend( , lty=v)`
  : Line styles

  `legend( , lwd=v)`
  : Line widths

  `legend( , pch=v)`
  : Plotting characters (character vector)

`title(main, sub)`
: Adds a title `main` to the top of the current plot in a large font and (optionally) a sub-title `sub` at the bottom in a smaller font.

`axis(side, ...)`
: Adds an axis to the current plot on the side given by the first argument (1 to 4, counting clockwise from the bottom.) Other arguments control the positioning of the axis within or beside the plot, and tick positions and labels. Useful for adding custom axes after calling `plot()` with the `axes=FALSE` argument.

Low-level plotting functions usually require some positioning information (e.g., x and y co-ordinates) to determine where to place the new plot elements. Coordinates are given in terms of *user coordinates* which are defined by the previous high-level graphics command and are chosen based on the supplied data.

Where `x` and `y` arguments are required, it is also sufficient to supply a single argument being a list with elements named `x` and `y`. Similarly a matrix with two columns is also valid input. In this way functions such as `locator()` (see below) may be used to specify positions on a plot interactively.

12.2.1 Mathematical annotation

In some cases, it is useful to add mathematical symbols and formulae to a plot. This can be achieved in R by specifying an *expression* rather than a character string in any one of `text`, `mtext`, `axis`, or `title`. For example, the following code draws the formula for the Binomial probability function:

```
> text(x, y, expression(paste(bgroup("(", atop(n, x), ")"), p^x, q^{n-x})))
```

More information, including a full listing of the features available can obtained from within R using the commands:

```
> help(plotmath)
> example(plotmath)
> demo(plotmath)
```

12.2.2 Hershey vector fonts

It is possible to specify Hershey vector fonts for rendering text when using the `text` and `contour` functions. There are three reasons for using the Hershey fonts:

- Hershey fonts can produce better output, especially on a computer screen, for rotated and/or small text.
- Hershey fonts provide certain symbols that may not be available in the standard fonts. In particular, there are zodiac signs, cartographic symbols and astronomical symbols.
- Hershey fonts provide cyrillic and japanese (Kana and Kanji) characters.

More information, including tables of Hershey characters can be obtained from within R using the commands:

```
> help(Hershey)
> demo(Hershey)
> help(Japanese)
> demo(Japanese)
```

12.3 Interacting with graphics

R also provides functions which allow users to extract or add information to a plot using a mouse. The simplest of these is the `locator()` function:

`locator(n, type)`
: Waits for the user to select locations on the current plot using the left mouse button. This continues until `n` (default 512) points have been selected, or another mouse button is pressed. The `type` argument allows for plotting at the selected points and has the same effect as for high-level graphics commands; the default is no plotting. `locator()` returns the locations of the points selected as a list with two components `x` and `y`.

`locator()` is usually called with no arguments. It is particularly useful for interactively selecting positions for graphic elements such as legends or labels when it is difficult to calculate in advance where the graphic should be placed. For example, to place some informative text near an outlying point, the command

```
> text(locator(1), "Outlier", adj=0)
```

may be useful. (`locator()` will be ignored if the current device, such as `postscript` does not support interactive pointing.)

`identify(x, y, labels)`
: Allow the user to highlight any of the points defined by `x` and `y` (using the left mouse button) by plotting the corresponding component of `labels` nearby (or the index number of the point if `labels` is absent). Returns the indices of the selected points when another button is pressed.

Sometimes we want to identify particular *points* on a plot, rather than their positions. For example, we may wish the user to select some observation of interest from a graphical display and then manipulate that observation in some way. Given a number of (x, y) coordinates in two numeric vectors `x` and `y`, we could use the `identify()` function as follows:

```
> plot(x, y)
> identify(x, y)
```

The `identify()` functions performs no plotting itself, but simply allows the user to move the mouse pointer and click the left mouse button near a point. If there is a point near the mouse pointer it will be marked with its index number (that is, its position in the `x`/`y` vectors) plotted nearby. Alternatively, you could use some informative string (such as a case name) as a highlight by using the `labels` argument to `identify()`, or disable marking altogether with the `plot = FALSE` argument. When the process is terminated (see above), `identify()` returns the indices of the selected points; you can use these indices to extract the selected points from the original vectors `x` and `y`.

12.4 Using graphics parameters

When creating graphics, particularly for presentation or publication purposes, R's defaults do not always produce exactly that which is required. You can, however, customize almost every aspect of the display using *graphics parameters*. R maintains a list of a large number of graphics parameters which control things such as line style, colors, figure arrangement and text justification among many others. Every graphics parameter has a name (such as '`col`', which controls colors,) and a value (a color number, for example.)

A separate list of graphics parameters is maintained for each active device, and each device has a default set of parameters when initialized. Graphics parameters can be set in two ways: either permanently, affecting all graphics functions which access the current device; or temporarily, affecting only a single graphics function call.

12.4.1 Permanent changes: The `par()` function

The `par()` function is used to access and modify the list of graphics parameters for the current graphics device.

`par()` Without arguments, returns a list of all graphics parameters and their values for the current device.

`par(c("col", "lty"))`
With a character vector argument, returns only the named graphics parameters (again, as a list.)

`par(col=4, lty=2)`
With named arguments (or a single list argument), sets the values of the named graphics parameters, and returns the original values of the parameters as a list.

Setting graphics parameters with the `par()` function changes the value of the parameters *permanently*, in the sense that all future calls to graphics functions (on the current device) will be affected by the new value. You can think of setting graphics parameters in this way as setting "default" values for the parameters, which will be used by all graphics functions unless an alternative value is given.

Note that calls to `par()` *always* affect the global values of graphics parameters, even when `par()` is called from within a function. This is often undesirable behavior—usually we want to set some graphics parameters, do some plotting, and then restore the original values so as not to affect the user's R session. You can restore the initial values by saving the result of `par()` when making changes, and restoring the initial values when plotting is complete.

```
> oldpar <- par(col=4, lty=2)
  ... plotting commands ...
> par(oldpar)
```

To save and restore *all* settable[1] graphical parameters use

```
> oldpar <- par(no.readonly=TRUE)
  ... plotting commands ...
> par(oldpar)
```

12.4.2 Temporary changes: Arguments to graphics functions

Graphics parameters may also be passed to (almost) any graphics function as named arguments. This has the same effect as passing the arguments to the `par()` function, except that the changes only last for the duration of the function call. For example:

```
> plot(x, y, pch="+")
```

produces a scatterplot using a plus sign as the plotting character, without changing the default plotting character for future plots.

Unfortunately, this is not implemented entirely consistently and it is sometimes necessary to set and reset graphics parameters using `par()`.

12.5 Graphics parameters list

The following sections detail many of the commonly-used graphical parameters. The R help documentation for the `par()` function provides a more concise summary; this is provided as a somewhat more detailed alternative.

Graphics parameters will be presented in the following form:

`name=value`
A description of the parameter's effect. *name* is the name of the parameter, that is, the argument name to use in calls to `par()` or a graphics function. *value* is a typical value you might use when setting the parameter.

Note that `axes` is **not** a graphics parameter but an argument to a few `plot` methods: see `xaxt` and `yaxt`.

[1] Some graphics parameters such as the size of the current device are for information only.

12.5.1 Graphical elements

R plots are made up of points, lines, text and polygons (filled regions.) Graphical parameters exist which control how these *graphical elements* are drawn, as follows:

`pch="+"` Character to be used for plotting points. The default varies with graphics drivers, but it is usually 'o'. Plotted points tend to appear slightly above or below the appropriate position unless you use `"."` as the plotting character, which produces centered points.

`pch=4` When `pch` is given as an integer between 0 and 25 inclusive, a specialized plotting symbol is produced. To see what the symbols are, use the command

```
> legend(locator(1), as.character(0:25), pch = 0:25)
```

Those from 21 to 25 may appear to duplicate earlier symbols, but can be coloured in different ways: see the help on `points` and its examples.

In addition, `pch` can be a character or a number in the range `32:255` representing a character in the current font.

`lty=2` Line types. Alternative line styles are not supported on all graphics devices (and vary on those that do) but line type 1 is always a solid line, line type 0 is always invisible, and line types 2 and onwards are dotted or dashed lines, or some combination of both.

`lwd=2` Line widths. Desired width of lines, in multiples of the "standard" line width. Affects axis lines as well as lines drawn with `lines()`, etc. Not all devices support this, and some have restrictions on the widths that can be used.

`col=2` Colors to be used for points, lines, text, filled regions and images. A number from the current palette (see `?palette`) or a named colour.

`col.axis`
`col.lab`
`col.main`
`col.sub` The color to be used for axis annotation, x and y labels, main and sub-titles, respectively.

`font=2` An integer which specifies which font to use for text. If possible, device drivers arrange so that `1` corresponds to plain text, `2` to bold face, `3` to italic, `4` to bold italic and `5` to a symbol font (which include Greek letters).

`font.axis`
`font.lab`
`font.main`
`font.sub` The font to be used for axis annotation, x and y labels, main and sub-titles, respectively.

`adj=-0.1` Justification of text relative to the plotting position. `0` means left justify, `1` means right justify and `0.5` means to center horizontally about the plotting position. The actual value is the proportion of text that appears to the left of the plotting position, so a value of `-0.1` leaves a gap of 10% of the text width between the text and the plotting position.

`cex=1.5` Character expansion. The value is the desired size of text characters (including plotting characters) relative to the default text size.

`cex.axis`
`cex.lab`
`cex.main`
`cex.sub` The character expansion to be used for axis annotation, x and y labels, main and sub-titles, respectively.

12.5.2 Axes and tick marks

Many of R's high-level plots have axes, and you can construct axes yourself with the low-level `axis()` graphics function. Axes have three main components: the *axis line* (line style controlled by the `lty` graphics parameter), the *tick marks* (which mark off unit divisions along the axis line) and the *tick labels* (which mark the units.) These components can be customized with the following graphics parameters.

`lab=c(5, 7, 12)`
The first two numbers are the desired number of tick intervals on the x and y axes respectively. The third number is the desired length of axis labels, in characters (including the decimal point.) Choosing a too-small value for this parameter may result in all tick labels being rounded to the same number!

`las=1` Orientation of axis labels. `0` means always parallel to axis, `1` means always horizontal, and `2` means always perpendicular to the axis.

`mgp=c(3, 1, 0)`
Positions of axis components. The first component is the distance from the axis label to the axis position, in text lines. The second component is the distance to the tick labels, and the final component is the distance from the axis position to the axis line (usually zero). Positive numbers measure outside the plot region, negative numbers inside.

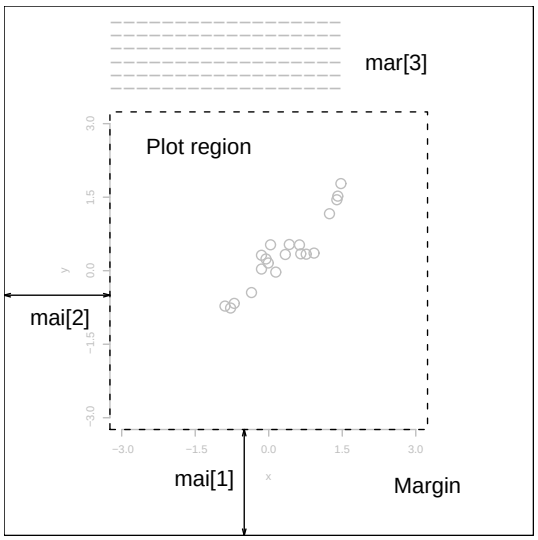
`tck=0.01` Length of tick marks, as a fraction of the size of the plotting region. When `tck` is small (less than 0.5) the tick marks on the x and y axes are forced to be the same size. A value of 1 gives grid lines. Negative values give tick marks outside the plotting region. Use `tck=0.01` and `mgp=c(1,-1.5,0)` for internal tick marks.

`xaxs="r"`
`yaxs="i"` Axis styles for the x and y axes, respectively. With styles `"i"` (internal) and `"r"` (the default) tick marks always fall within the range of the data, however style `"r"` leaves a small amount of space at the edges. (S has other styles not implemented in R.)

12.5.3 Figure margins

A single plot in R is known as a `figure` and comprises a *plot region* surrounded by margins (possibly containing axis labels, titles, etc.) and (usually) bounded by the axes themselves.

A typical figure is



Graphics parameters controlling figure layout include:

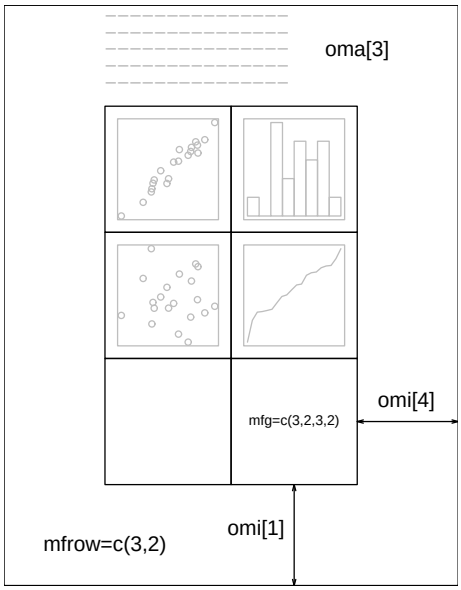
`mai=c(1, 0.5, 0.5, 0)`
: Widths of the bottom, left, top and right margins, respectively, measured in inches.

`mar=c(4, 2, 2, 1)`
: Similar to `mai`, except the measurement unit is text lines.

`mar` and `mai` are equivalent in the sense that setting one changes the value of the other. The default values chosen for this parameter are often too large; the right-hand margin is rarely needed, and neither is the top margin if no title is being used. The bottom and left margins must be large enough to accommodate the axis and tick labels. Furthermore, the default is chosen without regard to the size of the device surface: for example, using the `postscript()` driver with the `height=4` argument will result in a plot which is about 50% margin unless `mar` or `mai` are set explicitly. When multiple figures are in use (see below) the margins are reduced, however this may not be enough when many figures share the same page.

12.5.4 Multiple figure environment

R allows you to create an n by m array of figures on a single page. Each figure has its own margins, and the array of figures is optionally surrounded by an *outer margin*, as shown in the following figure.



The graphical parameters relating to multiple figures are as follows:

`mfcol=c(3, 2)`
`mfrow=c(2, 4)`
: Set the size of a multiple figure array. The first value is the number of rows; the second is the number of columns. The only difference between these two parameters is that setting `mfcol` causes figures to be filled by column; `mfrow` fills by rows.

  The layout in the Figure could have been created by setting `mfrow=c(3,2)`; the figure shows the page after four plots have been drawn.

  Setting either of these can reduce the base size of symbols and text (controlled by `par("cex")` and the pointsize of the device). In a layout with exactly two rows and columns the base size is reduced by a factor of 0.83: if there are three or more of either rows or columns, the reduction factor is 0.66.

`mfg=c(2, 2, 3, 2)`
: Position of the current figure in a multiple figure environment. The first two numbers are the row and column of the current figure; the last two are the number of rows and columns in the multiple figure array. Set this parameter to jump between figures in the array. You can even use different values for the last two numbers than the *true* values for unequally-sized figures on the same page.

`fig=c(4, 9, 1, 4)/10`
: Position of the current figure on the page. Values are the positions of the left, right, bottom and top edges respectively, as a percentage of the page measured from the bottom left corner. The example value would be for a figure in the bottom right of the page. Set this parameter for arbitrary positioning of figures within a page. If you want to add a figure to a current page, use `new=TRUE` as well (unlike S).

`oma=c(2, 0, 3, 0)`
`omi=c(0, 0, 0.8, 0)`
: Size of outer margins. Like `mar` and `mai`, the first measures in text lines and the second in inches, starting with the bottom margin and working clockwise.

Outer margins are particularly useful for page-wise titles, etc. Text can be added to the outer margins with the `mtext()` function with argument `outer=TRUE`. There are no outer margins by default, however, so you must create them explicitly using `oma` or `omi`.

More complicated arrangements of multiple figures can be produced by the `split.screen()` and `layout()` functions, as well as by the **grid** and **lattice** (`https://CRAN.R-project.org/package=lattice`) packages.

12.6 Device drivers

R can generate graphics (of varying levels of quality) on almost any type of display or printing device. Before this can begin, however, R needs to be informed what type of device it is dealing with. This is done by starting a *device driver*. The purpose of a device driver is to convert graphical instructions from R ("draw a line," for example) into a form that the particular device can understand.

Device drivers are started by calling a device driver function. There is one such function for every device driver: type `help(Devices)` for a list of them all. For example, issuing the command

```
> postscript()
```

causes all future graphics output to be sent to the printer in PostScript format. Some commonly-used device drivers are:

`X11()` For use with the X11 window system on Unix-alikes

`windows()`
For use on Windows

`quartz()` For use on OS X

`postscript()`
For printing on PostScript printers, or creating PostScript graphics files.

`pdf()` Produces a PDF file, which can also be included into PDF files.

`png()` Produces a bitmap PNG file. (Not always available: see its help page.)

`jpeg()` Produces a bitmap JPEG file, best used for `image` plots. (Not always available: see its help page.)

When you have finished with a device, be sure to terminate the device driver by issuing the command

```
> dev.off()
```

This ensures that the device finishes cleanly; for example in the case of hardcopy devices this ensures that every page is completed and has been sent to the printer. (This will happen automatically at the normal end of a session.)

12.6.1 PostScript diagrams for typeset documents

By passing the `file` argument to the `postscript()` device driver function, you may store the graphics in PostScript format in a file of your choice. The plot will be in landscape orientation unless the `horizontal=FALSE` argument is given, and you can control the size of the graphic with the `width` and `height` arguments (the plot will be scaled as appropriate to fit these dimensions.) For example, the command

```
> postscript("file.ps", horizontal=FALSE, height=5, pointsize=10)
```

will produce a file containing PostScript code for a figure five inches high, perhaps for inclusion in a document. It is important to note that if the file named in the command already exists,

it will be overwritten. This is the case even if the file was only created earlier in the same R session.

Many usages of PostScript output will be to incorporate the figure in another document. This works best when *encapsulated* PostScript is produced: R always produces conformant output, but only marks the output as such when the `onefile=FALSE` argument is supplied. This unusual notation stems from S-compatibility: it really means that the output will be a single page (which is part of the EPSF specification). Thus to produce a plot for inclusion use something like

```
> postscript("plot1.eps", horizontal=FALSE, onefile=FALSE,
             height=8, width=6, pointsize=10)
```

12.6.2 Multiple graphics devices

In advanced use of R it is often useful to have several graphics devices in use at the same time. Of course only one graphics device can accept graphics commands at any one time, and this is known as the *current device*. When multiple devices are open, they form a numbered sequence with names giving the kind of device at any position.

The main commands used for operating with multiple devices, and their meanings are as follows:

`X11()` [UNIX]

`windows()`
`win.printer()`
`win.metafile()`
[Windows]

`quartz()` [OS X]

`postscript()`
`pdf()`

`png()`

`jpeg()`

`tiff()`

`bitmap()`
`...` Each new call to a device driver function opens a new graphics device, thus extending by one the device list. This device becomes the current device, to which graphics output will be sent.

`dev.list()`
Returns the number and name of all active devices. The device at position 1 on the list is always the *null device* which does not accept graphics commands at all.

`dev.next()`
`dev.prev()`
Returns the number and name of the graphics device next to, or previous to the current device, respectively.

`dev.set(which=`*k*`)`
Can be used to change the current graphics device to the one at position *k* of the device list. Returns the number and label of the device.

`dev.off(`*k*`)`
Terminate the graphics device at point *k* of the device list. For some devices, such as `postscript` devices, this will either print the file immediately or correctly complete the file for later printing, depending on how the device was initiated.

`dev.copy(device, ..., which=k)`
`dev.print(device, ..., which=k)`
: Make a copy of the device k. Here `device` is a device function, such as `postscript`, with extra arguments, if needed, specified by '`...`'. `dev.print` is similar, but the copied device is immediately closed, so that end actions, such as printing hardcopies, are immediately performed.

`graphics.off()`
: Terminate all graphics devices on the list, except the null device.

12.7 Dynamic graphics

R does not have builtin capabilities for dynamic or interactive graphics, e.g. rotating point clouds or to "brushing" (interactively highlighting) points. However, extensive dynamic graphics facilities are available in the system GGobi by Swayne, Cook and Buja available from

`http://www.ggobi.org/`

and these can be accessed from R via the package **rggobi** (`https://CRAN.R-project.org/package=rggobi`), described at `http://www.ggobi.org/rggobi`.

Also, package **rgl** (`https://CRAN.R-project.org/package=rgl`) provides ways to interact with 3D plots, for example of surfaces.

13 Packages

All R functions and datasets are stored in *packages*. Only when a package is loaded are its contents available. This is done both for efficiency (the full list would take more memory and would take longer to search than a subset), and to aid package developers, who are protected from name clashes with other code. The process of developing packages is described in Section "Creating R packages" in *Writing R Extensions*. Here, we will describe them from a user's point of view.

To see which packages are installed at your site, issue the command

```
> library()
```

with no arguments. To load a particular package (e.g., the **boot** (`https://CRAN.R-project.org/package=boot`) package containing functions from Davison & Hinkley (1997)), use a command like

```
> library(boot)
```

Users connected to the Internet can use the `install.packages()` and `update.packages()` functions (available through the `Packages` menu in the Windows and OS X GUIs, see Section "Installing packages" in *R Installation and Administration*) to install and update packages.

To see which packages are currently loaded, use

```
> search()
```

to display the search list. Some packages may be loaded but not available on the search list (see Section 13.3 [Namespaces], page 78): these will be included in the list given by

```
> loadedNamespaces()
```

To see a list of all available help topics in an installed package, use

```
> help.start()
```

to start the HTML help system, and then navigate to the package listing in the `Reference` section.

13.1 Standard packages

The standard (or *base*) packages are considered part of the R source code. They contain the basic functions that allow R to work, and the datasets and standard statistical and graphical functions that are described in this manual. They should be automatically available in any R installation. See Section "R packages" in *R FAQ*, for a complete list.

13.2 Contributed packages and CRAN

There are thousands of contributed packages for R, written by many different authors. Some of these packages implement specialized statistical methods, others give access to data or hardware, and others are designed to complement textbooks. Some (the *recommended* packages) are distributed with every binary distribution of R. Most are available for download from CRAN (`https://CRAN.R-project.org/` and its mirrors) and other repositories such as Bioconductor (`https://www.bioconductor.org/`) and Omegahat (`http://www.omegahat.org/`). The *R FAQ* contains a list of CRAN packages current at the time of release, but the collection of available packages changes very frequently.

13.3 Namespaces

All packages have *namespaces*, and have since R 2.14.0. Namespaces do three things: they allow the package writer to hide functions and data that are meant only for internal use, they prevent functions from breaking when a user (or other package writer) picks a name that clashes with one in the package, and they provide a way to refer to an object within a particular package.

For example, `t()` is the transpose function in R, but users might define their own function named `t`. Namespaces prevent the user's definition from taking precedence, and breaking every function that tries to transpose a matrix.

There are two operators that work with namespaces. The double-colon operator `::` selects definitions from a particular namespace. In the example above, the transpose function will always be available as `base::t`, because it is defined in the `base` package. Only functions that are exported from the package can be retrieved in this way.

The triple-colon operator `:::` may be seen in a few places in R code: it acts like the double-colon operator but also allows access to hidden objects. Users are more likely to use the `getAnywhere()` function, which searches multiple packages.

Packages are often inter-dependent, and loading one may cause others to be automatically loaded. The colon operators described above will also cause automatic loading of the associated package. When packages with namespaces are loaded automatically they are not added to the search list.

14 OS facilities

R has quite extensive facilities to access the OS under which it is running: this allows it to be used as a scripting language and that ability is much used by R itself, for example to install packages.

Because R's own scripts need to work across all platforms, considerable effort has gone into make the scripting facilities as platform-independent as is feasible.

14.1 Files and directories

There are many functions to manipulate files and directories. Here are pointers to some of the more commonly used ones.

To create an (empty) file or directory, use `file.create` or `create.dir`. (These are the analogues of the POSIX utilities `touch` and `mkdir`.) For temporary files and directories in the R session directory see `tempfile`.

Files can be removed by either `file.remove` or `unlink`: the latter can remove directory trees.

For directory listings use `list.files` (also available as `dir`) or `list.dirs`. These can select files using a regular expression: to select by wildcards use `Sys.glob`.

Many types of information on a filepath (including for example if it is a file or directory) can be found by `file.info`.

There are several ways to find out if a file 'exists' (a file can exist on the filesystem and not be visible to the current user). There are functions `file.exists`, `file.access` and `file_test` with various versions of this test: `file_test` is a version of the POSIX `test` command for those familiar with shell scripting.

Function `file.copy` is the R analogue of the POSIX command `cp`.

Choosing files can be done interactively by `file.choose`: the Windows port has the more versatile functions `choose.files` and `choose.dir` and there are similar functions in the **tcltk** package: `tk_choose.files` and `tk_choose.dir`.

Functions `file.show` and `file.edit` will display and edit one or more files in a way appropriate to the R port, using the facilities of a console (such as RGui on Windows or R.app on OS X) if one is in use.

There is some support for *links* in the filesystem: see functions `file.link` and `Sys.readlink`.

14.2 Filepaths

With a few exceptions, R relies on the underlying OS functions to manipulate filepaths. Some aspects of this are allowed to depend on the OS, and do, even down to the version of the OS. There are POSIX standards for how OSes should interpret filepaths and many R users assume POSIX compliance: but Windows does not claim to be compliant and other OSes may be less than completely compliant.

The following are some issues which have been encountered with filepaths.

- POSIX filesystems are case-sensitive, so `foo.png` and `Foo.PNG` are different files. However, the defaults on Windows and OS X are to be case-insensitive, and FAT filesystems (commonly used on removable storage) are not normally case-sensitive (and all filepaths may be mapped to lower case).
- Almost all the Windows' OS services support the use of slash or backslash as the filepath separator, and R converts the known exceptions to the form required by Windows.

- The behaviour of filepaths with a trailing slash is OS-dependent. Such paths are not valid on Windows and should not be expected to work. POSIX-2008 requires such paths to match only directories, but earlier versions allowed them to also match files. So they are best avoided.
- Multiple slashes in filepaths such as `/abc//def` are valid on POSIX filesystems and treated as if there was only one slash. They are *usually* accepted by Windows' OS functions. However, leading double slashes may have a different meaning.
- Windows' UNC filepaths (such as `\\server\dir1\dir2\file` and `\\?\UNC\server\dir1\dir2\file`) are not supported, but they may work in some R functions. POSIX filesystems are allowed to treat a leading double slash specially.
- Windows allows filepaths containing drives and relative to the current directory on a drive, e.g. `d:foo/bar` refers to `d:/a/b/c/foo/bar` if the current directory *on drive* `d:` is `/a/b/c`. It is intended that these work, but the use of absolute paths is safer.

Functions `basename` and `dirname` select parts of a file path: the recommended way to assemble a file path from components is `file.path`. Function `pathexpand` does 'tilde expansion', substituting values for home directories (the current user's, and perhaps those of other users).

On filesystems with links, a single file can be referred to by many filepaths. Function `normalizePath` will find a canonical filepath.

Windows has the concepts of short ('8.3') and long file names: `normalizePath` will return an absolute path using long file names and `shortPathName` will return a version using short names. The latter does not contain spaces and uses backslash as the separator, so is sometimes useful for exporting names from R.

File *permissions* are a related topic. R has support for the POSIX concepts of read/write/execute permission for owner/group/all but this may be only partially supported on the filesystem (so for example on Windows only read-only files (for the account running the R session) are recognized. Access Control Lists (ACLs) are employed on several filesystems, but do not have an agreed standard and R has no facilities to control them. Use `Sys.chmod` to change permissions.

14.3 System commands

Functions `system` and `system2` are used to invoke a system command and optionally collect its output. `system2` is a little more general but its main advantage is that it is easier to write cross-platform code using it.

`system` behaves differently on Windows from other OSes (because the API C call of that name does). Elsewhere it invokes a shell to run the command: the Windows port of R has a function `shell` to do that.

To find out if the OS includes a command, use `Sys.which`, which attempts to do this in a cross-platform way (unfortunately it is not a standard OS service).

Function `shQuote` will quote filepaths as needed for commands in the current OS.

14.4 Compression and Archives

Recent versions of R have extensive facilities to read and write compressed files, often transparently. Reading of files in R is to a vey large extent done by *connections*, and the `file` function which is used to open a connection to a file (or a URL) and is able to identify the compression used from the 'magic' header of the file.

The type of compression which has been supported for longest is `gzip` compression, and that remains a good general compromise. Files compressed by the earlier Unix `compress` utility can also be read, but these are becoming rare. Two other forms of compression, those of the

`bzip2` and `xz` utilities are also available. These generally achieve higher rates of compression (depending on the file, much higher) at the expense of slower decompression and much slower compression.

There is some confusion between `xz` and `lzma` compression (see `https://en.wikipedia.org/wiki/Xz` and `https://en.wikipedia.org/wiki/LZMA`): R can read files compressed by most versions of either.

File archives are single files which contain a collection of files, the most common ones being 'tarballs' and zip files as used to distribute R packages. R can list and unpack both (see functions `untar` and `unzip`) and create both (for `zip` with the help of an external program).

Appendix A A sample session

The following session is intended to introduce to you some features of the R environment by using them. Many features of the system will be unfamiliar and puzzling at first, but this puzzlement will soon disappear.

`Start R appropriately for your platform (see Appendix B [Invoking R], page 85).`
: The R program begins, with a banner.

  (Within R code, the prompt on the left hand side will not be shown to avoid confusion.)

`help.start()`
: Start the HTML interface to on-line help (using a web browser available at your machine). You should briefly explore the features of this facility with the mouse.

  Iconify the help window and move on to the next part.

```
x <- rnorm(50)
y <- rnorm(x)
```
: Generate two pseudo-random normal vectors of x- and y-coordinates.

`plot(x, y)`
: Plot the points in the plane. A graphics window will appear automatically.

`ls()` See which R objects are now in the R workspace.

`rm(x, y)` Remove objects no longer needed. (Clean up).

`x <- 1:20` Make $x = (1, 2, \ldots, 20)$.

`w <- 1 + sqrt(x)/2`
: A 'weight' vector of standard deviations.

```
dummy <- data.frame(x=x, y= x + rnorm(x)*w)
dummy
```
: Make a *data frame* of two columns, x and y, and look at it.

```
fm <- lm(y ~ x, data=dummy)
summary(fm)
```
: Fit a simple linear regression and look at the analysis. With `y` to the left of the tilde, we are modelling y dependent on x.

```
fm1 <- lm(y ~ x, data=dummy, weight=1/w^2)
summary(fm1)
```
: Since we know the standard deviations, we can do a weighted regression.

`attach(dummy)`
: Make the columns in the data frame visible as variables.

`lrf <- lowess(x, y)`
: Make a nonparametric local regression function.

`plot(x, y)`
: Standard point plot.

`lines(x, lrf$y)`
: Add in the local regression.

`abline(0, 1, lty=3)`
: The true regression line: (intercept 0, slope 1).

`abline(coef(fm))`
: Unweighted regression line.

```
abline(coef(fm1), col = "red")
```
Weighted regression line.

```
detach()
```
Remove data frame from the search path.

```
plot(fitted(fm), resid(fm),
     xlab="Fitted values",
     ylab="Residuals",
     main="Residuals vs Fitted")
```
A standard regression diagnostic plot to check for heteroscedasticity. Can you see it?

```
qqnorm(resid(fm), main="Residuals Rankit Plot")
```
A normal scores plot to check for skewness, kurtosis and outliers. (Not very useful here.)

```
rm(fm, fm1, lrf, x, dummy)
```
Clean up again.

The next section will look at data from the classical experiment of Michelson to measure the speed of light. This dataset is available in the `morley` object, but we will read it to illustrate the `read.table` function.

```
filepath <- system.file("data", "morley.tab" , package="datasets")
filepath
```
Get the path to the data file.

```
file.show(filepath)
```
Optional. Look at the file.

```
mm <- read.table(filepath)
mm
```
Read in the Michelson data as a data frame, and look at it. There are five experiments (column `Expt`) and each has 20 runs (column `Run`) and `sl` is the recorded speed of light, suitably coded.

```
mm$Expt <- factor(mm$Expt)
mm$Run <- factor(mm$Run)
```
Change `Expt` and `Run` into factors.

```
attach(mm)
```
Make the data frame visible at position 3 (the default).

```
plot(Expt, Speed, main="Speed of Light Data", xlab="Experiment No.")
```
Compare the five experiments with simple boxplots.

```
fm <- aov(Speed ~ Run + Expt, data=mm)
summary(fm)
```
Analyze as a randomized block, with 'runs' and 'experiments' as factors.

```
fm0 <- update(fm, . ~ . - Run)
anova(fm0, fm)
```
Fit the sub-model omitting 'runs', and compare using a formal analysis of variance.

```
detach()
rm(fm, fm0)
```
Clean up before moving on.

We now look at some more graphical features: contour and image plots.

```
x <- seq(-pi, pi, len=50)
y <- x
```
x is a vector of 50 equally spaced values in $-\pi \le x \le \pi$. y is the same.

```
f <- outer(x, y, function(x, y) cos(y)/(1 + x^2))
```
f is a square matrix, with rows and columns indexed by x and y respectively, of values of the function $\cos(y)/(1 + x^2)$.

```
oldpar <- par(no.readonly = TRUE)
par(pty="s")
```
Save the plotting parameters and set the plotting region to "square".

```
contour(x, y, f)
contour(x, y, f, nlevels=15, add=TRUE)
```
Make a contour map of f; add in more lines for more detail.

```
fa <- (f-t(f))/2
```
`fa` is the "asymmetric part" of f. (`t()` is transpose).

```
contour(x, y, fa, nlevels=15)
```
Make a contour plot, ...

```
par(oldpar)
```
... and restore the old graphics parameters.

```
image(x, y, f)
image(x, y, fa)
```
Make some high density image plots, (of which you can get hardcopies if you wish), ...

```
objects(); rm(x, y, f, fa)
```
... and clean up before moving on.

R can do complex arithmetic, also.

```
th <- seq(-pi, pi, len=100)
z <- exp(1i*th)
```
`1i` is used for the complex number i.

```
par(pty="s")
plot(z, type="l")
```
Plotting complex arguments means plot imaginary versus real parts. This should be a circle.

```
w <- rnorm(100) + rnorm(100)*1i
```
Suppose we want to sample points within the unit circle. One method would be to take complex numbers with standard normal real and imaginary parts ...

```
w <- ifelse(Mod(w) > 1, 1/w, w)
```
... and to map any outside the circle onto their reciprocal.

```
plot(w, xlim=c(-1,1), ylim=c(-1,1), pch="+",xlab="x", ylab="y")
lines(z)
```
All points are inside the unit circle, but the distribution is not uniform.

```
w <- sqrt(runif(100))*exp(2*pi*runif(100)*1i)
plot(w, xlim=c(-1,1), ylim=c(-1,1), pch="+", xlab="x", ylab="y")
lines(z)
```
The second method uses the uniform distribution. The points should now look more evenly spaced over the disc.

```
rm(th, w, z)
```
Clean up again.

```
q()
```
Quit the R program. You will be asked if you want to save the R workspace, and for an exploratory session like this, you probably do not want to save it.

Appendix B Invoking R

Users of R on Windows or OS X should read the OS-specific section first, but command-line use is also supported.

B.1 Invoking R from the command line

When working at a command line on UNIX or Windows, the command '`R`' can be used both for starting the main R program in the form

`R` [*options*] [`<`*infile*] [`>`*outfile*],

or, via the `R CMD` interface, as a wrapper to various R tools (e.g., for processing files in R documentation format or manipulating add-on packages) which are not intended to be called "directly".

At the Windows command-line, `Rterm.exe` is preferred to `R`.

You need to ensure that either the environment variable `TMPDIR` is unset or it points to a valid place to create temporary files and directories.

Most options control what happens at the beginning and at the end of an R session. The startup mechanism is as follows (see also the on-line help for topic '`Startup`' for more information, and the section below for some Windows-specific details).

- Unless `--no-environ` was given, R searches for user and site files to process for setting environment variables. The name of the site file is the one pointed to by the environment variable `R_ENVIRON`; if this is unset, *`R_HOME`*`/etc/Renviron.site` is used (if it exists). The user file is the one pointed to by the environment variable `R_ENVIRON_USER` if this is set; otherwise, files `.Renviron` in the current or in the user's home directory (in that order) are searched for. These files should contain lines of the form '*`name=value`*'. (See `help("Startup")` for a precise description.) Variables you might want to set include `R_PAPERSIZE` (the default paper size), `R_PRINTCMD` (the default print command) and `R_LIBS` (specifies the list of R library trees searched for add-on packages).
- Then R searches for the site-wide startup profile unless the command line option `--no-site-file` was given. The name of this file is taken from the value of the `R_PROFILE` environment variable. If that variable is unset, the default *`R_HOME`*`/etc/Rprofile.site` is used if this exists.
- Then, unless `--no-init-file` was given, R searches for a user profile and sources it. The name of this file is taken from the environment variable `R_PROFILE_USER`; if unset, a file called `.Rprofile` in the current directory or in the user's home directory (in that order) is searched for.
- It also loads a saved workspace from file `.RData` in the current directory if there is one (unless `--no-restore` or `--no-restore-data` was specified).
- Finally, if a function `.First()` exists, it is executed. This function (as well as `.Last()` which is executed at the end of the R session) can be defined in the appropriate startup profiles, or reside in `.RData`.

In addition, there are options for controlling the memory available to the R process (see the on-line help for topic '`Memory`' for more information). Users will not normally need to use these unless they are trying to limit the amount of memory used by R.

R accepts the following command-line options.

`--help`
`-h` Print short help message to standard output and exit successfully.

`--version`
Print version information to standard output and exit successfully.

`--encoding=`*`enc`*
: Specify the encoding to be assumed for input from the console or `stdin`. This needs to be an encoding known to `iconv`: see its help page. (`--encoding` *`enc`* is also accepted.) The input is re-encoded to the locale R is running in and needs to be representable in the latter's encoding (so e.g. you cannot re-encode Greek text in a French locale unless that locale uses the UTF-8 encoding).

`RHOME`
: Print the path to the R "home directory" to standard output and exit successfully. Apart from the front-end shell script and the man page, R installation puts everything (executables, packages, etc.) into this directory.

`--save`
`--no-save`
: Control whether data sets should be saved or not at the end of the R session. If neither is given in an interactive session, the user is asked for the desired behavior when ending the session with *`q()`*; in non-interactive use one of these must be specified or implied by some other option (see below).

`--no-environ`
: Do not read any user file to set environment variables.

`--no-site-file`
: Do not read the site-wide profile at startup.

`--no-init-file`
: Do not read the user's profile at startup.

`--restore`
`--no-restore`
`--no-restore-data`
: Control whether saved images (file `.RData` in the directory where R was started) should be restored at startup or not. The default is to restore. (`--no-restore` implies all the specific `--no-restore-*` options.)

`--no-restore-history`
: Control whether the history file (normally file `.Rhistory` in the directory where R was started, but can be set by the environment variable `R_HISTFILE`) should be restored at startup or not. The default is to restore.

`--no-Rconsole`
: (Windows only) Prevent loading the `Rconsole` file at startup.

`--vanilla`
: Combine `--no-save`, `--no-environ`, `--no-site-file`, `--no-init-file` and `--no-restore`. Under Windows, this also includes `--no-Rconsole`.

`-f` *`file`*
`--file=`*`file`*
: (not `Rgui.exe`) Take input from *file*: '`-`' means `stdin`. Implies `--no-save` unless `--save` has been set. On a Unix-alike, shell metacharacters should be avoided in *file* (but spaces are allowed).

`-e` *`expression`*
: (not `Rgui.exe`) Use *expression* as an input line. One or more `-e` options can be used, but not together with `-f` or `--file`. Implies `--no-save` unless `--save` has been set. (There is a limit of 10,000 bytes on the total length of expressions used in this way. Expressions containing spaces or shell metacharacters will need to be quoted.)

`--no-readline`
: (UNIX only) Turn off command-line editing via **readline**. This is useful when running R from within Emacs using the ESS ("Emacs Speaks Statistics") package. See Appendix C [The command-line editor], page 92, for more information. Command-line editing is enabled for default interactive use (see `--interactive`). This option also affects tilde-expansion: see the help for `path.expand`.

`--min-vsize=`*N*
`--min-nsize=`*N*
: For expert use only: set the initial trigger sizes for garbage collection of vector heap (in bytes) and *cons cells* (number) respectively. Suffix '`M`' specifies megabytes or millions of cells respectively. The defaults are 6Mb and 350k respectively and can also be set by environment variables `R_NSIZE` and `R_VSIZE`.

`--max-ppsize=`*N*
: Specify the maximum size of the pointer protection stack as *N* locations. This defaults to 10000, but can be increased to allow large and complicated calculations to be done. Currently the maximum value accepted is 100000.

`--max-mem-size=`*N*
: (Windows only) Specify a limit for the amount of memory to be used both for R objects and working areas. This is set by default to the smaller of the amount of physical RAM in the machine and for 32-bit R, 1.5Gb[1], and must be between 32Mb and the maximum allowed on that version of Windows.

`--quiet`
`--silent`
`-q`
: Do not print out the initial copyright and welcome messages.

`--slave`
: Make R run as quietly as possible. This option is intended to support programs which use R to compute results for them. It implies `--quiet` and `--no-save`.

`--interactive`
: (UNIX only) Assert that R really is being run interactively even if input has been redirected: use if input is from a FIFO or pipe and fed from an interactive program. (The default is to deduce that R is being run interactively if and only if `stdin` is connected to a terminal or `pty`.) Using `-e`, `-f` or `--file` asserts non-interactive use even if `--interactive` is given.

  Note that this does not turn on command-line editing.

`--ess`
: (Windows only) Set `Rterm` up for use by `R-inferior-mode` in ESS, including asserting interactive use (without the command-line editor) and no buffering of `stdout`.

`--verbose`
: Print more information about progress, and in particular set R's option `verbose` to `TRUE`. R code uses this option to control the printing of diagnostic messages.

`--debugger=`*name*
`-d` *name*
: (UNIX only) Run R through debugger *name*. For most debuggers (the exceptions are `valgrind` and recent versions of `gdb`), further command line options are disregarded, and should instead be given when starting the R executable from inside the debugger.

`--gui=`*type*
`-g` *type*
: (UNIX only) Use *type* as graphical user interface (note that this also includes interactive graphics). Currently, possible values for *type* are '`X11`' (the default) and,

[1] 2.5Gb on versions of Windows that support 3Gb per process and have the support enabled: see the `rw-FAQ` Q2.9; 3.5Gb on most 64-bit versions of Windows.

provided that 'Tcl/Tk' support is available, 'Tk'. (For back-compatibility, 'x11' and 'tk' are accepted.)

`--arch=`*name*
: (UNIX only) Run the specified sub-architecture.

`--args`
: This flag does nothing except cause the rest of the command line to be skipped: this can be useful to retrieve values from it with `commandArgs(TRUE)`.

Note that input and output can be redirected in the usual way (using '<' and '>'), but the line length limit of 4095 bytes still applies. Warning and error messages are sent to the error channel (`stderr`).

The command `R CMD` allows the invocation of various tools which are useful in conjunction with R, but not intended to be called "directly". The general form is

```
R CMD command args
```

where *command* is the name of the tool and *args* the arguments passed on to it.

Currently, the following tools are available.

`BATCH`
: Run R in batch mode. Runs `R --restore --save` with possibly further options (see `?BATCH`).

`COMPILE`
: (UNIX only) Compile C, C++, Fortran ... files for use with R.

`SHLIB`
: Build shared library for dynamic loading.

`INSTALL`
: Install add-on packages.

`REMOVE`
: Remove add-on packages.

`build`
: Build (that is, package) add-on packages.

`check`
: Check add-on packages.

`LINK`
: (UNIX only) Front-end for creating executable programs.

`Rprof`
: Post-process R profiling files.

`Rdconv`
`Rd2txt`
: Convert Rd format to various other formats, including HTML, LaTeX, plain text, and extracting the examples. `Rd2txt` can be used as shorthand for `Rd2conv -t txt`.

`Rd2pdf`
: Convert Rd format to PDF.

`Stangle`
: Extract S/R code from Sweave or other vignette documentation

`Sweave`
: Process Sweave or other vignette documentation

`Rdiff`
: Diff R output ignoring headers etc

`config`
: Obtain configuration information

`javareconf`
: (Unix only) Update the Java configuration variables

`rtags`
: (Unix only) Create Emacs-style tag files from C, R, and Rd files

`open`
: (Windows only) Open a file via Windows' file associations

`texify`
: (Windows only) Process (La)TeX files with R's style files

Use

`R CMD` *`command`* `--help`

to obtain usage information for each of the tools accessible via the `R CMD` interface.

In addition, you can use options `--arch=`, `--no-environ`, `--no-init-file`, `--no-site-file` and `--vanilla` between `R` and `CMD`: these affect any R processes run by the tools. (Here `--vanilla` is equivalent to `--no-environ --no-site-file --no-init-file`.) However, note that `R CMD` does not of itself use any R startup files (in particular, neither user nor site `Renviron` files), and all of the R processes run by these tools (except `BATCH`) use `--no-restore`. Most use `--vanilla` and so invoke no R startup files: the current exceptions are `INSTALL`, `REMOVE`, `Sweave` and `SHLIB` (which uses `--no-site-file --no-init-file`).

`R CMD` *`cmd args`*

for any other executable *`cmd`* on the path or given by an absolute filepath: this is useful to have the same environment as R or the specific commands run under, for example to run `ldd` or `pdflatex`. Under Windows *cmd* can be an executable or a batch file, or if it has extension `.sh` or `.pl` the appropriate interpreter (if available) is called to run it.

B.2 Invoking R under Windows

There are two ways to run R under Windows. Within a terminal window (e.g. `cmd.exe` or a more capable shell), the methods described in the previous section may be used, invoking by `R.exe` or more directly by `Rterm.exe`. For interactive use, there is a console-based GUI (`Rgui.exe`).

The startup procedure under Windows is very similar to that under UNIX, but references to the 'home directory' need to be clarified, as this is not always defined on Windows. If the environment variable `R_USER` is defined, that gives the home directory. Next, if the environment variable `HOME` is defined, that gives the home directory. After those two user-controllable settings, R tries to find system defined home directories. It first tries to use the Windows "personal" directory (typically `C:\Documents and Settings\username\My Documents` in Windows XP). If that fails, and environment variables `HOMEDRIVE` and `HOMEPATH` are defined (and they normally are) these define the home directory. Failing all those, the home directory is taken to be the starting directory.

You need to ensure that either the environment variables `TMPDIR`, `TMP` and `TEMP` are either unset or one of them points to a valid place to create temporary files and directories.

Environment variables can be supplied as '*`name=value`*' pairs on the command line.

If there is an argument ending `.RData` (in any case) it is interpreted as the path to the workspace to be restored: it implies `--restore` and sets the working directory to the parent of the named file. (This mechanism is used for drag-and-drop and file association with `RGui.exe`, but also works for `Rterm.exe`. If the named file does not exist it sets the working directory if the parent directory exists.)

The following additional command-line options are available when invoking `RGui.exe`.

`--mdi`
`--sdi`
`--no-mdi` Control whether `Rgui` will operate as an MDI program (with multiple child windows within one main window) or an SDI application (with multiple top-level windows for the console, graphics and pager). The command-line setting overrides the setting in the user's `Rconsole` file.

`--debug` Enable the "Break to debugger" menu item in `Rgui`, and trigger a break to the debugger during command line processing.

Under Windows with `R CMD` you may also specify your own `.bat`, `.exe`, `.sh` or `.pl` file. It will be run under the appropriate interpreter (Perl for `.pl`) with several environment variables set appropriately, including `R_HOME`, `R_OSTYPE`, `PATH`, `BSTINPUTS` and `TEXINPUTS`. For example, if you already have `latex.exe` on your path, then

```
R CMD latex.exe mydoc
```

will run LaTeX on `mydoc.tex`, with the path to R's `share/texmf` macros appended to `TEXINPUTS`. (Unfortunately, this does not help with the MiKTeX build of LaTeX, but `R CMD texify mydoc` will work in that case.)

B.3 Invoking R under OS X

There are two ways to run R under OS X. Within a `Terminal.app` window by invoking `R`, the methods described in the first subsection apply. There is also console-based GUI (`R.app`) that by default is installed in the `Applications` folder on your system. It is a standard double-clickable OS X application.

The startup procedure under OS X is very similar to that under UNIX, but `R.app` does not make use of command-line arguments. The 'home directory' is the one inside the R.framework, but the startup and current working directory are set as the user's home directory unless a different startup directory is given in the Preferences window accessible from within the GUI.

B.4 Scripting with R

If you just want to run a file `foo.R` of R commands, the recommended way is to use `R CMD BATCH foo.R`. If you want to run this in the background or as a batch job use OS-specific facilities to do so: for example in most shells on Unix-alike OSes `R CMD BATCH foo.R &` runs a background job.

You can pass parameters to scripts via additional arguments on the command line: for example (where the exact quoting needed will depend on the shell in use)

```
R CMD BATCH "--args arg1 arg2" foo.R &
```

will pass arguments to a script which can be retrieved as a character vector by

```
args <- commandArgs(TRUE)
```

This is made simpler by the alternative front-end `Rscript`, which can be invoked by

```
Rscript foo.R arg1 arg2
```

and this can also be used to write executable script files like (at least on Unix-alikes, and in some Windows shells)

```
#! /path/to/Rscript
args <- commandArgs(TRUE)
...
q(status=<exit status code>)
```

If this is entered into a text file `runfoo` and this is made executable (by `chmod 755 runfoo`), it can be invoked for different arguments by

```
runfoo arg1 arg2
```

For further options see `help("Rscript")`. This writes R output to `stdout` and `stderr`, and this can be redirected in the usual way for the shell running the command.

If you do not wish to hardcode the path to `Rscript` but have it in your path (which is normally the case for an installed R except on Windows, but e.g. OS X users may need to add `/usr/local/bin` to their path), use

```
#! /usr/bin/env Rscript
...
```

At least in Bourne and bash shells, the `#!` mechanism does **not** allow extra arguments like `#! /usr/bin/env Rscript --vanilla`.

One thing to consider is what `stdin()` refers to. It is commonplace to write R scripts with segments like

```
chem <- scan(n=24)
2.90 3.10 3.40 3.40 3.70 3.70 2.80 2.50 2.40 2.40 2.70 2.20
5.28 3.37 3.03 3.03 28.95 3.77 3.40 2.20 3.50 3.60 3.70 3.70
```

and `stdin()` refers to the script file to allow such traditional usage. If you want to refer to the process's `stdin`, use `"stdin"` as a `file` connection, e.g. `scan("stdin", ...)`.

Another way to write executable script files (suggested by François Pinard) is to use a *here document* like

```
#!/bin/sh
[environment variables can be set here]
R --slave [other options] <<EOF

   R program goes here...

EOF
```

but here `stdin()` refers to the program source and `"stdin"` will not be usable.

Short scripts can be passed to `Rscript` on the command-line *via* the `-e` flag. (Empty scripts are not accepted.)

Note that on a Unix-alike the input filename (such as `foo.R`) should not contain spaces nor shell metacharacters.

Appendix C The command-line editor

C.1 Preliminaries

When the GNU **readline** library is available at the time R is configured for compilation under UNIX, an inbuilt command line editor allowing recall, editing and re-submission of prior commands is used. Note that other versions of **readline** exist and may be used by the inbuilt command line editor: this used to happen on OS X.

It can be disabled (useful for usage with ESS[1]) using the startup option `--no-readline`.

Windows versions of R have somewhat simpler command-line editing: see '`Console`' under the '`Help`' menu of the GUI, and the file `README.Rterm` for command-line editing under `Rterm.exe`.

When using R with **readline** capabilities, the functions described below are available, as well as others (probably) documented in `man readline` or `info readline` on your system.

Many of these use either Control or Meta characters. Control characters, such as *`Control-m`*, are obtained by holding the `CTRL` down while you press the `m` key, and are written as *`C-m`* below. Meta characters, such as *`Meta-b`*, are typed by holding down `META`[2] and pressing `b`, and written as *`M-b`* in the following. If your terminal does not have a `META` key enabled, you can still type Meta characters using two-character sequences starting with *`ESC`*. Thus, to enter *`M-b`*, you could type `ESCb`. The *`ESC`* character sequences are also allowed on terminals with real Meta keys. Note that case is significant for Meta characters.

C.2 Editing actions

The R program keeps a history of the command lines you type, including the erroneous lines, and commands in your history may be recalled, changed if necessary, and re-submitted as new commands. In Emacs-style command-line editing any straight typing you do while in this editing phase causes the characters to be inserted in the command you are editing, displacing any characters to the right of the cursor. In *vi* mode character insertion mode is started by *`M-i`* or *`M-a`*, characters are typed and insertion mode is finished by typing a further `ESC`. (The default is Emacs-style, and only that is described here: for *vi* mode see the **readline** documentation.)

Pressing the `RET` command at any time causes the command to be re-submitted.

Other editing actions are summarized in the following table.

C.3 Command-line editor summary

Command recall and vertical motion

`C-p` Go to the previous command (backwards in the history).

`C-n` Go to the next command (forwards in the history).

`C-r text` Find the last command with the *text* string in it.

On most terminals, you can also use the up and down arrow keys instead of *`C-p`* and *`C-n`*, respectively.

---

[1] The 'Emacs Speaks Statistics' package; see the URL `http://ESS.R-project.org`

[2] On a PC keyboard this is usually the Alt key, occasionally the 'Windows' key. On a Mac keyboard normally no meta key is available.

Horizontal motion of the cursor

C-a	Go to the beginning of the command.
C-e	Go to the end of the line.
M-b	Go back one word.
M-f	Go forward one word.
C-b	Go back one character.
C-f	Go forward one character.

On most terminals, you can also use the left and right arrow keys instead of *C-b* and *C-f*, respectively.

Editing and re-submission

text	Insert *text* at the cursor.
C-f text	Append *text* after the cursor.
DEL	Delete the previous character (left of the cursor).
C-d	Delete the character under the cursor.
M-d	Delete the rest of the word under the cursor, and "save" it.
C-k	Delete from cursor to end of command, and "save" it.
C-y	Insert (yank) the last "saved" text here.
C-t	Transpose the character under the cursor with the next.
M-l	Change the rest of the word to lower case.
M-c	Change the rest of the word to upper case.
RET	Re-submit the command to R.

The final RET terminates the command line editing sequence.

The **readline** key bindings can be customized in the usual way *via* a ~/.inputrc file. These customizations can be conditioned on application R, that is by including a section like

```
$if R
  "\C-xd": "q('no')\n"
$endif
```

Appendix D Function and variable index

R

S

T

U

V

W

X

Appendix E Concept index

Appendix F References

D. M. Bates and D. G. Watts (1988), *Nonlinear Regression Analysis and Its Applications.* John Wiley & Sons, New York.

Richard A. Becker, John M. Chambers and Allan R. Wilks (1988), *The New S Language.* Chapman & Hall, New York. This book is often called the "*Blue Book*".

John M. Chambers and Trevor J. Hastie eds. (1992), *Statistical Models in S.* Chapman & Hall, New York. This is also called the "*White Book*".

John M. Chambers (1998) *Programming with Data.* Springer, New York. This is also called the "*Green Book*".

A. C. Davison and D. V. Hinkley (1997), *Bootstrap Methods and Their Applications*, Cambridge University Press.

Annette J. Dobson (1990), *An Introduction to Generalized Linear Models*, Chapman and Hall, London.

Peter McCullagh and John A. Nelder (1989), *Generalized Linear Models.* Second edition, Chapman and Hall, London.

John A. Rice (1995), *Mathematical Statistics and Data Analysis.* Second edition. Duxbury Press, Belmont, CA.

S. D. Silvey (1970), *Statistical Inference.* Penguin, London.

Wisconsin_clustering_computation.R

```
require(maptools)
require(Imap)
require(sp)
require(rgdal)
require(rgeos)
require(fields)

map_2002=readOGR("Ward_Election_Data_2002.shp",
 "Ward_Election_Data_2002")
map_2002=spTransform(map_2002, CRS("+proj=longlat
 +datum=WGS84"))
map_2002$r_share=map_2002$GOV_REP/(map_2002$GOV_REP +
 map_2002$GOV_DEM)
map_2002$pvi=map_2002$r_share -
 sum(map_2002$GOV_REP)/(sum(map_2002$GOV_REP) +
 sum(map_2002$GOV_DEM))
map_2002$pvi[which(is.nan(map_2002$pvi))]=0

map_2004=readOGR("wtm_8391_wards_2004_ed.shp",
 "wtm_8391_wards_2004_ed")
map_2004=spTransform(map_2004, CRS("+proj=longlat
 +datum=WGS84"))
map_2004$r_share=map_2004$PRES_REP/(map_2004$PRES_REP +
 map_2004$PRES_DEM)
map_2004$pvi=map_2004$r_share -
 sum(map_2004$PRES_REP)/(sum(map_2004$PRES_REP) +
 sum(map_2004$PRES_DEM))
map_2004$pvi[which(is.nan(map_2004$pvi))]=0

map_2006=readOGR("2006_election_data_by_ward.shp",
 "2006_election_data_by_ward")
map_2006=spTransform(map_2006, CRS("+proj=longlat
 +datum=WGS84"))
map_2006$r_share=map_2006$SEN_REP/(map_2006$SEN_REP +
 map_2006$SEN_DEM)
map_2006$pvi=map_2006$r_share -
 sum(map_2006$SEN_REP)/(sum(map_2006$SEN_REP) +
 sum(map_2006$SEN_DEM))
```

```
map_2006$pvi[which(is.nan(map_2006$pvi))]=0

map_2008=readOGR("2008_election_data_by_ward.shp",
 "2008_election_data_by_ward")
map_2008=spTransform(map_2008, CRS("+proj=longlat
 +datum=WGS84"))
map_2008$r_share=map_2008$PRESREP08/(map_2008$PRESREP08 +
 map_2008$PRESDEM08)
map_2008$pvi=map_2008$r_share -
 sum(map_2008$PRESREP08)/(sum(map_2008$PRESREP08) +
 sum(map_2008$PRESDEM08))
map_2008$pvi[which(is.nan(map_2008$pvi))]=0

map_2010=readOGR("WISELR_Wards_WTM8391_041712.shp",
 "WISELR_Wards_WTM8391_041712")
map_2010=spTransform(map_2010, CRS("+proj=longlat
 +datum=WGS84"))
map_2010$r_share=map_2010$GOVREP10/(map_2010$GOVREP10 +
 map_2010$GOVDEM10)
map_2010$pvi=map_2010$r_share -
 sum(map_2010$GOVREP10)/(sum(map_2010$GOVREP10) +
 sum(map_2010$GOVDEM10))
map_2010$pvi[which(is.nan(map_2010$pvi))]=0

map_2012=readOGR("Wards_111312_ED_110612.shp",
 "Wards_111312_ED_110612")
map_2012=spTransform(map_2012, CRS("+proj=longlat
 +datum=WGS84"))
map_2012$r_share=map_2012$PRES_REP12/(map_2012$PRES_REP12 +
 map_2012$PRES_DEM12)
map_2012$pvi=map_2012$r_share -
 sum(map_2012$PRES_REP12)/(sum(map_2012$PRES_REP12) +
 sum(map_2012$PRES_DEM12))
map_2012$pvi[which(is.nan(map_2012$pvi))]=0

map_2014=readOGR("Wards_Final_Geo_111312_2014_ED.shp",
 "Wards_Final_Geo_111312_2014_ED")
map_2014=spTransform(map_2014, CRS("+proj=longlat
 +datum=WGS84"))
map_2014$r_share=map_2014$GOVREP14/(map_2014$GOVREP14 +
```

```
 map_2014$GOVDEM14)
map_2014$pvi=map_2012$r_share -
 sum(map_2014$GOVREP14)/(sum(map_2014$GOVREP14) +
 sum(map_2014$GOVDEM14))
map_2014$pvi[which(is.nan(map_2014$pvi))]=0

pare_down_map_percentile_d=function(map, percentile) {
    barrier=quantile(map$pvi, percentile)
    return(map[which(map$pvi < barrier),])
}

pare_down_map_percentile_r=function(map, percentile) {
    barrier=quantile(map$pvi, percentile)
    return(map[which(map$pvi > barrier),])
}

nearest_neighbor_distance_median=function(mtrx) {
    array_thing=NULL

    for (i in 1:dim(mtrx)[1]) {
        array_thing=rbind(array_thing, min(mtrx[i,][-i]))
    }
    return(median(array_thing))
}

years=c(2002, 2004, 2006, 2008, 2010, 2012, 2014)

d_means=c(mean(map_2002[map_2002$pvi < 0,]$pvi),
          mean(map_2004[map_2004$pvi < 0,]$pvi),
          mean(map_2006[map_2006$pvi < 0,]$pvi),
          mean(map_2008[map_2008$pvi < 0,]$pvi),
          mean(map_2010[map_2010$pvi < 0,]$pvi),
          mean(map_2012[map_2012$pvi < 0,]$pvi),
          mean(map_2014[map_2014$pvi < 0,]$pvi))

d_meds=c(median(map_2002[map_2002$pvi < 0,]$pvi),
         median(map_2004[map_2004$pvi < 0,]$pvi),
         median(map_2006[map_2006$pvi < 0,]$pvi),
         median(map_2008[map_2008$pvi < 0,]$pvi),
         median(map_2010[map_2010$pvi < 0,]$pvi),
```

```
        median(map_2012[map_2012$pvi < 0,]$pvi),
        median(map_2014[map_2014$pvi < 0,]$pvi))

r_means=c(mean(map_2002[map_2002$pvi > 0,]$pvi),
          mean(map_2004[map_2004$pvi > 0,]$pvi),
          mean(map_2006[map_2006$pvi > 0,]$pvi),
          mean(map_2008[map_2008$pvi > 0,]$pvi),
          mean(map_2010[map_2010$pvi > 0,]$pvi),
          mean(map_2012[map_2012$pvi > 0,]$pvi),
          mean(map_2014[map_2014$pvi > 0,]$pvi))

r_meds=c(median(map_2002[map_2002$pvi > 0,]$pvi),
         median(map_2004[map_2004$pvi > 0,]$pvi),
         median(map_2006[map_2006$pvi > 0,]$pvi),
         median(map_2008[map_2008$pvi > 0,]$pvi),
         median(map_2010[map_2010$pvi > 0,]$pvi),
         median(map_2012[map_2012$pvi > 0,]$pvi),
         median(map_2014[map_2014$pvi > 0,]$pvi))

write.csv(as.data.frame(cbind(years, d_means, d_meds,
 r_means, r_meds)), "wi_means_and_medians.csv", row.names =
 FALSE)

pvi_vec = seq(from = .03, to = .45, by = .03)
data_frame=NULL
row=NULL

for (i in pvi_vec) {
    row=cbind(row,
 nearest_neighbor_distance_median(rdist.earth(coordinates(pa
 re_down_map_percentile_d(map_2002, i)))))
    row=cbind(row,
 nearest_neighbor_distance_median(rdist.earth(coordinates(pa
 re_down_map_percentile_d(map_2004, i)))))
    row=cbind(row,
 nearest_neighbor_distance_median(rdist.earth(coordinates(pa
 re_down_map_percentile_d(map_2006, i)))))
    row=cbind(row,
 nearest_neighbor_distance_median(rdist.earth(coordinates(pa
 re_down_map_percentile_d(map_2008, i)))))
```

```
    row=cbind(row,
nearest_neighbor_distance_median(rdist.earth(coordinates(pa
re_down_map_percentile_d(map_2010, i)))))
    row=cbind(row,
nearest_neighbor_distance_median(rdist.earth(coordinates(pa
re_down_map_percentile_d(map_2012, i)))))
    row=cbind(row,
nearest_neighbor_distance_median(rdist.earth(coordinates(pa
re_down_map_percentile_d(map_2014, i)))))
    data_frame=rbind(data_frame, row)
    row=NULL
}

data_frame=t(data_frame)
data_frame=as.data.frame(data_frame)
rownames(data_frame)=years
names(data_frame)=pvi_vec

write.csv(data_frame, "d_lean_by_quantile.csv")

pvi_vec = seq(from = .55, to = .97, by = .03)
data_frame=NULL
row=NULL

for (i in pvi_vec) {
    row=cbind(row,
nearest_neighbor_distance_median(rdist.earth(coordinates(pa
re_down_map_percentile_r(map_2002, i)))))
    row=cbind(row,
nearest_neighbor_distance_median(rdist.earth(coordinates(pa
re_down_map_percentile_r(map_2004, i)))))
    row=cbind(row,
nearest_neighbor_distance_median(rdist.earth(coordinates(pa
re_down_map_percentile_r(map_2006, i)))))
    row=cbind(row,
nearest_neighbor_distance_median(rdist.earth(coordinates(pa
re_down_map_percentile_r(map_2008, i)))))
    row=cbind(row,
nearest_neighbor_distance_median(rdist.earth(coordinates(pa
re_down_map_percentile_r(map_2010, i)))))
```

```
    row=cbind(row,
nearest_neighbor_distance_median(rdist.earth(coordinates(pa
re_down_map_percentile_r(map_2012, i)))))
    row=cbind(row,
nearest_neighbor_distance_median(rdist.earth(coordinates(pa
re_down_map_percentile_r(map_2014, i)))))
    data_frame=rbind(data_frame, row)
    row=NULL
}

data_frame=t(data_frame)
data_frame=as.data.frame(data_frame)
rownames(data_frame)=years
names(data_frame)=1 - pvi_vec

write.csv(data_frame, "r_lean_by_quantile.csv")
```

A Unified Method of Evaluating Electoral Systems and Redistricting Plans
Author(s): Andrew Gelman and Gary King
Reviewed work(s):
Source: *American Journal of Political Science*, Vol. 38, No. 2 (May, 1994), pp. 514-554
Published by: Midwest Political Science Association
Stable URL: http://www.jstor.org/stable/2111417
Accessed: 18/10/2012 16:44

*A Unified Method of Evaluating Electoral Systems and Redistricting Plans**

Andrew Gelman, *Department of Statistics, University of California, Berkeley*
Gary King, *Department of Government, Harvard University*

We derive a unified statistical method with which one can produce substantially improved definitions and estimates of almost any feature of two-party electoral systems that can be defined based on district vote shares. Our single method enables one to calculate more efficient estimates, with more trustworthy assessments of their uncertainty, than each of the separate multifarious existing measures of partisan bias, electoral responsiveness, seats-votes curves, expected or predicted vote in each district in a legislature, the probability that a given party will win the seat in each district, the proportion of incumbents or others who will lose their seats, the proportion of women or minority candidates to be elected, the incumbency advantage and other causal effects, the likely effects on the electoral system and district votes of proposed electoral reforms such as term limitations, campaign spending limits, and drawing majority-minority districts, and numerous others. To illustrate, we estimate the partisan bias and electoral responsiveness of the U.S. House of Representatives since 1900 and evaluate the fairness of competing redistricting plans for the 1992 Ohio state legislature.

1. Introduction

We introduce a unified and relatively simple statistical model with which one can evaluate electoral systems and redistricting plans in almost any way for virtually any legislature with two major parties and predominately single-member districts. This model is useful (1) for understanding an election that has already taken place; (2) for predicting a future elec-

*We have written a computer program to implement the methods derived in this article. The program is called *JudgeIt* and is available from the ICPSR, via "gopher" or "anonymous FTP" from Haavelmo.Harvard.Edu, or by contacting us. *JudgeIt* has been used in redistricting processes in many states and won the 1992 Research Software Award from the American Political Science Association. All data and information necessary to replicate the empirical analyses in this article are available from the ICPSR in a Class V data set listed under our names. Section 6.2 was drawn from Gary King's experience as a consultant to the State of Ohio in 1991–92. We thank Jim Alt, Neal Beck, Mike Lewis-Beck, and Doug Rivers for many helpful comments on this paper and the National Science Foundation for research grant SBR-9223637. Please address correspondence to Gary King, Department of Government, Harvard University, Littauer Center North Yard, Cambridge, Massachusetts 02138; email: gk@isr.Harvard.Edu; phone: (617)495-2027.

American Journal of Political Science, Vol. 38, No. 2, May 1994, Pp. 514–54

tion, possibly subject to a new redistricting plan; and (3) for evaluating a past election under specified counterfactual conditions (e.g., supposing no incumbents had run for reelection). For each of these general situations, our method enables one to make virtually any prediction about or characterization of the pattern of district votes in an electoral system (e.g., the proportion of African American legislators who would lose their seats under a proposed redistricting plan). Some of the electoral system summaries that can be produced by this method include:

- the seats-votes curve (Edgeworth 1898; Butler 1951; Schrodt 1981; Niemi and Fett 1986; Gelman and King 1990b);
- partisan bias and electoral responsiveness (Tufte 1973; Grofman 1983; King and Browning 1987; Brady 1988);
- the expected or predicted vote in each legislative district (Lewis-Beck and Rice 1992; Cain 1985; Born 1985);
- the probability that a given party will win the seat in each district;
- incumbency advantage and other causal effects (Erikson 1971; Alford and Brady 1988; Gelman and King 1990a);
- the expected proportion of incumbents, or others, who will lose their seats (Mayhew 1974; Fiorina 1977; Ferejohn 1977; Jacobson 1987);
- the expected number of women or minority candidates to be elected;
- the likely effects on the electoral system and district votes of proposed electoral reforms, such as term limitations (Benjamin and Malbin 1992; Rothstein and Gilmour 1992), campaign spending limits (Jacobson 1980), and drawing majority-minority districts (Grofman, Handley, and Niemi 1992);
- the contributions of incumbency advantage, or other aspects of the electoral system, to electoral system phenomena such as divided government (Mayhew 1991; Fiorina 1992; Campbell 1992; King and Gelman 1991).

In providing estimates of these theoretical concepts, we apply the most important insight of the field of statistics in this century to the study of legislative elections—the distinction between the data one observes and the theoretical concepts to be estimated with the data. Too often in the legislative elections literature, scholars define a theoretical concept as identical to a measure of it. For example, scholars have defined the vulnerability of incumbents *as* their electoral margin in the last election. Certainly the latter has something to do with the former, but the two are

not the same.[1] Our model enables one to define any theoretical concept related to legislative elections and to make the best use of available data to provide empirical estimates.

Moreover, as soon as one makes the fundamental distinction between theory and data, the indispensable role of quantitative estimates of uncertainty (such as standard errors, confidence intervals, or margins of error) becomes absolutely clear. Perhaps since the distinction between theoretical concepts and data is not always made in legislative elections research, consistent reporting of standard errors is not yet routine. Our model automatically produces these estimates for virtually every quantity calculated and should therefore make reporting easier as well.

We also provide means of evaluating the fit of the model to the data. Since our model cannot be estimated with existing statistical software, we have made available, as an accompaniment to this paper, a general purpose computer program that implements this model. We have evaluated the performance of this model in hundreds of thousands of districts, in dozens of election years in the U.S. Congress and numerous state legislatures. We have also had some limited experience applying the model to data from foreign countries.

This paper had its origins in our attempts to generalize the models and methods developed in a series of articles by King and Browning (1987), King (1989a), Gelman and King (1990a, 1990b), and King and Gelman (1991). The most recent of these models is technically sophisticated and quite computationally intensive; we believed that adding additional features would make it more realistic but would unfortunately produce an even more complicated model. We were right about the former and wrong about the latter: adding additional information produced a more realistic model for which much of the algebra fell out, leaving a surprisingly simple model. We simplified the model further by eliminating features that did not materially affect the substantive conclusions, based on our analyses of congressional and state legislative elections. The resulting model turned out to be useful not only for our original goal but for numerous other applications in the academic literature as well.

Section 2 introduces the basic model; section 3 discusses issues of preliminary estimation. We derive the distribution of votes in actual, predictive, and counterfactual situations derived from this model in section 4. In section 5, we show how to estimate, along with a standard

[1]It is not difficult to imagine an incumbent who wins elections by small margins but, perhaps due to high levels of racial polarization, consistently and predictably wins reelection; similarly, an incumbent who won with a large margin in the last election would be quite vulnerable if he or she were convicted of a felony.

error, any feature or prediction of interest. Examples appear in section 6. Section 7 concludes. Appendixes A–C provide technical details.

2. A Model of District-based Electoral Systems

2.1. The Model

In order to generally distinguish between the data (actual election results) and theoretical quantities of interest, we begin by predicting what could happen, or what would have happened, if the election were held again under specified conditions. More specifically, we define *hypothetical election results* as the set of all possible election outcomes that could have occurred if all political conditions up to the start of the campaign were held constant and the campaign were run again.

To define hypothetical elections formally, we need a probability model to encompass our uncertainty and allow a range of reasonable possibilities for the hypothetical outcomes. This avoids the unreasonable assumption that election outcomes are *exactly* determined and can be forecast without error, given variables measured before the start of the campaign. The model presented here allows us to calculate the (posterior) probability distribution of these hypothetical election results. We think of the observed election result as just one of the possible hypothetical election results that could have occurred. Any specific theoretical quantity of interest and its standard error can then be calculated directly from the distribution of hypothetical election results.

Although our model applies to any two-party system, we use the labels "Democratic" and "Republican" to fix ideas more clearly. We also use a state legislative election as the running example with which to introduce this model. It should be clear, however, that the model applies much more widely. Following the algebraic presentation of the model, we discuss in more detail a substantive interpretation of the model in section 2.2, and the explanatory variables we recommend including in section 2.3. A discussion of alternative arrangements for the treatment of uncontested districts appears in Appendix A.

Notation. We assume a legislature comprising n single-member districts, denoting v_i as the Democratic proportion of the two-party vote in each district i, and v as the set of votes for all districts $(v_1, v_2, \ldots, v_n)$. The votes v will be predicted by k explanatory variables, which can together be written as an $n \times k$ matrix, X. The first column of X should be all ones, corresponding to an intercept term in a regression; the remaining columns should be substantive explanatory variables, which we discuss in section 2.3. The matrix X is always known, and the votes vector v is

Table 1. Model Structure

District Number	Actual Election Results	Hypothetical Replications of Each District Election			
		1	2	...	m
1	v_1	$v_1^{(\text{hyp})1}$	$v_1^{(\text{hyp})2}$	...	$v_1^{(\text{hyp})m}$
2	v_2	$v_2^{(\text{hyp})1}$	$v_2^{(\text{hyp})2}$	...	$v_2^{(\text{hyp})m}$
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
n	v_n	$v_n^{(\text{hyp})1}$	$v_n^{(\text{hyp})2}$	...	$v_n^{(\text{hyp})m}$
Quantity of interest	Q	$Q^{(\text{hyp})1}$	$Q^{(\text{hyp})2}$	...	$Q^{(\text{hyp})m}$

known when evaluating an election that has occurred, but unknown for prediction.

For hypothetical elections, we define a known matrix $X^{(\text{hyp})}$ of explanatory variables, and an unknown vector $v^{(\text{hyp})}$ of hypothetical district-level Democratic vote proportions. The vector $X^{(\text{hyp})}$ is either equal to X or, to evaluate counterfactual assumptions, is defined as needed.[2] The goal of the analysis is inference about $v^{(\text{hyp})}$, given $X^{(\text{hyp})}$ (and given v if available). For the next brief subsection, we refer to the individual hypothetical replication j of the election in district i as $v_i^{(\text{hyp})j}$.

Conceptual overview. The goal of our model is to calculate a joint probability distribution for all quantities of interest, such as those listed in section 1. From this, all point estimates and standard errors can be calculated. One way to portray our method of calculating this, as well as the results of the model we are about to describe, is Table 1 (see Rubin 1987). Each row in the table refers to a district, with the district number in the first column and the actual election result in the second. If the problem is one of prediction, instead of evaluation, the actual election result is obviously not known. The remaining columns depict m hypothetical replications of each district election. Thus, in district 2, the actual Democratic proportion of the two-party vote is v_2 (which might be, say, 0.56) and again is known only if we are evaluating and not predicting. The first hypothetical replication, numbered 1, of the election in district 2 is $v_2^{(\text{hyp})1}$ (which might be, say, 0.52 or 0.57), an example of what might happen if all conditions up to the start of the general election campaign were the same, but the campaign and balloting were run again. The sec-

[2]For example, one might set the incumbency codes to all zeros to study the likely effect of term limits.

ond hypothetical replication in district 2 is denoted $v_2^{(\text{hyp})2}$, which is another draw from the same probability distribution characterizing these hypothetical elections. In this way, we model the uncertainty in electoral results by this variation across hypothetical elections.

The last row in the table is a "quantity of interest," denoted Q for the actual election result and $Q^{(\text{hyp})j}$ for hypothetical election replication j. This summary statistic is calculated from its corresponding column of real or hypothetical data. This could be virtually anything, but to fix ideas imagine that it is the proportion of incumbents who win reelection. Since for most purposes, incumbency is one of the conditions that, at the start of a general election campaign, we assume to be fixed across hypothetical replications, $Q^{(\text{hyp})j}$ is calculated by taking the proportion of the district votes in the column greater than 0.5 (for Democratic incumbents) or less than 0.5 (for Republican incumbents) among only the rows that have incumbents.

Once we have a set of hypothetical election results for each district election, and we have decided which summary statistic Q we wish to calculate and have calculated it, we can use a simple procedure to calculate an overall point estimate and standard error. The point estimate is just the average, and the standard error is the square root of the variance, across the row of m summary statistics in the table calculated for each hypothetical election result:

$$\overline{Q} = \frac{1}{m}\sum_{j=1}^{m} Q^{(\text{hyp})j}; \tag{1}$$

$$\text{Var}(Q) = \frac{1}{m-1}\sum_{j=1}^{m} (Q^{(\text{hyp})j} - \overline{Q})^2. \tag{2}$$

The only task remaining, then, is to estimate the (posterior) probability distribution, with which we can generate these hypothetical election results and, in turn, calculate our point estimates and standard errors. Once we have this distribution, our actual method of calculation and estimation will correspond closely to the procedure illustrated in Table 1.

The probability model. We model the district vote outcomes with a random components linear regression of v on X,

$$v = X\beta + \gamma + \varepsilon, \tag{3}$$

where β is a vector of k parameters that must be estimated from data, and γ and ε are two vectors of independent error terms. Strictly speaking, this independence assumption is imposed as a definitional feature of our

model, not assumed as a characteristic of the world; testing whether definitions such as this are "true" makes little sense, since ε is *defined* as the part of the error term that is independent for each district vote. The variable ε is a traditional random error term; γ is the "random component" error term, which helps correct for the fact that the X variables do not completely describe the state of the electoral system at the start of the campaign due to the omission of relevant variables and measurement error in the variables included. (We provide a detailed interpretation of γ and ε in section 2.2.) For each district i, the error terms are assigned independent normal distributions,

$$\begin{aligned} \gamma_i &\sim N(0, \sigma_\gamma^2) \\ \varepsilon_i &\sim N(0, \sigma_\varepsilon^2), \end{aligned} \tag{4}$$

with variances σ_γ^2 and σ_ε^2 that must be estimated. Because γ_i and ε_i are independent, equations (3) and (4) are, for some purposes, equivalent to a linear regression of y on X with a single error term of variance $\sigma_\gamma^2 + \sigma_\varepsilon^2$. For mathematical convenience, we reparameterize by defining a parameter for the total variance and a parameter for the proportion of variance due to γ:

$$\sigma^2 = \sigma_\gamma^2 + \sigma_\varepsilon^2$$

$$\lambda = \frac{\sigma_\gamma^2}{\sigma_\gamma^2 + \sigma_\varepsilon^2}.$$

The vector $v^{(\text{hyp})}$ of hypothetical vote proportions is determined by an analogous probability model:

$$v^{(\text{hyp})} = X^{(\text{hyp})}\beta + \delta^{(\text{hyp})} + \gamma + \varepsilon^{(\text{hyp})}, \tag{5}$$

where $\varepsilon^{(\text{hyp})}$ is a new vector of n independent error terms with variance σ_ε^2, and $\delta^{(\text{hyp})}$ is a known constant used to model statewide partisan swing. The hypothetical outcome, $v^{(\text{hyp})}$, which generates the numerical results for the columns in Table 1, differs from the actual v in three ways:

1. The explanatory variable matrix X is replaced by $X^{(\text{hyp})}$, to recognize that we may wish to specify different conditions under which the hypothetical election may be run (such as no incumbents running).
2. A constant, $\delta^{(\text{hyp})}$, is added to allow a statewide partisan swing to be specified. One can specify either $\delta^{(\text{hyp})}$ or a corresponding value for the expected average district vote, $E(\bar{v}^{(\text{hyp})})$.[3]

[3]This is true, since $\delta^{(\text{hyp})} = E(\bar{v}^{(\text{hyp})}) - (1/n)\,\Sigma_{i=1}^{n}(X^{(\text{hyp})}\beta)_i$.

3. The new error term, $\varepsilon^{(\text{hyp})}$, models the fact that, even if the variables in X were unchanged, we would not expect $v^{(\text{hyp})}$ to be identical to v. Across many hypothetical elections, γ remains unchanged, while ε varies.[4]

The parameters of this model to be estimated—σ^2, λ, and β—are not usually of primary interest in evaluating electoral systems and redistricting plans (although β is in some cases of interest in evaluating causal effects). Instead, we define all the quantities of interest, including the seats-votes curve, district vote predictions, and the like, in terms of the distribution of hypothetical election outcomes $v^{(\text{hyp})}$, given the average district vote $\bar{v}$ or the actual election outcomes v when available (which, in turn, depend on the parameters). The specific method for calculating this conditional distribution from this model is described in section 4.

2.2. *Interpretation of the Model*

Aggregate partisan swing. The parameter $\delta^{(\text{hyp})}$, or the corresponding value of $E(\bar{v}^{(\text{hyp})})$, in this model is a notational convenience that allows us to vary the average district vote in a hypothetical (or predicted) election, without affecting the relative positions of the districts. This partitioning reflects the common result that it is often quite easy to predict which districts will vote more Republican than others, but it is harder to forecast exactly what the average vote will be across districts; put differently, given the average vote across districts, it is easy to predict the vote proportions within each district.

According to the model, expected votes differ across legislative districts at any one time (as governed by $X\beta + \gamma$). Over time, the districts swing along with the statewide mean (due to the scalar parameter $\delta^{(\text{hyp})}$) but only on average (due to the random error term ε). Another way to put this is that districts move on average with the statewide mean, but, for any given statewide issue producing a swing toward a party, any district may move with or against the statewide trend (as indicated by a scatter plot of district votes in two successive years; see, e.g., Figures 1 and 2 in King and Gelman 1991). This allows us to account for local differences in the appeal of the candidates and other factors simultaneously with statewide swings.

The stochastic model is interpreted slightly differently for prediction and evaluation: for prediction, we ask how many seats *will* the Democrats

[4]Another way to think of this model is that, for each district i, v_i and $v_i^{(\text{hyp})}$ are not independent, due to the random variable γ_i they share. Given the explanatory variables X, $X^{(\text{hyp})}$, and $\delta^{(\text{hyp})}$, and the parameters β, σ^2, and λ, we can combine equations (3) and (5) to find that the theoretical correlation between v_i and $v_i^{(\text{hyp})}$ is λ.

win with an average of 45% of the votes, say, and in evaluation we ask, how many seats *would* they have won if essentially the same election campaign had been run again.[5] The only difference between evaluation and prediction is that we observe one of the possible hypothetical election outcomes for the former (the actual vote, the second column in Table 1) and do not observe any for the latter. Even after taking into account the information in our model expressed through the explanatory variables X, observing the election outcome will in general help us to some degree in characterizing the conditional distribution of hypothetical election outcomes; thus, under our model, inference about observed election systems will generally be sharper than predictions.

Error terms. To understand the division between γ and ε (or, equivalently, the relative values of σ_γ^2 and σ_ε^2), consider the two extreme cases. If $\sigma_\gamma^2 = 0$, then $\lambda = 0$, and there is no *systematic* difference among districts except that described by $X\beta$. Consider two districts, i and j, with identical values of X but different outcomes in the election: $v_i = 0.6$ and $v_j = 0.7$. Without γ, all the difference between v_i and v_j must be attributed to the random error ε, and as a result, the distributions of hypothetical election results in districts i and j will be the same. For real electoral systems, the explanatory variables in X cannot be perfect, and it is usually wrong to ignore the additional information that the Democrats did better in district j in the observed election.

At the other extreme, if $\sigma_\varepsilon^2 = 0$, then $\lambda = 1$, and every hypothetical election outcome $v^{(\text{hyp})}$ must be *identical* to the observed v, except for the constant shift, $\delta^{(\text{hyp})}$. This is just the uniform partisan swing assumption (Butler 1951), which requires that individual districts move in lockstep with statewide swings. As distinct from the *assumption* of uniform partisan swing, there is also a uniform partisan swing *method,* which is only strictly appropriate when the corresponding assumption is correct. Unfortunately, this assumption does not apply to any known electoral system (again, see Figures 1 and 2 in King and Gelman 1991); in characterizing an electoral system, only the variation that persists in future election results deserves to be labeled "systematic" in the electoral system. In principle, the uniform partisan swing method can work in some cases even though the assumption is false, but it never produces honest estimates of uncertainty and is always less statistically efficient than the method we describe here.

[5]Of course, the model can address numerous other substantive questions; this one highlights the difference between what will happen in prediction and what would have happened in evaluation.

A final way of understanding the separation of γ and ε is by imagining that an extremely large number of explanatory variables were available in X—not just previous election results but also campaign spending, the effectiveness of every election campaign, lists of all campaign events, the results of all the campaign polls, the weather on election day, and so on. In this case, v could be predicted with essentially zero error. In this extreme situation, we would *not* want to assume that X is unchanged in hypothetical elections. It would make more sense to split X into two sets of variables: variables like past election results, incumbency, and party support in each district, which would not change if the election campaign were to be held again; and variables like campaign poll results and the weather on election day that would change. Between these extremes, some dividing line must be drawn that then *defines* the set of hypothetical elections and thus the electoral system.

In the ideal world in which X contains all possible variables, we would lump the first set of variables with γ as the systematic component and the second set with ε as the stochastic component. When, as in practice, the variables in X are not complete, we can attribute the residuals in the regression to the unobserved variables, and the residual variance may be partitioned *statistically,* as we do, into that due to γ and ε.[6]

2.3. *Explanatory Variables*

Whether we seek to predict the future or evaluate the past, our model requires explanatory variables that help predict votes v from explanatory variables X. (Section 4.2 continues the discussion of explanatory variables in the context of defining counterfactual scenarios.) Our model uses these explanatory variables to help statistically partition election results into systematic and stochastic components. Since the purpose is not estimating causal effects, the rules for choosing explanatory variables are different than usual. The main immediate goal is to choose variables that would help in forecasting future votes.[7]

When used for forecasting, our model will perform only as well as the variables chosen, and we take the main insight about the choice of

[6]The distinction between γ and ε can also be understood from the perspective of various philosophies about the role of random error in statistical analyses; see King (1989b, 1991b).

[7]By comparison, Gelman and King (1990a) did not control for any consequences of incumbency in their estimation of incumbency advantage. In the current application, we are just trying to predict, and we have no objection to controlling for variables such as campaign spending that are determined after, and in part as a consequence of, the incumbent's decision of whether to run for reelection. We should be careful, of course, not to interpret the coefficients as causal estimates.

variables from the large existing forecasting literature (see Lewis-Beck and Rice 1992 for a review). For a given set of explanatory variables, our model will not produce better forecasts on average, for example, but it will give more accurate estimates of the uncertainty of those forecasts.

On the basis of prior research, it is clear that the variables in X should certainly include *past legislative election results,* when available.[8] When including a previous legislative election, it also makes sense to include variables for *incumbency,* $inc(t)_i$ (defined as 1 if a Democratic incumbent is running; 0 if no incumbent is running; and -1 if a Republican incumbent is running), and *uncontestedness,* $unc(t)_i$ (defined as 1 if a Democrat is running uncontested; 0 if the district election is contested; and -1 if a Republican is running uncontested), for district i in election t. For election t, the variable $inc(t)$ is the set of all district values, $inc(t)_1, \ldots, inc(t)_n$, and similarly for $unc(t)$.

Uncontestedness is important because uncontested elections do not fit any linear model unless explicitly controlled for (or adjusted as in Appendix A). An incumbency status variable is important if there is a large incumbency advantage because without such an indicator, election results from a mixture of incumbent-controlled and open districts would not fit the assumptions of our linear model with independent error terms. In addition, including the incumbency variable usually improves the predictive power of the model. In years following redistricting, incumbency status is often unclear, but it is still better to estimate the variable roughly, for each district, than to ignore it. If one is concerned about assuming that incumbency status will remain the same after redistricting, perhaps because of nonpartisan population base changes, one could include interaction effects so that the effect of incumbency would vary by population base or other variables.[9]

[8]When predicting an election just following a redistricting, the past election results will have to be reaggregated into the new districts, if possible. However, it usually makes little sense to do this without taking into account differences between the districts. By far the most important problem in creating voting data for new districts is combining districts with different incumbency status. To deal with this problem, we suggest the following procedure: (1) estimate the incumbency advantage from historical voting data using the method described in Gelman and King (1990b); (2) obtain legislative voting data at the precinct (or election district) level; (3) subtract the incumbency advantage from the incumbent candidate in precincts represented by incumbents and give those votes to the opposition party candidate (thus creating an estimate of what the electoral data at the precinct level would have been in an open-seat election); (4) aggregate the precincts up to the new districts; and finally, (5) add back the incumbency advantage based on the incumbency status in the new districts. One could do better still by also correcting for differences in candidate quality or other factors, if it really seems worth the effort.

[9]The explanatory variable here could be the proportion of the people who *move* into or out of the district, for current districts, or proportion of people who *are moved* into or

As discussed in Gelman and King (1990a), a variable for *party control* in the district is also useful: $party(t)_i$ (defined as 1 if the sitting incumbent is a Democrat; -1 if the sitting incumbent is a Republican; and 0 otherwise—seat controlled by a third party). This variable controls for possible bimodality in the partisan strengths of districts.[10] It is especially important to include the party variable when the incumbency variable is inaccessible; without correcting for at least one of the two, most recent U.S. electoral systems would almost certainly be bimodal (Mayhew 1974). When using a statewide variable, such as the race for state comptroller, for prediction, it is of course unnecessary to include incumbency, as it affects all districts equally, but it may be desirable to adjust the district-level votes for comptroller by correcting for the "friends and neighbors" effect due to the local popularity of the statewide candidates.

Uncontestedness, incumbency, and party control should also, of course, be used when predicting future elections. Party control should always be available, but uncontestedness and the decision of the incumbent about whether to run will be more difficult to gather. We have found, however, that redistricters almost always have this information, so it is in principle not difficult if one can gain access.[11]

Statewide or nationwide election results, broken down by legislative districts, are also useful (when available). Other plausibly useful variables

---

out of the district by redrawing the district lines, for forecasts following redistricting. However, in our experience, nonpartisan base changes have little effect and are not especially useful for forecasting.

[10]A regression model fit to a distribution with two clusters of partisans will wrongly fit a single linear model to both clusters, and as a result, will overestimate the "regression effect." Or to put it another way, a distribution with two clusters, corresponding to the two parties, does not fit a linear model without including different intercepts and a common slope. See King (1991a) for a graph of this and an explanation, and Gelman and King (1990b) for a description of the procedures we have developed to accommodate this problem even with bimodality.

[11]In most states, these data are produced by state computers and officially available to the public, but in practice, obtaining relevant information on the computer can require some inside contact. If data gathering becomes difficult, it makes the most sense in the framework of our model to make a guess about the incumbency status and uncontestedness of each district in the predicted election, and then draw conclusions conditional on the guess. Uncertainty in future electoral conditions may be expressed by considering several possibilities for incumbency and uncontestedness and evaluating the predictions for each. A standard assumption to start with is that all incumbents run, and incumbency status in new districts is determined by their residence (the actual figure is that about 85% of state legislators run for reelection, and 90% among U.S. House members; the decision to run again is independent of the legislator's share of the vote in the previous election), all contested seats stay contested, and all uncontested seats likewise remain so. One should also remember that following redistricting, more seats tend to be contested, and fewer incumbents run.

are campaign contributions received before the start of the campaign (perhaps transformed to the log scale to better fit the linear model), demographic characteristics of the voting age population in the district, party registration figures (or relative proportions of votes cast in Democratic and Republican presidential primaries), and measures of candidate quality.

3. Preliminary Estimation

The purpose of the model (equations 3–5) is to estimate a quantity of interest Q from $v^{(\text{hyp})}$, for an existing, future, or counterfactual election, given a matrix of explanatory variables, $X^{(\text{hyp})}$, a shift parameter, $\delta^{(\text{hyp})}$ (or the corresponding value of the expected average district vote, $E(\bar{v}^{(\text{hyp})})$), and the actual election outcomes, if available. We can do so using equation (5), as long as we know the parameters, β, σ^2, and λ. Our approach is to estimate the model parameters using the regression equation (3), using current and past elections as data. Unlike in many applications of regression modeling in the social sciences, the parameters β, σ^2, and λ are not themselves the goal of our analysis but rather intermediate quantities, used for estimating the distribution of hypothetical election quantities.

When the current election, v, has been observed, we estimate β by simply regressing v on X, yielding $\hat{\beta}$ and the usual least squares variance matrix, $\hat{\Sigma}_\beta$. For prediction, when v has not yet occurred, we run the same regression in the most recent year for which we have electoral results.[12] The parameter σ^2 is estimated in the above regression by the usual least squares estimate of the "standard error of the regression," $\hat{\sigma}^2 = e'e/(n - k)$, where e is the vector of residuals from the regression of v on X; n is the number of observations; and k is the number of columns in X.

The remaining parameter to be estimated, λ, is needed when using the model for evaluation and counterfactual evaluation. To estimate λ, we use the Democratic proportion of the two-party vote in the election following the one used to define v and regress it on v and the original explanatory variables, X. Since we treat our analysis as conditional on uncontestedness and incumbency status, we also include the values of these two variables for the next election period. The regression coefficient on v is our estimate of λ, which may be thought of as an intuitive estimate of the proportion of variation due to γ, by directly estimating

[12]For our later calculations, we summarize our knowledge of β as a Bayesian multivariate normal posterior distribution (assuming a "noninformative" improper uniform prior distribution; see Box and Tiao 1973), with mean vector $\hat{\beta} = (X'X)^{-1}X'y$ and covariance matrix $\hat{\Sigma}_\beta = \hat{\sigma}^2(X'X)^{-1}$.

how much the actual vote predicts the next election over and above the predictive power of previous values of the explanatory variables.[13] Rare individual estimates of λ above one or below zero are truncated to within this range. If the next election result is not available, λ can be estimated from a recent pair of election years.[14]

Pooling estimates across election years. In practice, more accurate estimates of σ^2 and λ may be formed by averaging the estimates obtained in separate regressions from several election years from the same legislature. Since pooling in this way reduces the uncertainty in the estimates to essentially zero, we use the pooled point estimate instead of its distribution. Thus, we act as if we know $\sigma^2 = \hat{\sigma}^2$ and $\lambda = \hat{\lambda}$ for all further calculations.[15]

Since they are conditional on the explanatory variables, X, the estimates of σ^2 and λ should be pooled only for elections for which roughly the same information is available about X. Typically, election years will be divided into two classes: those immediately following redistricting periods (and the first election in the data set), and all the others. In the former, the explanatory variables will not usually include the vote in the previous election. It makes sense to calculate two estimates of each σ^2 and λ by pooling within each type of election.[16]

We do not advocate pooling the estimates of β, since these parameters are usually more volatile over time. In general, we want to estimate an electoral system using the data closest at hand, pooling only for the hyperparameters that seem to vary slowly over the years.

Empirically, our estimates of σ for U.S. legislative elections have been around 0.06, and almost always between about 0.02 and 0.12,

[13]The regression estimate is formally justified by the fact that λ is the coefficient of v in the expected value of the distribution for $v^{(\text{hyp})}$ in equation (7).

[14]Another method of estimating λ is based on the fact that the theoretical correlation between v and $v^{(\text{hyp})}$ is exactly λ. Although we do not observe $v^{(\text{hyp})}$, we can use the empirical correlation (truncated to be positive) between the residuals of the regressions predicting v from X in two successive years. We have found that the two estimators yield roughly the same estimates of λ in almost all empirical examples.

[15]To use a distribution, instead of assuming the pooled point estimate has no uncertainty, one would need the posterior distribution of σ^2. This is equivalent to the distribution of $\hat{\sigma}^2 \times N(n-k)/\chi^2_{N(n-k)}$, where $\chi^2_{N(n-k)}$ is a chi-square random variable with $N(n-k)$ degrees of freedom, and N is the number of election years used in the pooling. Since N is usually quite large (number of districts multiplied by the number of election years), this distribution has a very small variance, making our assumption in the text empirically reasonable. Our extensive empirical analyses, not presented here, confirm this point in practice.

[16]If precinct data are available and it is possible to follow the procedure in note 8, then all election years can be pooled for estimating σ^2 and λ.

indicating that our explanatory variables account for the district-level vote to within plus or minus about six percentage points. Estimates of λ have generally ranged from 0.2 and 0.9 (most commonly near 0.6), indicating that the proportion of the variation, not explained by X, that recurs in the next election varies greatly with the electoral system and the explanatory variables used. We also find that results are usually very robust to even moderate changes in estimates of λ and σ.

4. The Distribution of Hypothetical Votes

In this section, we give the distribution of $v^{(\text{hyp})}$, so that we can generate multiple hypothetical election results, as in Table 1, and eventually point estimates and standard errors of our quantities of interest. For prediction, v is unknown, so the goal is the unconditional distribution, $P(v^{(\text{hyp})})$. For evaluating an election that has already occurred, we use the information in v, which is available, and derive the distribution of $v^{(\text{hyp})}$ given v, $P(v^{(\text{hyp})}|v)$. In either case, the distribution for $v^{(\text{hyp})}$ is implicitly conditional on $X^{(\text{hyp})}$ and $\delta^{(\text{hyp})}$ as well as σ^2 and λ (since we take σ^2 and λ as fixed after their estimation in section 3). We describe how to calculate quantities of interest and their uncertainty from these distributions in section 5.

4.1. The Predictive Distribution of Future Elections

We give the predictive distribution first because it is a simpler special case. Predictive uncertainty has two components: first, the fundamental variability represented by the parameter σ^2, the variance of the district election results in v, conditional on the explanatory variables X, and second, the uncertainty due to our estimation of β, as modeled with the distribution in equation (10). If we had an infinite number of electoral districts, estimation uncertainty would drop to zero, but σ^2 would not change. We include the variance due to estimating β in our analysis, which in practice is a relatively smaller addition to the σ^2 from the model, although it does vary across districts, unlike σ^2.[17]

The technical derivation appears in Appendix B, ultimately producing the following predictive distribution of hypothetical election results:

$$P(v^{(\text{hyp})}) = N(v^{(\text{hyp})}|X^{(\text{hyp})}\hat{\beta} + \delta, X^{(\text{hyp})}\Sigma_\beta X^{(\text{hyp})\prime} + \sigma^2 I). \quad (6)$$

This is the unconditional Normal distribution of $v^{(\text{hyp})}$, from which we can calculate the hypothetical election results, as in Table 1, and then the

[17]Another possible source of variance is uncertainty in X and δ, which we do not model mathematically. Instead, if these explanatory variables are unknown, we would just compare several analyses, with the uncertain variables set at different reasonable values.

distributions of various summaries of the electoral system. The parameters λ and γ do not appear in equation (6) and are therefore unnecessary for prediction.[18] This is a familiar result in econometrics (e.g., Goldberger 1991, 175–76) and indicates that "predicted values" can be calculated as usual with the possible addition of a statewide swing parameter: $X^{(\text{hyp})}\hat{\beta} + \delta$. The uncertainty of these predictions is given by the distribution and variance in equation (6).

4.2. Evaluating an Existing Electoral System under Actual or Counterfactual Conditions

We now present the distribution of hypothetical votes for historical elections. Since the distribution of hypothetical votes for evaluating counterfactual conditions produces actual evaluations as a special case, we save space by including only the general result here.

When designing counterfactual scenarios, one should be careful in specifying $X^{(\text{hyp})}$ and deciding exactly what conditions to be held constant. For example, what if a term limitation initiative had swept the nation before the last election, and all the incumbent members of Congress were forced to retire? How would this have affected the electoral system?[19] The only practical difficulty is precisely defining the conditions under which the election would be held—in our model, the explanatory variables, $X^{(\text{hyp})}$ and statewide swing $\delta^{(\text{hyp})}$ (or $E(\bar{v}^{(\text{hyp})})$). One possible definition of this particular counterfactual scenario is as follows. Obviously, no incumbents run for reelection, so the column of $X^{(\text{hyp})}$ indicating incumbency status should have all zeros. With all seats open, it is also reasonable to assume that all the districts are contested. Now, instead of setting $E(\bar{v}^{(\text{hyp})})$, we can just set the aggregate partisan swing, $\delta^{(\text{hyp})}$, to zero, making the assumption that nothing else systematic changes but incumbency and contestedness. With all these variables set, the hypothetical elections can be modeled as in this section.

In general, when designing counterfactual scenarios, it is important to control only for variables that happened *before* the intervention of interest (see King 1991b). For example, when supposing that no incumbents will run for reelection, measures of campaign contributions and

[18]To account for the uncertainty from estimating σ^2 explicitly, the normal distribution would be changed to a Student's t distribution with $n - k$ degrees of freedom. However, pooling the estimate of σ^2 across elections, as we recommend, makes the degrees of freedom large, and the normal distribution in this situation approximates the t quite well.

[19]In the absence of the term limit threat, King and Gelman (1991) asked what the electoral system would have been like had the incumbency advantage suddenly disappeared, in order to understand the effect of the incumbency advantage on postwar U.S. House elections.

candidate quality, for both incumbents and challengers, should be discarded (or appropriately modified). Campaign contributions and the decision of opposing candidates of quality to run for election are both largely decided after the incumbent's reelection intentions are known and in part the consequence of incumbency advantage.

The technical details of our derivation appear in Appendix C. The result, which appears complicated at first, is the following Normal probability distribution of hypothetical district election results given observed votes:

$$P(v^{(\text{hyp})}|v) = N(v^{(\text{hyp})}|\lambda v + (X^{(\text{hyp})} - \lambda X)\hat{\beta} + \delta^{(\text{hyp})}, \\ (1-\lambda^2)\sigma^2 I + (X^{(\text{hyp})} - \lambda X)\Sigma_\beta (X^{(\text{hyp})} - \lambda X)'). \tag{7}$$

A good way to understand this result is to focus on the special case where we are analyzing the actual election, and therefore $X^{(\text{hyp})} = X$ and $\delta^{(\text{hyp})} = 0$. In this case, the expected Democratic proportion of the two-party vote in district i from equation (7) simplifies to

$$E(v_i^{(\text{hyp})}|v) = \lambda v + (1-\lambda)X\beta, \tag{8}$$

a weighted average of the observed vote, v, and the vote predicted from past elections, $X\beta$.

Considering only v (the correct estimate if $\lambda = 1$) produces the discredited uniform partisan swing assumption, while the other extreme of using only $X\beta$ (if $\lambda = 0$) wrongly ignores the information from the current election, which is obviously relevant to estimating the current electoral system even after taking into account one's predictions. To put it another way, λ is a parameter that determines which estimator of the form $\lambda v + (1-\lambda)X\beta$ is most effective at estimating the district-level results of another election, $v^{(\text{hyp})}$, from the same electoral system (i.e., if everything were the same up to the start of the campaign and the campaign and election were run again). Thus, in characterizing the expected value of the distribution $P(v^{(\text{hyp})}|v)$, we clearly want to use our prediction $X\beta$, and we also want to use the systematic and persistent aspects of v not predicted by $X\beta$. The question is how much to weight v and $X\beta$; λ is the weight and our method of estimating λ provides an answer.

We also use the entire distribution, including the variance of the predictive equation, to generate hypothetical election results like those in Table 1 and ultimately to obtain expressions of the uncertainty associated with various summaries of the electoral system.

5. Calculating Summaries of Electoral Systems

Once the distribution of hypothetical election results has been obtained, using either equation (6) for prediction or equation (7) for actual

or counterfactual evaluation, we calculate various summaries of interest of the electoral system and their standard errors. These may include, for example, district-level vote and seat forecasts, partisan bias, electoral responsiveness, the number of incumbents who are reelected, and others listed in section 1.

The general methodology we introduce to accomplish these tasks is called "Bayesian simulation," although it could as easily have been called "approximating conditional probability distributions by drawing random numbers." It is also intuitively explained in Table 1. To explain the exact mathematical procedure intuitively, section 5.1 shows how to calculate one feature of the electoral system—forecasts of district vote proportions—that can be calculated analytically. Section 5.2 demonstrates how to calculate these and other summaries with Bayesian simulation. We then turn to other features of an electoral system for which Bayesian simulation works, but no analytic solution is available. No new assumptions are introduced in this section; we show only how to re-express the simulations of district-level vote proportions in more useful formats. Bayesian simulation should be widely applicable to many political science problems, beyond those we describe here.

5.1. District-level Vote and Seat Predictions: Analytic Solution

We can predict district vote proportions using our estimate of the expected vote in each district, $E(v_i^{(\text{hyp})})$, the average over the range of possible hypothetical election results, $v_i^{(\text{hyp})}$. We calculate this from the expected value of the multivariate normal distribution from either equation (6) for prediction or equation (7) for actual or counterfactual evaluation. For example, the vector of district-level predictions for the Democratic proportion of the two-party vote is $X^{(\text{hyp})}\hat{\beta} + \delta^{(\text{hyp})}$. The variables in $X^{(\text{hyp})}$ are chosen at the start, and $\hat{\beta}$ is estimated as part of the procedure. Only the constant, $\delta^{(\text{hyp})}$, needs to be specified. For most purposes, $\delta^{(\text{hyp})}$ will be set to zero, but to simulate statewide swings it will also be useful to set it to other values. For example, we could set $\delta^{(\text{hyp})}$ so that the average district vote takes on several specified values, such as those of the last several elections, to see how the predictions depend on the statewide partisan swing.

To calculate standard errors for the predictions, we can use the standard deviation of the same multivariate normal distribution. However, in this case, and most others, one can calculate two types of standard errors, based on either the predictive uncertainty or only the uncertainty in the expected value. For district-level vote predictions, total predictive uncertainty is probably most relevant. However, if one wishes a summary of only the uncertainty in where the expected vote is, then only the estimation uncertainty should be used.

The standard errors based on the total predictive uncertainty are merely the square roots of the diagonal elements of the variance matrix in equation (6), $X^{(\text{hyp})}\Sigma_\beta X^{(\text{hyp})\prime} + \sigma^2 I$ (or equation 7 for actual or counterfactual evaluation). For standard errors based only on the estimation uncertainty in the expected district vote, we would take the square root of the diagonal elements of only the first term in the variance matrix, such as $X^{(\text{hyp})}\Sigma_\beta X^{(\text{hyp})\prime}$ for prediction.

Another useful way to summarize the uncertainty for district-level predictions is to focus not on votes but on the probability that the election in district i goes for the Democrat (or Republican),

$$P(\text{Democrat wins}) = P(v_i^{(\text{hyp})} > 0.5) = \Phi\left[\frac{E(v_i^{(\text{hyp})}) - 0.5}{\text{var}(v_i^{(\text{hyp})})^{1/2}}\right], \tag{9}$$

where Φ is the standard normal cumulative distribution function. When aggregated over all districts, these probabilities provide a better measure of Democratic strength than counting the number of districts with $E(v_i^{(\text{hyp})}) > 0.5$. For example, if all the districts in a state have $E(v_i^{(\text{hyp})}) = 0.49$ and $\text{var}(v_i^{(\text{hyp})}) = 0.07^2$, then, for each district, $P(\text{Democrat wins}) = \Phi(-0.01/0.07) = 0.44$. However, we certainly would not expect every Republican candidate in the state to win; with enough districts, even small probabilities will produce at least some Democratic wins.

5.2. An Introduction to Bayesian Simulation

Bayesian simulation allows us to summarize the electoral system using observable quantities of immediate interest. In contrast, just reporting regression coefficients or some other model estimate may require convoluted reasoning to translate into something with substantive political meaning.

For many summaries, one can calculate exact analytic solutions directly from the expected values in either equation (6) or equation (7), as was done in section 5.1. In general, though, simulation is the easiest approach to computing estimates and standard errors of summaries of the distributions we have derived. It is frequently easier, computationally faster, and simpler to understand than the analytic solution.

For example, suppose one has a distribution $f(y)$ and wishes to calculate its expected value. The formal "analytic" method would be to compute the integral, $E(Y) = \int_{-\infty}^{\infty} yf(y)\,dy$, which is difficult or impossible in many cases, especially if y represents a large vector of dependent random variables. Fortunately, if random draws can be obtained from this distribution (a task that is often easy, especially with modern computers), it

is possible to use sampling theory to approximate the expected value to an arbitrary degree of precision. That is, one merely takes a large number of random draws from this distribution and averages them. The desired degree of precision can be attained by merely increasing the number of draws. We apply this same logic to our distributions and theoretical features with any degree of complexity.

Simulating hypothetical elections from our model requires three steps:

1. Create a large number, m, of simulated hypothetical election results, $v^{(\text{hyp})}$, by taking random draws from the multivariate normal distribution (in equation 6 or 7). Each election result is a vector of length n, $v^{(\text{hyp})} = (v_1^{(\text{hyp})}, \ldots, v_n^{(\text{hyp})})$ and is portrayed as a column in Table 1.
2. For each simulated vector of hypothetical election results $v^{(\text{hyp})}$, calculate the desired descriptive summary (e.g., vote in district 4, $v_4^{(\text{hyp})}$; expected seats given average district vote set at 0.5, $E(\bar{s} \mid \bar{v} = 0.5)$; or the proportion of districts with between 0.25 and 0.5 probability of a Democratic win). This is $Q^{(\text{hyp})j}$ in Table 1.
3. Calculate the mean and standard deviation of the m summaries (as in equations 1 and 2); these constitute the point estimate and its standard error, respectively.

If it were possible to use an infinite number of simulations ($m \to \infty$), this procedure would yield exact standard errors and point estimates. For most purposes, $m = 100$ suffices to produce sufficiently precise approximations.

Perhaps the only unfamiliar technical issue involved here is how to draw values randomly from one of the two multivariate normal distributions. Fortunately, the solution to this problem is quite simple. To create a single simulated hypothetical election outcome (one of the m required in the description, above):[20]

1. Draw one random vector, β°, from the posterior distribution for β (see equation 10).
2. If the current election, v, has been observed, insert β° into the distribution for $\gamma \mid \beta$ (see equation 15), and draw a vector, γ°, from

[20]One might think that draws from the posterior distribution of $v^{(\text{hyp})}$ might easily be obtained directly from the appropriate normal distribution (equation 6 for prediction or equation 7 for observed elections), using some method like the Cholesky decomposition applied to the covariance matrix. However, that approach requires inverting an n-dimensional matrix, a task that is not desirable or even possible here because the covariance matrixes are singular.

this distribution. In prediction, for which v has not yet occurred, skip this step.

3. Finally, for prediction, insert β° into the distribution $P(v^{(\text{hyp})}|\beta)$ (see equation 11), or for evaluation or counterfactual evaluation insert β° and γ° into the distribution $P(v^{(\text{hyp})}|\gamma, \beta, v)$ (see equation 14), set $\delta^{(\text{hyp})}$ as you choose, and draw a single random vector, $v^{(\text{hyp})}$, from the appropriate distribution.[21]

We now apply Bayesian simulation to calculate several more complicated quantities of interest, along with their standard errors.

5.3. The Distribution of Seats Given Votes

Following the notation of Gelman and King (1990b) and King and Gelman (1991), we label $\bar{v}$ and $\bar{s}$ as the average district vote and the average proportion of seats obtained by the Democrats in the current election, and $\bar{v}^{(\text{hyp})}$ and $\bar{s}^{(\text{hyp})}$ for the corresponding hypothetical quantities. To obtain the seats-votes curve, we estimate the distribution of $\bar{s}^{(\text{hyp})}$, given $\bar{v}^{(\text{hyp})}$, for a range of values of $\bar{v}^{(\text{hyp})}$. Of course, our predictions and estimates are only as good as our model, and it is only reasonable to expect the model to hold over a plausible range of aggregate outcomes. For example, if $\bar{v}^{(\text{hyp})}$ is currently 0.5, and has been near that for many years, we would not use our model to predict what would happen in the case of $\bar{v} = 0.2$; such a change would imply an electoral transition beyond the scope of the historical data on which the model is based. We typically calculate the seats-votes curve for a range of votes centered at the current value, such as every percentage point in the range $[\bar{v}^{(\text{hyp})} \pm 0.15]$, or perhaps at a conventional range such as [0.4, 0.6].

Whether for a predicted or hypothetical election, we can easily obtain the distribution of $\bar{s}^{(\text{hyp})}$, given $\bar{v}^{(\text{hyp})}$, for each of a range of values, $\bar{v}^{(1)}, \ldots, \bar{v}^{(l)}$, as follows:

1. Using the appropriate distribution (prediction of a future election or actual or counterfactual evaluation of a past election), simulate a set of m election results, $v^{(\text{hyp})}$, given some arbitrary value, $\bar{v}^{(0)}$ (or the corresponding value of $\delta^{(\text{hyp})}$), for the average district vote.
2. For $j = 1, 2, \ldots, l$:
   a. Add the constant, $\bar{v}^{(j)} - \bar{v}^{(0)}$, to each component of each simulated vector, v, produced in step 1. No new simulations are required.

[21]The distributions in equations (11) and (14) have diagonal covariance matrices, so this last step requires draws from independent normal distributions.

b. Calculate $\bar{s}^{(\text{hyp})}$ for each simulated election; the set of these m values estimates the distribution of $\bar{s}^{(\text{hyp})}$, given $\bar{v}^{(\text{hyp})} = \bar{v}^{(j)}$. In particular, the sample mean and standard deviation of the m values provide estimates of $E(\bar{s}^{(\text{hyp})} | \bar{v}^{(\text{hyp})})$ and $\sqrt{\text{var}(\bar{s}^{(\text{hyp})} | \bar{v}^{(\text{hyp})})}$, respectively.

One can then plot the seats-votes curve with dotted lines at two standard deviations above and below the seats-votes curve, to indicate the region where, according to the model, 95% of predicted elections will fall.

In addition to summarizing our uncertainty in any given prediction, the standard errors make our predictions testable: we can run the model on a past election, make a prediction, and then see whether about 95% of the election outcomes fall in the 95% region. Repeating this over many past elections, we can see whether our assessments of uncertainty are accurate.

5.4. *Bias and Responsiveness*

We can calculate bias and responsiveness (along with standard errors) with the same simulations used for the seats-votes curve. We define four quantities of interest:

- Responsiveness as $\bar{v}^{(\text{hyp})}$ ranges from 0.45 to 0.55: the average difference, $[E(\bar{s}^{(\text{hyp})} | \bar{v}^{(\text{hyp})} = 0.55) - E(\bar{s}^{(\text{hyp})} | \bar{v}^{(\text{hyp})} = 0.45)]$, divided by the vote swing, $0.55 - 0.45$. If the electoral system one is studying regularly produces average district votes at or near this interval, then this is a reasonable estimate to evaluate the responsiveness of the electoral system. For example, a responsiveness of 2.3 means that a 1% increase in the average district vote for a party across districts will translate into an expected 2.3% increase in seats for the same party in the legislature (plus or minus the estimated standard error). The responsiveness is of interest to people concerned that redistricting plans will be drawn to protect incumbents from vote swings. In addition, scholars at least since Mayhew (1974) and Erikson (1971) have been interested in the responsiveness of the U.S. House and other legislatures.
- Responsiveness at a chosen value v_0 (such as the actual average district vote): the average difference, $[E(\bar{s}^{(\text{hyp})} | \bar{v}^{(\text{hyp})} = v_0 + 0.01) - E(\bar{s}^{(\text{hyp})} | \bar{v}^{(\text{hyp})} = v_0 - 0.01)]$, divided by the vote swing, 0.02. One must be careful with this measure, however, since it tracks with v_0; even if the electoral system is unchanging, a change in $\bar{v}$ moves the system to a new point on the seats-votes curve, and a new local responsiveness. On the other hand, setting v_0 to the observed $\bar{v}$ is an unquestionably realistic point.

- Average partisan bias between $\bar{v}^{(\text{hyp})} = 0.45$ and $\bar{v}^{(\text{hyp})} = 0.55$. We define partisan bias as the deviation from partisan symmetry. For example, if one party is able to translate 55% of the average district vote into 75% of the seats in the legislature, then it would be symmetric for the other party, if it were to receive 55% of the average district vote, to also receive 75% of the seats. We follow King and Browning (1987), King (1989a), and Gelman and King (1990b) and define partisan bias as the proportion of the seats in the legislature the Democrats receive over and above what is fair. For example, if partisan bias is -0.05, then the Democrats receive 5% fewer seats in the legislature than they should under the symmetry standard (and the Republicans receive 5% more seats than they should).
- Partisan bias at $\bar{v}^{(\text{hyp})} = 0.5$: the average value of $\bar{s}^{(\text{hyp})}$, given $\bar{v}^{(\text{hyp})} = 0.5$, minus 0.5; that is, $E(\bar{s}^{(\text{hyp})} | \bar{v}^{(\text{hyp})} = 0.5) - 0.5$. This is interpreted in the same manner as above, but is most appropriate for very competitive electoral systems. This partisan bias statistic is the expected proportion of the seats over 0.5 that the Democrats receive when they receive exactly half the average district vote.

In all these cases, we calculate the summary for each simulated hypothetical election result, $v^{(\text{hyp})}$, and then use the average of the m summaries as the point estimate and their standard deviation as the estimate of uncertainty (the standard error).

The two definitions of bias presented above will generally produce very similar results because it is rare to see an electoral system whose districts are all clustered around 0.5. In most situations, we prefer bias averaged from 0.45 to 0.55 because it allows for a realistic range of average district vote, $\bar{v}$. For the same reason, we prefer responsiveness measured over a 10% vote swing; again, except at the extremes, swing has little effect on the estimates.

We allow responsiveness to be defined at any value of the average district vote; 0.5 is often a natural reference point, but it is frequently more reasonable to use a value nearer to typical election outcomes. For example, the average Democratic share of the vote in U.S. House districts has fluctuated around 0.55 for decades, and many southern state legislatures have never seen the average Democratic vote fall below 0.6.

In contrast, we define partisan bias as a departure of the seats-votes curve from partisan symmetry, which only makes sense when centered at $\bar{v} = 0.5$. In a state in which the Democrats have never achieved more than $\bar{v} = 0.2$, for example, we would not trust any model to make assumptions about what might happen if they suddenly received half the vote.

This is especially important when considering minority representation (see King, Bruce, and Gelman 1993).

5.5. Other Summaries

The flexibility of our Bayesian simulation-based method allows us to estimate any summary of the predicted or hypothetical elections, not just the specific quantities discussed above.

For example, suppose we are evaluating a redistricting plan and are interested in the likely reelection rate: the proportion of incumbents, of both parties, who will win their bids for reelection. For purposes of the prediction, we shall assume that we know which incumbents will run for reelection, and that the statewide partisan swing, $\delta^{(\text{hyp})}$, is zero. The following procedure, again using the m simulated predicted elections $v^{(\text{hyp})}$, creates an estimate and standard error for the reelection rate.

1. Simulate m vectors, $v^{(\text{hyp})}$, of hypothetical elections, given the specified values of $X^{(\text{hyp})}$ and $\delta^{(\text{hyp})}$.
2. For each set of hypothetical election results, compute the proportion of incumbents who win reelection. That is, count the number of hypothetical districts with incumbents whose vote, $v_i^{(\text{hyp})}$, exceeds 0.5 (for Democrats) or is less than 0.5 (for Republicans) and divide by the number of incumbents running.
3. The resulting set of m proportions approximates the distribution of the reelection rate. If we wish, we can summarize the distribution with estimates of its expected value and corresponding standard error by calculating the mean and standard deviation of the m values.

To include a more realistic assessment of uncertainty, one should repeat the above analysis with a range of plausible values of $\delta^{(\text{hyp})}$ (or $E(\bar{v}^{(\text{hyp})}) = \delta^{(\text{hyp})} + \Sigma_{i=1}^{n}(X^{(\text{hyp})}\beta)_i$), rather than merely setting $\delta^{(\text{hyp})} = 0$. Reasonable values of average district vote may be obtained by examining a graph of $\bar{v}$ in recent elections, or more formally using a forecasting procedure. Repeating the above analysis with k different values of $\delta^{(\text{hyp})}$ yields km simulated vectors $v^{(\text{hyp})}$, from which means and standard deviations of all quantities can be estimated, as before.

For another, more complicated, example, suppose we are interested in the number of "marginal seats" in an electoral system, which we shall define (for convenience) as seats with at least one chance in four of changing parties if the election were to be repeated with the same pattern of uncontested seats and incumbents. We shall assume, to be specific, that in the hypothetical repeated election, a statewide partisan swing of up to two percentage points from the current value, $\bar{v}$, could occur in

either direction. We can estimate the number of marginal seats, so defined, as follows.

1. Perform the preliminary estimation as described in section 4.
2. Simulate β° and, if v has been observed, γ°, using steps 1 and 2 of section 5.2.
3. Perform the following steps 100 times, to obtain 100 simulated vectors, $v^{(\text{hyp})}$.
   a. Simulate one hypothetical election result, $v^{(\text{hyp})}$, with average district vote equal to the current $\bar{v}$, following step 3 of section 5.2.
   b. Add the shift, $\delta^{(\text{hyp})}$, to all districts in the vector $v^{(\text{hyp})}$, where $\delta^{(\text{hyp})}$ is drawn at random from a uniform distribution with range $[-0.02, +0.02]$.
4. For each district, count the number of times the seat changed parties (i.e., that $v_i^{(\text{hyp})} > 0.5$ if the observed $v_i < 0.5$ and vice versa). If a district changed parties in at least 25 of 100 simulations, it is "marginal." Compute the proportion of marginal districts out of the n districts.
5. Repeat steps 2–4 m times, to create m simulations of the electoral system and m corresponding simulated proportions of marginal seats in each hypothetical election. Our point estimate of the proportion of marginal seats is the average of the m proportions, and their standard error is the corresponding standard deviation over the m simulations.

6. Examples

6.1. The U.S. House of Representatives since 1900

In this section, we estimate partisan bias and electoral responsiveness for the House of Representatives in nonsouthern states during this century. To begin, we define X as including a constant term along with $inc(t)$, $unc(t)$, party(t), $v(t-1)$, $inc(t-1)$, and $unc(t-1)$. In years following redistricting (i.e., those ending in "2"), we do not include the variables corresponding to time $(t-1)$, since they are not available without a great deal of work matching and disaggregating districts between redistrictings. We ignore the fact that some states performed additional redistrictings in other years.[22]

[22]We imputed uncontested district election results at 0.25 and 0.75 and discarded the few cases of multimember districts and districts with third-party victories. Had there been many anomalous district elections, we would have to use more complicated methods to handle the missing data, but it did not seem worth the effort here.

We estimate σ^2 and λ for as many of the years as possible. (The parameter λ cannot be estimated for the final election before a redistricting.) We then pooled estimates for years following redistricting and also for all other years. For years following redistricting, our estimates are $\hat{\sigma} = 0.091$ and $\hat{\lambda} = 0.663$, and for all other years, $\hat{\sigma} = 0.065$ and $\hat{\lambda} = 0.560$. Recall that σ is a proportion, so 0.065 means that we can forecast future district-level results to within about 6.5% of the vote (plus the error in estimating β). The relative values of these parameters are reasonable: in years following redistricting, compared to other years, we have less information (fewer variables) in X; as a result, predictive uncertainty, as measured by σ, is higher. In addition, when estimating hypothetical election results, the variables in X, representing past elections, are less useful immediately following redistricting, and so λ is higher, as it should be. (The parameter λ is highest when $v^{(\text{hyp})}$ is estimated solely from v and is lowest when only X is used.)

Figure 1 plots partisan bias (with $E(\bar{v}^{(\text{hyp})})$ from 0.45 to 0.55). A standard error bar appears at the top of the graph. Most small year-to-year changes in bias are within a standard error or two, indicating stochastic variation, rather than a systematic change in the underlying electoral system. However, the long-term systemic trends are quite substantial. The evidence demonstrates unambiguous bias in favor of the Republican party from the 1930s until the mid-1960s. After that, the trend toward Democratic bias is consistent and sustained. The pattern during the first two decades of this century is less clear, and the standard errors are much larger. These substantive results are close to those produced by the more complicated, and computationally intensive, model of King and Gelman (1991) for the period they examined, 1946–86. The explanation for the trend since the mid-1960s is the growth in incumbency advantage from about 2% to nearly 12% by the late 1980s.

Figure 2 displays the estimates of electoral responsiveness (with $E(\bar{v}^{(\text{hyp})})$ set to the actual average district vote, $\bar{v}$) of the U.S. House during the same period, evaluated at the observed average district vote. For comparison, the figure also presents this same summary calculated with the assumption of uniform partisan swing. On average over many years, the two methods produce roughly the same results, but the uniform partisan swing estimate is much more variable, an indication that it is fitting transient year-to-year changes instead of the long-term systematic patterns of interest. This relative "statistical inefficiency" reflects the fact that uniform partisan swing uses only the information in v, whereas our model uses information in v as well as in the predicted values $X\beta$ (see equation 8). As a consequence, uniform partisan swing might yield reasonable estimates on average over a century, but it can be far from

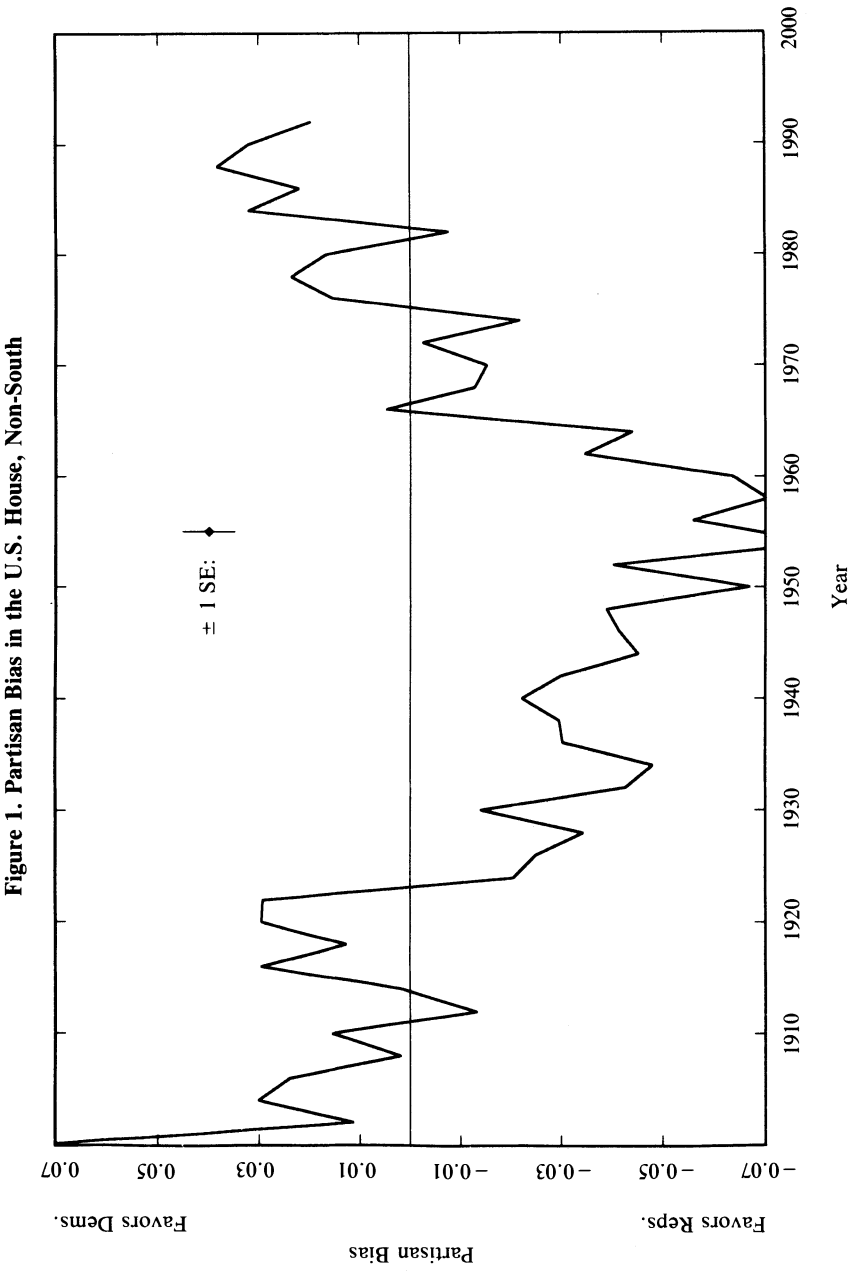


Figure 1. Partisan Bias in the U.S. House, Non-South

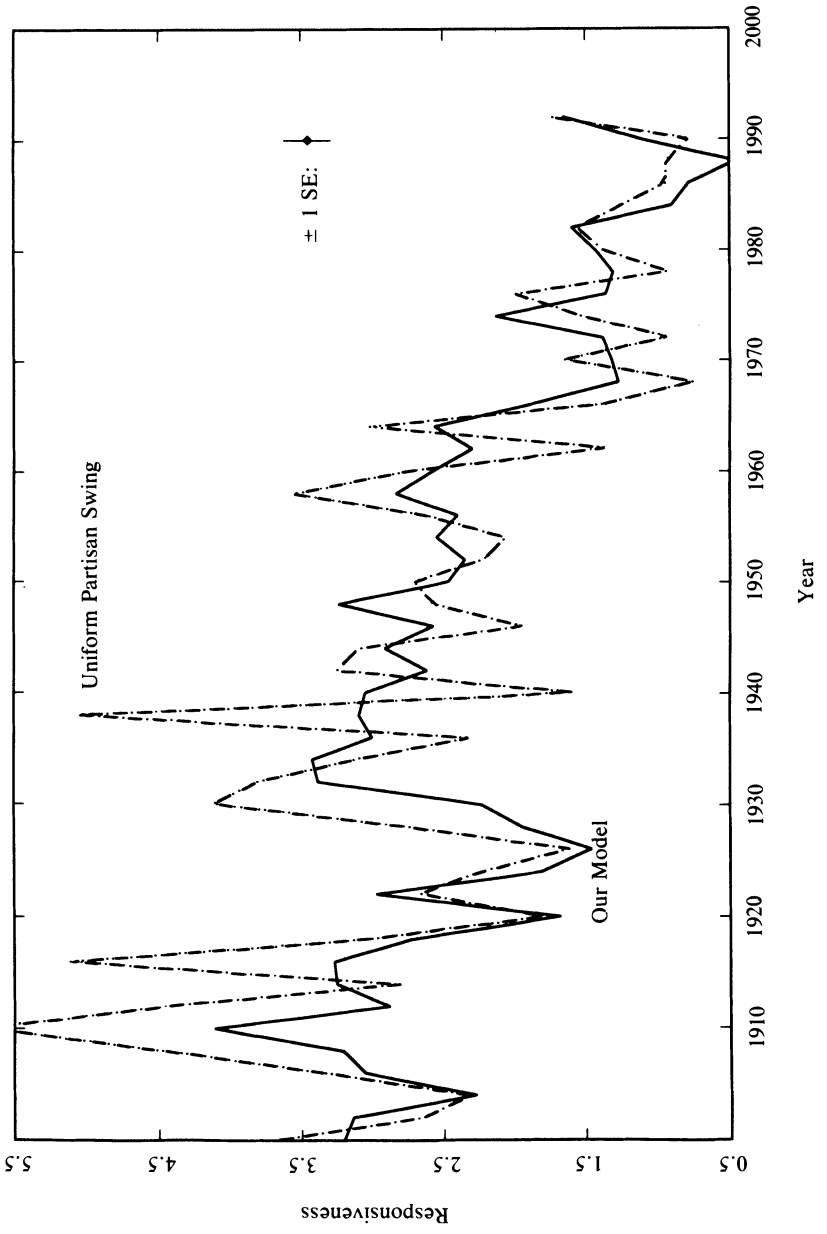


Figure 2. Responsiveness in the U.S. House, Non-South

accurate for any one year, or even as an average over a small number of years. In addition, the implied standard errors of uniform partisan swing are all zero, which corresponds to the incredible assumption that, given v and $\bar{v}^{(\text{hyp})}$, the hypothetical election outcomes, $v^{(\text{hyp})}$, are exactly known and implies that all summaries of the electoral system are known perfectly. Our method is estimated one year at a time. Thus, the smoothness of the resulting estimates is not an artifact but is due to a lower level of estimation error resulting from the much larger quantity of information extracted from the same data in each election used to make the desired inference.

Evaluating predictive uncertainty estimates. To illustrate the accuracy of forecasts made with our model, and most important, the accuracy of the uncertainty estimates implied, we predict the U.S. House of Representatives in 1988. As an example, we use as explanatory variables district election results of 1986 and incumbency and uncontestedness of each district for 1986 and 1988, variables that are typically available months before the election. In state legislative elections, our forecasts will frequently be better because better explanatory variables, such as votes for statewide offices broken down into legislative districts, are usually available. We could also improve forecasts of these U.S. House data with other variables, such as campaign spending, scandals, candidate quality, and the like, but our primary purpose here is to evaluate the model's uncertainty estimates, not to produce the best possible predictions. As usual, we only predict the *relative* district election results; for convenience, we shall assume that the nationwide partisan swing, $\delta^{(\text{hyp})}$, is zero.

To apply our model, we first estimate the parameter β by predicting 1986 from 1984, which is, in fact, the same estimation done above for evaluating the electoral system. We then draw from the distribution of predicted 1988 elections, using the estimated β and the pooled estimate of σ^2 obtained above.

For each district, i, we label the means of the simulated predictions as $\hat{v}_i$ and the actual election result as v_i. We would like to evaluate our model by comparing the prediction errors, $v_i - \hat{v}_i$, to the standard errors of the district election predictions.[23] Figure 3 plots the prediction errors versus the observed election outcomes. A solid line appears at the nationwide partisan swing, and dotted lines are displayed at plus and minus

[23]Before making this comparison, we correct for the nationwide partisan swing, which we do not try to predict with our model. From 1986 to 1988, the average Democratic district vote swung from 57.2% to 56.6%.

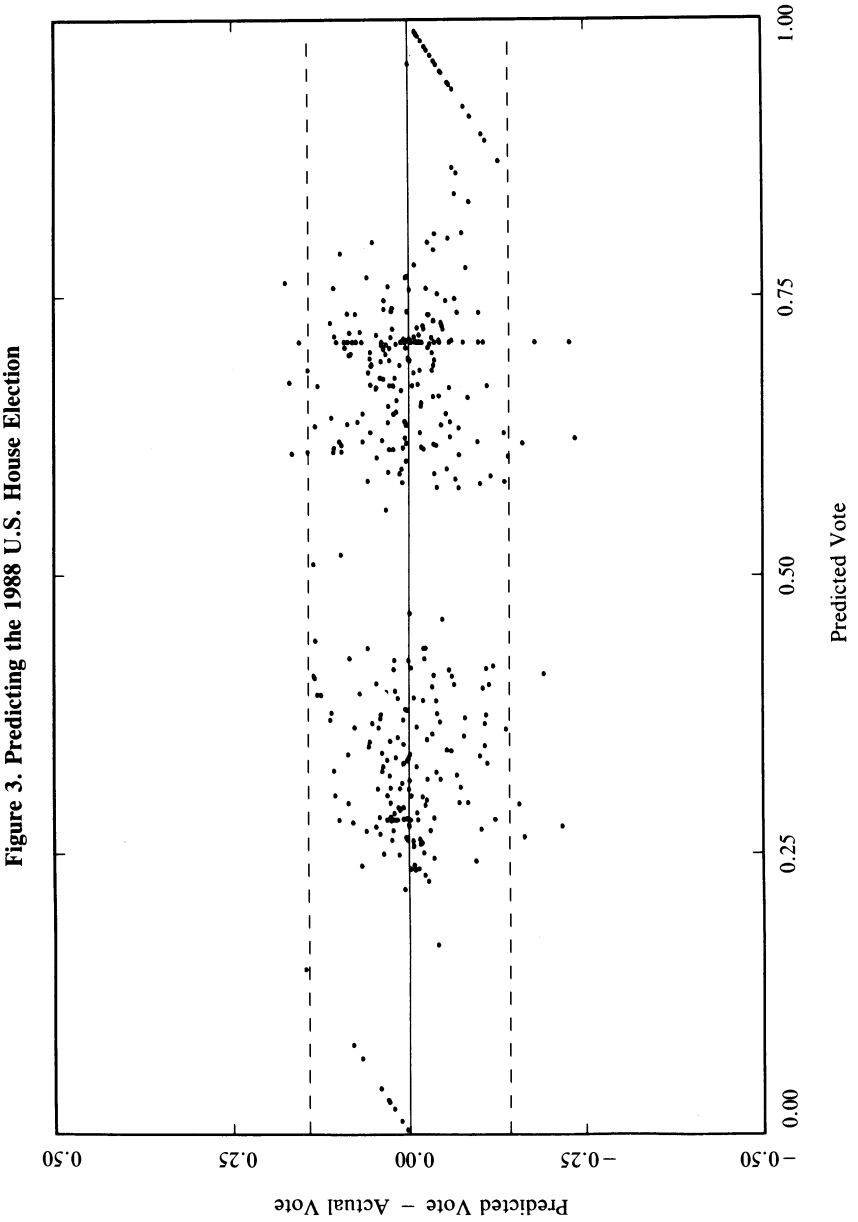


Figure 3. Predicting the 1988 U.S. House Election

twice the average predictive standard error. If the uncertainty estimates are exactly correct, 95% of the prediction errors would fall within the dotted lines on average; the actual result in this case was 96.4%.[24]

We also evaluated the accuracy of our uncertainty estimates by comparing the average of the squared actual prediction errors from the prediction model with the estimated prediction variance from the model (the average of the diagonal elements of the variance in equation 7). The average of the observed squared prediction errors was 0.065^2, which is slightly smaller than the average predicted variance, 0.071^2.

The fact that the predictive standard errors in Figure 3 and the predictive error variance just calculated are so close to their theoretical values is very strong confirmation of our model and of the accuracy of the uncertainty estimates it produces. Moreover, our extensive analyses of thousands of other elections and with different sets of explanatory variables give the same consistent support for our model and its uncertainty estimates.

Some may view the uncertainty estimates reported here as large. In part, this is because almost all existing election forecasters calculate uncertainty estimates incorrectly, and so those presented here are among the first correct estimates published (see Beck 1992; Greene 1993). They can be reduced by including explanatory variables based on more detailed knowledge of the district election results, such as campaign spending, scandals, candidate quality, relative campaign effectiveness, and the like. In our experience, the state of knowledge about elections is unlikely to reduce our uncertainty by more than half the illustrative values we report here. Most important for present purposes is that the figures reported here demonstrate that our model produces accurate uncertainty estimates. With this model and computer program, scholars can process correctly whatever explanatory variables they are innovative enough to collect.[25]

[24]If the absolute values of the prediction errors were consistently too low, our model would be overstating our predictive uncertainty, yielding overly cautious predictions. High absolute prediction errors would imply that the predictions are less accurate than expected, yielding overconfident predictions. The former wastes information; the latter risks incorrect inferences.

[25]In an earlier version of this paper, Figure 3 had four points that appeared to be true outliers because they stood more than four standard deviations away from their predicted values. In fact, these turned out to be coding rather than prediction errors, and our model does fit quite well. In practice, the existence of a few real outliers, were they to occur, would not greatly affect most of the electoral summaries we calculate, such as bias, responsiveness, number of incumbents to lose reelection, etc.

6.2. Redistricting the Ohio Legislature

Legislatures and courts are often called upon to evaluate competing redistricting plans. We show here how our method can be applied to this problem by estimating partisan bias for elections held in the lower house of the Ohio state legislature from 1968 to 1990 and by predicting partisan bias for each of three redistricting plans proposed for the legislature in 1992.

For the current districts and the proposed redistricting plans, we use the following explanatory variables: incumbency, uncontestedness, previous vote for the state House, all available statewide races broken down into current or proposed legislative districts, and the proportion of African Americans of the voting age population in the district. We estimate the parameters, β, σ^2, and λ with these explanatory variables for current districts, the latter two pooled over the years.

During the period 1968–90 to which our data apply, elections have been held under three redistricting plans. The long solid line in Figure 4 demonstrates the consequences of these redistricting plans by tracing the degree and direction of partisan bias (estimated from 0.45 to 0.55 average district vote) in the Ohio State House of Representatives. In 1971 the redistricting process was controlled by the Democratic party, which managed to implement a substantial gerrymander, one that has been found to have caused the largest effect on partisan bias ever noted in the academic literature (Gelman and King 1990b). The Democrats controlled the redistricting process in 1981 and made few changes from the 1971 plan; partisan bias did not change significantly. As a result, the electoral system in the Ohio State House favored the Democratic party consistently over the last 20 years, giving them 3%–4% more seats (and the Republicans 3%–4% fewer seats) than they would have been expected to receive in an electoral system without bias.

In the 1992 apportionment process, the Republicans controlled three of the five seats on the Apportionment Board. They delegated the mapping process almost entirely to James Tilling, secretary to the board and staff director of the Senate Republicans. According to his account, Tilling knew the current districts heavily favored the Democrats, and he set out to rectify this situation. However, he reports having decided to draw fair districts, rather than turning a Democratic gerrymander into a Republican gerrymander. His reasoning was that this would produce a fair result for the voters of Ohio, but it would also benefit his party, in comparison to the existing districts. Of course, it would not benefit the Republicans as much as it could have, so the Republican party produced a competing

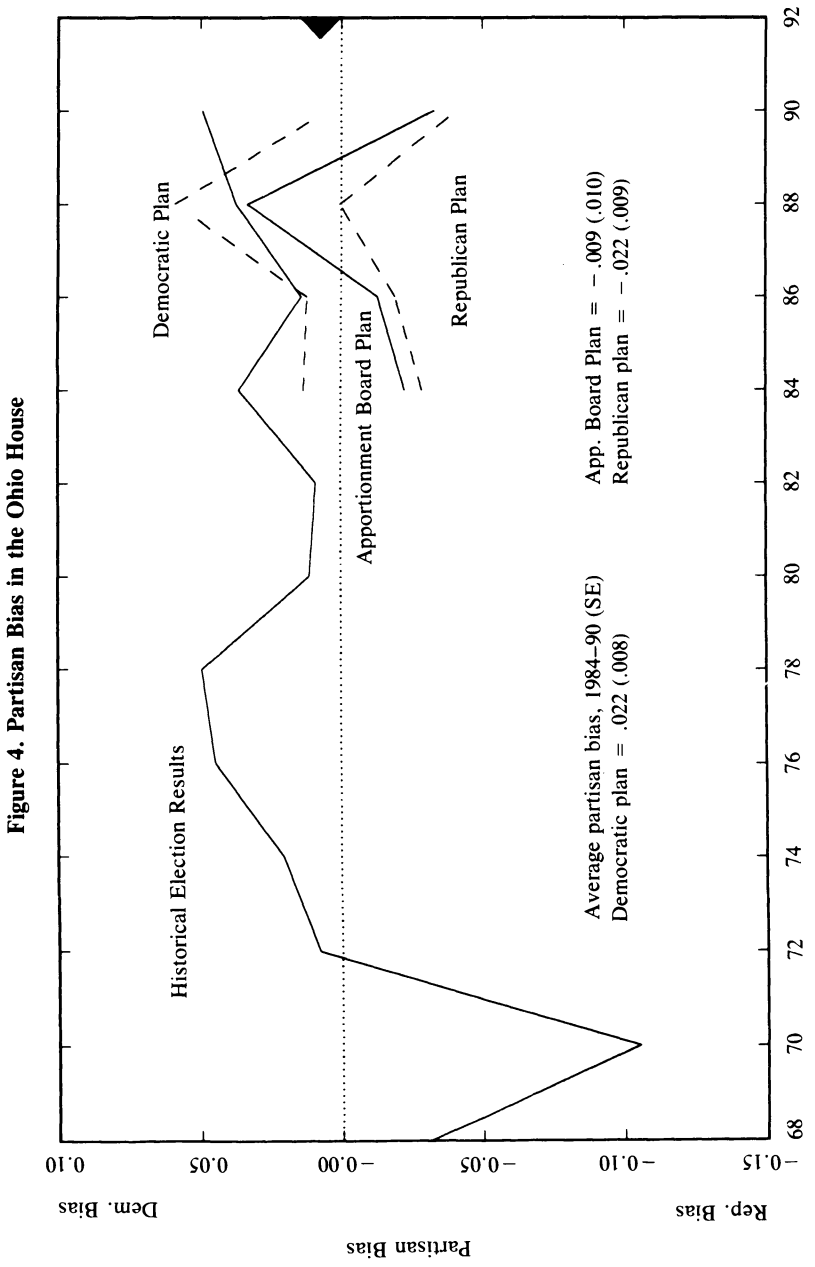


Figure 4. Partisan Bias in the Ohio House

plan and tried to get it adopted. The Democrats tried the same, but the Apportionment Board stuck with Tilling's plan.[26]

Figure 4 also displays the predicted 1992 results for the three proposed redistricting plans using corresponding explanatory variables based on reaggregated electoral results from 1984 to 1990. We use the variability in our explanatory variables from 1984 to 1990 in this way to portray our uncertainty in their actual values, and the actual electoral conditions, that could obtain in 1992. Thus, our predictions for partisan bias in 1992 are plotted on the right side of Figure 4 at each of these four election years.

As can be clearly seen, the partisans' plans favor their own parties, and the Apportionment Board plan lay in between. A summary of the results, with standard errors of these summaries, also appears on the figure.[27] On average, both the Democratic and Republican plans would bias the electoral system by about 2.2% in their favor. The Tilling plan, adopted by the Apportionment Board, is forecast to have approximately zero partisan bias. The standard errors indicate that our estimates are fairly precise; both partisan plans are clearly distinguishable from zero bias, whereas the Tilling plan cannot be distinguished from zero. Interestingly, even the Democratic plan was slightly less biased toward the Democrats than the actual results from the late 1980s. Perhaps the Democrats wanted to draw a plan that would be reasonable enough to be accepted by the courts; both parties also appeared more concerned about other issues, such as protecting the particular incumbents currently in office.

We finished the analyses in this paper well before the 1992 elections and widely distributed the results of an earlier draft; they were also presented in court and are on the public record. After the election, we estimated partisan bias from the actual election results. This result, a partisan

[26]In response to a suit brought by the Apportionment Board, the Ohio state court declared the districts constitutional according to Ohio law. The Democratic party then brought suit before a three-judge federal panel charging the board with creating a political gerrymander under the guise of the Voting Rights Act. However, the plaintiffs offered no evidence about partisan bias and, after the information in Figure 4 was presented at the trial, effectively withdrew their partisan gerrymandering claim. The Court invalidated the Apportionment Board plan, essentially on grounds of reverse discrimination. This decision was stayed by the U.S. Supreme Court, meaning that the 1992 elections for the Ohio House were held under the Ohio Apportionment Board's redistricting plan. The Supreme Court later reversed the lower court, leaving the districts in place for the rest of the decade (*Voinovich v. Quilter* 1993).

[27]The standard error of each *point* is of course larger than the standard error of the *average* reported in the figure.

bias of 0.007 (with a standard error of 0.010) differs from the predicted bias by a trivial amount, not statistically distinguishable from our prediction of zero. This result appears as a solid triangle at the right of Figure 4. This result therefore provides further evidence in support of the model and its uncertainty estimates. Moreover, an analysis of the individual district forecasts produced correct seat predictions in every district.

7. Discussion

Instead of summarizing the contributions of this paper, we focus in this concluding section on how it differs from the previous method proposed in the literature, published in King and Gelman (1991). The most important improvement in our model is the introduction of explanatory variables, which add considerable additional information to the analysis and give us substantial leverage and statistical efficiency for most problems. Including explanatory variables also enables us to integrate the study of counterfactual evaluations directly and to make predictions based on stable and verifiable relationships in existing legislative elections data. Any information lost by introducing various simplifications is more than compensated for by the information in these explanatory variables. If explanatory variables are not available, then the King and Gelman (1991) method should be used instead of the one we propose here (e.g., Jackman, forthcoming).

Because our previous model had no explanatory variables, it had to model the well-known bimodal pattern of district vote outcomes, and it did so by fitting it to a mixture of normal distributions. By requiring the inclusion of either partisan control or incumbency status as explanatory variables, we avoid this complication: conditional on these explanatory variables, voting data are unimodal and reasonably modeled by a single normal distribution and the random components regression model we introduce.

We model district vote proportions directly, without the logit transformation used in King and Gelman (1991). The loss is very slight because contested district vote proportions above 0.8 or below 0.2 are rare. This greatly simplifies the analysis, eliminating the third-order Taylor series approximations previously used to compute expected values.

Finally, we singly impute votes in uncontested districts (Appendix A), avoiding the complication of estimating a distribution for what could have happened had the elections been contested, as in King and Gelman (1991).

Since our approach can be evaluated in new legislative elections data, as we did in our examples here, it will always outperform methods of forecasting and evaluation based on plausible but unchecked assump-

tions. One example is the use of "baseline" votes, the votes for a low-visibility office elected statewide and broken down into legislative districts, for forecasting or, in combination with uniform partisan swing, for evaluations (as suggested by Backstrom, Robins, and Eller 1978).[28] Since the baseline vote can be used as an additional predictor (another column in X in our model) and since our method enables the relationship between this variable and the legislative elections of interest to be estimated rather than assumed, our approach is guaranteed to do better. In addition, our approach has the advantage of using the same model to make district vote predictions and statewide evaluations, guaranteeing internally consistent results.

While one can use this model (and the associated computer program) for many different types of analyses of legislative elections, there are a number of topics for future research that remain worthy of study. The model does not deal with multimember districts, primaries, nonpartisan elections, or multiparty elections. It is also not yet equipped to handle other types of electoral systems, such as the various types of proportional representation, more common in other countries.

Manuscript submitted 5 October 1992
Final manuscript received 10 August 1993

APPENDIX A
Vote in Uncontested Districts

Depending on the goal of the analysis, one may consider adjusting the vote, v_i, in uncontested districts. This is especially important in evaluating a redistricting plan, since one wants to evaluate the effect on actual or potential voters; decisions taken away from the voters, such as when candidate decisions make voting for one party impossible, require special treatment. Adjustment is particularly important when studying electoral systems with a small number of districts, such as a congressional delegation in one state, or with a large number of uncontested seats.

Adjustment is used to make the votes reflect what the election would have been like had the district election been contested. For prediction of election outcomes, any reasonable adjustment preserves the same result: a win for the single contesting candidate. In contrast, when we summarize an electoral system by the seats-votes curve, bias, and responsiveness, or the expected proportion of marginal seats, the assignment of votes to uncontested districts has an effect because it alters v and, thus, the average district vote, without changing the number of seats won by either party. Imputing values other than zero or one tends to reduce $\bar{v}$ for the party that wins more uncontested seats, thus increasing the apparent

[28]Statewide measures of partisan bias and electoral responsiveness, and some other issues, are discussed by King and Browning (1987), Cain (1985), Grofman (1983), King (1989a), Niemi (1985), and Niemi and Fett (1986).

partisan bias in their favor. That is, imputation reduces the number of apparently "wasted" votes in uncontested districts.

For many applications, we believe some imputation for uncontested districts is essential. One can examine the votes in districts before and after being uncontested, or look at district totals in statewide races and observe partisan strength that is generally closer to 25% and 75% than zero and one. In defining the properties of the electoral system or evaluating the effects of redistricting, it seems reasonable to be interested in how many seats a party receives as a function of its aggregate statewide support. The fact that a seat is uncontested is an indication of strong support in a district, but certainly not 100% support.

There is, of course, room for debate on what values to impute. Three relatively simple options include keeping the vote proportions at zero and one; imputing common values such as 0.25 and 0.75 for Republican and Democratic uncontested districts, respectively; or imputing values from elections in the same districts but different years.[29]

After considerable experimentation, we suggest the following method. First, we temporarily impute values 0.25 and 0.75 for Republican and Democratic uncontested seats. Second, we regress the vote v on the same explanatory variables X as in our model (equation 3). If the legislative district vote for one or more years is included among the explanatory variables, we temporarily impute 0.25 and 0.75 for these too, even though the focus of this regression is creating imputations for the dependent variable. Finally, we create an imputation for each uncontested district i by calculating $x_i\beta + \varepsilon_i$, where x_i is row i of the explanatory variable matrix X, and ε_i is a single random draw from a normal distribution with mean zero and variance estimated as part of this regression. We then use these imputations for uncontested districts, along with the actual vote in the contested districts, in all subsequent analyses.

This procedure produces more accurate and efficient imputations and is logically consistent with the rest of our model, while incorporating the uncertainty in making these imputations in any final estimate or summary. The uncertainty is introduced in our method by the addition of a random ε_i for each district instead of using only the fitted values $(x_i\beta)$, a version of Rubin's (1987) "multiple imputation" method for nonresponse in surveys.[30]

APPENDIX B
Deriving the Predictive Distribution of Future Elections

The predictive distribution of $v^{(\text{hyp})}$ discussed in section 4.1 can be derived by classical methods in econometrics. However, we present a different, Bayesian derivation of this result here in order to introduce this method in a relatively familiar case. As a result, explaining the distributions we derive for actual and counterfactual evaluation in Appendix C, which require Bayesian analyses, will be considerably easier.[31]

[29]King and Gelman (1991) arrive at the approximate value 0.75 in congressional elections to signify the strength of a party running uncontested, by taking the average of all election results in districts, in the elections preceding uncontestedness (but with the same incumbency status). That article also demonstrates how to use an entire distribution for uncontested elections to incorporate the uncertainty in this process; however, we have found that single imputation, as we recommend in this article, is simpler and does not substantially alter substantive inferences.

[30]A single imputation is sufficient in this context, since we are usually interested in summaries that average over all the districts. Independent single imputation across uncontested districts is roughly equivalent to a multiple imputation for the entire system.

[31]Bayesian analysis is based on the following principles: (1) all unknown quantities,

To simplify our derivations, we introduce the following statistical notation for the multivariate normal distribution. If a random vector, z, is normally distributed with mean vector μ, and covariance matrix Σ, we write its probability density function as

$$P(z) = N(z|\mu, \Sigma).$$

For example, the posterior distribution for β from the preliminary estimation in section 3 may be written as

$$P(\beta) \sim N(\beta|\hat{\beta}, \hat{\Sigma}_\beta). \tag{10}$$

The derivation of $P(v^{(hyp)})$ requires three essential steps, which will be repeated for the derivations in Appendix C. In step 1, we identify the distribution of $v^{(hyp)}$ given γ and β, which is given by equations (4) and (5):

$$P(v^{(hyp)}|\gamma, \beta) = N(v^{(hyp)}|X^{(hyp)}\beta + \gamma + \delta^{(hyp)}, (1 - \lambda)\sigma^2 I).$$

We cannot use this distribution directly because it conditions on two unknowns, γ and β. (Recall that λ and σ^2 are assumed known after estimation.) The remaining two steps in this derivation average over the distributions of these two parameters.

In step 2, we average over the distribution of γ, which is simply $P(\gamma|\beta) = N(\gamma|0, \lambda\sigma^2 I)$, from equation (4).

$$\begin{aligned} P(v^{(hyp)}|\beta) &= \int_{-\infty}^{\infty} P(v^{(hyp)}|\gamma, \beta)P(\gamma|\beta)d\gamma \\ &= N(v^{(hyp)}|X^{(hyp)}\beta + \delta^{(hyp)}, \sigma^2 I). \end{aligned} \tag{11}$$

Finally, in step 3, we average over the uncertainty in β (equation 10), using the same rule from probability theory:

$$P(v^{(hyp)}) = \int_{-\infty}^{\infty} P(v^{(hyp)}|\beta)P(\beta)d\beta \tag{12}$$

$$= \int_{-\infty}^{\infty} N(v^{(hyp)}|X^{(hyp)}\beta + \delta^{(hyp)}, \sigma^2 I)N(\beta|\hat{\beta}, \Sigma_\beta)d\beta$$

$$= N(v^{(hyp)}|X^{(hyp)}\hat{\beta} + \delta, X^{(hyp)}\Sigma_\beta X^{(hyp)\prime} + \sigma^2 I). \tag{13}$$

The last line here is the final unconditional distribution of $v^{(hyp)}$, which we use to calculate the hypothetical election results, as in Table 1, and then the distributions of various summaries of the electoral system.

APPENDIX C
Deriving the Distribution under Actual or Counterfactual Electoral Conditions

We now derive the distribution of $P(v^{(hyp)}|v)$ in a manner directly analogous to that in Appendix B. In step 1, we identify the distribution of $v^{(hyp)}$ given γ and β and, unlike prediction, observed votes v. Actually, once γ and β are known, v gives no additional

---

including $v^{(hyp)}$, β, and so on, are treated as random variables, and they have a joint probability distribution; (2) if we have two variables (or sets of variables), A and B, and we have observed B, then we are interested in the conditional distribution of A, given B: $P(A|B) = P(A, B)/P(B)$; (3) if two variables (or sets of variables), A and B, are both unknown, then the distribution of A alone is $P(A) = \int P(A, B)dB = \int P(A|B)P(B)dB$.

information about $v^{(\text{hyp})}$ because ε and $\varepsilon^{(\text{hyp})}$ are independent, and so the distribution of $v^{(\text{hyp})}$ is still given by equation (5):

$$P(v^{(\text{hyp})}|\gamma, \beta, v) = N(v^{(\text{hyp})}|X^{(\text{hyp})}\beta + \gamma + \delta^{(\text{hyp})}, (1 - \lambda)\sigma^2 I). \tag{14}$$

Step 2 requires averaging over γ, but in this case the distribution can be improved by conditioning on v—the key difference between this derivation and that for prediction in Appendix B. We condition on v with Bayes's theorem:

$$\begin{aligned} P(\gamma|\beta, v) &= \frac{P(v|\gamma, \beta)P(\gamma|\beta)}{P(v|\beta)} \\ &\propto N(v|X^{(\text{hyp})}\beta + \gamma, (1 - \lambda)\sigma^2 I)N(\gamma|0, \lambda\sigma^2 I) \\ &= N(\gamma|\lambda(v - X^{(\text{hyp})}\beta), \lambda(1 - \lambda)\sigma^2 I). \end{aligned} \tag{15}$$

To complete step 2, we use this result to integrate out γ:

$$\begin{aligned} P(v^{(\text{hyp})}|\beta, v) &= \int_{-\infty}^{\infty} P(v^{(\text{hyp})}|\gamma, \beta, v)P(\gamma|\beta, v)d\gamma \\ &= N(v^{(\text{hyp})}|\lambda v + (X^{(\text{hyp})} - \lambda X)\beta + \delta^{(\text{hyp})}, (1 - \lambda^2)\sigma^2 I). \end{aligned} \tag{16}$$

Finally, we integrate out β for step 3:

$$\begin{aligned} P(v^{(\text{hyp})}|v) &= \int_{-\infty}^{\infty} P(v^{(\text{hyp})}|\beta, v)P(\beta|v)d\beta \\ &= N(v^{(\text{hyp})}|\lambda v + (X^{(\text{hyp})} - \lambda X)\hat{\beta} + \delta^{(\text{hyp})}, \\ &\quad (1 - \lambda^2)\sigma^2 I + (X^{(\text{hyp})} - \lambda X)\Sigma_\beta(X^{(\text{hyp})} - \lambda X)'). \end{aligned}$$

REFERENCES

Alford, John, and David W. Brady. 1988. "Partisan and Incumbent Advantage in U.S. House Elections, 1846–1986." Working paper 11. Rice University.

Backstrom, Charles, Leonard Robins, and Scott Eller. 1978. "Issues in Gerrymandering: An Exploratory Measure of Partisan Gerrymandering Applied to Minnesota." *Minnesota Law Review* 62:1121–59.

Beck, Nathaniel. 1992. "Forecasting the 1992 Presidential Election: The Message Is in the Confidence Interval." *Public Perspective* 3:32–34.

Benjamin, Gerald, and Michael J. Malbin. 1992. *Limiting Legislative Terms*. Washington, DC: Congressional Quarterly Press.

Born, Richard. 1985. "Partisan Intentions and Election Day Realities in the Congressional Redistricting Process." *American Political Science Review* 79:305–19.

Box, George E. P., and George C. Tiao. 1973. *Bayesian Inference in Statistical Analysis*. Reading, MA: Addison-Wesley.

Brady, David W. 1988. *Critical Elections and Congressional Policy Making*. Stanford: Stanford University Press.

Butler, David E. 1951. Appendix. In *The British General Election of 1950*, ed. H. G. Nicholas. London: Macmillan.

Cain, Bruce. 1985. "Assessing the Partisan Effects of Redistricting." *American Political Science Review* 79:320–33.

Campbell, James E. 1992. "Divided Government, Partisan Bias, and Turnout in Congres-

sional Elections: Do Democrats Sit in the 'Cheap Seats'?" Photocopy, Louisiana State University.

Edgeworth, Francis Y. 1898. "Miscellaneous Applications of the Calculus of Probabilities." *Journal of the Royal Statistical Society* 61:534–44.

Erikson, Robert S. 1971. "The Advantage of Incumbency in Congressional Elections." *Polity* 3:395–405.

Ferejohn, John A. 1977. "On the Decline of Competition in Congressional Elections." *American Political Science Review* 71:166–76.

Fiorina, Morris. 1977. *Congress: Keystone of the Washington Establishment*. New Haven: Yale University Press.

———. 1992. *Divided Government*. New York: Macmillan.

Gelman, Andrew, and Gary King. 1990a. "Estimating Incumbency Advantage without Bias." *American Journal of Political Science* 34:1142–64.

———. 1990b. "Estimating the Electoral Consequences of Legislative Redistricting." *Journal of the American Statistical Association* 85:274–82.

Goldberger, Arthur S. 1991. *A Course in Econometrics*. Cambridge: Harvard University Press.

Greene, Jay P. 1993. "Forewarned before Forecast: Presidential Election Forecasting Models and the 1992 Election." *PS* 26:17–21.

Grofman, Bernard. 1983. "Measures of Bias and Proportionality in Seats-Votes Relationships." *Political Methodology* 9:295–327.

Grofman, Bernard, Lisa Handley, and Richard Niemi. 1992. *Minority Representation and the Quest for Voting Equality*. New York: Cambridge University Press.

Hanushek, Eric, and John Jackson. 1977. *Statistical Methods for Social Scientists*. New York: Academic Press.

Jackman, Simon. N.d. "Measuring Electoral Bias: Australia, 1949–1993." *British Journal of Political Science*. Forthcoming.

Jacobson, Gary C. 1980. *Money in Congressional Elections*. New Haven: Yale University Press.

———. 1987. "The Marginals Never Vanished: Incumbency and Competition in Elections to the U.S. House of Representatives." *American Journal of Political Science* 31: 126–41.

———. 1990. *The Electoral Origins of Divided Government: Competition in U.S. House Elections, 1946–1988*. Boulder, CO: Westview Press.

King, Gary. 1989a. "Representation through Legislative Redistricting: A Stochastic Model." *American Journal of Political Science* 33:787–824.

———. 1989b. *Unifying Political Methodology*. New York: Cambridge University Press.

———. 1991a. "Stochastic Variation: A Comment on Lewis-Beck and Skalaban's 'The R-Square.'" *Political Analysis* 2:185–200.

———. 1991b. "'Truth' Is Stranger than Prediction, More Questionable than Causal Inference." *American Journal of Political Science* 35:1047–53.

King, Gary, and Robert X. Browning. 1987. "Democratic Representation and Partisan Bias in Congressional Elections." *American Political Science Review* 81:1251–76.

King, Gary, John Bruce, and Andrew Gelman. 1993. "Standards of Racial Fairness in Legislative Redistricting." In *Classifying by Race*, ed. Paul E. Peterson.

King, Gary, and Andrew Gelman. 1991. "Systemic Consequences of Incumbency Advantage in U.S. House Elections." *American Journal of Political Science* 35:110–38.

Lewis-Beck, Michael S., and Tom W. Rice. 1992. *Forecasting Elections*. Washington, DC: Congressional Quarterly Press.

Mayhew, David R. 1974. "Congressional Elections: The Case of the Vanishing Marginals." *Polity* 6:295–317.

———. 1991. *Divided We Govern: Party Control, Lawmaking, and Investigations, 1946–1990*. New Haven: Yale University Press.

Niemi, Richard G., and Patrick Fett. 1986. "The Swing Ratio: An Explanation and an Assessment." *Legislative Studies Quarterly* 11:75–90.

Rothstein, Paul, and John B. Gilmour. 1992. "Term Limitation in a Dynamic Model of Partisan Balance." Photocopy, Washington University, St. Louis.

Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Schrodt, Philip A. 1981. "A Statistical Study of the Cube Law in Five Electoral Systems." *Political Methodology* 7:31–53.

Tufte, Edward. 1973. "The Relationship between Seats and Votes in Two-Party Systems." *American Political Science Review* 67:540–54.

American Economic Review 2008, 98:1, 113–144
http://www.aeaweb.org/articles.php?doi=10.1257/aer.98.1.113

Optimal Gerrymandering: Sometimes Pack, But Never Crack

By John N. Friedman and Richard T. Holden*

Standard intuitions for optimal gerrymandering involve concentrating one's extreme opponents in "unwinnable" districts ("packing") and spreading one's supporters evenly over "winnable" districts ("cracking"). These intuitions come from models with either no uncertainty about voter preferences or only two voter types. In contrast, we characterize the solution to a problem in which a gerrymanderer observes a noisy signal of voter preferences from a continuous distribution and creates N districts of equal size to maximize the expected number of districts she wins. Under mild regularity conditions, we show that cracking is never optimal—one's most ardent supporters should be grouped together. Moreover, for sufficiently precise signals, the optimal solution involves creating a district that matches extreme "Republicans" with extreme "Democrats," and then continuing to match toward the center of the signal distribution. (*JEL* D72)

One of the more curious features of American democracy is that electoral boundaries are drawn by political parties. In order to ensure a notion of equal representation, the Constitution of the United States provides that "Representatives and direct Taxes shall be apportioned among the several States which may be included within this Union, according to their respective Numbers."[1] Since populations change over time, the Constitution also provides a time frame according to which representation shall be adjusted—"… within every subsequent Term of ten Years, in such Manner as they shall by Law direct"—where "they" represents the states. In practice, this leaves the process of redistricting to state legislatures and governors.

History has shown that political parties act in their own interests; redistricting is no exception, and the advantages gained can be large. From Massachusetts's Elbridge Gerry in 1812 (after whom the term "Gerrymander" was coined) to the recent actions of Texas Representative Tom DeLay, American politicians have used the redistricting process to achieve partisan political ends. Most recently, the much publicized Republican redistricting in Texas in 2003 caused four Democratic congressmen to lose their seats and would have been even more extreme but for the Voting Rights Act, which effectively protected nine Democratic incumbents. Other particularly stark current examples include Florida, Michigan, and Pennsylvania—states that are evenly divided, but whose delegations to the 109th Congress collectively comprised 39 Republicans and 20 Democrats. Democrats are also familiar with the practice; although President George W. Bush won Arkansas by more than 10 points in 2004, the state's delegation to the 109th Congress,

* Friedman: University of California at Berkeley, Evans Hall, 5th floor, Berkeley, CA 94720 (e-mail: jfriedman@post.harvard.edu); Holden: Massachusetts Institute of Technology, Sloan School of Management, E52-410, 50 Memorial Drive, Cambridge, MA 02142 (e-mail: rholden@mit.edu). We owe special thanks and a large intellectual debt to Paul Milgrom. We would also like to thank three anonymous referees, Philippe Aghion, Alberto Alesina, Doug Bernheim, Steve Coate, Eddie Dekel, Rosalind Dixon, Allan Drazen, Glenn Ellison, Drew Fudenberg, Luis Garicano, Matt Gentzkow, Ed Glaeser, Christine Jolls, Kevin Murphy, Barry Nalebuff, Torsten Persson, Jesse Shapiro, Andrei Shleifer, and Jeremy Stein for helpful suggestions and participants in seminars at UC Berkeley, University of Chicago, Cornell University, Harvard Law School, Harvard University, MIT, Northwestern University, Stanford University, University of Pennsylvania, Yale University, and the NBER political economy group meetings. Friedman acknowledges the support of the NSF and NBER grant NIA T32-AG00186.

[1] Article I, section 2, clause 3.

bolstered by the Democratic state legislature's redistricting in 2001, contained three Democrats and one Republican.

Although gerrymandering using unequal district sizes is unlawful, partisan gerrymandering remains legal, though controversial. In *Davis v. Bandemer* (1986), the Supreme Court declared partisan gerrymandering inimical to norms of fair and equal representation; but the majority was unable enunciate a workable test for where redistricting stops and gerrymandering begins. Nearly two decades later, despite numerous attempts to find such a standard, four members of the court (Chief Justice Rehnquist and Justices O'Connor, Scalia, and Thomas) found in *Vieth v. Jubelirer* (2004) (a 4-1-4 decision) that the test laid down in *Bandemer* was not practicable, in that it gave no guidance to legislatures and lower courts, and, absent such a test, partisan redistricting was not justiciable.[2]

In the wake of this decision and the controversial Texas redistricting in 2003, there has been renewed interest in legislative reform to change the partisan nature of redistricting. Currently, two states, Iowa (since 1980) and Arizona (since 2000), include nonpartisan commissions in their decennial redistricting processes, but only Arizona completely excludes political bodies. More than 20 states have considered similar amendments in the past decade, though, and movements advocating such changes seem to be gaining momentum.

Recently, three states, California, Florida, and Ohio, held referenda that proposed that panels of retired judges take charge of the redistricting process. None of these passed. But despite the great impact of gerrymandering on the American political system and the surge of recent interest in reform, few authors have attempted to understand the basic incentives at work.

In this paper, we view the issue of redistricting through the lens of an economist concerned with the endogenous formation of political institutions. In particular, we frame the issue as a maximization problem by the gerrymanderer where the choice variables are the allocations of voters to districts. In contrast, most previous analyses model the problem as a trade-off between "biasedness"—the degree to which an evenly divided population would elect an uneven slate of legislators—and "responsiveness"—the sensitivity of the share of seats held by a party to the share of voters supportive of that party (Guillermo Owen and Bernard Grofman 1988; Katerina Sherstyuk 1998; Gary W. Cox and Jonathan N. Katz 2002). In these models, the gerrymanderer optimally concentrates those least likely to vote for her in districts that are "thrown away" or "packed," and spreads remaining voters evenly over the other districts, which are "smoothed" or "cracked." A major limitation of these models is that they are not micro-founded; the gerrymanderer chooses properties of the redistricting plan, as a whole, rather than the placement of voters into districts. Since there is no one-to-one mapping from these aggregate characteristics to individual district profiles, there is no guarantee that the solution from these models is actually optimal.

Thomas W. Gilligan and John G. Matsusaka (1999) take an alternative approach, instead analyzing a micro-founded model in which individuals with known party affiliations vote for those parties with probability one. Since one party wins a district comprising $n + 1$ of its supporters and n opponents with certainty, the optimal strategy is to make as many districts like this as possible. Indeed, if one party holds a bare majority of the population, then they win all districts! Though the assumptions of observability and deterministic voting simplify the analysis greatly, they clearly do so at some cost.

Kenneth W. Shotts (2002) considers the impact of majority-minority districting. He develops a model with a continuum of voters whose identities are perfectly known to the gerrymanderer, and imposes a constraint he calls the "minimum density constraint." This requires the gerrymanderer to put a positive measure of all voter types in each district. This is a reduced form way of

[2] "… the legacy of the plurality's test is one long record of puzzlement and consternation," Scalia J.

modelling the constraint that districts be contiguous and the fact that in practice the gerrymanderer receives a noisy signal of voter preferences.

We analyze a model in which there is a continuum of voter preferences, and where the gerrymanderer observes a noisy signal of these preferences. We show that the optimal strategy always involves concentrating one's most ardent supporters together. Intuitively, since district composition determines the median voter, smoothing districts makes inefficient use of extreme Republicans as right-of-the-median voters in many districts, rather than having them be the median in some districts. This contrasts with the "cracking" intuition, which calls for the creation of identical profiles among districts the gerrymanderer expects to win. When the signal a gerrymanderer receives is sufficiently precise, we obtain a sharper characterization. The optimal strategy creates districts by matching increasingly extreme blocks of voters from opposite tails of the signal distribution. Intuitively, extreme Democrats can be best neutralized by matching them with a slightly larger mass of extreme Republicans.

This analysis is a first step toward a more complete understanding of the phenomenon of gerrymandering. There are important issues this paper does not address. Most notably, we abstract from geographical considerations, such as the legal requirement of contiguity (see Section I below, however), as well a preference for compactness or the recognition of communities of interest. Second, we focus exclusively on partisan incentives, to the exclusion of the motivations of incumbents (i.e., incumbent gerrymandering). Finally, we do not model the constraints imposed by the Voting Rights Act. Of course, this does not mean that racial and partisan gerrymandering are distinct phenomena. Given that race is a component of the signal of voter preference observed by the gerrymanderer, there may be circumstances where they are essentially the same practice. Ultimately this is an empirical question, which depends on the joint distribution of voter preferences and voter characteristics. (These issues are further explored in Section VI).

The remainder of the paper is organized as follows, Section I details the legal and institutional backdrop against which redistricting takes place. In Section II we present some basic examples to illustrate the primary intuitions of the solution to our more general model, which we present in Section III along with comparative statics. Section IV reports the result of a number of numerical examples of the model in order to illustrate further the optimal strategy and its comparative statics. In Section V we explore a number of extensions to the basic model, including alternative partisan objective functions, the effects of gerrymandering on policy outcomes, candidate specific advantages, and uncertain voter turnout. Finally, Section VI contains some concluding remarks and suggests directions for future work.

I. Institutional Background[3]

The process of redistricting was politicized in America as early as 1740 (in favor of the Quaker minority in the colony of Pennsylvania). Until the landmark Supreme Court decision *Baker v. Carr* in 1962, the major legal constraint on gerrymandering was that districts be contiguous. Many states, particularly in the South, had not redrawn Congressional districts after each decennial Census. Since population growth was much greater in urban areas, this inertia served to dilute the urban vote—often poor and black—and enhance the political power of rural white voters who traditionally supported the Democratic Party. After the 1960 Census, the population disparities between congressional districts had become as great as 3 to 1 in Georgia (and as extreme as 1,000 to 1 for state legislature seats in some states). The decision in *Baker* declared that challenges to such districting plans were justiciable, and two years later the Court clarified its

[3] This section details the legal and political backdrop against which gerrymandering occurs today. Readers uninterested in, or already familiar with, this material may wish to skip directly to the analysis in Section II.

position on the standard for unlawful redistricting plans, stating in *Wesberry v. Sanders* that only congressional districts with populations "as nearly equal as possible" were acceptable under the Equal Protection clause.[4] Furthermore, federal district courts were empowered, as part of their remedial discretion, to draw district boundaries themselves should a state prove either unable or unwilling to produce a satisfactory plan.

Consensus over the practical implications of the Court's decisions solidified over the next 15 years. Though federal district courts initially experimented with strict upper bounds on the maximum population deviation across districts, by the late 1970s states were subject to a more flexible set of criteria, in which concerns such as the compactness of districts or the preservation of "communities of interest" justified small deviations in representation. As of 1980, though, contiguity and population equality across districts were the principle constraints on redistricting.

In the 1990s, debates around gerrymandering shifted to the issue of "race conscious" redistricting. While it had long been clear that intentional dilution of the voting strength of racial minorities violated the Equal Protection clause, it was less clear that states could draw boundaries such that racial minorities could elect their preferred candidates (Samuel Issacharoff, Pamela S. Karlan, and Richard H. Pildes 2002). In a number of cases, culminating in *Shaw v. Reno* (1993), the Court found that redistricting plans would be held to the same strict scrutiny with respect to race as other state actions. In practice, this means that, once plaintiffs demonstrate that racial concerns were a "predominant factor" in the design of a districting plan, the plan is illegal unless the state can justify the use of race and show that such factors were considered only when necessary. This places a heavy burden on the states. Some federal courts initially interpreted these decisions as requiring states to ensure minority representation through the creation of majority-minority districts, but the Supreme Court declared that this practice would violate Section 2 of the Voting Rights Act. In more recent cases, the Court has continued to downplay the importance of racial considerations; for instance, litigation surrounding the 1991 North Carolina redistricting ended when the Court ruled, in *Easley v. Cromartie* (2001), that partisan concerns, not racial concerns, "predominated" in the construction of the heavily black and Democratic 12th district, and thus the plan was legal.

The history of attempts to ban partisan gerrymandering have proven less successful still. In *Davis v. Bandemer*, the Supreme Court attempted to limit the impact of partisan concerns in redistricting processes by stating that such claims were, in theory, justiciable (though they did not decide the merits). Though the years following this decision saw many attempts to define the level and shape of such a standard, there was little agreement, and no claim of partisan gerrymandering ever succeeded. In *Vieth v. Jubelirer*, four members of the Court found that such attempts were doomed. While *Bandemer* is still good law, the future justiciability of partisan gerrymandering claims seems far from assured.

The current reality of political redistricting reflects the past 40 years of case history. States now use increasingly powerful computers to aid in the creation of districts, and, accordingly, *Baker*'s "as nearly equal as possible" population requirement is extremely strict. A Pennsylvania redistricting plan was struck down in 2002 for having one district with 19 more people than another without justification! On the other hand, the law does allow for some slight deviations, provided there is adequate justification. In Iowa, for instance, congressional districts must comprise whole counties; the current maximum population deviation of the Iowa redistricting plan is 131 people, but the legislature rejected an earlier plan with a 483-person deviation. Such cases are not common, though. The current Texas districting plan is more representative and has,

[4] See *Wesberry v. Sanders* 376 US 1 (1964). The court applied a similar standard to districts for statewide legislative bodies in *Reynolds v. Sims* 377 US 533 (1964), and for general purpose local governments in *Avery v. Midland County* 390 US 474 (1968).

to integer rounding, equal population in each district.

As previously mentioned, districts must be contiguous. This requirement first appears in the Apportionment Act of 1842, though it was standard long before then. While technology has tightened the population equality constraint, computers have effectively loosened the contiguity requirement, as legislators can now draw districts more finely than ever before. In the 1970s, districting plans were laborious to create and difficult to change, as each required hours of drawing on large floor-maps using dry-erase markers; now lawmakers use Census TIGERLine files to create and analyze many alternative districting schemes both quickly and accurately. Contiguity has been stretched to the limit in such recent cases. Florida's 19th, 22nd, and 23rd districts, shown in Figure 1, are one such case. The 22nd comprises a coastal strip not more than several hundred meters wide in some places but 90 miles long, while tentacles from the 19th and 23rd intertwine to divide the voters of West Palm Beach and Fort Lauderdale. Even more striking is the shape of the Illinois 4th (shown in Figure 2), drawn to include large Hispanic neighborhoods in the North and South of Chicago but not much in between. Each of these districts is, in some places, no more than one city block wide, and such necks are often narrower than 50 meters.

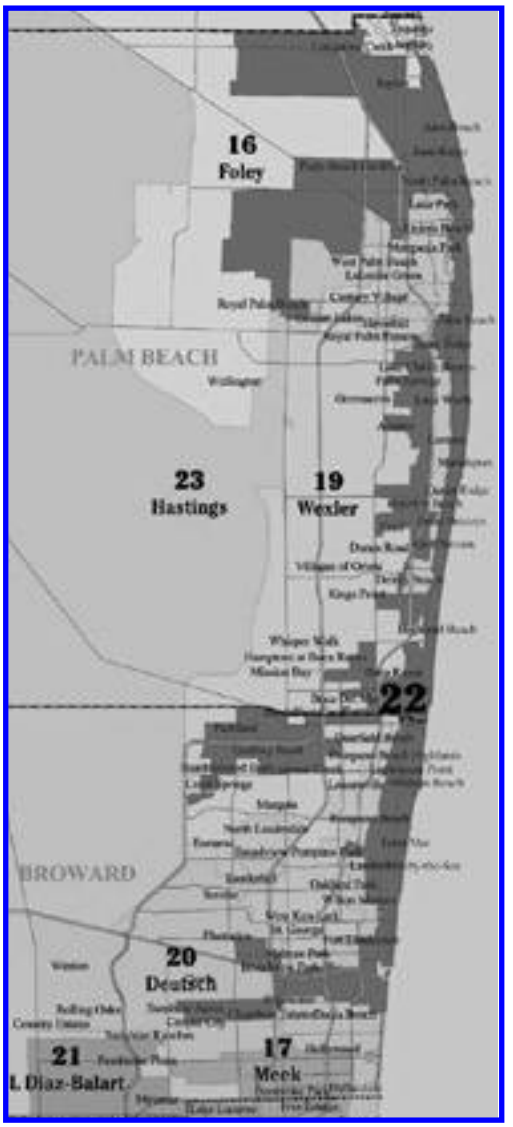


FIGURE 1.
FLORIDA 16TH TO 23RD CONGRESSIONAL DISTRICTS

State law governs procedures for redrawing district boundaries. In most states, redistricting plans are standard laws, proposed by the members of the legislature and subject to approval by the legislatures and the governor. Arizona and Iowa delegate redistricting to independent commissions, though in Iowa legislators must still approve the plan and may edit proposed schemes after several have been rejected. In 2001, for instance, the legislature rejected the first proposed plan along partisan lines.[5] Arizona and Iowa also instruct their redistricting commissions to make districts "compact," respect the boundaries of existing "communities of interest," and use geographic features and existing political boundaries to delineate districts "to the extent practicable." Finally, Arizona mandates that "competitive districts should be favored where to do so would create no significant detriment" to other objectives.[6] No other states have explicitly defined redistricting goals along these lines.

[5] "Senate Rejects Districts," *Des Moines Telegraph Herald*, May 3, 2001.
[6] See Arizona Proposition 106, and 1981 Iowa Acts, 2nd Extraordinary Session, Ch. 1.

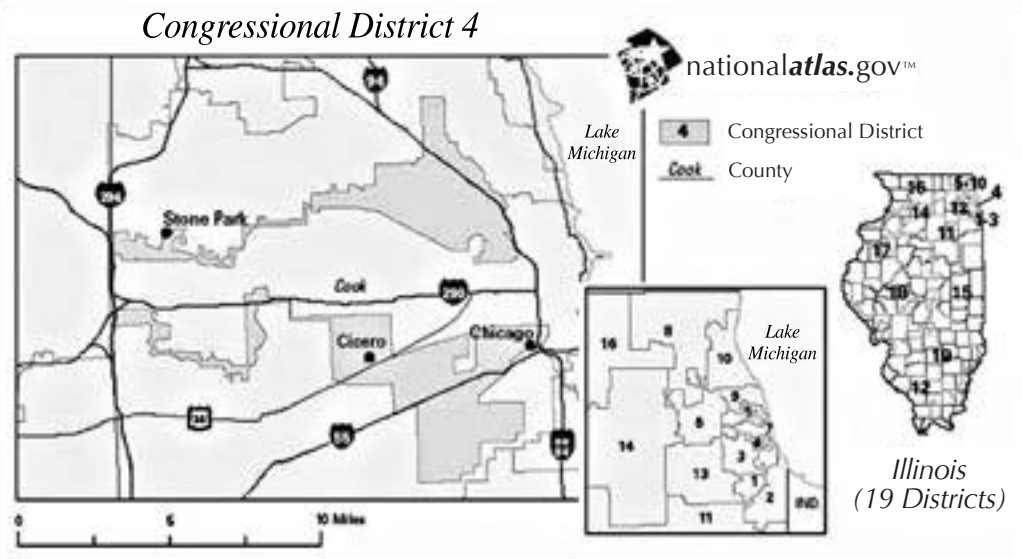


FIGURE 2. ILLINOIS 4TH CONGRESSIONAL DISTRICT

There are three key messages to understand from the backdrop against which gerrymandering takes place. First, contiguity may well not be a binding constraint because of the fine lines gerrymanderers use to create districts. Second, other spatial/geographic concerns such as compactness and communities of interest have found little legal traction. As such, they are really not constraints on gerrymanderers. Third, the Supreme Court has consistently considered partisan and racial gerrymandering to be analytically distinct—*Cromartie* even going so far as to allow racial gerrymandering if it is not deemed the predominant motive. The first two of these points suggest that spatial/geographic considerations are not first-order concerns. Accordingly, our model omits them. The third rests on the premise that signals of voting propensity and race are sufficiently uncorrelated that an optimal gerrymandering strategy does not conflate the two issues. This is a point to which we return later in the paper.

II. Some Simple Examples

In order to illustrate the intuition behind the theory in this paper, we now provide some very simple examples that capture the basic features of the more general model in Section III. In these examples, for simplicity, voters have single-peaked preferences. In the general model, voter preferences satisfy single-crossing—an arguably less restrictive condition. For instance, when voters have a convex loss function over the distance of their bliss point from the actual policy, then single-peakedness implies that single-crossing is satisfied.

A. *Example 1*

Consider the problem faced by a gerrymanderer in a state in which a population of voters has single-peaked preferences that are symmetric about a policy β, within a one-dimensional policy space. We assume that each voter has bliss point β, and that, across the population, β is

distributed uniformly on $[-1, 1]$. These assumptions imply that, in a two-party election, each voter supports the candidate located closest to her on the ideological spectrum. To begin, we assume that the gerrymanderer can directly observe β for each voter. We assume that all candidates—the right-wing "Republican" candidates and the left-wing "Democrats"—locate symmetrically about zero, and so the percent of votes captured by the Republican candidate in any election is simply the proportion of voters to the right of zero.

The gerrymanderer—suppose she is a Republican—must break up the population into equal-sized districts in which different elections take place with the goal of maximizing the expected number of seats won by her party. Since we abstract from geographic concerns here, the gerrymanderer can match any pieces of the population into a district. Suppose, for simplicity, that the gerrymanderer must form two districts, so that each district must comprise a one-half mass of voters. Since all voters for whom $\beta \geq 0$ support the Republican candidate with certainty, Republicans win all districts containing one-quater or more mass of such voters.[7] From Gilligan and Matsusaka (1999), the optimal gerrymander makes exactly half of the voters in each district have preferences $\beta \geq 0$; in this basic setup, Republicans win each district with certainty. It does not matter which right-wing voters go into each district.

B. *Example 2*

We now add some noise to the preferences in Example 1. Suppose that, after candidates are positioned, an aggregate preference shock A affects the population so that preferences are now single-peaked about $\hat{\beta} = \beta - \text{A}$. The gerrymanderer observes only β and not A or $\hat{\beta}$. Suppose that A is distributed uniformly on $[-1, 1]$. While voters for whom $\beta > 0$ now vote for the right-wing candidate in expectation, only those for whom $\beta = 1$ support the Republican candidate with certainty; a voter with $\beta = 0.5$, for instance, prefers the right-wing candidate only if $A < 0.5$, which happens 75 percent of the time.

In this example we can make a sharper prediction about the form of the optimal gerrymander. Half of the voters in each district should have $\beta > 0$, but it now matters which of these voters go into which district. The optimal gerrymander groups all extreme voters for whom $\beta \in [0.5, 1]$ into one district (denoted as District 1) and more moderate right wingers with $\beta \in [0, 0.5]$ into District 2. These blocks of right-wing voters are then grouped with any mass of voters for whom $\beta < 0$; since the preference of the median voter in each district ($\mu_1 = 0.5$ in District 1 and $\mu_2 = 0$ in District 2) is already determined, the composition of the left-wing voters does not matter. Republican candidates now win District 1 with probability 0.75 and District 2 with probability 0.5. Any other distribution of right-wing voters between the two districts (with one-quarter mass to each) would dilute the power of the extreme right-wing voters by wasting some in District 2, since that median voter would still have $\beta = 0$ while the preferences of the median voter of District 1 would fall. Only by concentrating the most extreme right-wing voters together can the gerrymanderer make the most effective use of her supporters.

C. *Example 3*

Finally, suppose (in addition to the setup in the second example) that individual preferences are measured with noise by the political parties. That is, let the gerrymanderer observe only s, a signal of preferences, instead of β itself. Across the population, let s be distributed uniformly on $[-1, 1]$, and let $\beta | s$ be distributed uniformly on $[s - 0.5, s + 0.5]$, with an independent draw

[7] For the sake of simplicity, we resolve all "ties" in this example in favor of the Republican candidate. Voters with $\beta = 0$ support the right-wing candidate, and if the candidates have equal vote shares, the Republican wins.

of β for each voter with a given signal s. Suppose the gerrymanderer creates districts as above (grouping voters for whom $s \in [0.5, 1]$ into District 1 and $s \in [0, 0.5]$ into District 2), and, furthermore, groups the most extreme left-wing voters into District 1 and the others in District 2. Because the measurement of preferences is noisy, the median voter in District 1 falls to $\mu_1 = 0$; the Republicans gain no advantage over proportional representation. Intuitively, the Republicans are "cutting it too close" in District 1. Although District 1 contains the most extreme right-wing voters, there are only one-quarter mass of them, and so the most left-wing voter with a right-wing signal is the median voter. Since some of those right-wing voters end up with more moderate preferences than their signal suggested, the median voter in the district is a moderate.

Instead, consider a gerrymander who groups all voters with $s \in [p, 1]$ into District 1 and $s \in [0, p]$ into District 2. Because of the intuition developed in the second example, this districting scheme still keeps the most extreme right-wing voters together. Now, though, the Republicans have more than just a bare majority of supporters in District 1, reducing the problem caused by preference mismeasurement above.

To complete this optimal districting, the gerrymanderer must allocate the left-wing voters. Her problem here is exactly opposite that faced with the right-wing voters: she must decide how best to *neutralize* the voting power of the extreme left-wingers. The key to this problem is that, since the majority of District 2 voters are left-wingers (assuming $p < ½$), μ_2 is far more sensitive to the allocation of these voters than μ_1. Thus, the optimal gerrymanderer should concentrate those least likely to vote for the Republican candidate into District 1, where they affect the median voter least.

Combining these insights, consider a districting plan such that voters for whom $s \in [-1, -1 + p] \cup [p, 1]$ make up District 1 and the rest are placed in District 2. The particular distributional assumptions made above imply that

$$\mu_1 = p + \sqrt{1 - 2p} - \frac{1}{2} \text{ and } \mu_2 = p - \frac{1}{2}.$$

The optimal gerrymander sets $p^* = ⅜$; Republican candidates win $1⅛$ seats in expectation. By including more right-wingers in District 1, μ_1 becomes less sensitive to the mismeasurement of preferences, and thus increases quite a bit, while μ_2, which depended less on the precision of the signal, does not decrease by as much. Furthermore, the right-wing voters of District 1 determine that $\mu_1 = \sqrt{¼} - ⅛ = ⅜$, and so the inclusion of the most extreme left-wingers has no effect. If, for instance, the gerrymanderer had included these least favorable voters into District 2 and placed voters with $s \in [-1 + p, -1 + 2p]$ into District 1, μ_2 would fall while μ_1 would not change.

These three simple examples illustrate how key features of an optimal partisan gerrymander differ from the standard "throwing away" and "smoothing" intuitions. First, it is not best to "smooth" extreme and moderate right-wing voters across many districts; rather, one should concentrate the most extreme right-wingers into a single district in order to not waste them all as right-of-median voters. Second, it is not efficient to "pack" those least likely to vote for one's candidate into a district that is "thrown away"; instead, these extreme left-winger voters are best countered by matching them with a greater number of extreme right-wingers.

We now turn to our model, which provides a more general characterization of the optimal partisan gerrymander, but the intuitions brought out in our examples are still prominent. Indeed, under certain regularity conditions, the optimal districting scheme has exactly the same form as in the final example above, matching increasingly extreme slices of voters from opposite sides of the signal distribution for the population.

III. The Model

A. *Overview*

There are two parties, D and R, which can be interpreted as the Democratic Party and the Republican Party. One of these parties (without loss of generality, we assume it to be R) is the gerrymanderer and creates districts. There is a unit mass of voters with preferences over a one-dimensional policy space. The gerrymanderer does not observe a voter's preferences, but, instead, receives a noisy signal of them. Also, she observes the posterior distribution of policy preferences conditional on the signal. We will sometimes refer to the marginal distribution of the signal as the "signal distribution." Thus, her problem is to create N voting districts by allocating voters from the signal distribution. Her objective is to maximize the expected number of districts won. The probability that each party wins a district is determined by the median voter in that district. The only constraints we place on the gerrymanderer are that: (a) each voter must be allocated to one and only one district; and (b) all districts must contain an identical mass of voters.

B. *Statement of the Problem*

There is a unit mass of voters who differ in their political preference over two candidates who locate on the real line such that $D < R$. We assume that this location happens prior to observing any signals about voters preferences. Denote the payoff to voter i of candidate x being elected as $u_i(x)$.

DEFINITION 1: *Voter preferences satisfy Single-Crossing if, for any two voters i and j such that* $i < j$ *and any two candidate locations* $D < R$, *the following hold: (i)* $u_j(D) > u_j(R) \Rightarrow u_i(D) > u_i(R)$ *and (ii)* $u_i(R) > u_i(D) \Rightarrow u_j(R) > u_j(D)$.

We assume voters have preferences satisfying single-crossing. Let $\beta_i = u_i(R) - u_i(D)$, for each voter type $i \in \mathbb{R}$. Without loss of generality, we reorder the voters so that β is monotonic. From this point on, the indexing of voters will reflect this reordering.[8]

These preferences are not observed by the gerrymanderer, who instead receives a noisy signal, $s \in \mathbb{R}$. Let the joint distribution of β and s be given by $F(\beta, s)$, which is assumed to have full support on $\mathbb{R}^2$. Let player R be the gerrymanderer. Let R have a Bayesian posterior $G(\beta|s)$ for the distribution of preferences given an observed signal. We refer to this distribution as the "conditional preference" distribution. We assume that both F and G are absolutely continuous. Define the marginal distribution of s as

$$h(s) = \int f(\beta, s)\, d\beta.$$

Since there is a continuum of voters, we can interpret h not only as characterizing a single draw from the population of voters, but also the mass of voters in the population. We refer to h as the "signal distribution." R allocates mass from this distribution in order to form districts. Normalize the median of s in the population to zero.

Since preferences satisfy single-crossing, the median voter determines a Condorcet winner (Paul Rothstein 1991). As a reduced form representation of electoral uncertainty, we assume that, in each election, *after* R observes the signal s, there is an aggregate shock decreasing all

[8] In the Appendix, we offer a result, which is of independent interest, that under single-crossing preferences the probability that a voter votes Republican is increasing in her type.

preferences by A. Thus, if the median voter in district n has preferences such that $\beta = \mu_n$, she votes for the Republican candidate if and only if $A \leq \mu_n$, which occurs with probability $B(\mu_n)$, where $B(\cdot)$ denotes the c.d.f of A. We assume that A can take any value in $\mathbb{R}$ with positive probability, so that B is strictly increasing.[9] One can also think of A as an "electoral breakpoint" such that voters positioned above (to the right) of the realization of the breakpoint vote for the Republican candidate, while those on the left vote democratic. Importantly, once the breakpoint is determined, all uncertainty is resolved and the position of voters relative to A determines for whom they vote with certainty. The uncertainty about whom a particular voter will vote for comes from the fact that A is stochastic.

Our assumptions about the location of candidates imply: (a) that all candidates of a given party and state locate in the same place; and (b) that this location takes place before receiving signals of voter preferences. In essence, these assumptions imply that there is nothing "local" about an election. Though perhaps counterintuitive, research suggests that this may not be far from the truth. Stephen Ansolabehere, James M. Snyder, and Charles Stewart III (2001) argue that, while district-to-district competition may exert some influence on the candidate platforms, the effect is "minor compared to the weight of the national parties." Allowing for state-to-state differences would surely leave even less variation in local platforms. Similarly, David S. Lee, Enrico Moretti, and Matthew J. Butler (2004) demonstrate that exogenous shifts in electoral preferences do not affect the menu of candidates offered to voters, perhaps because politicians have no way to credibly commit to campaign promises. We discuss the effects of certain departures from this assumption in Section V.

R divides the population into N equal-sized districts to maximize the expected number of seats won in the election. Let $\psi_n(s)$ denote the mass of voters from the population placed in district n. Formally, R solves the program

$$\max_{\{\psi_n(s)\}_{n=1}^{N}} \left\{ \frac{1}{N} \sum_{n=1}^{N} B(\mu_n) \right\} \tag{1}$$

$$\text{s.t.} \int_{-\infty}^{\infty} \psi_n(s)\, ds = \frac{1}{N}, \forall n \qquad \sum_{n=1}^{N} \psi_n(s) = h(s), \forall s \qquad 0 \leq \psi_n(s) \leq h(s), \forall n, s,$$

where

$$\mu_n = \hat{\beta} \qquad \text{s.t.} \int_{-\infty}^{\infty} G(\hat{\beta}|s)\psi_n(s)\, ds \equiv \Gamma_n(\hat{\beta}) = \frac{1}{2N}. \tag{2}$$

It will be useful to define the following for notational purposes:

$$\gamma_n(\beta) = \frac{\partial \Gamma_n(\beta)}{\partial \beta}. \tag{3}$$

Given a district profile $\psi_n(s)$, equation (2) determines μ_n with certainty. Though R could not identify any single voter as the median voter in a district, there is nothing stochastic about the preference parameter of the median voter.[10]

[9] This implies that the shock is independent of voter type. It may be the case that more "extreme" types are less affected by such shocks. This could be explored in future work.

[10] This model structure is isomorphic to the inclusion of further levels of uncertainty between signals and preferences. For instance, suppose that the gerrymanderer believed that, with 50 percent probability, preferences had a conditional distribution $G_1(\beta|s)$, and otherwise they were conditionally distributed as $G_2(\beta|s)$. Equation (2) would then

C. *Characterization of the Optimum*

No Cracking.—In order to analyze the problem, it is necessary to place some structure on the conditional distribution of preferences. The first restriction we require is that the signal be informative in the following sense.

CONDITION 1 (Informative Signal Property): *Let* $\partial G(\beta|s)/\partial s = z(\beta|s)$. *Then,*

$$\frac{z(\beta|s')}{z(\beta|s)} < \frac{z(\beta'|s')}{z(\beta'|s)}, \forall s' > s, \beta' > \beta.$$

This property is similar to the Monotone Likelihood Ratio Property (MLRP) due to Samuel Karlin and Herman Rubin (1956) (see also Paul R. Milgrom 1981). In fact, if a higher signal simply shifts the mean of the conditional preference distribution, then this property is equivalent to MLRP.[11] When this is the case, the condition essentially states that higher and higher signals (more right-wing) are more and more likely to come from voters who have underlying preferences that are farther to the right. Many common distributions satisfy it, including: the normal, exponential, uniform, chi-square, Poisson, binomial, noncentral t, and noncentral *F*. If a higher signal also changes the shape of the conditional distribution, then this property, like MLRP, becomes less intuitive. Condition 1 does, in general, imply first-order stochastic dominance,[12] and as such rules out cases where observing a higher signal makes *both* the probability of the voter being extreme left-wing *and* the probability of being extreme right-wing increase.

The second condition we require is a form of unimodality.

CONDITION 2 (Central Unimodality): $g(\beta|s)$ *is a unimodal distribution where the mode lies at the median.*

Also note that, without loss of generality, we can "rescale" s such that $s = \max_\beta g(\beta|s)$. Though many distributions that satisfy Condition 1 are unimodal, some are not, and we rule these out. Furthermore, Condition 1 implies that the mode of $g(\beta|s)$ must lie below the mode of $g(\beta|s')$ if $s < s'$. We can thus "relabel" the signals such that the mode of $g(\beta|s)$ lies at s. The two properties in Condition 2, taken together, intuitively imply that, conditional on signal s, preferences are distributed "near" s and not elsewhere.

Step 1: Slicing

LEMMA 1: *Suppose Condition 1 holds, and consider two districts, i and j, such that* $\mu_i < \mu_j$. *Consider any two voter types,* $s_1', s_2' \in \psi_i$ *(i.e., in district i). Then, any districting plan such that* $s \in \psi_j$ *for any* $s \in [s_1', s_2']$ *cannot be optimal, except perhaps on a set of measure zero.*

become $\int_{-\infty}^{\infty} \tfrac{1}{2}[G_1(\mu_n|s) + G_2(\mu_n|s)]\psi_n(s)\,ds = \tfrac{1}{2N}$, which is isomorphic to our original problem, if instead $G(\beta|s) = \tfrac{1}{2}[G_1(\mu_n|s) + G_2(\mu_n|s)]$.

[11] To see this, note that, if changing s shifts only the mean of the conditional preference distribution, then $G(\beta|s) = G(\beta|(s' - s))$. Therefore, $z(\beta|s) = -g(\beta|s)$, and hence Condition 1, imply MLRP.

[12] MLRP always implies this as well.

Lemma 1 shows that we can restrict attention, without loss of generality, to a much smaller strategy space. Districts are constructed from vertical slices of h—either whole slices (as in districts 1, 2, and 3 in the figure below), or a slice shared between districts that have the same median ("parfaits") (as in districts 4 and 5). Furthermore, in the optimal gerrymander, the voters in higher-median districts must lie outside—that is, have more extreme preferences—those in lower-median districts. The intuitions here are very similar to those discussed in the examples above. Extreme right-wing voters should be concentrated to maximize their voting strength—that is, the optimal districting scheme places an unbroken mass of voters with higher signals into the higher-median district rather than alternate smaller slices into all districts.

Step 2: No Parfaits

LEMMA 2: *Suppose that Conditions 1 and 2 hold. If* $j \neq i$, *then* $\mu_j \neq \mu_i$.

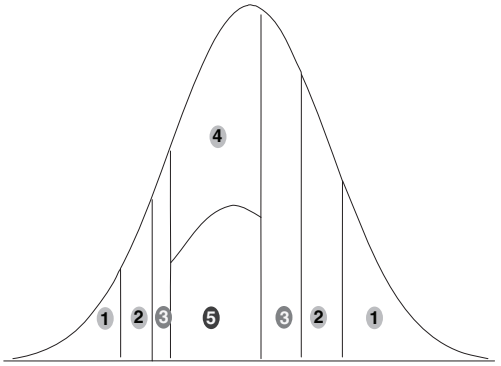


FIGURE 3. SLICES AND PARFAITS

This penultimate step rules out parfaits, as defined above. Parfaits appeared stable above because the split equated both the medians and the sensitivity of the median to changes across the two districts. But this is not so. One can reallocate mass between two such districts to maintain the equality of medians and make one district more sensitive to change than the other. Then, a profitable deviation exists which lowers the less sensitive median by some but increases the other by more. Hence, parfaits cannot be optimal.

Once again, the driving intuition in this case in that of concentrating extreme voters together to maximize their electoral power. In a way, parfaits are the least efficient use of extreme voters, and so it cannot be surprising that they are not optimal. Thus, the optimal gerrymander must contain *only* vertical slices of the signal distribution h that do not violate the ordering restriction from Lemma 1.

Step 3: No Intermediate Slices

LEMMA 3: *Suppose Condition 1 holds and consider three districts j, i, and k such that* $\mu_j > \mu_i > \mu_k$. *Now, fix* $h(s)$ *and N. Then, for a sufficiently precise signal, there does not exist a voter type* $s^* \in \psi_j$ *such that* $s' > s^* > s''$ *where* $s' \in \psi_i$ *and* $s'' \in \psi_k$, *except perhaps on a set of measure zero.*

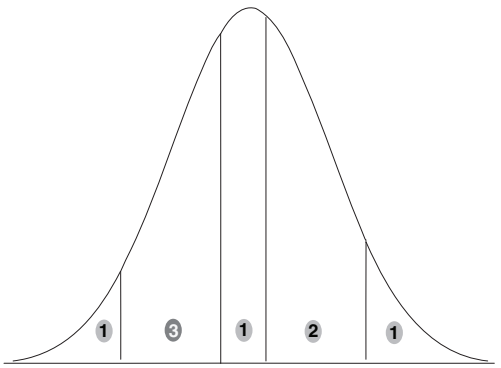


FIGURE 4.
AN EXAMPLE OF A STRATEGY RULED OUT BY LEMMA 3

This final step expands Lemma 2 by showing that voters in a higher-median district cannot lie within the set of all voters in lower-median districts. That is, by ruling out cases like that in Figure 4, it shows that optimal districts must comprise either a single slice or two slices matching mass from opposite tails of the distribution. The intuition is very similar to that of Lemma 2, that lower medians (such as those

in Districts 2 and 3 in Figure 4) are more positively affected by the inclusion of moderate instead of extreme left-wing voters. On the other hand, the higher medians (such as that of District 1) are hardly lowered by the substitution of extreme left-wingers. In order for these arguments to hold, though, the signal distribution must have high enough quality. If it does not, then intermediate slices are possible.

PROPOSITION 1: *Suppose that Conditions 1 and 2 hold, and that the signal distribution is of sufficiently high quality (as defined in Lemma 3). Consider a districting plan with N districts labelled such that $\mu_j > \mu_i$ if and only if $j < i$. This plan is optimal if and only if it can be characterized by "breakpoints" $\{u_n\}_{n=1}^{N-1}$ and $\{l_n\}_{n=1}^{N-1}$ (ordered such that $u_1 > u_2 > \cdots > u_{N-1} > l_{N-1} \geq l_{N-2} \geq \cdots l_1 \geq -\infty$) such that*

$$\psi_1 = \begin{cases} h(s) & \text{if } s < l_1 \text{ or } s > u_1 \\ 0 & \text{otherwise} \end{cases},$$

$$\psi_n = \begin{cases} h(s) & \text{if } l_{n-1} < s < l_n \text{ or } u_{n-1} > s > u_n \\ 0 & \text{otherwise} \end{cases} \quad \text{for } 1 < n < N,$$

and

$$\psi_N = \begin{cases} h(s) & \text{if } s > l_{N-1} \text{ or } s < u_{N-1} \\ 0 & \text{otherwise} \end{cases}.$$

At this point, we have established that cracking is not optimal, although some form of packing may still be. That is, we have not yet ruled out the type of strategy depicted in Figure 5. We will now provide conditions under which packing is not optimal—and show that matching of extreme supporters with extreme opponents is.

No Packing.—We now offer a result which shows that if the signal quality is sufficiently high, the optimal strategy cannot involve packing, by which we mean concentrating one's most ardent opponents into a single district—a notion we immediately make precise.

PROPOSITION 2: *Suppose Conditions 1 and 2 hold and the signal is of sufficiently high quality. Then, there exists n, and $s < s'$, such that $\mu_n > \mu_N$ and $s \in \psi_n$, $s' \in \psi_N$.*

To understand the intuition for this result, first consider a potential deviation from a districting plan that "packs," as in Figure 5: R could take the most left-wing voters from District 3 into District 1, and then "slide" Districts 2 and 3 to the right, thereby gaining in Districts 2 and 3 but losing ground in District 1. Now, consider how this strategy changes in value as we remove noise from the signal. As the signal becomes more precise, the cost of the proposed change in District 1 decreases, since the voters R removes from District 1 are less likely to be actually right-of-median. (The voters R

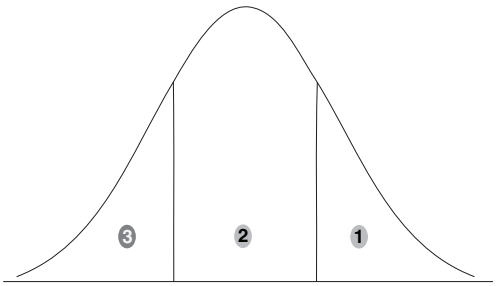


FIGURE 5. PROPOSITION 2 RULES OUT THIS STRATEGY

adds can be so far to the left that they are *always* left-of-median.) But the gains in Districts 2 and 3 stay roughly constant, since the entire districts are sliding to the right. At some point, when the signal is precise enough, the steady gains must begin to outweigh the shrinking loss. In the limit, as the signal becomes perfect, there is no cost to R in District 1 from this deviation, and R seeks to match an infinitesimally larger slice of right-wing voters with left-wing voters in each district, as in Example 2 in Section II.

Figure 6 is an example of a potentially optimal strategy. District 1 comprises a slice of extreme Republicans and a slice of extreme Democrats, and this slicing proceeds toward the center of the signal distribution. The slices from the right tail of the signal distribution contain more mass than the matched slice from the left tail, lest Republicans "cut it too close" in accounting for the noisy measurement of preferences. This follows the intuition developed in the third example in Section II.

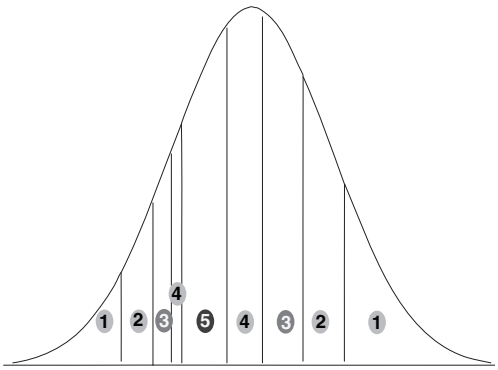


FIGURE 6. AN EXAMPLE OF THE OPTIMAL STRATEGY

We are unable to offer an analytical solution for the "breakpoints" $\{u_i\}_{n=1}^{N-1}$ and $\{l_i\}_{n=1}^{N-1}$. However, they are easily computed numerically, given a signal distribution (as Section IV demonstrates). We also conjecture that as the spread of the noise distribution increases, the ratio of mass in upper slices to lower slices increases—limiting to the case where districts are comprised of whole slices, rather than matching ones. This is certainly the case in a wide variety of numerical examples we have explored, and we are yet to find a counterexample. It does, however, remain a conjecture.

D. *Comparison with Received Literature*

Previous work has considered two types of models which are both special cases of our model. The approach most similar to ours is that of Gilligan and Matsusaka (1999), in which voters always vote for a given party and their preferences are known with certainty to the gerrymanderer. Our model simplifies to this case (as shown in the first example in Section II) if the conditional preference distribution limits to a point-mass at the true preference (so that preferences are observable) and if the breakpoint distribution $B(\cdot)$ is a point mass (so that voters are either Democrats or Republicans). As such, our model is more general and captures an important intuition—that more noise leads the gerrymanderer to create a larger buffer. Furthermore, our model has a continuum of preferences, and therefore is instructive not only as to the optimal number of Republicans and Democrats in a district, but also as to which types of Republicans and Democrats should be combined.

The second approach to modelling gerrymandering—one perhaps more popular than that of Gilligan and Matsusaka—is a binary signal model with noise. In such a model (e.g., Owen and Grofman 1988), the optimal strategy involves "packing" some districts and "cracking" others. Owen and Grofman refer to this as a "bipartisan gerrymander," since there are Democratic districts (those thrown away) and Republican districts (the others). For instance, if 60 percent of the population have signal r and 40 percent signal d, then the optimal strategy involves creating a certain number of districts that contain only those with signal d, and spreading the r voters uniformly over the remaining districts. This result is also a special case of our model, with additional assumptions, as shown in Proposition 3.

PROPOSITION 3: *Suppose $s \in \{d, r\}$ and that Conditions 1 and 2 hold. Suppose further that $B(\cdot)$ is unimodal, with mode greater than d and less than r. Then, the optimal gerrymander involves creating some districts with all voters of type d, and others with a constant proportion of r and d, and possibly one "odd district" with a nonzero but less-than-half proportion of r (from integer rounding problems). When $N \to \infty$, the optimal solution is a pure "bipartisan gerrymander."*

Thus, our model nests the solution of "bipartisan gerrymandering," but the conclusions of such a model are very sensitive to several extreme assumptions. Furthermore, the intuitions this special case highlights are very misleading. For instance, suppose that there are three signals: *r*, *d*, and *i* (Independents). As Proposition 2 shows, the optimal strategy matches increasingly extreme segments from the right and left tails (in this case Republicans and Democrats) into the same districts. The district where Republicans have the lowest chance of winning is not one that contains many Democrats, but rather one that contains many Independents. That is, these least Republican districts contain voters from the middle of the signal distribution, not the extreme left tail. It is also clear that "smoothing" is not a robust intuition. It is true only in the special case of a binary signal, because there is no heterogeneity among potential Republican voters.

E. *Comparative Statics*

In this subsection, we consider how the value of being the gerrymanderer responds to changes in the underlying distribution of voter preferences and signals. We also consider how this value changes as the number of districts to be created changes.

Our first comparative static shows that more precise signals are always better for the gerrymanderer.

DEFINITION 2: *Consider two conditional preference distributions g and g′. The distribution g provides a More Precise signal than g′ if there exists a conditional distribution $c(s'|s)$ such that*

$$\int g(\beta|s')c(s'|s)\,ds' = g(\beta|s).$$

PROPOSITION 4: *The expected number of districts won by the gerrymanderer is increasing in the precision of the signal.*

This result shows that the gerrymanderer wins more districts in expectation as the signal received becomes more precise. Intuitively, as the gerrymanderer receives a better signal, the need for a large "buffer" of voters in a district declines. Instead, she can construct districts of a given median with a smaller proportion of voters from the right hand tail, leaving more right-wingers for other districts. Mathematically, the gerrymanderer could always lower the quality of the signal, while the reverse operation is not possible. Thus, it cannot be that a lower quality signal is better.

Our second comparative static result shows that the gerrymanderer does better as the distribution of voters becomes more spread out.

PROPOSITION 5: *Consider two joint distributions $F(\beta, s)$ and $\hat{F}(\beta, s)$, with marginal distributions of β given by $F(\beta)$ and $\hat{F}(\beta)$, such that $\hat{F}(\beta)$ is a symmetric spread of $F(\beta)$. Then, the expected number of districts won by the gerrymanderer is higher for $\hat{F}$ than for F.*

Intuitively, suppose that all signals have the same variance of preferences conditional on the signal. But, if the breakpoint is more likely to be near the center of the preference distribution, there is less uncertainty as to the voting patterns of extreme voters. For instance, suppose the breakpoint is normally distributed. If a voter has either $\beta = -0.5$ or $\beta = 0.5$, she will vote Republican either 31 percent or 69 percent of the time, quite a bit of uncertainty; but if a voter has either $\beta = 1.5$ or $\beta = 2.5$, she will vote Republican either 93 percent or 99 percent of the time. Extreme voters are thus more valuable to the gerrymanderer. Since an increase in the variance of the voter preference distribution increases the share of extreme voters in the population, the expected number of seats won increases.

The final comparative static concerns the number of districts.

PROPOSITION 6: *Suppose that the number of districts increases by an integer multiple (that is, doubles or triples). Then, the expected percentage of districts won by the gerrymanderer strictly increases.*

In previous analyses in this literature, proportional increases in the number of districts has little import; if twice the number of districts are required, the existing districts are split into equal parts, and so the voter profiles of the districts do not change. Our model implies that such parfaits are inefficient. Instead, the gerrymanderer can do better by slicing within previous districts, grouping the most and least Republican voters from an old district into one new district, and giving the all less extreme voters to the other.

IV. Numerical Examples

In order to illustrate the characterization of the optimal gerrymandering strategy and its comparative statics, we report the results of a number of numerical examples in this section. The examples all assume that there are five districts and that the gerrymanderer is Republican. In these examples, we assume that the joint distribution of preferences and signals, $F(\beta, s)$, is multivariate normal with parameters $\mu_\beta = \mu_s = 0$ and covariance matrix $\sum$, with

$$\sum = \begin{pmatrix} \sigma_\beta^2 & \rho\sigma_\beta\sigma_s \\ \rho\sigma_\beta\sigma_s & \sigma_s^2 \end{pmatrix}.$$

This implies that both the signal distribution and the conditional preference distribution are themselves normal. Note that this assumption satisfies Conditions 1 and 2. In this base case, we assume a distribution of $F(\beta, s)$ such that $\beta \sim N(0,5)$ and $\rho = 0.5$. Furthermore, we assume that $\sigma_s = \rho\sigma_\beta$ so that $G(\beta|s) \sim N\big(s, \sigma^2_{\beta|s} = \sigma^2_\beta(1-\rho)\big)$. In all examples, we let $B \sim N(0,1)$ and set $N = 5$. Note that these assumptions imply that, nominally, half the voters are Republicans and half are Democrats—without gerrymandering, each party would win 2.5 seats, in expectation.

Panel A of Table 1 highlights a number of features of the optimal strategy. First, the highest median district (District 1) consists of 62 percent from a slice from the right tail of the distribution and 38 percent from a slice from the left tail. These upper slices get progressively larger for the lower median districts. While District 4 comprises a whole slice, Districts 1 through 3 are formed by matching slices from the right and left tails. (District 5 consists of a whole slice containing those voters remaining after removing the first four districts from the signal distribution, and so the fraction in the upper and lower slice is not relevant.) Second, note that the probability of winning District 1 is very high—87.5 percent. This means that those in the left-most part of the distribution have very little chance of gaining representation. Third, no districts are "thrown

TABLE 1—NUMERICAL EXAMPLES OF OPTIMAL GERRYMANDERING

Panel A. Baseline example

		District				
		1	2	3	4	5
	Upper Slice	0.62	0.73	0.91	1	n/a
	Lower Slice	0.38	0.27	0.09	0	n/a
	Prob (win)	87.5%	74.8%	65.7%	41.7%	13.7%

Panel B. Signal coarseness

		Probability of winning district				
Signal variance	E[Districts won]	1	2	3	4	5
0.50	3.46	97.4%	86.9%	74.3%	56.6%	30.9%
2.50	2.83	87.5%	74.8%	65.7%	41.7%	13.7%
4.50	2.53	68.2%	61.9%	55.7%	41.8%	25.9%

Panel C. Spread of voter preferences

		Probability of winning district				
Preference variance	E[Districts won]	1	2	3	4	5
3.0	2.55	71.0%	62.3%	55.6%	41.2%	25.1%
5.0	2.83	87.5%	74.8%	65.7%	41.7%	13.7%
25.0	3.78	100.0%	97.1%	90.6%	73.9%	16.4%

Panel D. Partisan bias of the population

			Probability of winning district				
% Republican	E[Won]	"Value"	1	2	3	4	5
30%	2.04	0.58	49.4%	47.0%	40.7%	27.8%	10.2%
40%	2.44	0.48	87.0%	73.0%	52.3%	25.1%	6.2%
50%	2.83	0.33	87.5%	74.8%	65.7%	41.7%	13.7%
60%	3.24	0.20	87.8%	76.1%	67.3%	58.6%	34.5%
70%	3.67	0.12	90.2%	79.6%	71.7%	65.0%	59.1%

away"; the gerrymanderer has more than a 13 percent chance of winning even the district least favorable to her. If she had "thrown away" the district—that is, put those with the lowest signal into it—then, in this example, she would win it only 1.4 percent of the time. Finally, the number of districts won in expectation in this case is 2.8, compared with a non-gerrymandered equal representation benchmark of 2.5. Hence, in this case, the ability to be the gerrymanderer leads to a 13 percent increase in the expected number of districts won.

Panel B illustrates how a change in the spread of the conditional preference distribution affects the gerrymanderer. In accordance with our comparative static results, the gerrymanderer does worse as the quality of her signal deteriorates. This is reflected in a lower probability of winning each district, and hence a lower overall value to being the gerrymanderer. For instance, note that when the signal is very coarse, $\sigma^2_{\beta|s} = 4.5$, the gerrymanderer wins only 2.54 districts in expectation—barely more than the 2.5 won under proportional representation. Also, in the $\sigma^2_{\beta|s} = 0.5$ case, the gerrymanderer has a 31 percent chance of winning district 5—if she "threw it away" that would be just 0.2 percent. Finally, although the expected districts won, and hence the value function is monotonic in $\sigma^2_{\beta|s}$ (as we have shown analytically), the probability of winning each district is not monotonic. Intuitively, as the signal becomes more informative, the gerrymanderer can cut the districts finer, but the probability of winning the votes of those with the lowest signals decreases. These two effects work in opposite directions, which leads to the potential nonmonotonicity of the probability of winning districts with "low" medians (here Districts 4 and 5).

Panel C shows how a change in the spread of the voter preferences affects the gerrymanderer. As voter preferences become more spread out, the gerrymanderer does better, as our comparative

static results showed. There is a monotonic increase in the probability of winning Districts 1–4 as voter preferences become more spread out, since fewer extreme voters are necessary to provide a solid margin of victory (in expectation). A similar nonmonotonicity, as discussed above, is at work here with the probability of winning District 5.

Panel D reports how changes in the mean affect gerrymandering. A natural interpretation of a change in the mean is that it is a change in the number of nominal Republicans/Democrats. With the mean at zero, there are 50 percent nominal Republicans. As the mean increases, the share of nominal Republicans increases, and vice versa. Note that as the proportion of nominal Republicans increases, the expected number of seats won increases, and the value to being the gerrymanderer decreases. This value represents the difference in expected seats won compared to proportional representation.

V. Extensions

In this section, we discuss some extensions to the basic model.

A. *Majority Power, Risk Aversion, and District-Specific Objectives*

Our analysis thus far has considered a gerrymanderer whose payoff function is equal to the expected number of districts won. This is likely a good approximation for congressional districting, where the uncertainty over the eventual party balance in the House of Representatives makes each district in a given state equally important. But in state legislatures, other objectives may play an important role. For instance, a party might derive great benefit from remaining in the majority, in which case the gerrymanderer's value function would include a positive discontinuity at 50 percent of the seats. The marginal benefit to the gerrymanderer from each seat won might also be diminishing as she wins more seats, in which case the objective function would become concave. Finally, some districts may be more important than others, since different incumbents may be more valuable to the party than others. The next proposition shows that Propositions 1 and 2 characterize the optimum in all of these cases.

PROPOSITION 7: *Suppose that the gerrymanderer constructs districts so as to maximize*

$$E\left[V\left(\frac{1}{N}\sum_{n=1}^{N} w_n d_n\right)\right],$$

where $d_n = 1$ *if the Republicans win district n and* $d_n = 0$ *otherwise; V is any strictly increasing function; and* $\{w_n\}_{n=1}^{N}$ *are a strictly positive set of weights which add to* 1. *Then Propositions* 1 *and* 2 *characterize the optimal partisan gerrymander.*

Proposition 7 shows that our earlier analysis is robust to most any plausible gerrymanderer objective function. The key to this result is the fact that the domain of the underlying objective function comprises only a discrete subset of values, since one of the parties must actually win each seat in the election. Taking an expectation over this underlying function smooths out the problem, so that increasing the probability of winning any one district, holding the others constant, has a linear impact on the expected value of the redistricting scheme. Our earlier assumption of a linear objective function made this marginal impact the same across all districts. Extending our results to this broader case, where the slope of each impact may vary across districts, merely adds a constant in our proofs, but the linearity ensures the proofs still go through.

The only restriction we must place on the objective function is that the gerrymander must gain from winning another district. If, at some point, V were flat or decreasing, so that the gerrymanderer was indifferent or averse to winning, our result would not hold. Similarly, we require that the weights $\{w_n\}_{n=1}^{N}$ be bounded away from zero, lest the gerrymanderer not care at all about a certain district.

Though Propositions 1 and 2 still hold, the effect of the optimal redistricting plan will vary as the underlying objective function changes. For instance, suppose the objective function were linear but for a positive discontinuity at winning a majority. Under normal circumstances, where the gerrymander possesses a commanding popular majority in the state, redistricters would now be risk averse and thus seek to win fewer districts but hold the majority with greater probability. Practically, such a change would mean grouping larger numbers of Republican voters (the right-hand "slice") into a small majority of the districts. On the other hand, if the gerrymanderer faces a hostile population (perhaps due to the inequities of gerrymanders past), the party would become risk-loving. The other two alternative objective functions we mentioned above—concavity and unequal weighting among districts—manifest themselves in more straightforward ways in district composition, with incumbents making some districts more secure at the expense of others.

Risk-aversion also provides a simple rationale for ruling out cracking. As previously noted, a districting plan determines the probability of winning each district; and in the previous sections we have considered the mean of these probabilities. However, a celebrated theorem of Siméon-Denis Poisson (1837) allows us to analyze the variance as well. Substantially generalizing the work of Bernoulli, Poisson showed that the variance of nonidentical independent trials $p_1,\ldots,p_n$ is

$$Var(x) = n\bar{p}(1-\bar{p}) - n\sigma_p^2,$$

where $\bar{p} = (\sum_{i=1}^{n} p_i)/n$ and σ_p^2 is the variance of $p_1,\ldots,p_n$. It is immediate that, fixing $\bar{p}$, the variance is reduced by "spreading out" $(p_1,\ldots,p_n)$. That is, the *maximum* variance of the number of successes (i.e., districts won) is achieved when $p_1 = p_2 = \cdots = p_n$. Further, Wassily Hoeffding (1956) showed that, fixing $\bar{p}$, any increasing concave function of the number of successes is minimized when $p_1 = p_2 = \cdots = p_n$. These theorems show that cracking is suboptimal for a risk-averse gerrymanderer, since cracking involves making a number of districts have the same median voter type, and hence the same probability of winning. Under a pack-and-crack strategy, probabilities of winning districts are as follows:

$$p_1^c = \cdots = p_k^c > p_{k+1}^p > \cdots > p_N^p, \tag{4}$$

where superscripts p and c denote packed districts and cracked districts, respectively. The district winning probabilities under the strategy of Propositions 1 and 2 is

$$p_1 > \cdots > p_N. \tag{5}$$

Now, consider a deviation toward (5) from the pack-and-crack strategy which generates (4). In particular, suppose two cracked districts are altered so that $\hat{p}_1^c > p_1^c$ and $\hat{p}_2^c < p_2^c$, with $\hat{p}_1^c + \hat{p}_2^c = p_1^c + p_2^c$. Proposition 2 tells us that there exists such a deviation with $\hat{p}_1^c + \hat{p}_2^c > p_1^c + p_2^c$, but to apply combinatoric theorems with the expected number of successes constant, we address the case where $\hat{p}_1^c + \hat{p}_2^c = p_1^c + p_2^c$. By Poisson's Theorem the variance of the number of districts won under pack-and-crack is $N\bar{p}(1-\bar{p}) - N \cdot Var(p_1^c,\ldots,p_N^p)$. Under the proposed deviation, the variance is $N\bar{p}(1-\bar{p}) - N \cdot Var(\hat{p}_1^c,\ldots,\hat{p}_N^p)$. To show that the number of districts won under the

deviation is lower, we require $Var(p_1^c,\ldots,p_N^p) < Var(\hat{p}_1^c,\ldots,\hat{p}_N^p)$. That is, $\frac{1}{N}\sum_{i=1}^{N}(p_i - \bar{p})^2 < \frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - \bar{p})^2$. Removing common terms, this becomes $(p_1^c - \bar{p})^2 + (p_2^c - \bar{p})^2 < (\hat{p}_1^c - \bar{p})^2 + (\hat{p}_2^c - \bar{p})^2$, or, equivalently, $(p_2^c - \bar{p})^2 - (\hat{p}_2^c - \bar{p})^2 < (\hat{p}_1^c - \bar{p})^2 - (p_1^c - \bar{p})^2$. Since $\hat{p}_1^c > p_1^c = p_2^c > \hat{p}_2^c$, the inequality holds.

Cracking, therefore, not only lowers the mean number of districts won, it also increases the risk borne by the gerrymanderer.

It is important to note that, since the aggregate shock affects all districts, the probabilities of winning districts are not independent trials. As we show, however, in Proposition 8 below, the analysis leading to Propositions 1 and 2 applies to the case where there are district-specific shocks. Therefore, treating the trials as we have here as independent is arguably a more general approach.

Applying Hoeffding's Theorem to the kind of deviational argument just made, a deviation such as the one above is preferred by a gerrymanderer whose payoff function is increasing and concave in the number of districts won. Thus, pack-and-crack is suboptimal for any gerrymanderer whose payoff is an increasing concave function of the number of districts won.

B. *Policy Consequences*

Our analysis has thus far considered only a districting scheme's impact on party representation in the legislature. In this section, we consider the potential distance between the median voter's preference and the actual outcome under the optimal partisan gerrymander.[13] We have in mind a setting where district medians determine the preferences of legislators, who then vote on policy alternatives. To illustrate this, we consider the case where voter preferences are perfectly observable (i.e., $\beta = s$). Let each voter have a most preferred policy given by the c.d.f. $H(s)$ with continuous p.d.f. $h(s)$. Assume that the median voter is given by $H(s_m) = \frac{1}{2}$. Let the ideal policy of the median voter in district d be s_m^d. Ordering these median voters within a district as $s_m^1 \geq \cdots \geq s_m^{(N+1)/2} \geq \cdots s_m^N$, we have what we will refer to as the "representative median voter" $s_m^{(N+1)/2}$. We take this to be the preference of the median legislator. For simplicity, we assume that N is odd—although nothing important hinges on this.

The question we ask here is: what is the difference in preferences between the representative median voter and the population median voter under the optimal gerrymander? That is, what is the magnitude of $|H(s_m^{(N+1)/2}) - H(s_m)|$?

If the gerrymanderer maximizes $H(s_m^{(N+1)/2})$, then—since the signal is perfect—Proposition 2 tells us that this is achieved by combining a mass of voters with the highest bliss points with an (infinitesimally smaller) mass of voters with the lowest bliss points, and then continuing to match into the center of the distribution. Under this gerrymander, the median voter in the median district is the left-most voter in the right-hand slice of district $(N + 1)/2$. It is immediate that, under this gerrymander, $\lim_{N\to\infty} H(s_m^{(N+1)/2}) = \lim_{N\to\infty}(N + 1)/4N = \frac{1}{4}$, and hence $|H(s_m^{(N+1)/2}) - H(s_m)| = \frac{1}{4}$. Therefore (for states with large numbers of districts[14]), under the optimal gerrymander, a minority constituting just 25 percent of the population can constitute a winning coalition.

Interestingly, the "dominance of the 25 percent majority" under representative systems was conjectured in the seminal work of James M. Buchanan and Gordon Tullock (1962, 221–22).

This analysis of policy consequences could be extended to the case of a noisy signal. We conjecture that the "buffer" of voters required by the gerrymanderer to equate median-like outcomes becomes larger as the signal quality decreases, and hence $|H(s_m^{(N+1)/2}) - H(s_m)|$ decreases

[13] We are grateful to an anonymous referee for suggesting this, as well as details of the approach.

[14] For a state with 53 districts (e.g., California), $H(s_m^{(N+1)/2}) = 0.255$, and for a state with 5 districts is 0.3.

monotonically in the quality of the signal. We have found this to be the case in a large number of numerical examples—but it remains a conjecture.

C. *Candidate Effects*

Another empirical regularity of congressional races is the seemingly large electoral advantage enjoyed by incumbents—fewer than 3 percent of incumbents are defeated in the typical election cycle. There are three possible causes for this edge. First, an incumbent may simply reflect the preferences of her constituents, or may generally be of high quality. In this case, incumbency is simply a proxy for match quality between a representative and her district, and one can say that incumbency, per se, has no effect. Second, the incumbent may be more well known to her constituents in a variety of ways, and thus more easily elected; a (Republican) gerrymanderer would respond to this type of incumbent advantage by maintaining Republican incumbent districts as constant as possible, while matching Democratic incumbents to new and unfamiliar (though not necessarily different, from a signal profile perspective) districts. Indeed, such tactics were a key part of the Republican gerrymander of Texas in 2003. This effect is primarily a geographic concern, though, and is thus somewhat orthogonal to the predictions of our model.

A third source of advantage for an incumbent may be, broadly speaking, her résumé of congressional experience and the resulting low quality of opponents, an edge which would follow her no matter the make-up of her district. Stephen Ansolabehere, James M. Snyder, and Charles Stewart III (2000) use the decennial redrawing of district boundaries to estimate that this third channel accounts for one-third to one-half of the incumbency advantage, on average, though there is surely much individual heterogeneity in the magnitude of the effect. The conclusions of our model would change in the presence of large incumbent effects of this third type, which would, in effect, make the distribution of the electoral breakpoint district-specific. For instance, suppose that a particular Democratic incumbent was universally well liked and assured of election regardless of the composition of her district. It would then be optimal for a Republican gerrymanderer to "throw away" her district by including in it the most extreme Democrats.

We can model this extension by assuming that incumbent n (from district n) has an electoral advantage ζ_n such that voters support the incumbent if $\beta - A + \zeta_n > 0$. Republicans have positive ζ's, and Democratic incumbents have negative ζ's. Furthermore, suppose that this advantage is independent of the voters in the incumbent's district. As the intuition above suggests, our Lemma 3, and thus Proposition 1, fail with this addition. But, as the following proposition shows, Lemmas 1 and 2 still hold.

PROPOSITION 8: *Suppose that incumbent n in district n has an additional electoral advantage ζ_n, and that $F(\beta, s)$ satisfies Conditions 1 and 2. Then, Lemmas 1 and 2 hold, while Lemma 3, in general, does not.*

Though the ordering of the slices would be somewhat different, the main force of our results still hold. Optimal districts comprise only vertical slices, and such slices may not "interlock," as in Lemma 2. This model does generate the familiar prescription of districts that are "thrown away," but it does not generate "smoothing" across Republican voters, as in standard model. Of course, such a deviation depends on the magnitude of a quite particular effect of incumbency which, in practice, may be quite limited. Even the most well-liked politicians may have trouble attracting votes from affiliates of the opposite party; would Rep. Tom Delay still get elected if his district contained the poor inner cities of Houston instead of Sugarland? Nevertheless, this is the only extension from our model we discuss that does generate "throwing away" districts, and it perhaps deserves further study.

D. *Voter Turnout*

In our model, we have implicitly assumed that everyone votes; obviously, in a system with non-compulsory voting, voter turnout is a real and important issue. In theory, voter turnout could vary with any aspect of the individual or district; research on electoral participation suggests two sets of factors that might affect turnout. First, the literature has identified a number of individual attributes—including education, age, marriage status, occupation, and ideological extremism—which affect the probability of voting (see Orley Ashenfelter and Stanley Kelly, Jr., 1975; Raymond Wolfinger and Steven Rosenstone 1980; John G. Matsusaka and Filip Palda 1993; Edward L. Glaeser, Giacomo A. M. Ponzetto, and Jesse Shapiro 2005). These factors do not have a direct impact on our results, since voter turnout exogenous to the creation of districts will not affect the predictions.

The political science literature has also found a number of district-specific effects. For instance, Kamhon Kan and C. C. Yang (2001) find that turnout is higher when the perceived differences between candidate ideological platforms are higher and when voters "fear" one candidate more than the other. But this type of effect will not change our characterization of the optimal strategy either, since all voters in a district would turn out more or less, depending on the particulars of district construction. Similarly, Ebonya Washington (2006) finds that black candidates increase turnout both among black and white voters, and the difference is not statistically significant.

A final class of models of endogenous turnout allows the probability of voting to depend on district-specific characteristics, but affects different voters within a district in different ways. For instance, people might be more or less likely to vote if their policy bliss point is closer to one of the candidate's platform. Alternatively, moderate voters might be more or less likely to turn out if grouped in the same district with extreme voters from their own party, or extreme voters from different parties. Such models can change the structure of the optimal gerrymander; for instance, if extreme voters of one party make moderate voters from the other party less likely to vote, the matching of extreme democrats with extreme republicans may fail. Of course, the structure of the optimal strategy in our model could just as easily be reinforced if the opposite were true, and incensed Republican moderates turned out to oppose the more extreme Democrats with whom our strategy would match them. Since there is little evidence of either the presence or the direction of these effects, we do not explicitly model these factors here, but such efforts might be a plausible direction for future work.

VI. Conclusion

This paper shows that existing intuitions for optimal partisan gerrymandering are rather misleading—and are the consequence of simplifying assumptions. We have analyzed a more general model with a continuum of voter preferences and noisy signals of those preferences. The model nests major models in the literature as special cases. Smoothing supporters evenly is always suboptimal. When the signal the gerrymanderer receives is precise enough, the optimal strategy involves matching extreme Republicans with extreme Democrats. This characterization of the optimal partisan gerrymander is robust to a number of extensions, including alternative partisan objective functions.

The primary import of our paper is to suggest a reexamination of widely held intuitions about the effects of partisan gerrymandering. These intuitions are not simply academic speculations, but give rise to conventional wisdom about partisan gerrymandering which is not wholly accurate. For instance, traditional models imply that groups that have very different preferences from the gerrymanderer do not fare so badly—that is, although gerrymandering makes them worse off than proportional representation, they are assured of a lower bound of representation due to

the gerrymanderer's "throwing away" some districts. Our model has very different implications. Instead, because of the "matching slices" strategy, they are combined into districts with a larger group of voters who have extremely different preferences from them, and so they have *very little* representation as a result of gerrymandering. Thus, our model suggests that the negative consequences of partisan gerrymandering for minority representation in government may be far worse than currently thought.

A natural question that follows from this analysis is to ask: who are the voters in the opposite tail of the distribution to the gerrymanderers? To illustrate this connection, suppose that the gerrymanderer is a Republican and that African Americans are highly represented in the far-left tail of the signal distribution (i.e., they have characteristics that make them very likely to vote for Democrats). In this case, under the optimal gerrymander, African Americans would be placed in districts such that they receive very little representation. Data from the 2000 US Census and the 2000 presidential election suggest that African Americans do, indeed, constitute the far-left tail, and so an implementation of the optimal strategy, as characterized in this paper, would be severely disadvantageous to that population. The unmistakable implication of these facts is that partisan gerrymandering (when practiced by Republicans) and racial gerrymandering are basically synonymous *in effect*. Since the 1960s, however, the Supreme Court has adopted a test based on intent, rather than effect.

A further implication of our analysis is that gerrymandering can be very valuable, and indeed is more valuable today than ever before. Technological advances have allowed gerrymanderers to gain better information about voters—in our model, a less coarse signal distribution in the sense of Blackwell—and draw boundaries with a finer pen. One would therefore expect parties to use an increasingly large amount of resources in order to become the gerrymanderer. Since the practice itself probably lowers social welfare (see Stephen Coate and Brian Knight (2006) for an illuminating analysis of socially optimal districting), spending resources on it merely exacerbates the social loss associated with partisan gerrymandering. This implies that the welfare loss from gerrymandering is linked to such technologies, and has grown over time.

There are two clear directions for future work. The first involves empirical investigations of gerrymandering in light of the theory developed here. The structure provided by our characterization of the optimal gerrymandering strategy is important for such empirical work. Previous empirical work on gerrymandering (see, for instance, Andrew Gelman and Gary King 1990, 1994) assumes a nonmicrofounded structural model which may give inaccurate estimates of the degree of gerrymandering. The second set of open issues involves the regulation of gerrymandering. Enriching the model to capture spatial considerations would make it possible to analyze the impact of constraints such as compactness. Although there is a body of work that attempts to deal with spatial considerations, the underlying models of gerrymandering they employ are, as we have discussed, insufficiently rich to capture the core intuitions of the optimal strategy.

Ultimately, the effect of gerrymandering is an empirical question. As our model highlights, the impact of it depends on the particulars of the signal and preference distribution. One thing this paper demonstrates, however, is that empirical investigations alone can be misleading. Without understanding the optimal strategy for a gerrymanderer, one cannot properly assess the impact of partisan gerrymandering.

Appendix

A. *Monotonicity of Voting*

We remarked in a footnote in the text that, under the assumption of single-crossing preferences, the probability that a voter votes Republican is increasing in her type. This is not of direct relevance to the other results in the paper, but may be of independent interest.

DEFINITION 3: *Let X and Y be subsets of $\mathbb{R}$, and let $K : X \times Y \to \mathbb{R}$. We say that K is Totally Positive of order n ("TP_n") if $x_1 < \cdots < x_n$ and $y_1 < \cdots < y_n$ imply*

$$\begin{vmatrix} K(x_1,y_1) & \cdots & K(x_1,y_m) \\ \vdots & & \vdots \\ K(x_m,y_1) & \cdots & K(x_m,y_m) \end{vmatrix} \geq 0$$

for each $m = 1,\ldots,n$.

THEOREM 1 (Karlin 1968): *Let K be TP_n on $X \times Y$ and let μ be a σ-finite measure on X. If $f : \mathbb{R} \to \mathbb{R}$ has at most $k \leq n-1$ sign changes, then for $y \in Y$,*

$$f^*(y) = \int f(x)K(x,y)\,d\mu(x)$$

has at most k sign changes. Furthermore, if f^* has exactly k sign changes, then f and f^* have the same pattern of sign changes.

Total positivity of order two is familiar in economics and has had wide applications in the theory of moral hazard, as well as mechanism and market design.

REMARK 1: *Suppose that $K(x,y)$ is a probability density function, denoted $f(x|y)$, with respect to a σ-finite measure μ such that $\int f(x|y)\mu(dx) = 1$. Then, if $f(x|y)$ is TP_2, then $f(x|y)$ satisfies the MLRP.*[15]

Karlin's Theorem (commonly referred to as the Variation Diminishing Property (VDP)) allows us to observe that voters of higher type (higher i) are more likely to vote Republican provided $g(\beta|s)$ is TP_2. To see this, recall that, since voter preferences satisfy single-crossing (combined with our reordering), $\beta_i = u_i(R) - u_i(D)$ is a monotonic function with one sign change. The stochastic objective $f^*(s) = \int \beta g(\beta|s)\,d\beta$ is then also monotonic. Let k be an arbitrary constant and consider $f^*(s) - k = \int(\beta - k)g(\beta|s)\,d\beta$. Since $\beta - k$ has only one sign change, the VDP implies that $f^*(s) - k$ has only one sign change. This immediately implies monotonicity of $f^*(s)$. Monotonicity of $f^*(s)$ implies that for any two signals of voter types, $i > j$, the probability that type i votes Republican is greater than the probability that type j does.

B. *Proofs*

PROOF OF LEMMA 1:

The maximization problem can be described by the Lagrangian

$$L = \sum_{n=1}^{N} B(\mu_n) - \sum_{n=1}^{N} \lambda_n \left[\int_{-\infty}^{\infty} \psi_n(s)\,ds - \frac{1}{N} \right], \tag{6}$$

[15] For the classic reference to likelihood ratios and their applications to economics, see Milgrom (1981).

in addition to the boundary constraints. Note that the first-order necessary conditions imply

$$\Delta\psi_n(s)\left(b(\mu_n)\frac{\partial\mu_n}{\partial\psi_n(s)} - \lambda_n\right) = 0 \text{ for } n = i,j, \qquad s = s_1, s_2. \tag{7}$$

Now, consider districts i and j, and suppose that $\mu_i < \mu_j$.

Throughout, whenever we speak of removing voters of type s, we refer to an interval $[s - \varepsilon/2, s + \varepsilon/2]$. Denote the derivative of the objective function with respect to a switch of voters of type s from district j to district i as $\phi_{ji}(s)$. Then, for any ε, the change in the value of the objective function is

$$\Delta V(s) = \int_{s-\varepsilon/2}^{s+\varepsilon/2} \phi_{ji}(s')\,ds'.$$

Note that, as $\varepsilon \to 0$, the change in the value of the objective function from such a move approaches the derivative of the objective function at s multiplied by ε, since

$$\lim_{\varepsilon\to 0}\int_{s-\varepsilon/2}^{s+\varepsilon/2} \phi_{ji}(s')\,ds' = \phi_{ji}(s)\varepsilon.$$

The derivative of the objective function from moving voters of type s from district j and adding them to district i is

$$\phi_{ji}(s) = \left(b_i\frac{\partial\mu_i}{\partial\psi_i(s)} - \lambda_i\right) - \left(b_j\frac{\partial\mu_j}{\partial\psi_j(s)} - \lambda_j\right).$$

Implicitly differentiating (2), which determines the medians, yields

$$0 = \int_{-\infty}^{\infty} g(\mu_i|s)\psi_i(s)\,ds\partial\mu_i + G(\mu_i|s)\partial\psi_j(s);$$

$$\frac{\partial\mu_i}{\partial\psi_i(s)} = -\frac{G(\mu_i|s)}{\int_{-\infty}^{\infty} g(\mu_i|s)\psi_i(s)\,ds} \tag{8}$$

$$\equiv -\frac{G(\mu_i|s)}{\gamma_i(\mu_i)}. \tag{9}$$

Hence, the change in the value of the objective function is

$$\Delta V(s) = \varepsilon\left(\frac{b(\mu_j)}{\gamma_j(\mu_j)}G(\mu_j|s) - \frac{b(\mu_i)}{\gamma_i(\mu_i)}G(\mu_i|s) + \lambda_j - \lambda_i\right). \tag{10}$$

Note that if $\phi_{ji}(s) > \phi_{ji}(s')$, then $\Delta V(s) > \Delta V(s')$ for any $\varepsilon > 0$. While equation (10) need not be positive for all s in district n, it must be, $\forall s' \in \psi_j$ and $s \in \psi_i$, that $\phi_{ji}(s) \geq \phi_{ji}(s')$. Note that $\partial\phi_{ji}(s)/\partial s > 0$ is equivalent to $z(\mu_j|s)/z(\mu_i|s) < b(\mu_i)\gamma_j(\mu_j)/b(\mu_j)\gamma_i(\mu_i)$, and since the left-hand side is monotonically increasing in s from Condition 1, $\phi_{ji}(s)$ cannot be convex. If $s_1, s_2 \in \psi_i$, then, for any point $s' \in [s_1, s_2]$, $\phi_{ji}(s') > \min[\phi_{ji}(s_1), \phi_{ji}(s_2)]$. Thus, $s' \notin \psi_j$, if $\varepsilon > 0$.

This implies that any two districts j and i (where, without loss of generality $\mu_j > \mu_i$) cannot share voters of the same type except on a set of measure zero.

This also implies that districts must comprise vertical slices. Suppose that there exists an interval voter of types $[s - a, s + a]$ such that all voters of type $s' \in [s - a, s + a]$ are in both districts j and i. This contradicts the statement above that if $s_1, s_2 \in \psi_i$, then, for any point $s' \in [s_1, s_2]$, $s' \notin \psi_j$, if $\varepsilon > 0$.

PROOF OF LEMMA 2:

Suppose, by way of contradiction, that there exist districts j and i such that $\mu_j = \mu_i$, and that there exist intervals of positive measure about types s_1 and s_2 (with $s_1 > s_2$), which are in both districts. Consider moving a small mass from an interval about s_1 into district j and a comparable mass of voters around s_2 back into district i. The first-order conditions imply that the net gain, which must equal zero, is proportional to

$$\frac{b(\mu_j)}{\gamma_j(\mu_j)}[\mathrm{G}(\mu_j|s_2) - G(\mu_j|s_1)] - \frac{b(\mu_i)}{\gamma_i(\mu_i)}[G(\mu_i|s_2) - G(\mu_i|s_1)] \tag{11}$$

for $\varepsilon > 0$. Since $\mu_i = \mu_j$, we know that $b(\mu_j) = b(\mu_i)$ and $G(\mu_i|s_2) - G(\mu_i|s_1) = G(\mu_j|s_2) - G(\mu_j|s_1)$. Therefore, it must be that $\gamma_i(\mu_i) = \gamma_j(\mu_j)$.

Consider, again, the districts j and i with $\mu_i = \mu_j$. By Lemma 1, those voters in districts j and i must make up one or two complete vertical slices of $h(s)$. Since F has full support and the two aforementioned slices contain a positive interval of voter types, there must exist four voter types $s_1 < s_2 < \mu_j < s_3 < s_4$ such that $G(\mu_j|s_1) - G(\mu_j|s_2) = G(\mu_j|s_3) - G(\mu_j|s_4)$ and $\psi_i(s_1) > 0$, $\psi_i(s_4) > 0$, $\psi_j(s_2) > 0$, and $\psi_j(s_3) > 0$. In words, one district contains some of the inner type of voters, while the other district contains some of the more extreme types of voters relative to the district medians.

Now, consider a perturbation in which an equal mass of voters around type s_1 and around type s_4 are transferred to district j from district i, and similarly an equal mass of voters around type s_2 and around type s_3 are transferred from district j to i. By construction, both μ_j and μ_i remain unchanged, as does the value function; but $\gamma_i(\mu_i)$ and $\gamma_j(\mu_j)$ have changed. By definition,

$$\frac{\partial \gamma_i(\mu_i)}{\partial \psi(s)} = g(\mu_i|s),$$

and so the derivative of $\gamma_i(\mu_i)$ for perturbations of this type is

$$\partial\gamma_i(\mu_i) = \varepsilon\left(\frac{\partial \gamma_i(\mu_i)}{\partial \psi(s_2)} - \frac{\partial \gamma_i(\mu_i)}{\partial \psi(s_1)} + \frac{\partial \gamma_i(\mu_i)}{\partial \psi(s_3)} - \frac{\partial \gamma_i(\mu_i)}{\partial \psi(s_4)}\right)$$

$$= \varepsilon(\mathrm{g}(\mu_i|s_2) - g(\mu_i|s_1) + g(\mu_i|s_3) - g(\mu_i|s_4)).$$

But, by Condition 2, the modes of the lower signals lie below μ_i. Thus, we know that $\mathrm{g}(\mu_i|s_2) > \mathrm{g}(\mu_i|s_1)$, and similarly that $\mathrm{g}(\mu_i|s_3) > \mathrm{g}(\mu_i|s_4)$, and so $\partial\gamma_i(\mu_i) > 0$, for $\varepsilon > 0$. By similar reasoning, $\partial\gamma_j(\mu_j) < 0$. After performing such a perturbation, the new districting arrangement has $\mu_j = \mu_i$, while $\gamma_i(\beta) \neq \gamma_j(\beta)$. This now violates the condition above, which holds that for two districts that share a positive mass of voters and for which $\mu_j = \mu_i$, it must be that $\gamma_i(\beta) = \gamma_j(\beta)$. This new arrangement is not optimal, but the value function is unchanged from the old districting plan, and so the old plan cannot be optimal either—a contradiction.

PROOF OF LEMMA 3:

Suppose, by way of contradiction, that such a case existed. Without loss of generality, from Lemma 1, we can assume that districts i and k each comprise one whole slice. It also must be that $s^* < s'$ for all $s' \in \psi_i$ and that $s^* > s''$ for all $s'' \in \psi_k$. Denote $\bar{s}_i = sup\{s \in \psi_i\}$, $\bar{s}_k = sup\{s \in \psi_k\}$, $\underline{s}_i = inf\{s \in \psi_i\}$, and $\underline{s}_k = inf\{s \in \psi_k\}$. Of course, $\bar{s}_i > \underline{s}_i > s^* > \bar{s}_k > \underline{s}_k$.

The Lagrangian from equation (6) implies that, if $s \in \psi_j$, then

$$\varepsilon(-a_j G(\mu_j|s) - \lambda_j) \geq \varepsilon \max_n(-a_n G(\mu_n|s) - \lambda_n)$$

for all districts n, and, hence,

$$-a_j G(\mu_j|s) - \lambda_j \geq \max_n(-a_n G(\mu_n|s) - \lambda_n), \forall \varepsilon > 0,$$

where $a_n = b(\mu_n)/\gamma_n(\mu_n)$. These a_n coefficients represent the sensitivity of the median of district n to changes. For each district n, denote these expressions by η_n. We know that

$$\eta_i(\bar{s}_i) \geq \eta_j(\bar{s}_i) \text{ and } \eta_j(s^*) \geq \eta_i(s^*),$$

which implies that

$$a_j \leq a_i \frac{G(\mu_i|s^*) - G(\mu_i|\bar{s}_i)}{G(\mu_j|s^*) - G(\mu_j|\bar{s}_i)}. \tag{12}$$

Equation (12) states that district j must not be too sensitive compared to district i. Were this so, a profitable deviation would exist by shifting district i down to include s^* and giving voters of type $\bar{s}_i$ to district j. Similar arguments imply that

$$a_j \geq a_k \frac{G(\mu_k|\underline{s}_k) - G(\mu_k|s^*)}{G(\mu_j|\underline{s}_k) - G(\mu_j|s^*)}, \tag{13}$$

which has the interpretation that district j must be sensitive enough relative to district k so that shifting district k up to include s^* is not profitable. Of course, (12) and (13) can hold simultaneously only if the right-hand side of (12) is greater than or equal to the right-hand side of (13). This requires

$$\frac{a_k}{a_i} = \frac{b(\mu_k)\gamma_i(\mu_i)}{b(\mu_i)\gamma_k(\mu_k)} \leq \frac{G(\mu_i|s^*) - G(\mu_i|\bar{s}_i)}{G(\mu_k|\underline{s}_k) - G(\mu_k|s^*)} \frac{G(\mu_j|\underline{s}_k) - G(\mu_j|s^*)}{G(\mu_j|s^*) - G(\mu_j|\bar{s}_i)}. \tag{14}$$

Now, consider what happens to this ratio as we increase the precision of the signal (which can be thought of here as shrinking the conditional preference distribution G into the median). Since district i contains voters closer in signal to the median of district j, the ratio $[G(\mu_j|\underline{s}_k) - G(\mu_j|s^*)]/[G(\mu_j|s^*) - G(\mu_j|\bar{s}_i)]$ will shrink, going to 0 in the limit. On the other hand, both $G(\mu_i|s^*) - G(\mu_i|\bar{s}_i)$ and $G(\mu_k|\underline{s}_k) - G(\mu_k|s^*)$ rise to 1, since $\underline{s}_k < \mu_k < s^* < \mu_i < \bar{s}_i$. Thus, the right-hand side of (14) shrinks to 0 as the precision of the signal increases. Note, however, that the ratio a_k/a_i is bounded away from 0, since $\gamma_i(\mu_i)/\gamma_k(\mu_k)$ will limit to 1 (by the definition of $\gamma(\mu)$) and $b(\mu_k)/b(\mu_i)$ is bounded away from 0 since the medians μ_i and μ_k are bounded and

the c.d.f. B is strictly increasing. Thus, for sufficiently high signal quality, the inequality in (14) cannot hold—a contradiction.

PROOF OF PROPOSITION 1:

Apply Lemmas 1–3.

PROOF OF PROPOSITION 2:

Suppose not. Consider the districting plan that entirely packs. That is, consider the districting plan described by $N-1$ cutoffs $\{\tau_n\}_{n=1}^{N-1}$ (where $\tau_1 > \tau_2 > \cdots > \tau_{N-1}$) such that $s \in \psi_n$ if and only if $s \in [\tau_n, \tau_{n-1}]$. (For notational ease, suppose that $\tau_0 = \infty$ and $\tau_N = -\infty$.) Consider the marginal gain from moving voters of type τ_n from district n to district $n+1$ and moving voters from the far-left tail to district 1. Following the first-order condition in equations (7) and (9) (contained in the Appendix in the proof of Lemma 1 p. 137), the impact on μ_n for $n > 1$ is

$$\Delta\mu_n = \varepsilon\left(\frac{b(\mu_n)}{\gamma_n(\mu_n)}[G(\mu_n|\tau_n) - G(\mu_n|\tau_{n-1})]\right) > 0,$$

since $\tau_n < \mu_n < \tau_{n-1}$ and, therefore, $G(\mu_n|\tau_n) > 0.5 > G(\mu_n|\tau_{n-1})$. We use ε here to denote the small positive mass of voters moved in each shift, as we discuss in detail in the proof of Lemma 1. The impact on μ_1 will be

$$\Delta\mu_1 = \varepsilon\left(\frac{b(\mu_1)}{\gamma_1(\mu_1)}[G(\mu_1|\tau_1) - G(\mu_1|\tau_N)]\right) < 0,$$

where, for these purposes,

$$G(\mu_1|\tau_N) = \lim_{s\to-\infty} G(\mu_1|s) = 1.$$

Note further that, by the definition of τ_1 and μ_1, $G(\mu_1|\tau_1) > 0.5$.

Now, consider increasing the signal quality, which is to say decreasing the spread of the conditional distribution of β given s about the center of that distribution. Note that $G(\beta|s)$ is centered around s by Condition 2, and so, if $G(\mu_n|s) > 0.5$, then $\partial G(\mu_n|s)/\partial\sigma^2_{\beta|s} < 0$, so that $G(\mu_n|s)$ increases as the signal quality increases. (When we shrink $\sigma^2_{\beta|s}$, we refer to a reduction in the spread of the distribution around the median and mode of s, rather than the mean, so as to maintain Condition 2.) If $G(\mu_n|s) < 0.5$, then $\partial G(\mu_n|s)/\partial\sigma^2_{\beta|s} > 0$. The term $\gamma_1(\mu_1)$ will also increase, but it is (by definition) bounded above by the marginal distribution of β in the population. Thus, we know that, at least for high enough signal quality,

$$\frac{\partial\Delta\mu_n}{\sigma^2_{\beta|s}} > 0 \quad \forall n,$$

which implies that

$$\frac{\partial}{\sigma^2_{\beta|s}} \sum_{n=1}^{N}\Delta\mu_n = \varepsilon\frac{b(\mu_1)}{\gamma_1(\mu_1)} > 0.$$

The aggregate impact on the expected number of seats won from the proposed deviation becomes more positive or less negative as the signal quality increases. Finally, note that

$$\lim_{\sigma^2_{\beta|s}\to 0} G(\mu_1|\tau_1) = 1,$$

so that

$$\lim_{\sigma^2_{\beta|s}\to 0} \Delta\mu_1 = 0,$$

while

$$\lim_{\sigma^2_{\beta|s}\to 0} \Delta\mu_n > 0,\ n \neq 1,$$

and therefore

$$\lim_{\sigma^2_{\beta|s}\to 0} \sum_{n=1}^{N} \Delta\mu_n > 0.$$

Since the sum converges to the limit as $\sigma^2_{\beta|s}$ decreases, we know that there exists $\underline{\sigma}^2$ such that $\sum_{n=1}^{N}\Delta\mu_n > 0$ whenever $\sigma^2_{\beta|s} < \underline{\sigma}^2$.

PROOF OF PROPOSITION 3:

Suppose not. The choice variable for each district can be summarized by ψ_n, the proportion of R in the district. Then, there exist two districts j and i such that $\psi_j \neq \psi_i$ and $\psi_n > 0$ for $n = \{j, i\}$. Without loss of generality, let $\mu_j > \mu_i$. By Condition 1, $G(\beta|r)$ first-order stochastically dominates $G(\beta|d)$, and so $\psi_j > \psi_i$.

In order that there be no profitable deviations, it must that $\partial\mu_i/\partial\psi_i = \partial\mu_j/\partial\psi_j$. But, in general,

$$\frac{\partial^2\mu}{\partial\psi_n^2} = \frac{\dfrac{\partial\mu_n}{\partial\psi_n}}{\gamma(\mu)^2}\left\{\begin{array}{l}\big([g(\mu|d) - g(\mu|r)]b(\mu) + [G(\mu|d) - G(\mu|r)]b'(\mu)\big) \\ \quad \times [\psi_n(g(\mu|r) - g(\mu|d)) + g(\mu|d)] \\ \quad - b(\mu)[G(\mu|d) - G(\mu|r)][\gamma'(\mu) + g(\mu|r) - g(\mu|d)]\end{array}\right\},$$

which is positive when $\mu < 0$ and negative when $\mu > 0$. Since $\mu > 0 \Leftrightarrow \psi > 0.5$, the concavity of μ implies that one could never have $\psi_j > \psi_i \geq 0.5$, since then $\partial\mu_i/\partial\psi_i > \partial\mu_j/\partial\psi_j$, and so R could do better by increasing i and decreasing j. It also implies that there cannot be $0.5 > \psi_j \geq \psi_i$, since then $\partial\mu_i/\partial\psi_i < \partial\mu_j/\partial\psi_j$ and the opposite deviation would improve R's representation. Thus, there can be only one "odd district" with $0 < \psi < 0.5$, and all districts with $\psi > 0.5$ must have equal proportions of r and d.

Suppose that $N \to \infty$. Note that there can be only one odd district. Let the mass of voters in this district have Lebesgue measure τ. Since each district must have an equal mass of voters, $\tau = 1/N$. Clearly, $\lim_{N\to\infty}\tau = 0$.

PROOF OF PROPOSITION 4:

First, note that signal precision provides a partial ordering on conditional preference distribution. If the signal contains no information, the expected number of seats won by the gerrymanderer

is the population share. If the signal is perfectly precise such that $s = \beta$, it is possible (see Proposition 1) to create districts such that only the lowest median district has a median equal to the population median, while all others lie above. Hence, the gerrymanderer wins more seats in expectation with a perfect signal. Now, consider any two conditional preference distributions g and g' such that g provides a more precise signal than g'. The gerrymanderer must win at least as many seats in expectation under g than g' since the value function has the Blackwell Property. That is, she could construct a distribution c such that from g she could generate g'.

PROOF OF PROPOSITION 5:

Fix the optimal districting plan under $F(\beta, s)$ and consider the construction of the highest median district (without loss of generality, District 1) with median μ_1 given by $\int_{s \in \psi_1} G(\mu_1 | s) h(s)\, ds = \frac{1}{2N}$, comprising an upper and lower slice. Let the upper slice contain w_1 share of the voters in the district. Suppose that, under $\hat{F}(\beta, s)$, the gerrymanderer sets $\hat{\mu}_1 = \mu_1$. This can be achieved with at least as small an upper slice $\hat{w}_1 \leq w_1$, since the Republican voters (who make up more than half of the district) are at least as likely to vote Republican as before. If $\hat{w}_1 < w_1$, then note that all other districts $2, \ldots, N$ have a higher medians even if we set $\hat{w}_i = w_i$ for all i, that is, without reoptimizing their construction. If $\hat{w}_1 = w_1$, then repeat this procedure until finding a district n^* such that $\hat{w}_{n^*} < w_{n^*}$. By assumption that $\hat{F}$ has greater symmetric spread than F, this must be true for at least one district. Hence the value function under $\hat{F}(\beta, s)$ is higher than under $F(\beta, s)$. This reasoning must hold for any such pair of distributions.

PROOF OF PROPOSITION 6:

Consider an increase from N districts to mN, where m is an integer. By replication, the gerrymanderer could do at least as well with mN districts as with N—but this replication involves creating parfaits. From Lemma 2, this is a suboptimal strategy. Hence, the value function under the optimal strategy must be higher.

PROOF OF PROPOSITION 7:

Suppose that the objective function is now

$$E\left[V\left(\frac{1}{N}\sum_{n=1}^{N} w_n d_n\right)\right],$$

and suppose that V is a strictly increasing function. We can rewrite this expression as the sum of $V(D)$, where $D = 0, \ldots, N$, weighted by the combinatorial probability that the Republicans win exactly D districts. Note that this expression can be factored into two parts: those outcomes where R wins some district n, and those where R loses district n. Since the probability of winning a district is just $B(\mu_n)$, this expression is just

$$B(\mu_n)K_n + (1 - B(\mu_n))L_n,$$

where $K_n = \mathrm{E}[V | d_n = 1]$, the expected value if the Republican candidate wins in district n; and $L_n = \mathrm{E}[V | d_n = 0]$, the expected value if the Democrat wins in district n. Now, fix the districting scheme and consider the marginal benefit from a small deviation x in district n, which is

$$\frac{\partial E[V]}{\partial x} = b(\mu_n)(K_n - L_n)\frac{\partial \mu_n}{\partial x}.$$

The conditions from (7) must still hold for these new first-order conditions, but note that this expression is identical to the value derived in equation (7) but for the term $(K_n - L_n)$, which is fixed for all deviations from a districting plan. Thus, the "sensitivities" $\{a_n\}_{n=1}^{N}$ (as in Lemma 3) are now differently scaled, but the constant does not affect any proofs. Propositions 1 through 6 hold.

PROOF OF PROPOSITION 8:

Suppose candidates are each associated with an electoral benefit ζ_n such that voters support them if $\beta - A + \zeta_n > 0$. In this case, the Republican candidate wins district n if and only if $\mu_n + \zeta_n > A$, which occurs with probability $B(\mu_n + \zeta_n)$. The marginal benefit to R from a small deviation x in district n would be

$$\frac{\partial V}{\partial x} = b(\mu_n + \zeta_n)\frac{\partial \mu_n}{\partial x}.$$

Since the district-specific constant $b(\mu_n + \zeta_n)$ cancels out in Lemma 1, the proof still holds. Lemma 2 is similarly unaffected, as the constant does not affect the proofs. In Lemma 3, the ratio $a_k/a_i = b(\mu_k + \zeta_k)\gamma_i(\mu_i + \zeta_i)/b(\mu_i + \zeta_i)\gamma_k(\mu_k + \zeta_k)$ is no longer bounded away from 0, because $\gamma_i(\mu_i + \zeta_i)$ need not limit to 1 as the precision of the signal increases.

REFERENCES

Ansolabehere, Stephen, James M. Snyder, and Charles Stewart III. 2000. "Old Voters, New Voters, and the Personal Vote: Using Redistricting to Measure the Incumbency Advantage." *American Journal of Political Science*, 44(1): 17–34.

Ansolabehere, Stephen, James M. Snyder, and Charles Stewart III. 2001. "Candidate Positioning in the U.S. House Elections." *American Journal of Political Science*, 45(1): 36–59.

Ashenfelter, Orley, and Stanley Kelly, Jr. 1975. "Determinants of Participation in Presidential Elections." *Journal of Law and Economics*, 18(3): 695–733.

Buchanan, James M., and Gordon Tullock. 1962. *The Calculus of Consent: Logical Foundations of Constitutional Democracy*. Ann Arbor: University of Michigan Press.

Coate, Stephen, and Brian Knight. 2007. "Socially Optimal Districting: A Theoretical and Empirical Exploration." *Quarterly Journal of Economics*, 122(4): 1409–71.

Cox, Gary W., and Jonathan N. Katz. 2002. *Elbridge Gerry's Salamander: The Electoral Consequences of the Apportionment Revolution*. Cambridge, MA: Cambridge University Press.

Downs, Anthony. 1957. *An Economic Theory of Democracy*. New York: Harper and Row.

Gelman, Andrew, and Gary King. 1990. "Estimating the Electoral Consequences of Legislative Redistricting." *Journal of the American Statistical Association*, 85(410): 274–82.

Gelman, Andrew, and Gary King. 1994. "Enhancing Legislative Redistricting Through Legislative Redistricting." *American Political Science Review*, 88(3): 541–59.

Gilligan, Thomas W., and John G. Matsusaka. 1999. "Structural Constraints on Partisan Bias under the Efficient Gerrymander." *Public Choice*, 100(1–2): 65–84.

Glaeser, Edward L., Giacomo A. M. Ponzetto, and Jesse M. Shapiro. 2005. "Strategic Extremism: Why Republicans and Democrats Divide on Religious Values." *Quarterly Journal of Economics*, 120(4): 1283–1330.

Hoeffding, Wassily. 1956. "On the Distribution of the Number of Successes in Independent Trials." *Annals of Mathematical Statistics*, 27(3): 713–21.

Issacharoff, Samuel, Pamela S. Karlan, and Richard H. Pildes. 2002. *The Law of Democracy: Legal Structure of the Political Process*. New York: Foundation Press.

Kan, Kamhon, and C. C. Yang. 2001. "On Expressive Voting: Evidence from the 1988 U.S. Presidential Election." *Public Choice*, 108(3-4): 295–312.

Karlin, Samuel, and Herman Rubin. 1956. "The Theory of Decision Procedures for Distributions with the Monotone Likelihood Ratio." *Annals of Mathematical Statistics*, 27(2): 272–99.

Lee, David S., Enrico Moretti, and Matthew J. Butler. 2004. "Do Voters Affect or Elect Policies? Evidence from the U.S. House." *Quarterly Journal of Economics*, 119(3): 807–59.

Matsusaka, John G., and Filip Palda. 1993. "The Downsian Voter Meets the Ecological Fallacy." *Public Choice*, 77(4): 855–918.

Milgrom, Paul R. 1981. "Good News and Bad News: Representation Theorems and Applications." *Bell Journal of Economics*, 12(2): 380–91.

Owen, Guillermo, and Bernard Grofman. 1988. "Optimal Partisan Gerrymandering." *Political Geography Quarterly*, 7(1): 5–22.

Poisson, Siméon-Denis. 1837. *Recherches sur la probabilité des jugements en matière criminelle et en matière civile, précedées des regles générales du calcul des probabilités.* Paris: Bachelier.

Rothstein, Paul. 1991. "Representative Voter Theorems." *Public Choice*, 72(2–3): 193–212.

Sherstyuk, Katerina. 1998. "How to Gerrymander: A Formal Analysis." *Public Choice*, 95(1–2): 27–49.

Shotts, Kenneth W. 2002. "Gerrymandering, Legislative Composition, and National Policy Outcomes." *American Journal of Political Science*, 46(2): 398–414.

Washington, Ebonya. 2006. "How Black Candidates Affect Turnout." *Quarterly Journal of Economics*, 121(3): 973–98.

Wolfinger, Raymond, and Steven Rosenstone. 1980. *Who Votes?* New Haven: Yale University Press.

Luc Anselin

Local Indicators of Spatial Association—LISA

The capabilities for visualization, rapid data retrieval, and manipulation in geographic information systems (GIS) have created the need for new techniques of exploratory data analysis that focus on the "spatial" aspects of the data. The identification of local patterns of spatial association is an important concern in this respect. In this paper, I outline a new general class of local indicators of spatial association (LISA) and show how they allow for the decomposition of global indicators, such as Moran's I, into the contribution of each observation. The LISA statistics serve two purposes. On one hand, they may be interpreted as indicators of local pockets of nonstationarity, or hot spots, similar to the G_i and G_i^* statistics of Getis and Ord (1992). On the other hand, they may be used to assess the influence of individual locations on the magnitude of the global statistic and to identify "outliers," as in Anselin's Moran scatterplot (1993a). An initial evaluation of the properties of a LISA statistic is carried out for the local Moran, which is applied in a study of the spatial pattern of conflict for African countries and in a number of Monte Carlo simulations.

1. INTRODUCTION

The increased availability of large spatially referenced data sets and the sophisticated capabilities for visualization, rapid data retrieval, and manipulation in geographic information systems (GIS) have created a demand for new techniques for spatial data analysis of both an exploratory and a confirmatory nature (Anselin and Getis 1992; Openshaw 1993). Although many methods are available in the toolbox of the geographical analyst, only few of those are appropriate to deal explicitly with the "spatial" aspects in these large data sets (Anselin 1993b).

In the analysis of spatial association, it has long been recognized that the as-

The research of which this paper is an outgrowth was supported in part by grants SES 88-10917 (to the National Center for Geographic Information and Analysis, NCGIA) and SES 89-21385 from the U.S. National Science Foundation, and by grant GA-AS 9212 from the Rockefeller Foundation. Earlier versions were presented at the NCGIA Workshop on Exploratory Spatial Data Analysis and GIS, Santa Barbara, Calif., February 25–27, 1993, and at the GISDATA Specialist Meeting on GIS and Spatial Analysis, Amsterdam, The Netherlands, December 1–5, 1993. The comments by Arthur Getis and two anonymous referees on an earlier draft are greatly appreciated.

Luc Anselin is research professor of regional science at the Regional Research Institute of West Virginia University, where he is also professor of economics, adjunct professor of geography, and adjunct professor of agricultural and resource economics.

Geographical Analysis, Vol. 27, No. 2 (April 1995) © Ohio State University Press
Submitted 1/94. Revised version accepted 6/94.

sumption of stationarity or structural stability over space may be highly unrealistic, especially when a large number of spatial observations are used. Spatial structural instability or spatial drift has been incorporated in a number of modeling approaches. For example, discrete spatial regimes are accounted for in spatial analysis of variance (Griffith 1978, 1992; Sokal et al. 1993), and in regression models with spatial structural change (Anselin 1988, 1990). Continuous variation over space is the basis for the spatial expansion paradigm (Casetti 1972, 1986; Jones and Casetti 1992) and spatial adaptive filtering (Foster and Gorr 1986; Gorr and Olligschlaeger 1994). In exploratory spatial data analysis (ESDA), the predominant approach to assess the degree of spatial association still ignores this potential instability, as it is based on global statistics such as Moran's I or Geary's c (as in Griffith 1993). A focus on local patterns of association (hot spots) and an allowance for local instabilities in overall spatial association has only recently been suggested as a more appropriate perspective, for example, in Getis and Ord (1992), Openshaw (1993), and Anselin (1993b). Examples of techniques that reflect this approach are the various geographical analysis machines developed by Openshaw and associates (for example, Openshaw, Brundson, and Charlton 1991; and Openshaw, Cross, and Charlton 1990), the distance-based statistics of Getis and Ord (1992) (see also Ord and Getis 1994), and the Moran scatterplot (Anselin 1993a). Also, a few approaches have been suggested that are based on a geostatistical perspective, such as the pocket plot of Cressie (1991) and the interactive spatial graphics of Haslett et al. (1991).

In the current paper, I elaborate upon this general idea and outline a class of *local indicators of spatial association* (LISA). These indicators allow for the decomposition of global indicators, such as Moran's I, into the contribution of each individual observation. I suggest that this class of indicators may become a useful addition to the toolbox of ESDA techniques in that two important interpretations are combined: the assessment of significant local spatial clustering around an individual location, similar to the interpretation of the G_i and G_i^* statistics of Getis and Ord (1992); and the indication of pockets of spatial nonstationarity, or the suggestion of outliers or spatial regimes, similar to the use of the *Moran scatterplot* of Anselin (1993a).

In the remainder of the paper, I first outline the general principles underlying a LISA statistic, and suggest how it may be interpreted. I next show how a number of familiar global spatial autocorrelation statistics may be expressed in the form of a LISA. As an example of a LISA, I examine the *local Moran* more closely, first empirically, comparing it to the G_i^* statistic and the Moran scatterplot in an analysis of spatial pattern of conflict between African nations in the period 1966–78. This is followed by a series of simple Monte Carlo experiments, to provide further insight into the properties of the local Moran, its interpretation, and the relation between global and local spatial association. I close with some concluding remarks on future research directions.

2. LOCAL INDICATORS OF SPATIAL ASSOCIATION

Definition

As an operational definition, I suggest that a *local indicator of spatial association* (LISA) is any statistic that satisfies the following two requirements:

a. the LISA for each observation gives an indication of the extent of significant spatial clustering of similar values around that observation;
b. the sum of LISAs for all observations is proportional to a global indicator of spatial association.

More formally, but still in general terms, I express a LISA for a variable y_i, observed at location i, as a statistic L_i, such that

$$L_i = f(y_i, y_{J_i}), \tag{1}$$

where f is a function (possibly including additional parameters), and the y_{J_i} are the values observed in the neighborhood J_i of i.

The values y used in the computation of the statistic may be the original (raw) observations, or, more appropriately, some standardization of these in order to avoid scale dependence of the local indicators, similar to the practice often taken for global indicators of spatial association. For example, in Moran's I, as well as in its local version discussed in the next section, the observations are taken as deviations from their mean.

The neighborhood J_i for each observation is defined in the usual fashion, and may be formalized by means of a spatial weights or contiguity matrix, **W**. The columns with nonzero elements in a given row of this matrix indicate the relevant *neighbors* for the observation that corresponds to the row, that is, the elements of J_i. Examples of criteria that could be used to define neighbors are first-order contiguity and critical distance thresholds. The spatial weights matrix may be row-standardized (such that its row elements sum to one) to facilitate interpretation of the statistics, but this is not required. However, when row standardization is carried out, the function $f(y_i, y_{J_i})$ typically corresponds to a form of weighted average of the values at all observations $j \in J_i$.

The L_i should be such that it is possible to infer the statistical significance of the pattern of spatial association at location i. More formally, this requires the operationalization of a statement such as

$$\text{Prob}\,[L_i > \delta_i] \leq \alpha_i, \tag{2}$$

where δ_i is a critical value, and α_i is a chosen significance or pseudo significance level, for example, as the result of a randomization test.

The second requirement of a LISA, that is, its relation to a global statistic, may be stated formally as

$$\sum_i L_i = \gamma\Lambda, \tag{3}$$

where Λ is a global indicator of spatial association and γ is a scale factor. In other words, the sum of the local indicators is proportional to a global indicator. For the latter, a statement such as

$$\text{Prob}\,[\Lambda > \delta] \leq \alpha, \tag{4}$$

indicates significant spatial association over the whole data set.

Identification of Local Spatial Clusters

Local spatial clusters, sometimes referred to as *hot spots*, may be identified as those locations or sets of contiguous locations for which the LISA is significant. Similar to the rationale behind the significance tests for the G_i and G_i^* statistics of Getis and Ord (1992), the general LISA can be used as the basis for a test on the null hypothesis of no local spatial association. However, in contrast to what holds for the G_i and G_i^* statistics, general results on the distribution of a generic LISA may be hard to obtain. This is similar to the problems encountered in

deriving distributions for global statistics, for which typically only approximate or asymptotic results are available.[1] An alternative is the use of a conditional randomization or permutation approach to yield empirical so-called pseudo significance levels (for example, as in Hubert 1987). The randomization is conditional in the sense that the value y_i at a location i is held fixed (that is, not used in the permutation) and the remaining values are randomly permuted over the locations in the data set. For each of these resampled data sets, the value of L_i can be computed. The resulting empirical distribution function provides the basis for a statement about the extremeness (or lack of extremeness) of the observed statistic, relative to (and conditional on) the values computed under the null hypothesis (the randomly permuted values). In practice, this is straightforward to implement, since for each location only as many values as there are in the neighborhood set need to be resampled. Note that this same approach can also easily be applied to the G_i and G_i^* statistics.

A complicating factor in the assessment of significance of LISAs is that the statistics for individual locations will tend to be correlated, as pointed out by Ord and Getis (1994) in the context of their G_i and G_i^* statistics. In general, whenever the neighborhood sets J_i and J_k of two locations i and k contain common elements, the corresponding L_i and L_k will be correlated. Due to this correlation, and the associated problem of multiple comparisons, the usual interpretation of significance will be flawed. Moreover, it is typically impossible to derive the exact marginal distribution of each statistic and the significance levels must be approximated by Bonferroni inequalities or following the approach suggested in Sidák (1967).[2] This means that when the overall significance associated with the multiple comparisons (correlated tests) is set to α, and there are m comparisons, then the individual significance α_i should be set to either α/m (Bonferroni) or $1-(1-\alpha)^{1/m}$ (Sidák). The latter procedure, which yields slightly sharper bounds, is suggested by Ord and Getis (1994), with $m = n$, that is, the number of observations.[3] Note that the use of Bonferroni bounds may be too conservative for the LISA of individual locations. For example, if m is indeed taken to equal the number of observations, then an overall significance of $\alpha = 0.05$ would imply individual levels of $\alpha_i = 0.0005$ in a data set with one hundred observations, possibly revealing only very few if any "significant" locations. However, since the correlation between individual statistics is due to the common elements in the neighborhood sets, only for a small number of locations k will the statistics actually be correlated with an individual L_i. For example, on a regular lattice using the queen criterion of contiguity, first-order neighbors (ignoring border and corner cells) will have four common elements in their neighborhood sets, second-order neighbors three, and higher-order neighbors none. Clearly, the number of common neighbors does not change with the number of observations, so that using the latter in the computation of the Bonferroni bounds may be overly conservative. Hence, while it is obvious that some correction to the individual significance levels is needed, the extent to which it is indeed necessary to take $m = n$ remains to be further investigated.

[1] With the exception of the results in Tiefelsdorf and Boots (1994), the general statement by Cliff and Ord (1981, p. 46) still holds: "except for very small lattices, exact evaluation of the distribution function is impractical and approximations must be found."

[2] An application of these procedures to the interpretation of the significance of a spatial correlogram was earlier suggested by Oden (1984).

[3] Note that the Sidák approach only holds when the statistics under consideration are multivariate normal, which is unlikely to be the case for the general class of LISA statistics [see also Savin (1980) for an extensive discussion of the relative merits of various notions of bounds].

Indication of Local Instability

The indication of local patterns of spatial association may be in line with a global indication, although this is not necessarily the case. In fact, it is quite possible that the local pattern is an aberration that the global indicator would not pick up, or it may be that a few local patterns run in the opposite direction of the global spatial trend. The second requirement in the definition of a LISA statistic is imposed to allow for the decomposition of a global statistic into its constituent parts. This is of interest to assess the extent to which the global statistic is representative of the average pattern of local association. If the underlying process is stable throughout the data, then one would expect the local indications to show little variation around their average. In other words, local values that are very different from the mean (or median) would indicate locations that contribute more than their expected share to the global statistic. These may be outliers or high leverage points and thus would invite closer scrutiny. This interpretation is roughly similar to the use of Cressie's (1991) pocket plots in geostatistics. By imposing the requirement that the L_i sum to a magnitude that is proportional to a global statistic, their distribution around the mean $\gamma\Lambda/n$ can be evaluated. Extreme L_i can be identified as outliers in this distribution, for example, as those values that are more than two standard deviations from the mean (the two-sigma rule) or more than 1.5 times the interquartile range larger than the third quartile (for example, in a box plot).

This second interpretation of the LISA statistics is similar to the use of a Moran scatterplot to identify outliers and leverage points for Moran's I (Anselin 1993a). In general, it may be more appropriate than the interpretation of locations as hot spots suggested in the previous section when an indicator of global spatial association is significant. In this respect, it is important to note that the Getis-Ord G_i and G_i^* statistics were suggested to detect significant spatial clustering at a local level when global statistics do not provide evidence of spatial association (Getis and Ord 1992, p. 201). Indeed, in their example, Getis and Ord find no global autocorrelation for SIDS cases in North Carolina counties, while several significant local clusters are indicated. However, the opposite case often occurs as well, that is, a strong and significant indication of global spatial association may hide totally random subsets, particularly in large data sets. For example, in an analysis of the 1930 elections in Weimar Germany, O'Loughlin, Flint, and Anselin (1994) found that a highly significant Moran's I at the level of 921 electoral districts in effect hides several distinct local patterns of spatial clustering and complete spatial randomness for six regional subsets. In such an instance, the distribution of the L_i statistic as indicator of local spatial clustering will be affected by the presence of global spatial association. However, the second interpretation of LISA statistics, as indications of outliers or leverage points in the computation of a global statistic is not affected. I return to this issue in section 5.

3. LISA FORM OF FAMILIAR SPATIAL AUTOCORRELATION STATISTICS

Local Gamma

A broad class of spatial association statistics may be based on the general index of matrix association or Γ index, originally outlined in Mantel (1967). The application of the Γ index to spatial autocorrelation in a wide range of contexts is described in a series of papers by Hubert and Golledge (for example, Hubert 1985; Hubert, Golledge, and Costanzo 1981; Hubert et al. 1985; Costanzo,

Hubert, and Golledge 1983).[4] Such an index consists of the sum of the cross products of the matching elements a_{ij} and b_{ij} in two matrices of similarity, say **A** and **B**, such that

$$\Gamma = \sum_i \sum_j a_{ij} b_{ij}. \tag{5}$$

Measures of spatial association are obtained by expressing spatial similarity in one matrix (for example, a contiguity or spatial weights matrix) and value similarity in the other. Different measures of value similarity yield different indices for spatial association. For example, using $a_{ij} = x_i x_j$ yields a Moran-like measure, setting $a_{ij} = (x_i - x_j)^2$ yields a Geary-like index, while taking $a_{ij} = |x_i - x_j|$ results in an indicator equivalent to the one suggested by Royaltey, Astrachan, and Sokal (1975) [see, for example, Anselin (1986) for details on the implementation].

Since the Γ index is a simple sum over the subscript i, a local Gamma index for a location i may be defined as

$$\Gamma_i = \sum_j a_{ij} b_{ij}. \tag{6}$$

Similar to what holds for the global Γ measure, different measures of value similarity will yield different indices of local spatial association. It is easy to see that the Γ_i statistics sum to the global measure Γ. It is possible that the distribution of the individual Γ_i can be approximated using the principles outlined by Mielke (1979) and Costanzo et al. (1983), though this is likely to be complex, and beyond the current scope. On the other hand, the implementation of a conditional permutation approach is straightforward. This allows the individual Γ_i to be interpreted as indicators of significant local spatial clusters. The second interpretation of the LISA statistic, as a diagnostic for outliers or leverage points can be carried out by comparing the distribution of the Γ_i to Γ/n.

Local Moran

As a special case of the local Gamma, a local Moran statistic for an observation i may be defined as

$$I_i = z_i \sum_j w_{ij} z_j, \tag{7}$$

where, analogous to the global Moran's I, the observations z_i, z_j are in deviations from the mean, and the summation over j is such that only neighboring values $j \in J_i$ are included. For ease of interpretation, the weights w_{ij} may be in row-standardized form, though this is not necessary, and by convention, $w_{ii} = 0$.

It can be easily seen that the corresponding global statistic is indeed the familiar Moran's I. The sum of local Morans is

$$\sum_i I_i = \sum_i z_i \sum_j w_{ij} z_j, \tag{8}$$

while Moran's I is

[4] Note that this statistic also forms the basis for the derivation of the distribution of Moran's I and Geary's c statistics in Cliff and Ord (1981, p. 23 and chapter 2). In Getis (1991), this index is applied to integrate spatial association statistics and spatial interaction models into a common framework.

$$I = (n/S_0) \sum_i \sum_j w_{ij} z_i z_j / \sum_i z_i^2, \tag{9}$$

or

$$I = \sum_i I_i / \left[S_0 \left(\sum_i z_i^2 / n \right) \right], \tag{10}$$

where $S_0 = \sum_i \sum_j w_{ij}$. Using the same notation as Cliff and Ord (1981, p. 45), and taking $m_2 = \sum_i z_i^2/n$ as the second moment (a consistent, but not unbiased estimate of the variance), the factor of proportionality between the sum of the local and the global Moran is, in the notation of (3),

$$\gamma = S_0 m_2. \tag{11}$$

Note that for a row-standardized spatial weights matrix, $S_0 = n$, so that $\gamma = \sum_i z_i^2$, and for standardized variables (that is, with the mean subtracted and divided by the standard deviation), $m_2 = 1$, so that $\gamma = S_0$. Also, the same type of results obtain if instead of (7) each local indicator is divided by m_2, which is a constant for all locations. In other words, the local Moran would then be computed as

$$I_i = (z_i/m_2) \sum_j w_{ij} z_j, \tag{12}$$

The moments for I_i under the null hypothesis of no spatial association can be derived using the principles outlined by Cliff and Ord (1981, pp. 42–46) and a reasoning similar to the one by Getis and Ord (1992, pp. 190–92). For example, for a randomization hypothesis, the expected value turns out to be

$$E[I_i] = -w_i/(n-1), \tag{13}$$

with w_i as the sum of the row elements, $\sum_j w_{ij}$, and the variance is found as

$$\begin{aligned} \text{Var}\,[I_i] &= w_{i(2)}(n - b_2)/(n-1) \\ &\quad + 2w_{i(kh)}(2b_2 - n)/(n-1)(n-2) - w_i^2/(n-1)^2, \end{aligned} \tag{14}$$

with $b_2 = m_4/m_2^2$, $m_4 = \sum_i z_i^4/n$ as the fourth moment, $w_{i(2)} = \sum_{j \neq i} w_{ij}^2$, and $2w_{i(kh)} = \sum_{k \neq i} \sum_{h \neq i} w_{ik} w_{ih}$. The details of the derivation are given in Appendix A.

A test for significant local spatial association may be based on these moments, although the exact distribution of such a statistic is still unknown. This is further explored in section 5. Alternatively, a conditional randomization approach may be taken, as outlined earlier. Given the structure of the statistic in (12), it follows that only the quantity $\sum_j w_{ij} z_j$ needs to be computed for each permutation (since the z_i/m_2 remains constant). Note that the randomization method applied to (12) will yield the same empirical reference distribution as when applied to the Getis and Ord G_i and G_i^* statistics. Hence, inference based on this nonparametric approach will be identical for the two statistics. This easily follows from considering which elements in the statistics change for each permutation of the data. For example, the G_i statistic for an observation i is defined as

$$G_i = \sum_j w_{ij}(d)\, z_j / \sum_j z_j, \tag{15}$$

where $w_{ij}(d)$ are the elements in a distance-based weights matrix [for details, see Getis and Ord (1992)]. The only aspect of equation (15) that changes with each permutation is the numerator, since the denominator does not depend on the spatial allocation of observations. Clearly, this is the same term as the varying part of the numerator in (12). In other words, the pseudo significance levels (that is, the inference) generated with a permutation approach applied to the I_i statistic will be identical to that for a G_i or G_i^* statistic.[5]

The interpretation of the local Moran as an indicator of local instability follows easily from the relation between local and global statistics expressed in equation (11). Specifically, the average of the I_i will equal the global I, up to a factor of proportionality. Extreme contributions may thus be identified by means of simple rules, such as the two-sigma rule, or by identifying outliers in a box plot. Note that this notion of extremeness does not imply that the corresponding I_i are significant in the sense outlined earlier, but only indicates the importance of observation i in determining the global statistic. This similarity to the identification of outliers, leverage and influence points in the Moran scatterplot (Anselin 1993a) will be further examined in the empirical illustration.

Local Geary

Using the same principles as before, a local Geary statistic for each observation i may be defined as

$$c_i = \sum_j w_{ij}(z_i - z_j)^2, \tag{16}$$

or as

$$c_i = (1/m_2) \sum_j w_{ij}(z_i - z_j)^2, \tag{17}$$

using the same notation as before. Using expression (17) (without loss of generality), the summation of the c_i over all observations yields

$$\sum_i c_i = n \left[\sum_i \sum_j w_{ij}(z_i - z_j)^2 / \sum_i z_i^2 \right]. \tag{18}$$

In comparison, Geary's familiar c statistic is

$$c = [(n-1)/2S_0] \left[\sum_i \sum_j w_{ij}(z_i - z_j)^2 / \sum_i z_i^2 \right]. \tag{19}$$

Thus, the factor of proportionality between the sum of the local and the global Geary statistic is, in the notation of (3),

$$\gamma = 2nS_0/(n-1). \tag{20}$$

Clearly, for row-standardized weights, since $S_0 = n$, this factor becomes

[5] See also Ord and Getis (1994) for a discussion of the relationship between their statistics and Moran's I.

$2n^2/(n-1)$. The c_i statistic is interpreted in the same way as the local Gamma and the local Moran.

4. ILLUSTRATION: SPATIAL PATTERNS OF CONFLICT IN AFRICA

A geographical perspective has received much interest in recent years in the analysis of international interactions in general, and of international conflict in particular [see, for example, the review by Diehl (1992)]. Measures of spatial association, such as Moran's I, have been applied to quantitative indices for various types of conflicts and cooperation between nation-states, such as those contained in the COPDAB data base (Azar 1980). For such indices of international conflict and cooperation, both O'Loughlin (1986) and Kirby and Ward (1987) found significant patterns of spatial association indicated by Moran's I. The importance of spatial effects in the statistical analysis of conflict and cooperation was confirmed in a study of the interactions among forty-two African nations, over the period 1966–78, reported in a series of papers by O'Loughlin and Anselin (O'Loughlin and Anselin 1991, 1992; Anselin and O'Loughlin 1990, 1992). For an index of total conflict in particular, there was strong evidence of both positive spatial autocorrelation (as indicated by Moran's I, by a Γ index of spatial association, and by the estimates in a mixed regressive, spatial autoregressive model), as well as of spatial heterogeneity in the form of two distinct spatial regimes (as indicated by Getis-Ord G_i^* statistics and the results of a spatial Chow test on the stability of regression coefficients). This phenomenon is thus particularly suited to illustrate the LISA statistics suggested in this paper. The illustration focuses on the two interpretations of the LISA statistics, as indicators of local spatial clusters and as diagnostics for local instability. It is approached from the perspective of exploratory spatial data analysis and the substantive interpretation of the models is not considered here [see O'Loughlin and Anselin (1992) for a more extensive discussion].

The spatial pattern of the index for total conflict is illustrated in the quartile map in Figure 1, with the darkest shade corresponding to the highest quartile [for details on the data sources, see Anselin and O'Loughlin (1992)]. The suggestion of spatial clustering of similar values that follows from a visual inspection of this map is confirmed by a strong positive and significant Moran's I of 0.417, with an associated standard normal z-value of 4.35 ($p < 0.001$), and a Geary c index of 0.584, with associated standard normal z-value of -2.90 ($p < 0.002$).[6] These statistics are computed for a row-standardized spatial weights matrix based on first-order contiguity (common border), given the importance of borders in the study of international conflict (Diehl 1992).

Identification of Local Spatial Clusters

I first focus on a comparison of the identification of local spatial clusters provided by the Getis-Ord G_i^* statistic (as a standardized z-value) and the local Moran I_i indicator presented in equation (12). Note that the former, while being a statistic for local spatial association, is not a LISA in the terminology of section 2, since its individual components are not related to a global statistic of spatial association. This requirement is not needed for the identification of significant local spatial clusters, but it is important for the second interpretation of a LISA, as a diagnostic of local instability in measures of global spatial

[6] All computations were carried out with the *SpaceStat* software for spatial data analysis (Anselin 1992); the map was created with the *Idrisi* software (Eastman 1992), using the *SpaceStat-Idrisi* interface; other graphics were produced by means of the *SPlus* statistical software.

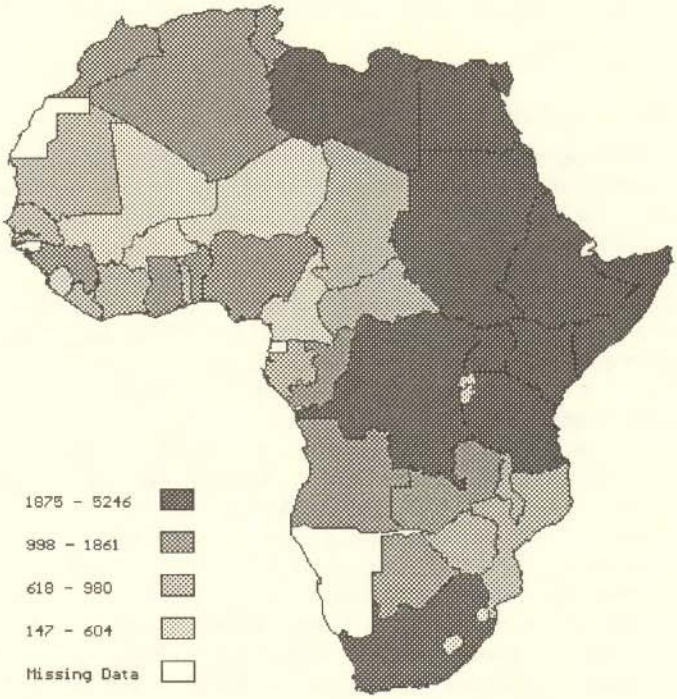


FIG. 1. Total Conflict Index for African Countries (1966–78)

association (for example, in the presence of significant global association), which is discussed in the next section.

Using the same row-standardized weights matrix as for the global measures given earlier, the results for the indicators of local spatial association are reported in the third and fifth columns of Table 1, for each of the forty-two countries in the example. The standardized z-value for I_i, computed by subtracting the expected value (13) and dividing by the standard deviation [the square root of (14)], is listed in the sixth column. Two indications of significance are given, one based on an approximation by the normal distribution, p_n (in the seventh column of Table 1) and one derived from conditional randomization, using a sample of 10,000 permutations, p_r (in the last column of Table 1).[7] As mentioned earlier, the pseudo significance obtained by means of a conditional randomization procedure is identical for G_i^* and I_i. While this may suggest that the normal approximation shown to hold for G_i^* (listed in column four) and assessed in detail in Ord and Getis (1994) may be valid for the I_i statistic as well, this has not been demonstrated. In fact, evidence from some initial Monte Carlo experiments in section 5 seems to indicate otherwise.

Note that the two statistics measure different concepts of spatial association. For the G_i^* statistic, a positive value indicates a spatial clustering of high values, and a negative value a spatial clustering of low values, while for the I_i, a positive value indicates spatial clustering of *similar* values (either high or low), and negative values a clustering of *dissimilar* values (for example, a location with high values surrounded by neighbors with low values), as in the interpretation of the

[7] More precisely, the sample consists of the original observed value of the statistic and the values computed for 9,999 conditionally randomized data sets.

TABLE 1
Measures of Local Spatial Association

Id	Country	G_i^*	p	I_i	$z(I_i)$	p_n	p_r
1	Gambia	−0.984	0.1626	0.375	0.428	0.3342	0.4727
2	Mali	−1.699	0.0447	0.464	1.482	0.0692	0.0456
3	Senegal	−1.463	0.0717	0.257	0.623	0.2667	0.0270
4	Benin	−1.301	0.0966	0.194	0.484	0.3142	0.0612
5	Mauritania	−0.605	0.2726	0.097	0.269	0.3940	0.4111
6	Niger	−1.049	0.1471	0.231	0.774	0.2193	0.2404
7	Ivory Coast	−1.417	0.0782	0.290	0.788	0.2154	0.0611
8	Guinea	−1.449	0.0737	0.183	0.519	0.3020	0.0365
9	Burkina Faso	−1.751	0.0400	0.508	1.479	0.0695	0.0339
10	Liberia	−1.041	0.1490	0.186	0.398	0.3452	0.1333
11	Sierra Leone	−0.870	0.1921	0.265	0.444	0.3286	0.4006
12	Ghana	−1.103	0.1351	0.148	0.326	0.3721	0.0885
13	Togo	−0.991	0.1610	0.219	0.462	0.3219	0.1894
14	Cameroon	−1.133	0.1285	0.259	0.711	0.2387	0.1706
15	Nigeria	−1.173	0.1205	0.114	0.306	0.3798	0.0851
16	Gabon	−0.789	0.2150	0.204	0.349	0.3634	0.3139
17	CAR	1.174	0.1203	−0.442	−1.046	0.1477	0.0613
18	Chad	0.463	0.3218	−0.105	−0.225	0.4111	0.2125
19	Congo	−0.203	0.4198	0.011	0.079	0.4684	0.4734
20	Zaire	2.023	0.0216	0.710	2.591	0.0048	0.0404
21	Angola	1.235	0.1085	0.118	0.270	0.3936	0.0999
22	Uganda	3.336	**0.0004**	1.943	4.928	**0.0000**	0.0031
23	Kenya	3.503	**0.0002**	1.197	3.060	**0.0011**	0.0016
24	Tanzania	1.098	0.1360	0.272	0.973	0.1652	0.1898
25	Burundi	0.774	0.2194	−0.484	−0.872	0.1915	0.1040
26	Rwanda	1.457	0.0725	−0.752	−1.613	0.0534	0.0285
27	Somalia	1.183	0.1184	0.453	0.731	0.2324	0.1266
28	Ethiopia	2.627	0.0043	0.725	1.422	0.0775	0.0090
29	Zambia	0.753	0.2258	0.042	0.219	0.4134	0.1934
30	Zimbabwe	−0.200	0.4209	−0.010	0.033	0.4868	0.4041
31	Malawi	0.212	0.4161	−0.229	−0.388	0.3490	0.2088
32	Mozambique	−0.288	0.3868	0.017	0.114	0.4545	0.4728
33	South Africa	−0.868	0.1927	−0.183	−0.480	0.3156	0.1435
34	Lesotho	−0.298	0.3827	−0.419	−0.423	0.3361	0.2341
35	Botswana	0.041	0.4837	−0.004	0.039	0.4845	0.3691
36	Swaziland	−0.659	0.2548	0.017	0.063	0.4749	0.4128
37	Morocco	0.022	0.4913	−0.097	−0.111	0.4557	0.4995
38	Algeria	−0.363	0.3583	−0.010	0.040	0.4841	0.4139
39	Tunesia	0.579	0.2813	0.005	0.046	0.4818	0.1804
40	Libya	2.553	0.0053	0.804	2.300	0.0107	0.0133
41	Sudan	4.039	**0.0000**	2.988	9.898	**0.0000**	**0.0003**
42	Egypt	4.421	**0.0000**	6.947	10.679	**0.0000**	0.0058

global Moran's I. This explains the sign differences between the values in the third and fifth columns of Table 1 (for example, for the first sixteen countries in the table). Following the suggestion by Ord and Getis (1994), a Bonferroni bounds procedure is used to assess significance. With an overall α level of 0.05, the individual significance levels for each observation should be taken as 0.05/42, or 0.0012.[8] Given this conservative procedure, the normal approximation for both the G_i^* and the I_i show the same four countries to exhibit local

[8] For $\alpha = 0.10$, the corresponding individual significance level is 0.0024. Since normality was not demonstrated, the original Bonferroni bounds were used, rather than the slightly sharper Sidák procedure suggested in Ord and Getis (1994). This does not affect the interpretation of the results in Table 1, since the difference between the two only appears at the fifth significant digit. For example, for $\alpha = 0.05$, the Bonferroni bound is 0.001190, while the Sidák bounds are 0.001221.

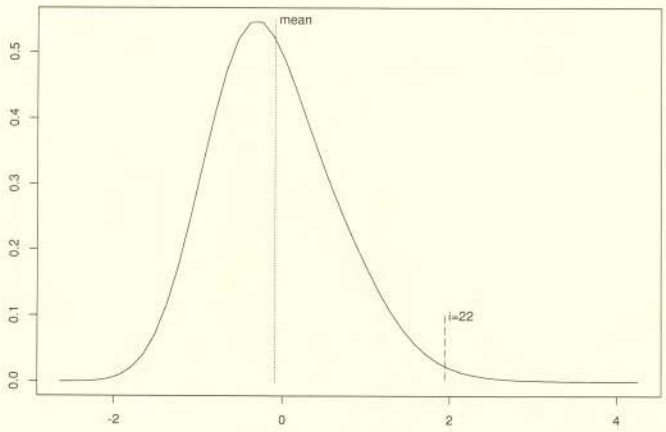


FIG. 2. Density of Randomized Local Moran for Uganda ($i = 22$)

spatial clustering (with the significance levels in bold type in Table 1). They are Uganda (22), Kenya (23), Sudan (41), and Egypt (42), which themselves form a cluster in the northeast of Africa, part of the so-called Shatterbelt.[9] This spatial clustering (or spatial autocorrelation) of both the G_i^* and the LISA statistics is a result of the way they are constructed, and should be kept in mind when visually interpreting a map of LISAs (or G_i^*).

The conditional randomization approach provides a still more conservative picture of (pseudo) significant local spatial clustering, with only Sudan meeting the Bonferroni bound for an overall $\alpha = 0.05$. For this country, two out of the 9,999 statistics computed from the randomized samples exceed the observed one, clearly labeling the latter as "extreme."[10] Of the other three previously significant countries, only Kenya comes close to the threshold (with a pseudo significance of 0.0016), but both Uganda (0.0031) and Egypt (0.0058) fall short of even the bounds for an overall $\alpha = 0.10$.

Some insight into the reasons for the differences in interpretation between the normal approximation and the randomization strategy can be gained from Figure 2, which shows the empirical distribution of the I_i for the 10,000 samples used in the computation of the pseudo significance for Uganda (22). This country was chosen since it has different significance indications between the two criteria, and it is not a boundary or corner location (it has five neighbors, which is about average for the sample). The density function in Figure 2 is smoothed, using a smoothing parameter of twice the interquartile distance. The sample average and the observed value are indicated on the figure (the latter with the label "$i = 22$"). The density under the curve for values larger than 1.943 (the observed value) is 0.0031, indicating its extremeness (but not significance according to the Bonferroni criterion). The distribution is clearly non-normal, and heavily skewed to the right (skewness is 0.7997). Its average of -0.0904 is smaller than the expected value under the null hypothesis for observation 22, which is -0.0244. In addition, its standard deviation of 0.6340 is more than 1.5 times the value that would be expected under the theoretical null distribution, or 0.3991 [the square root of expression (14)].

[9] The identification numbers in parentheses correspond to the labels in the Moran scatterplot of Figure 3.

[10] The Bonferroni bound for an overall significance level of $\alpha = 0.01$ would be 0.0002.

The differences between the empirical density in Figure 2 and (i) the theoretical moment and (ii) an approximation of the null distribution by a normal raise two important issues. First, the normal may not be an appropriate approximation, and higher order moments may have to be used, as in the approximation to the global Γ statistics in Costanzo, Hubert, and Golledge (1983). However, it may also be that the sample size and/or the number of neighbors in this example (respectively, forty-two and five) are too small for a valid approximation by the normal.[11] Secondly, and more importantly, the moments under the null hypothesis are derived assuming that each value is equally likely at any location, which is inappropriate in the presence of global spatial association. In other words, the theoretical moments in (13) and (14) do not reflect the latter. This is appropriate when the objective is to detect local spatial clusters in the absence of global spatial association [for example, as was the stated goal in Getis and Ord (1992)], but is not correct when global spatial association is present (as is the case in the example considered here). While the z-values for both G_i^* and I_i would suffer from this problem, the conditional randomization strategy does not, since it treats the observations *as if* they were spatially uncorrelated. This issue is revisited in section 5.

Indication of Local Instability

The second interpretation of a LISA is as a diagnostic for outliers with respect to a measure of global association, in this example Moran's I. The I_i statistics are compared to the insights provided by the Moran scatterplot, suggested by Anselin (1993a) as a device to achieve a similar objective, that is, to visualize local instability in spatial autocorrelation. Note that the Moran scatterplot is not a LISA in the sense of this paper, since no indication of significant local spatial clustering is obtained. The principle behind the interpretation of the Moran scatterplot is that many statistics for global association are of the form $x'\mathbf{A}x/x'x$, where x is a vector of observations (in deviations from the mean) and $\mathbf{A}$ is a matrix of known elements. In the case of Moran's I, the $\mathbf{A}$ is the row-standardized spatial weights matrix $\mathbf{W}$. Given this form for the statistic, it may be visualized as the slope of a linear regression of $\mathbf{W}x$ on x [see also Anselin (1980) for the interpretation of Moran's I as a regression coefficient]. A scatterplot of $\mathbf{W}x$ on x [similar to a spatial lag scatterplot in geostatistics, for example, as in Cressie (1991)], with the linear regression line superimposed, provides insight into the extent to which individual $(\mathbf{W}x_i, x_i)$ pairs influence the global measure, exert leverage, or may be interpreted as outliers, based on the extensive set of standard regression diagnostics (for example, Cook 1977; Hoaglin and Welsch 1978; Belsley, Kuh, and Welsch 1980).

The Moran scatterplot for the African conflict data is given as Figure 3, with the individual countries labeled as in Table 1. The $(\mathbf{W}x_i, x_i)$ pairs are given for standardized values, so that "outliers" may be easily visualized as points further than two units away from the origin. In Figure 3, both Sudan (41) and Egypt (42) have values for total conflict that are more than two standard deviations higher than the mean (on the horizontal axis of Figure 3), while Egypt also has values for the spatial lag that are twice the mean (vertical axis of Figure 3). The use of standardized values also allows the Moran scatterplots for different variables to be comparable. The four quadrants in Figure 3 correspond to the four types of spatial association. The lower left and upper right quadrants indicate spatial clustering of similar values: low values (that is, less than the mean) in

[11] See Getis and Ord (1992, pp. 191–92) for the importance of both sample size and the number of neighbors for the normal approximation of the G_i and G_i^* statistics.

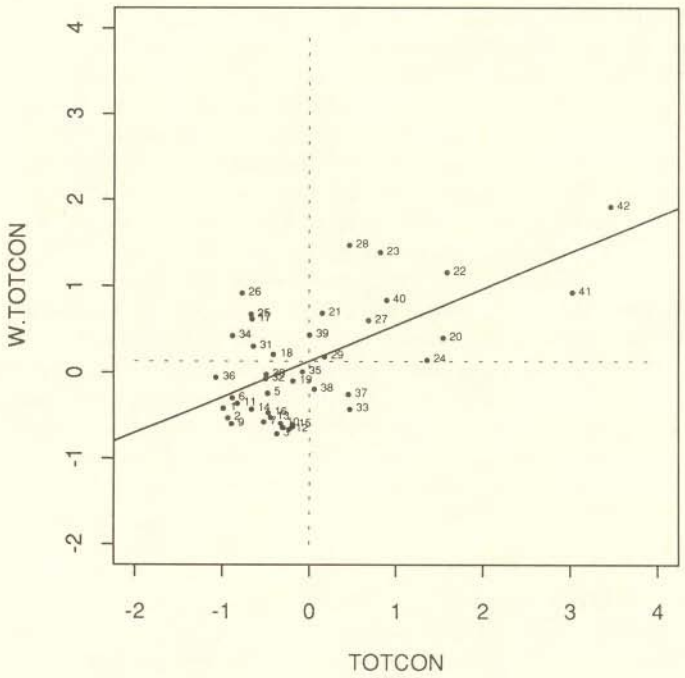


FIG. 3. Moran Scatterplot for Total Conflict ($I = 0.417$)

the lower left and high values in the upper right. Stated differently, the lower left pairs would correspond to negative values of the G_i and G_i^*, and the upper right pairs to positive values. With the I_i statistics, no distinction is possible between the two forms of association since both result in a positive sign. The upper left and lower right quadrants of Figure 3 indicate spatial association of dissimilar values: low values surrounded by high neighboring values for the former, and high values surrounded by low values for the latter. These correspond to I_i statistics with a negative sign. Since they are not cross-product statistics, the G_i and G_i^* statistics do not capture this form of spatial association.

While the overall pattern of spatial association is clearly positive, as indicated by the slope of the regression line (Moran's I), eleven observations show association between dissimilar values: eight in the upper left quadrant, also shown as light islands within the darkest clusters of Figure 1; and three in the lower right quadrant (Algeria, 38, Morocco, 37, and South Africa, 33), surrounded by countries in the first and second quartile in Figure 1. This may indicate the existence of different regimes of spatial association.

The application of regression diagnostics for leverage to the scatterplot suggests that two observations deserve closer scrutiny. The highly significant local spatial association for Sudan (41) and Egypt (42) finds a match with the indication of leverage provided by the diagonal elements of the hat matrix. These are respectively 0.247 (for Sudan) and 0.316 (for Egypt), both distinctly larger than the usual cutoff of $2k/n$ (where k is the number of explanatory variables in the regression, or 2 in this example), or 0.095.[12] The third largest hat value of 0.085

[12] The diagonal elements of the hat matrix $H = X(X'X)^{-1}X'$, with X as the matrix of observations on the explanatory variables in a regression, are well known indicators of leverage. See, for example, Hoaglin and Welsch (1978).

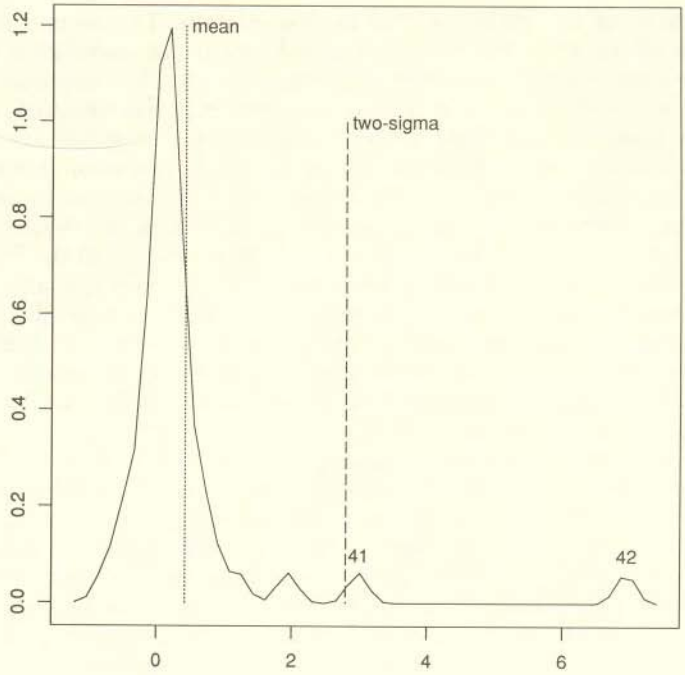


FIG. 4. Local Moran Outliers

for Uganda, 22, does not exceed this threshold. One may be tempted to conclude that the elimination of Egypt and Sudan from the sample would void the indication of spatial association, but this is not the case. Without both countries, Moran's I drops to 0.254, but its associated z-value is 2.53 (using the randomization null hypothesis), which is still highly significant at $p < 0.006$.

The distribution of I_i statistics for the sample can similarly be exploited to provide an indication of outliers or leverage points. In Figure 4, this is illustrated by means of a simple two-sigma rule. The mean of the distribution of the I_i is Moran's I, or 0.417, and twice the standard deviation from the mean corresponds to the value of 2.798. Clearly, this is exceeded by both Sudan (41), with a value of 2.988 for I_i, and by Egypt (42), with a value of 6.947. While this is obviously not a test in a strict sense, it provides useful insight into the special nature of these two observations. All four indicators are in agreement in this respect, that is, the G_i^* and I_i as measures of local spatial clusters, and the Moran scatterplot and I_i as indicators of outliers. The substantive interpretation of the special nature of these observations is beyond the scope of the exploratory data analysis. The role of the latter is to point them out and by doing so to aid in the suggestion of possible explanations or hypotheses. Alternatively, the indication of "strange" observations may point to data quality problems, such as coding mistakes, or, in the case of spatial analysis, problems with the choice of the spatial weights matrix.

5. MONTE CARLO EVIDENCE: GLOBAL AND LOCAL SPATIAL ASSOCIATION

Two issues raised by the results of the empirical illustration in the previous section are revisited here by means of some initial Monte Carlo experiments. The first pertains to the distribution of the local Moran I_i statistic under the

null hypothesis of no (global) spatial autocorrelation. The second issue is the distribution of the local statistic when global spatial autocorrelation is present, and its implication for the assessment of significance. This may also have relevance for the distribution of the G_i and G_i^* statistics in this situation, since their distribution under the null is also based on the absence of global association. As pointed out earlier, it is quite common to study local association in the presence of global association, for example, this is the case in the illustration presented in the previous section. This second issue also has relevance for the assessment of outliers or local instability, that is, the second interpretation of the local Moran. It is well known that many spatial processes that produce spatially autocorrelated patterns also generate spatial heterogeneity. For example, this is the case for the familiar spatial autoregressive process (Anselin 1990). The spatial heterogeneity indicated by LISAs, based on a null hypothesis of no spatial association may therefore be a natural characteristic of the spatial process, and not an indication of local pockets of nonstationarity.

Two sets of experiments were carried out, one based on the same spatial weights matrix as for the African example (with $n = 42$), the other on the weights matrix for a 9 by 9 regular grid, using the queen notion of contiguity (with $n = 81$). Both weights matrices were used in row-standardized form. For each of these configurations, 10,000 random samples were generated with increasing degrees of spatial autocorrelation, constructed by means of a simple spatial autoregressive transformation. More formally, given a vector ε of randomly generated standard normal variates, a spatially autocorrelated landscape was generated as a vector y:

$$y = (\mathbf{I} - \rho\mathbf{W})^{-1}\varepsilon, \tag{21}$$

where ρ is the autoregressive parameter, taking values of 0.0, 0.3, 0.6, and 0.9, and $\mathbf{I}$ is a n by n identity matrix. While the resulting samples will be spatially autocorrelated for nonzero values of ρ, there is no one-to-one match between the value of ρ and the global Moran's I. As is well known, the latter is capable of detecting many different forms of spatial association, and is not linked to a specific spatial process as the sole alternative hypothesis.

Distribution of the Local Moran under the Null Hypothesis

The distribution of the standardized z-values that correspond to the I_i statistic was considered in detail for two selected observations, the location corresponding to Uganda, $i = 22$, for the African weights matrix, and the location corresponding to the central cell, $i = 41$, for the regular lattice. Not only are the dimensions of the data sets different in the two examples ($n = 42$ and $n = 81$), but also the number of neighbors differ for the observations under consideration, as they are respectively 5 and 8. The moments of each distribution for the z-values, based on the 10,000 replications, are given in the first row of Table 2. While the mean and standard deviation are roughly in accordance with those for a standard normal distribution, the kurtosis and to a lesser extent the skewness are not. This is further illustrated by the density graph in Figure 5 (for $n = 81$), which clearly shows the leptokurtic nature of the distribution and the associated thicker tails (compared to a normal density). The density graph for the African case is very similar and is not shown. Instead, a quantile-quantile plot for the African example is given in Figure 6, to further illustrate the lack of normality. While there is general agreement in the central section of the two distributions (total agreement would be shown as a perfect

TABLE 2
Moments of Local Moran with Global Spatial Autocorrelation[a]

	$n = 42$				$n = 81$			
ρ	Mean	St.Dev.	Skew	Kurtosis	Mean	St.Dev.	Skew	Kurtosis
0.0	0.0032	0.9895	−0.2599	7.993	0.0236	1.0356	−0.1073	7.711
0.3	0.2491	1.0730	0.7417	7.635	0.2666	1.1733	0.9320	7.853
0.6	0.5833	1.2144	1.4748	7.454	0.6057	1.3958	1.7475	8.673
0.9	1.0782	1.3465	1.5357	5.850	0.8961	1.4690	2.4073	11.114

a. z-values for local Moran; 10,000 replications, using observation 22 for $n = 42$ and observation 41 for $n = 81$.

linear fit), at the tails, that is, where it matters in terms of significance, this clearly is not the case. A more rigorous assessment of the distribution, based on an asymptotic chi-squared test constructed around the third and fourth moments (Kiefer and Salmon 1983) strongly rejects the null hypothesis of normality in both cases.

This more extensive assessment confirms (in a controlled setting) the earlier suggestion implied by the discrepancy between the significance levels under the normal approximation and the conditional randomization in Table 1. Note that the African example in Table 1 exhibited significant global spatial autocorrelation, while the simulations here do not (by design). Further results are needed to see whether larger sample sizes or higher numbers of neighbors are needed before normality is obtained. However, from the initial impressions gained here it would seem that the normal approximation may be inappropriate, and that higher moments (given the values for skewness and kurtosis in Table 2) would be needed in order to obtain a better approximation [for example, as in Costanzo, Hubert, and Golledge (1983) for the Γ statistic].

The implications of these results for inference in practice are that even when no global spatial autocorrelation is present, the significance levels indicated by a normal approximation will result in an over-rejection of the null hypothesis for a given α_i Type I error. Clearly, a more conservative approach is warranted, although the exact nature of the corrections to the α_i awaits further investigation. In the meantime, a conditional randomization approach provides a useful alternative.

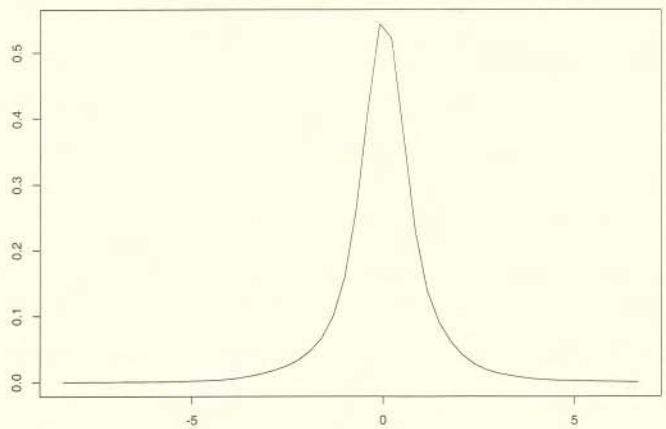


FIG. 5. Density of z-value for Local Moran ($n = 81$; 10,000 Replications)

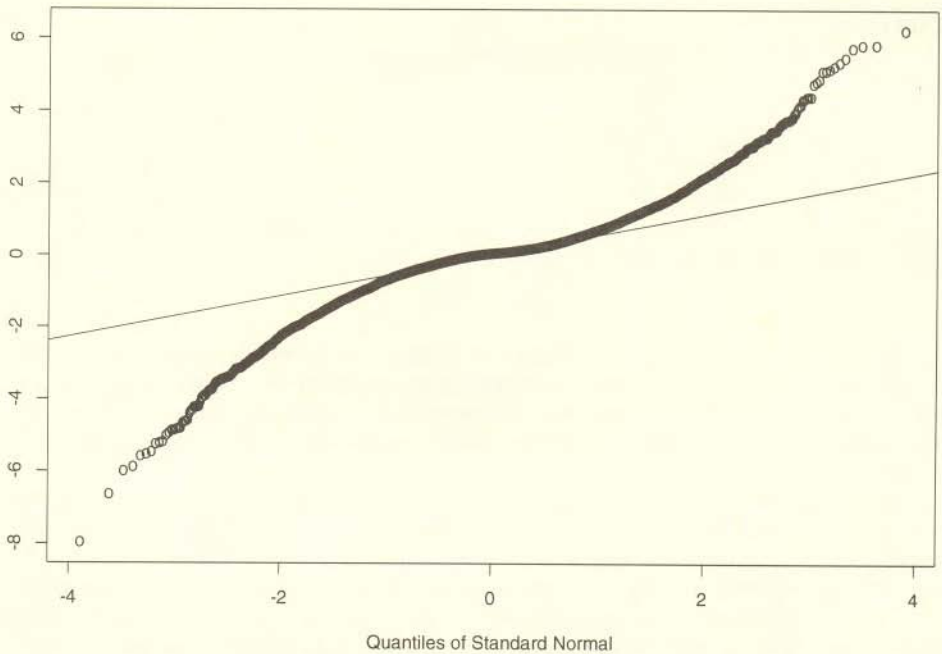


FIG. 6. Quantiles of z-values of Local Moran against the Normal Distribution ($n = 42$; 10,000 Replications)

Distribution of the Local Moran in the Presence of Global Spatial Autocorrelation

The presence of global spatial autocorrelation has a strong influence on the moments of the distribution of the local Moran, as indicated by the results in Table 2. Both mean and standard deviation increase with spatial autocorrelation, but the most significant effect seems to be on the skewness of the distribution. This is further illustrated by the box plots in Figure 7 (for the case with $n = 42$; the results for the larger sample size are similar). As ρ increases, the distribution becomes more and more asymmetric around the median, while both the interquartile range and the median itself increase as well. Clearly, in the presence of global spatial autocorrelation, the moments indicated by the expressions (13) and (14) become inappropriate estimates of the moments of the actual distribution. The same problem would seem to also affect the distribution for the Getis and Ord G_i and G_i^* statistics, since they are derived in a similar manner. Consequently, inference for tests on local spatial clusters that ignores this effect is likely to be misleading. The magnitude of the error cannot be derived from the initial Monte Carlo results reported here, and further investigation is needed, both empirical and analytical. In practice, inference based on the pseudo significance levels indicated by a conditional randomization approach seems to be the only viable alternative.

Evidence of Outliers in the Presence of Global Spatial Autocorrelation

A final issue to be examined is how the magnitude of global spatial autocorrelation affects the distribution of the I_i around the sample mean (the global Moran's I), which is used to detect outliers. In contrast to the earlier experi-

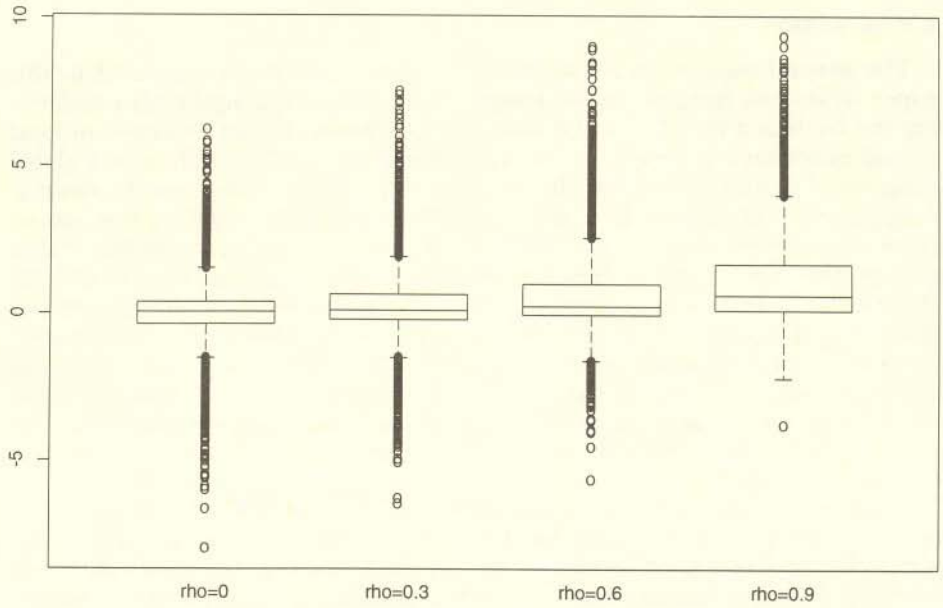


FIG. 7. Box Plots of Local Moran z-value with Spatial Autocorrelation ($n = 42$; 10,000 Replications)

ments, the focus is not on I_i for an individual location, but on how the spread of the statistics in each sample is affected by the strength of global spatial autocorrelation. In Table 3, the average over the 10,000 replications of twice the standard deviation around the mean in each replication is listed, as well as the average (over the 10,000 replications) number of outliers indicated by using the two sigma rule. With increasing global spatial autocorrelation, both the spread and the number of "outliers" increases. This implies that in the presence of a high degree of spatial autocorrelation, several extreme values of the I_i statistic are to be expected as a "normal" result of the heterogeneity induced by a spatial autoregressive process. In practice, this is not much different from the usual treatment of outliers, and without further evidence, it is not possible to state in a rigorous manner which extreme values are to be expected and which are unusual observations. However, as an exploratory device, the lack of symmetry of the distribution of the I_i around the global I, and/or the presence of very large values provides insight into the stability of the indication of global spatial association over the sample.

TABLE 3
Two-Sigma Rule with Global Spatial Autocorrelation[a]

	$n = 42$		$n = 81$	
ρ	2σ	Outliers	2σ	Outliers
0.0	1.0199	3	0.7538	5
0.3	1.1171	3	0.8675	6
0.6	1.3519	4	1.1280	7
0.9	1.7112	5	1.7017	9

a. 2σ computed as average 2σ over 10,000 replications; outliers are median number of observations more than 2σ from the mean in each sample.

6. CONCLUSION

The general class of local indicators of spatial association suggested in this paper serves two main purposes. Firstly, the LISA generalize the idea underlying the Getis and Ord G_i and G_i^* statistics to a broad class of measures of local spatial association. Secondly, by directly linking the local indicators to a global measure of spatial association, the decomposition of the latter into its observation-specific components becomes straightforward, thus enabling the assessment of influential observations and outliers. It is this dual property that distinguishes the class of LISA from existing techniques, such as the G_i and G_i^* statistics and the Moran scatterplot. The LISA presented here are easy to implement and lend themselves readily to visualization. They thus serve a useful purpose in an exploratory analysis of spatial data, potentially indicating local spatial clusters and forming the basis for a sensitivity analysis (outliers). While the former is more appropriate when no global spatial autocorrelation is present, the latter is particularly useful when there is spatial autocorrelation in the data.

A number of issues remain to be investigated further. The illustration in this paper primarily pertained to the local Moran I_i indices, but the extension to the wider class of LISA statistics can be carried out in a straightforward way. From both the empirical example and the initial simulation experiments, it follows that the null distribution of the local Moran cannot be effectively approximated by the normal, at least not for the small sample sizes employed here. Also, it seems that higher moments may be necessary in order to obtain a better approximation. Furthermore, the uncritical use of the null distribution in the presence of global spatial autocorrelation will give incorrect significance levels. The problem also pertains to the G_i and G_i^* statistics and would suggest that a test for global spatial autocorrelation should precede the assessment of significant local spatial clusters. However, such a two-pronged strategy raises the issue of pretesting and multiple comparisons, and would require an adjustment of the significance levels to reflect this. This further complicates the determination of a proper significance level for an individual LISA, given the built-in correlatedness of measures for adjoining locations. It is clear that some type of bounds procedure is needed, but which degree of correction is sufficient still remains to be addressed.

Finally, the conditional randomization approach suggested here seems to provide a reliable basis for inference for the LISA, both in the absence and in the presence of global spatial autocorrelation.

LITERATURE CITED

Anselin, L. (1980). *Estimation Methods for Spatial Autoregressive Structures*. Ithaca, N.Y.: Regional Science Dissertation and Monograph Series.

——— (1986). "MicroQAP, a Microcomputer Implementation of Generalized Measures of Spatial Association." Department of Geography, University of California, Santa Barbara, Calif.

——— (1988). *Spatial Econometrics: Methods and Models*. Dordrecht: Kluwer Academic.

——— (1990). "Spatial Dependence and Spatial Structural Instability in Applied Regression Analysis." *Journal of Regional Science* 30, 185–207.

——— (1992). *SpaceStat: A Program for the Analysis of Spatial Data*. National Center for Geographic Information and Analysis, University of California, Santa Barbara, Calif.

——— (1993a). "The Moran Scatterplot as an ESDA Tool to Assess Local Instability in Spatial Association." Paper presented at the GISDATA Specialist Meeting on GIS and Spatial Analysis, Amsterdam, The Netherlands, December 1-5 (West Virginia University, Regional Research Institute, Research Paper 9330).

——— (1993b). "Exploratory Spatial Data Analysis and Geographic Information Systems." Paper pre-

sented at the DOSES/Eurostat Workshop on New Tools for Spatial Analysis, Lisbon, Portugal, November 18-20 (West Virginia University, Regional Research Institute, Research Paper 9329).

Anselin, L., and A. Getis (1992). "Spatial Statistical Analysis and Geographic Information Systems." *The Annals of Regional Science* 26, 19–33.

Anselin, L., and J. O'Loughlin (1990). "Spatial Econometric Models of International Conflicts." In *Dynamics and Conflict in Regional Structural Change*, edited by M. Chatterji and R. Kuenne, pp. 325–45. London: Macmillan.

——— (1992). "Geography of International Conflict and Cooperation: Spatial Dependence and Regional Context in Africa." In *The New Geopolitics*, edited by M. Ward, pp. 39–75. Philadelphia, Penn.: Gordon and Breach.

Azar, E. (1980). "The Conflict and Peace Data Bank (COPDAB) Project." *Journal of Conflict Resolution* 24, 143–52.

Belsley, D, E. Kuh, and R. Welsch (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Wiley.

Casetti, E. (1972). "Generating Models by the Expansion Method: Applications to Geographical Research." *Geographical Analysis* 4, 81–91.

——— (1986). "The Dual Expansion Method: An Application for Evaluating the Effects of Population Growth on Development." *IEEE Transactions on Systems, Man and Cybernetics* SMC-16, 29–39.

Cliff, A., and J. K. Ord (1981). *Spatial Processes: Moels and Applications*. London: Pion.

Costanzo, C. M., L. J. Hubert, and R. G. Golledge (1983). "A Higher Moment for Spatial Statistics." *Geographical Analysis* 15, 347–51.

Cook, R. (1977). "Detection of Influential Observations in Linear Regression." *Technometrics* 19, 15–18.

Cressie, N. (1991). *Statistics for Spatial Data*. New York: Wiley.

Diehl, P. (1992). "Geography and War: A Review and Assessment of the Empirical Literature." In *The New Geopolitics*, edited by M. Ward, pp. 121–37. Philadelphia, Penn.: Gordon and Breach.

Eastman, R. (1992). *IDRISI Version 4.0*. Worcester, Mass.: Clark University Graduate School of Geography.

Foster, S. A., and W. Gorr (1986). "An Adaptive Filter for Estimating Spatially-Varying Parameters: Application to Modeling Police Hours in Response to Calls for Service." *Management Science* 32, 878–89.

Getis, A. (1991). "Spatial Interaction and Spatial Autocorrelation: A Cross-Product Approach." *Environment and Planning* A 23, 1269–77.

Getis, A., and K. Ord (1992). "The Analysis of Spatial Association by Use of Distance Statistics." *Geographical Analysis* 24, 189–206.

Gorr, W., and A. Olligschlaeger (1994). "Weighted Spatial Adaptive Filtering: Monte Carlo Studies and Application to Illicit Drug Market Modeling." *Geographical Analysis* 26, 67–87.

Griffith, D. A. (1978). "A Spatially Adjusted ANOVA Model." *Geographical Analysis* 10, 296–301.

——— (1992). "A Spatially Adjusted N-Way ANOVA Model." *Regional Science and Urban Economics* 22, 347–69.

——— (1993). "Which Spatial Statistics Techniques Should Be Converted to GIS Functions? In *Geographic Information Systems, Spatial Modelling and Policy Evaluation*, edited by M. M. Fischer and P. Nijkamp, pp. 101–14. Berlin: Springer Verlag.

Haslett, J., R. Bradley, P. Craig, A. Unwin, and C. Wills (1991). "Dynamic Graphics for Exploring Spatial Data with Applications to Locating Global and Local Anomalies." *The American Statistician* 45, 234–42.

Hoaglin, D., and R. Welsch (1978). "The Hat Matrix in Regression and ANOVA." *The American Statistician* 32, 17–22.

Hubert, L. J. (1985). "Combinatorial Data Analysis: Association and Partial Association." *Psychometrika* 50, 449–67.

——— (1987). *Assignment Methods in Combinatorial Data Analysis*. New York: Marcel Dekker.

Hubert, L. J., R. Golledge, and C. M. Costanzo (1981). "Generalized Procedures for Evaluating Spatial Autocorrelation." *Geographical Analysis* 13, 224–33.

Hubert, L. J., R. Golledge, C. M. Costanzo, and N. Gale (1985). "Measuring Association between Spatially Defined Variables: An Alternative Procedure." *Geographical Analysis* 17, 36–46.

Jones, J. P., and E. Casetti (1992). *Applications of the Expansion Method*. London: Routledge.

Kiefer, N., and M. Salmon (1983). "Testing Normality in Econometric Models." *Economics Letters* 11, 123–8.

Kirby, A., and M. Ward (1987). "The Spatial Analysis of Peace and War." *Comparative Political Studies* 20, 293–313.

Mantel, N. (1967). "The Detection of Disease Clustering and a Generalized Regression Approach." *Cancer Research* 27, 209–20.

Mielke, P. W. (1979). "On Asymptotic Non-Normality of Null Distributions of MRPP Statistics." *Communications Statistical Theory and Methods A* 8, 1541–50.

Oden, N. L. (1984). "Assessing the Significance of a Spatial Correlogram." *Geographical Analysis* 16, 1–16.

O'Loughlin, J. (1986). "Spatial Models of International Conflicts: Extending Current Theories of War Behavior." *Annals, Association of American Geographers* 76, 63–80.

O'Loughlin, J. and L. Anselin (1991). "Bringing Geography Back to the Study of International Relations: Dependence and Regional Context in Africa, 1966–1978." *International Interactions* 17, 29–61.

——— (1992). "Geography of International Conflict and Cooperation: Theory and Methods." In *The New Geopolitics*, edited by M. Ward, pp. 11–38. Philadelphia, Penn.: Gordon and Breach.

O'Loughlin, J., C. Flint, and L. Anselin (1994). "The Political Geography of the Nazi Vote: Context, Confession, and Class in the Reichstag Election of 1930." Annals, Association of American Geographers 84, 351–80.

Openshaw, S. (1993). "Some Suggestions concerning the Development of Artificial Intelligence Tools for Spatial Modelling and Analysis in GIS." In *Geographic Information Systems, Spatial Modelling and Policy Evaluation*, edited by M. M. Fischer and P. Nijkamp, pp. 17–33. Berlin: Springer Verlag.

Openshaw, S., C. Brundson, and M. Charlton (1991). "A Spatial Analysis Toolkit for GIS." *EGIS '91, Proceedings of the Second European Conference on Geographical Information Systems*, pp. 788–96. Utrecht: EGIS Foundation.

Openshaw, S., A. Cross, and M. Charlton (1990). "Building a Prototype Geographical Correlates Exploration Machine." *International Journal of Geographical Information Systems* 4, 297–311.

Ord, J. K., and A. Getis (1994). "Distributional Issues concerning Distance Statistics." Working paper.

Royaltey, H., E. Astrachan, and R. Sokal (1975). "Tests for Patterns in Geographic Variation." *Geographical Analysis* 7, 369–96.

Savin, N. E. (1980). "The Bonferroni and the Scheffé Multiple Comparison Procedures." *Review of Economic Studies* 67, 255–73.

Sidák, Z. (1967). "Rectangular Confidence Regions for the Means of Multivariate Normal Distributions." *Journal of the American Statistical Association* 62, 626–33.

Sokal, R., N. Oden, B. Thomson, and J. Kim (1993). "Testing for Regional Differences in Means: Distinguishing Inherent from Spurious Spatial Autocorrelation by Restricted Randomization." *Geographical Analysis* 25, 199–210.

Tiefelsdorf, M., and B. Boots (1994). "The Exact Distribution of Moran's *I*." *Environment and Planning A* (forthcoming).

APPENDIX A

The moments of the local Moran statistic can be derived using the results in Cliff and Ord (1981, pp. 42–46). Using (12), the expected value of I_i under the randomization hypothesis is

$$\mathrm{E}[I_i] = \left(\sum_j w_{ij}/m_2 \right) \mathrm{E}[z_i z_j].$$

The value of the expectations term is

$$\mathrm{E}[z_i z_j] = -m_2/(n-1),$$

based on equation (2.37) of Cliff and Ord (1981, p. 45). Consequently, the expected value of I_i becomes

$$\mathrm{E}[I_i] = -w_i/(n-1),$$

with w_i as the sum of the row elements, $\sum_j w_{ij}$. Obviously, in the case a row-standardized weights matrix is used, this sum will be one.

To obtain the second moment, the following expression must be evaluated:

$$\mathrm{E}[I_i^2] = (1/m_2^2)\mathrm{E}\left[z_i^2\left(\sum_j w_{ij}z_j\right)^2\right]$$

or

$$\mathrm{E}[I_i^2] = (1/m_2^2)\mathrm{E}\left[z_i^2\left(\sum_{j\neq i} w_{ij}^2 z_j^2 + \sum_{k\neq i}\sum_{h\neq i} w_{ik}w_{ih}z_k z_h\right)\right],$$

for which the following results are important, based on equation (2.39) of Cliff and Ord (1981, p. 46):

$$\mathrm{E}[z_i^2 z_j^2] = (nm_2^2 - m_4)/(n-1);$$

$$\mathrm{E}[z_i^2 z_k z_h] = (2m_4 - nm_2^2)/(n-1)\ (n-2)$$

with $m_4 = \sum_i z_j^4/n$ as the fourth moment. The first weights term in the expectation consists of the sum of all weights squared, or, $w_{i(2)} = \sum_{j\neq i} w_{ij}^2$, and the second is twice the sum of the cross products (avoiding identical subscripts), or, $2w_{i(kh)} = \sum_{k\neq i}\sum_{h\neq i} w_{ik}w_{ih}$. After combining terms, the second moment is found as

$$\mathrm{E}[I_i^2] = (1/m_2^2)\ [w_{i(2)}(nm_2^2 - m_4)/(n-1)$$

$$+2w_{i(kh)}(2m_4 - nm_2^2)/(n-1)\ (n-2)],$$

which simplifies somewhat after using $b_2 = m_4/m_2^2$, to

$$\mathrm{E}[I_i^2] = w_{i(2)}(n-b_2)/(n-1) + 2w_{i(kh)}(2b_2 - n)/(n-1)(n-2).$$

Consequently, the variance of I_i is

$$\mathrm{Var}[I_i] = w_{i(2)}(n-b_2)/(n-1) + 2w_{i(kh)}(2b_2 - n)/(n-1)(n-2)$$

Contagion Effects and Ethnic Contribution Networks

Wendy K. Tam Cho University of Illinois at Urbana-Champaign

Many political behavior theories explicitly incorporate the idea that context matters in politics. Nonetheless, the concept of spatial dependence—in particular, that behavior in geographic units is somehow related to and affected by behavior in neighboring areas—is not extensively explored. The study of campaign finance is no exception. Research in this area concentrates on the attributes of the individual donor, leaving context underexplored. Concepts such as contribution networks, for instance, are not rigorously tested. This article reexamines the impact of conventional socio-demographic covariates on campaign donation behavior by ethnic contributors and explicitly models spatial effects. The spatial analysis reveals that patterns of campaign donations are geographically clustered (exhibiting both spatial dependence, implying a neighborhood effect, and spatial heterogeneity, implying a regional effect), and that this clustering cannot be explained completely by socio-economic and demographic variables. While socio-demographic characteristics are important components of the dynamic underlying campaign contributions, there is also evidence consistent with a contagion effect whereby ethnic contribution networks are fueling funds to candidate coffers.

Context matters in politics. Politics, after all, is not a set of unrelated individual actions, but is instead an interrelated set of social phenomena. A corollary of this claim is that people are influenced by the context in which they find themselves. Indeed, it is not hard to imagine a plethora of circumstances under which colleagues and neighbors would be influential in the formation and solidification of political beliefs or would be the impetus behind the emergence of some type of political action. Although people can and do maintain relationships that span large distances, it is clear that one of the great sources of enduring and influential interactions is physical proximity. Despite easily-formed theories for spatial effects, the concept of "space"—in particular, that the behavior of people is somehow related to and affected by the behavior of those who reside in close proximity—has received too little attention in political science. The lack of inquiry seems especially strange since many classes of theories in political behavior focus on context and geography. Indeed, this discussion and these theories have spanned and evolved over many decades (Key 1949; Berelson, Lazarsfeld, and McPhee 1954; Putnam 1966; Huckfeldt 1979; Eulau 1986; Huckfeldt and Sprague 1987; 1992; Putnam 2000).

There is a line of research that has focused on various spatial dimensions, social networks, and neighborhood effects. For example, Putnam (1966), Huckfeldt, Plutzer, and Sprague (1993), Huckfeldt (1979) have conducted many studies on social interaction. Weatherford (1982) and Crenson (1978) have focused on the idea of social networks. The role of geography is clear in Baybeck (2001), Tir and Diehl (2002), and Baybeck and Huckfeldt (2002). As well, the policy diffusion literature has looked closely at the idea of how policy innovations adopted in one state may spread to neighboring states (see, e.g., Walker 1969; Gray 1973; Berry and Berry 1992). Finally, Johnston et al. (1997, 1998, 2000, 2001) have, on many occasions, examined the role of spatial context in British elections. All of these works emphasize the role of spatial context and the role of simple geography, though in a manner that is somewhat different than the methods employed here.

This article takes advantage of recent and significant advances in geographic information systems (GIS) and the proliferation of research methodologies and tools for spatial analysis. The confluence of these two factors has created conditions that are ripe for spatial analyses of political data, allowing us to broaden our examination and

Associate Professor, Department of Political Science and Department of Statistics, University of Illinois at Urbana-Champaign, 361 Lincoln Hall, 702 S. Wright St., Urbana, IL 61801-3696 (wendy@cho.pol.uiuc.edu).

Thanks to Luc Anselin, Michael Bailey, Bill Bernhard, Brian Gaines, Kristin Kanthak, Jim Kuklinski, Christopher Mooney, Peter Nardulli, Eric Patashnik, Paul Quirk, and participants at the Department of Psychology Quantitative seminar and the Department of Political Science American Politics seminar at the University of Illinois at Urbana-Champaign for very helpful comments. Thanks to Geoffrey Brewster and David Darmofal for valuable research assistance.

American Journal of Political Science, Vol. 47, No. 2, April 2003, Pp. 368–387

 ISSN 0092-5853

conceptualization of the spatial realm in politics, and to do so in a more systematic and expansive manner. There have been some significant studies utilizing these spatial methodologies in relating political phenomena to geography, especially in the field of conflict studies (see, e.g., Kirby and Ward (1987), Starr (2001), and O'Loughlin (1987). Most and Starr (1980, 1982, 1983, 1984) and Starr and Most (1976, 1978, 1983), in particular, have conducted a number of studies along these dimensions). There have also been some studies in American politics (see, e.g., Gimpel 1999; Rom, Peterson, and Scheve 1998; Saavedra 1998; Sui and Hugill 2002; Kohfeld and Sprague 2002; and Darmofal 2002) as well as comparative politics (see, e.g., Agnew 1987; Brustein 1990; O'Loughlin, Flint, and Anselin 1994; Shin 2001; O'Loughlin 2002; and Shin and Agnew 2002). This piece joins these articles in incorporating and emphasizing the role of geography and context by utilizing spatial econometric techniques to explain political phenomena, with a focus on individual political behavior and American politics.

Another important component behind the increasing ability to examine spatial phenomena is the growing availability of geo-coded data. The primary advantage that accrues from analyzing the spatial dimension is that we can move away from theories that incorporate only individual decision-making, whether across time or in a singular incident, in an isolated realm. That is, the individual need no longer be seen as an atomistic actor. Instead, we can consider theoretical frameworks that place the individual's actions in the context of his "neighborhood," where behavior can be compared to and observed in relation to the behavior of others in close proximity.

Perhaps not surprisingly, spatial analyses are important for both substantive as well as statistical reasons, and these two dimensions are inextricably linked in this context. On the substantive front, spatial models allow us to examine critically theories about the political behavior of individuals in the proper context. Aspatial models omit this spatial component and thus allow one to examine the individual primarily as an atomistic actor only. Statistically, if spatial processes underlie the behavior of interest but are not accounted for in the model, inferences will be inaccurate and coefficient estimates may be biased. Erroneously ignoring spatial dependence (in the form of a spatial lag) may create bias and inconsistency in the same way that we understand the omitted variable problem to affect OLS estimates (Anselin 1988, 1990). Alternatively, when the spatial error structure is ignored, simple inefficiency is apparent in the estimates but the standard errors are biased (Anselin and Griffith 1988). Hence, even if one were not interested specifically in the spatial effect but only in the aspatial effects, omitting the possibility of a spatial aspect from the model may affect the interpretation of the results, spatial and otherwise.

Given that many spatial theories have been proposed (but not tested or tested in limited settings only), the increasing availability of geo-coded data, and the statistical issues that arise, rigorous testing of spatial effects is a natural next step. This article examines spatial effects in the context of campaign contributions. I begin by positing why this form of political behavior may be particularly susceptible to spatial effects. Next, I describe the data gathering and merging process. The contributions data are from the Federal Election Commission (FEC). These data are merged to U.S. Census zip code data. I then present spatial models of campaign donations for 10 separate years. Finally, I conclude by discussing the impact of spatial as well as some aspatial effects, such as time and demographics, on the campaign contribution dynamic.

Spatial and Aspatial Theories of Campaign Donations

Although the idea that the patterns behind campaign contributions have a spatial component has scarcely been tested empirically, the reasoning behind why contributions would exhibit a spatial pattern is not lacking. Some of these reasons are spatial (i.e., attributable to geography), while others are aspatial (i.e., attributable to non-geographic components such as income). For instance, one reason why campaign donations would exhibit a spatial pattern is that campaigns are strategic but have limited resources, and so attempt to allocate these resources wisely. This may mean that a candidate will focus on specific media markets, bombarding the campaign battlegrounds while leaving air time in another part of the country relatively barren. Because this courting is geographically definable, donations may appear to be rolling in in geographic clusters rather than emerging as random, independent events across the United States.

As well, candidates may appeal to specific electoral groups. For instance, it is well known that minority groups (especially blacks and Latinos) tend to favor Democratic candidates. To the extent that these ethnic groups are segregated, whether voluntarily or not, geographic clustering of behavior may again appear. Similarly, Asian Americans tend to reside in clusters. If a candidate is especially attractive to or adept at courting minorities, his set of campaign donations will appear to have some spatial structure, even though the mechanism creating that structure is not a spatial process per se, but is, rather, connected to the dispersion of the minority population.

Alternatively, simple proximity to others exhibiting a certain type of behavior may also be a factor. Social networks may develop in response to "mobilization" so that active solicitation of donations by a candidate has spillover effects via the formation of networks (Putnam 2000; Weatherford 1982). Campbell et al. (1960) identify two factors, community identification and perceived community standards, that serve as the basis for an explanation of community influence. An idea behind this literature is that the initial impetus may be an individual action or a neighborhood fundraiser that then, through social interaction, diffuses to neighboring areas and emerges as a spatial pattern.

Lastly, since money is involved in campaign donations and financial donations are not obligatory, income level always emerges as an obvious explanatory variable. Indeed, research on the origins of campaign donations often focus on socioeconomic factors such as age, education, and income. Verba, Schlozman, and Brady (1995) single out income as "overwhelmingly, the dominant factor" in political contributions. According to their analysis, "[e]ducation, vocabulary, and civic skills play no role" (1995, 361). Gierzynski (2000) concurs, stating that "a look at individual contributors reveals a disproportionate representation of those of higher socioeconomic status" (107). The Brown, Powell, and Wilcox study found that contributors "are generally white, male, well-educated, affluent, and active in contributing at several levels of government" (1995, 49). Rosenstone and Hansen find that education is the most crucial resource in defining participation levels in various political acts with one exception, "[i]ncome—not education—is the most crucial resource for donations of money to political campaigns" (1993, 136).

Socioeconomic variables, especially income, are, then, our chief candidates for aspatial explanatory variables that might be producing the spatial patterning that we observe. Certainly, education and income levels are found in clusters throughout the U.S. Whether the sociodemographic variables are the sources of spatial patterning or if the patterns can be attributed to a more pure spatial process (such as a neighborhood effect) will be the focus of the modeling to follow, but it should be clear that there are many reasons why the campaign finance data may be spatially clustered.

In short, there are many theories behind the dynamic of campaign contributions, both with spatial and aspatial roots that would result in clear spatial patterning. Notably, spatial explanations do not take away from the aspatial findings that have been proposed previously, since both sets of findings can be true simultaneously. There can be a spatial component that complements the nonspatial components. Alternatively, we may find that the spatial explanations comprise a greater proportion of the overall explanation than we had previously thought, i.e., the nonspatial components become less significant or even disappear when viewed in light of the spatial components. In the past, the nonspatial theories have received more attention but not necessarily rightly or justifiably so. The bias results more from a paucity of research in the spatial realm than from a lack of theories.

The immediate goal here is to gain insight into how and why contribution patterns appear as they do across the country. Is there some type of spatial or time-related pattern to the data or are these levels of political behavior solely attributable to decision making that occurs outside simple geography. If the decision-making process is mostly a function of individual traits, then in a unit-level analysis of donation levels, covariates such as partisanship or income levels might be significant predictors, but the spatial parameters should not be significant in the model specifications that control for these covariates. On the other hand, if the contribution dynamic is primarily a diffusion process, driven by network or neighborhood effects, then the spatial lag will be significant, while the socioeconomic indicators will not be significant.

It may also be the case that the patterns can be explained by elite political mobilization, driven perhaps by candidate appearances. Since this analysis does not incorporate a variable such as candidate appearances, if the spatial patterning were the result of this unmeasured variable, the spatial error model would be a relevant spatial specification, and the fit of the spatial error model or evidence of remaining spatial error dependence should provide evidence for or against a mobilization theory. Alternatively, and perhaps most likely, the effect may be a combination of the (measured and unmeasured) spatial and aspatial sets of variables. So, there may be "neighborhood effects" as well as effects that are more directly and narrowly connected to individual characteristics and elite tactics. It is important to note here that the specific *mechanism* that produces the spatial patterns is unknown and not determinable via the spatial analyses that are employed here. What we can uncover are patterns that are consistent with the specific mechanisms that produce the contribution patterns that we observe.

The focus of this study is on ethnic contribution networks, specifically, Asian American contribution networks. Akin to the literature on campaign finance behavior, little is known about ethnic contribution networks. In the minority realm, as in the nonminority realm, research has focused on individual-level decision making. For Asian Americans, the impetus behind the contribution dynamic also has roots in socioeconomic factors. One

large difference is the importance of ethnic cues and ethnic candidates (Cho 2001, 2002). These factors are, again, based on individual traits, not on social context or contribution networks. In this sense, the research presented here complements and augments the vast literature that has amassed on political behavior and minority political behavior. Individual effects are considered, but alongside the context in which individuals find themselves.

Data Analysis

The data for this project are from the FEC (1980–1998). The database includes contributions to candidates for federal office as well as PACs and party organizations. The specific data for this article include a subset of these data: all contributions from Asian American donors.[1] The Asian American group is perhaps the only group that can be reasonably identified solely by name and so the only group that can be extracted reliably from the FEC data.[2] Since this smaller data set is still quite large (over 65,000 observations), there is not much lost in asymptotics.[3]

An important feature of the FEC data is that they are objective whereas surveys rely on self-reported accounts, which may limit the generalizability of the analysis.[4] There are two main drawbacks to the course taken here. First, the analysis is conducted at the level of a geographic unit rather than at the individual level. This does not take away from the spatial component and the ability to evaluate spatial effects, but only the inferences that we can gather about individuals. Second, because the FEC data are not rich in variables as surveys often are, we can observe and model the spatial patterning of contributions, but an extensive study of contributor motives is not possible.

To perform a spatial analysis, one needs data units that are geographic. Accordingly, the unit of aggregation here is the zip code, and all of the FEC data have been aggregated to the zip code level. The zip code level was chosen because the it is the lowest level of aggregation for which we can obtain data from both the Census and the FEC.[5] In this analysis, the dependent variable is the amount in contributions received from Asian ethnic groups (Chinese, Japanese, Korean, and Vietnamese).[6] The independent variables are from the U.S. Census (Census STF3b file), and have been merged to the FEC data. The independent variables include several measures of socioeconomic status, including income, education, and age. The income variable is the median income in the zip code, measured in $10,000s. The education variable is a 7-category variable that measures the mean educational attainment. The age variable is a 5-category variable that measures the mean age in the zip code. Also available from the census is total population in a zip code[7] and the percentage of the zip code that various groups (such as Asian Americans, blacks, and Latinos) comprise.

Indicators of Spatial Autocorrelation

The first step in a spatial analysis is to determine whether there is any spatial autocorrelation in the data at all.

[1]The full collection of FEC individual contribution reports (1980–1998) is very large, approximately 6 million records. The database was parsed using Asian name dictionaries (both first and last names).

[2]Since the FEC do not include any demographic variables, it is difficult to place many identifying characteristics on the individual donors. While it would be extremely interesting to contrast these findings on Asian Americans with whites, or blacks, or Latinos, such an analysis is not feasible with the FEC data, as none of these three groups can be reliably identified in the data.

[3]This smaller data set makes this problem manageable. Analyzing the entire FEC data set is not feasible at this point because of computing limitations associated with the massive size of the entire data set and the computational intensive nature of spatial analyses (Smirnov and Anselin 2001).

[4]Surveys, often the best sources of individual-level data, are of limited usefulness here, since they conflict markedly in their accounts of campaign contribution levels (Cho 2002). Another major drawback of surveys for the current task is that geographic identifiers are rarely available.

[5]The observations are not individual contributors. While zip codes for individuals are also available, the number of observations is a limiting factor. The greater problem, however, is that the FEC provides no demographic variables for individuals. The census, on the other hand, collects a large number of variables at the zip code level. Thus, the zip code level is an appropriate aggregation level because it is the lowest level of aggregation for which there is extensive data available for estimating the models and theories of interest.

[6]The pan-ethnic identity is certainly one of great contention (Espiritu 1992; Tam 1995), and so the use of the umbrella category always needs to be broached with caution. Many have argued that the pan-ethnic group rises to the occasion in contexts where they are treated by others as a homogeneous group (Espiritu 1992; Lien 2001). In these instances, they join together to fight a common cause or misconception where they have a joint stake. The case of campaign donations and the 1996 campaign finance scandal surrounding Asian donations is certainly a case in point, and so the use of a pan-ethnic category here is justifiable. To the extent that the pan-ethnic identity is not appropriate, the results that follow are conservative estimates of the possible diffusion processes at play. These processes are likely to be even stronger if we were to examine just one ethnic group as diffusion effects are more likely within a single ethnic group rather than across the often internally heterogeneous Asian American group.

[7]Total population is available, but to ensure some consistency in the range of variables, the population variable in the spatial models is population in 1000s of people.

Accordingly, we want to test the null hypothesis of spatial randomness against the alternative hypothesis of spatial autocorrelation. Spatial autocorrelation occurs when values of a certain variable are systematically related to their geographic locations. That is, there is some relationship between the levels of donations in neighboring area. Evidence of such a relationship would support the spatial theories. If the spatial autocorrelation statistic is statistically significant, however, further analysis needs to be conducted to determine the source of the autocorrelation. The Moran's I statistic (Moran 1948; Cliff and Ord 1973) is the most commonly employed method of assessing the significance and/or degree of spatial autocorrelation in the data (Cliff and Ord 1981). A positive and significant Moran's I indicates spatial clustering of contribution amounts. Specifically, the Moran's I statistic is

$$I = \frac{\sum_i \sum_j w_{ij}(y_i - \mu)(y_j - \mu)}{\sum_i (y_i - \mu)^2},$$

where w_{ij} is an element of a row-standardized spatial weights matrix, y is the contribution amount, and μ is the average contribution amount in the sample. The Moran's I statistic can be thought of as a counterpart to the familiar Durbin-Watson statistic used to detect autocorrelation in time-series data. Spatial autocorrelation occurs when the similarity of values of interest is related to the locations of the units, i.e.

$$\mathrm{Cov}(y_i, y_j) = E(y_i y_j) - E(y_i)E(y_j) \neq 0, \quad \forall\, i \neq j.$$

If spatial autocorrelation is present in the data, models that do not explicitly account for spatial effects are inadequate for adjudicating between spatial and nonspatial theories. If spatial randomness is rejected, the next recourse is to explore the processes that may have generated the observed spatial patterns.

We can see from Table 1 that the global Moran's I statistic is highly significant for every year of the FEC data.[8] However, note that even if the pattern seems to be spatially clustered, the pattern of contributions may, in fact, be spatially random, driven simply from clustering of demographic traits such as income. So far, we have only observed spatial patterns. We cannot yet make any claims about why these patterns occur, because we have not conducted any analysis of this type. We have simply surmised that the pattern of contributions, without controlling for any variables, is not random.[9] We will explore the source of the spatial dependence in the spatial regression models to follow.

TABLE 1 Global Moran's I Statistic

Year	Moran's I	Z-value	p-value
1980	0.2845	21.79	0.00
1982	0.1398	8.35	0.00
1984	0.1981	14.39	0.00
1986	0.0411	3.49	0.00
1988	0.1739	24.06	0.00
1990	0.3749	50.98	0.00
1992	0.2599	50.10	0.00
1994	0.2196	35.28	0.00
1996	0.1795	35.51	0.00
1998	0.2612	47.27	0.00

We can obtain a more detailed look at spatial autocorrelation by examining the local indicators of spatial autocorrelation (LISA) statistics (Anselin 1995). This local Moran statistic is closely related to the global Moran's I statistic. Specifically, the local Moran's I statistic is

$$I_i = \frac{z_i}{\sum z_i^2} \sum_j w_{ij} z_j \tag{1}$$

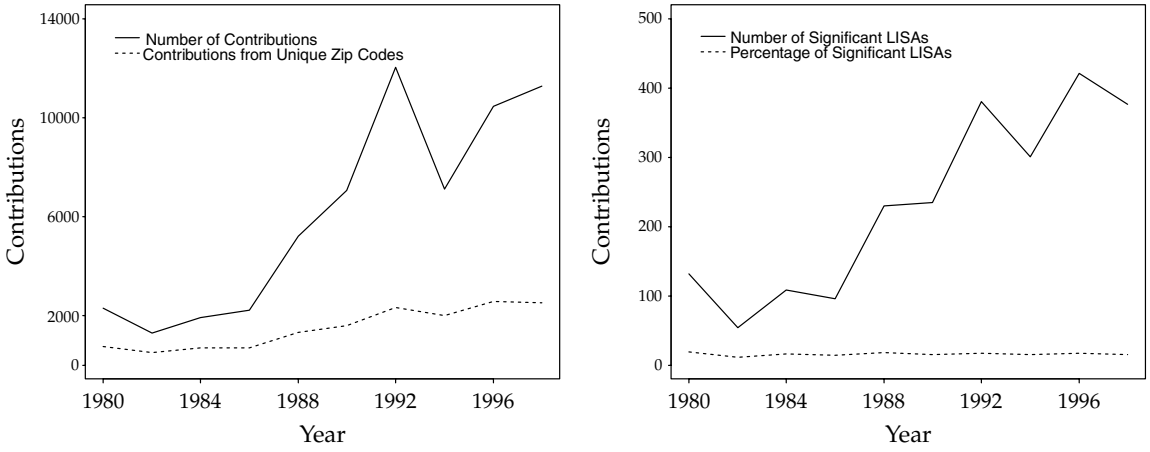
where z is the mean-deviated contribution amounts given by Asian Americans. Inference is based on a conditional randomization approach.[10] The average of the local Moran's I statistics is equal to the global Moran's I, to a factor of proportionality. Examining the local autocorrelation statistics allows us to identify observations that are "extreme contributions" to the global statistic by noting which values are, say, 2 or more standard deviations from the mean. These local indicators allow us, moreover, to

[8]The weights matrix is based on an inverse distance measure where the distance band is 50 miles. That is, the spatial lag for each zip code can be seen as the weighted average (with the w_{ij} being the weights) of its geographically-defined neighbors (those zip codes that fall within the distance band). The distance is measured from the centroid. Different weights matrices (but the same specification, as described above) are computed for each year, so the connectivity structure differs from year to year, where the change is dependent on the specific contributions in that year. In the 1998 data, the minimum number of neighbors is 0. Fifty-one observations have no neighbors. The maximum number of neighbors is 253. Only one observation has this many neighbors. The average number of neighbors is 86. The specification of the weights matrix is important in any spatial analysis. Accordingly, here, different specifications for the weights matrix were examined. For instance, a distance band of 100 miles was also employed. The results were basically identical to those that resulted from using the 50-mile band. K-nearest neighbor and contiguity definitions were briefly explored, but were not used, as it is difficult to reconcile these specifications with a substantive story or theory.

[9]Note as well that because the Moran's I statistic is sensitive to other forms of specification errors such as non-normality and heteroskedasticity (Anselin and Rey 1991), these results should be examined further. Both of these characteristics can affect the sensitivity of the results.

[10]Significance was based on a permutation approach with the number of permutations set at 999.

FIGURE 1 LISA Statistics and Rise in Contribution Activity
The plot on the left shows the rise in the sheer number of contributions and the rise in the number of sites where contributions originate. The plot on the right shows a rise in the number of significant LISA statistics each year and the relatively stable percentage of significant LISA statistics.



identify areas of interest that may have nonrandomly distributed values (high or low) in relation to their neighboring values. Rejection of the null hypothesis here indicates local clustering (either a high value surrounded by high values or a low value surrounded by low values) or local spatial outliers (a high value surrounded by low values or a low value surrounded by high values). Figures A-1–A-10 display plots of the LISA statistics for each of the years listed in Table 1.[11] The black dots indicate areas with significant LISA statistics. The grey dots indicate areas with insignificant LISA statistics.

Two observations are immediately obvious from this set of plots. The first observation is that the number of sites where contributions originate generally increases every two-year cycle. An examination of the data indicate that the sheer number of contributions generally increases every election cycle as well. We can see this graphically in the plot on the left in Figure 1. So, as a group, Asian Americans are becoming increasingly active in this form of political participation. The number of contributors is rising and their geographical diversity is growing. Second, more observations have significant LISA statistics at the end of the time cycle than at the beginning. In other words, with each passing election cycle, more observations are correlated with their neighboring values, giving one more reason to explore possible diffusion effects. We can see this graphically in the plot on the right of Figure 1. The dotted line in this plot indicates that despite the clear rise in both the sheer number of significant observations and the clear upward swing of the numeric base of correlated values, the percentage of all LISA statistics that are significant in a given year does not change dramatically over time.

If we make the leap to assume that the spatial autocorrelation is more likely to originate from donors who have resided in the U.S. for a longer period of time, because they have simply had more opportunities to integrate into a neighborhood structure, the patterns in the LISA statistics might implicate some themes in the literature. For instance, there may be evidence for the idea that newly arrived immigrants behave uniquely relative to those who have resided in the U.S. longer because their incentives and cost structure differ significantly (Cho 1999; Wong 2000). Relatedly, others have argued that one's stake in the political system and thus one's level of political participation rises concurrently with the amount of time an immigrant is in the U.S. (Uhlaner, Cain, and Kiewiet 1989). For Asian Americans, then, the continuous rapid flow of immigrants serves to supply a constant set of new immigrants to the mix as well as to expand the base of potential contributors. The rising number of significant LISA statistics imply that these contributors are spatially related. This phenomenal growth shows no sign of yielding. The flow of Asian immigrants into the U.S. has been nothing short of dramatic in the last few decades. The growth in the number of contributors has almost kept this same phenomenal pace.

[11] Alaska and Hawaii are included in the analysis, though not in the plots. The omission from the plots is purely a matter of aesthetics.

It is clear from these census figures and the plots in Figure 1 that campaign finance is an increasingly pertinent arena for those interested in the political behavior of the most rapidly growing group in the U.S. With the passing of each election cycle, a growing number and a broader geographic mix of Asian Americans are engaging in this form of political participation. Perhaps more importantly, and more indicative of the sophistication behind this form of behavior, the spatial patterns imply that the web that underlies this form of behavior is growing in size as well as complexity.

Spatial Models

Given that we have significant spatial autocorrelation in our data, both on a global scale as well as a more local scale, as indicated by the Moran's I statistic and the large number of significant LISA statistics every year, the next step is to determine whether these spatial effects are true spatial effects, or if they are spurious, in the sense that they can be attributed entirely to patterns in other variables such as income or education. If we control for all of these other factors and the spatial variable remains significant, then we have evidence that the pattern is consistent with a "neighborhood effect"(via a spatially lagged dependent variable), or perhaps an elite mobilization effect (via an unmeasured mobilization variable), but not an effect that is solely attributable to socioeconomic characteristics of these areas. In other words, if contribution amounts are determined solely by the structural factors included in the model as independent variables, no remaining spatial patterning of contribution amounts beyond those resulting from socio-demographic similarity of geographically-proximate areas should remain.

For the national data, the spatial dependence that appears in the contributions data may be modeled as a spatial lag model.[12] The robust Lagrange Multiplier diagnostics for each of the years, excluding 1986, indicate that the spatial lag route would be profitable. In the spatial lag model, an otherwise routine regression has an additional regressor that takes the form of a spatially lagged dependent variable, Wy. That is, the spatial lag model would take the form

$$y = \rho\, Wy + X\beta + \varepsilon$$

where W is an $N \times N$ spatial weights matrix, ρ is the spatial autoregressive coefficient, ε is the error term, and X and β have the usual interpretation in a regression model. The spatial lag can be seen as the weighted average (with the w_{ij} being the weights) of its geographically-defined neighbors. In this model specification, because the lag term is correlated with the error term, OLS should not be used, since it will be both biased and inconsistent (Ord 1975; Anselin 1988). Instead, the spatial lag model should be estimated via a maximum likelihood or instrumental variables formulation.

The spatial lag model is most consistent with contagion theories and diffusion processes. The explicit inclusion of the spatial lag term implies that the influence of a "neighbor's" (as defined by the weights matrix) contribution amount is not an artifact of measured and unmeasured independent variables, but that the contribution amounts in neighboring areas actually increases the likelihood of campaign contributions in its neighbors. Note that the evidence of a diffusion or contagion effect is indirect. The spatial regression models cannot identify the specific mechanism that produces the spatial effects. Instead, the value added is that *if* the observed phenomenon were actually characterized by a diffusion process, then we would expect to see these spatial imprints emerge. The discovery of spatial effects, then, behooves future research to place some emphasis on uncovering the mechanisms that would produce diffusion.

[12]The specification of the spatial model is, of course, chosen after examining the data and various diagnostics. The other large class of models involves modeling the spatial dependence as a spatial error model. In the spatial error model, the dependence is incorporated into the error structure so that $E[\varepsilon_i \varepsilon_j] \neq 0$, i.e. the off-diagonal elements of the error covariance matrix are non-zero and incorporate the structure of the spatial dependence. In this case, OLS is unbiased but is not efficient. So, the estimate of standard errors will be biased. The spatial error model would evaluate the extent to which the spatial patterns of campaign contributions not explained by the measured independent variables can be accounted for by clustering of error terms. In other words, the spatial error model captures the spatial effects of unmeasured independent variables. A satisfactory spatial error model implies that a spatially-lagged dependent variable is not necessary for explaining the observed spatial patterns. Instead, the patterns are explained by geographic patterning of measured and unmeasured independent variables. Whether a spatial lag or a spatial error formulation is employed is a decision that is based on diagnostics. In this particular study, the diagnostics indicated that no spatial effects remained in the 1986 national data after controlling for other covariates. The spatial effects were not detected in the model even though the Moran's I statistic for the 1986 data was significant. This is not unusual, as the Moran's I statistic is very sensitive to various forms of specification errors such as non-normality and heteroskedasticity (Anselin and Rey 1991). A little exploration into these data indicate the presence of both non-normality and heteroskedasticity. For the other years, the robust Lagrange Multiplier diagnostic for the spatial lag was significant. Somewhat atypically, the robust Lagrange Multiplier diagnostics for spatial error was also significant in some of the years (1988–1998 for the national data, and 1990, 1992, and 1998 for the western region data). Because the robust Lagrange Multiplier lag test statistic is larger than the robust Lagrange Multiplier error test statistic, a spatial lag model was pursued. Attempts to explore sources of spatial heterogeneity in the data helped to resolve this issue for some of the data, but not these aforementioned years.

These data also exhibit qualities consistent with spatial heterogeneity as indicated by spatial Chow tests on the overall coefficient stability across regimes. In particular, we can see from Table 2 that a spatial Chow test indicates that observations in the West differed significantly from observations in other states for each election cycle beginning in 1988. Given the evidence of distinct spatial regimes, a disaggregated modeling strategy is pursued for these data. That is, we analyze separate models for each region and examine each of these models separately for evidence of spatial dependence. In addition, spatial Chow tests for the non-western states also indicate that the states in the Northeast are significantly different from the other non-western states for each of the years from 1988–1998, except 1990. These pockets of distinctive behavior are not surprising given our substantive inclinations about the Asian American group. The bulk of the Asian American population resides in the West, and proportionally, the West bears more than its share of campaign contributors. The Northeast bears many of these same qualities, but to a lesser degree.

The models for the entire nation are reported in Table 2.[13] The models for the other regions are displayed in Tables 3 and 4. Spatial lag models are indicated by a value for the spatial lag variable, while spatial error models are indicated by a value for the spatial error variable.[14] In each of the tables, the column heading is the year indicator. The dependent variable is the amount in contributions received from Asian donors, and observations are zip codes.[15]

The results vary somewhat from year to year, but some consistent themes are evident as well. One implication of these patterns of change and stability is that the logic behind Asian American contributions is evolving, not static in this twenty-year time period. Although one might prefer and expect an overarching story, the lack of a single story throughout this time period is not unusual and should not necessarily be expected given the phenomenal growth and compositional change that has characterized this time period for the Asian American group. Asian Americans have been arriving in droves only since the late 1960s, after the Immigration and Nationality Act of 1965 eliminated racial quotas. One can hardly expect in some 15–30 years that they would have established deep-set grooves of political behavior in an American system that was, until just recently, largely foreign to them. Integrating into the political mainstream certainly does not occur instantaneously (see e.g., Reedy 1991 and Glazer and Moynihan 1972), and so we should not expect that a group's political presence would appear instantaneously with its physical presence. Moreover, we would expect the establishment of a pattern of political behavior to follow later yet.

The theme of change is one that has been uncovered by previous studies (Tam 1995; Wong 2000; Lien 2001). Indeed, as will see, several themes of the Asian American political behavior literature will be uncovered again, while some will appear to have changed. The difference is that spatial effects are explicitly considered here. The interplay between spatial effects and the traditional individual-level variables affects the results. In some instances, the effects will be evident simultaneously. In others, one effect may dominant the others or negatively impact the others.

Strikingly, the early patterns of contributions do not seem to be related to income. The lack of relationship here is surprising given the strong relationship between an individual's income level and campaign contributions that has been uncovered by more than one study (Rosenstone and Hansen 1993; Verba, Scholzman, and Brady 1995; Brown, Powell, and Wilcox 1995; Gierzynski 2000). In these data, the relationship between income and

[13] As previously discussed, non-normality and heteroskedasticity affect the estimation (Anselin and Rey 1991). Because the Koenker-Basset diagnostic for heteroskedasticity indicated that heteroskedasticity might be an issue for the 1980, and 1990–1998 national data, the model for these years is computed via instrumental variables (2SLS) with a groupwise heteroskedasticity variable. In the presence of a high degree of non-normality and especially for large data sets, 2SLS is the preferred strategy over the asymptotically more efficient maximum likelihood approach. The groupwise heteroskedasticity variable is a region variable (Northeast, South, Midwest, Central South, Mountain states, California, Hawaii, and the rest of the West). Inclusion of this variable alleviated the problem with heteroskedasticity in all of the years except 1990 and 1992. That is, the Koenker-Basset diagnostic does not indicate a problem with heteroskedasticity once this region variable is included.

[14] For the West, the Koenker-Basset diagnostic for heteroskedasticity indicated that heteroskedasticity might be an issue in these data for the years 1988–1998. Hence, the model for these years was computed via instrumental variables with a groupwise heteroskedasticity variable. In this case, the groupwise heteroskedasticity variable separated the regions of California (Bay Area, Los Angeles area, Central Valley, and the remaining parts of California), and the rest of the western region (Oregon, Washington, Hawaii, and Alaska). Inclusion of this variable alleviated the problem with heteroskedasticity in all of the models. In 1990, and 1998, the data indicate that a spatial error model may be appropriate, but the spatial error model diagnostics indicate additional spatial lag dependence, and the spatial lag models had slightly better fit statistics, and so Table 3 reports the spatial lag results. For the Northeast, each model was computed via instrumental variables with a groupwise heteroskedasticity variable. In the other regions, only the 1992 and 1996 data had indications of heteroskedasticity. The other models were computed via maximum likelihood.

[15] Only zip codes where some money originated from Asian Americans are included. The other zip codes are essentially "islands," and so there is no "neighboring behavior" to observe or analyze.

TABLE 2 Spatial Lag Models. Dependent Variable: Contribution Amount from Asian Americans at the Zip Code Level

	1980	1982	1984	1986	1988	1990	1992	1994	1996	1998
Constant	25.14	−2481.55	−3014.99	−7372.14**	−564.30	−1622.47**	−2865.79**	−2021.39**	−3853.76**	−1833.75**
	(418.09)	(2119.21)	(1798.39)	(2266.44)	(756.81)	(721.32)	(915.54)	(766.01)	(951.51)	(669.68)
Population	6.16**	10.17	9.20	42.40**	4.42	19.55**	20.13**	0.15	24.60**	10.96**
	(3.06)	(11.02)	(9.78)	(13.87)	(5.88)	(5.16)	(6.68)	(5.27)	(7.25)	(5.05)
Percent Asian	43.30**	40.01**	65.22**	73.95**	112.93**	68.71**	83.12**	100.87**	124.81**	164.87**
	(8.44)	(14.38)	(13.12)	(16.38)	(15.75)	(14.98)	(12.92)	(15.00)	(18.88)	(20.03)
Age	242.79	1590.89**	1803.25**	1931.35**	553.53**	454.60	979.77**	664.38**	1447.73**	522.42**
	(159.65)	(733.99)	(707.76)	(802.33)	(234.85)	(262.45)	(304.83)	(261.81)	(338.45)	(232.35)
Education	−95.54	−209.69	−86.92	439.88	−116.13	−0.47	36.33	261.54	−80.24	224.68
	(96.57)	(434.86)	(379.02)	(506.12)	(162.93)	(162.30)	(196.48)	(168.58)	(213.84)	(150.52)
Income	40.46	259.37	148.76	444.82**	205.30**	281.94**	348.60**	141.54**	500.78**	153.30**
	(38.00)	(154.13)	(136.19)	(176.11)	(66.04)	(67.81)	(82.49)	(65.67)	(99.39)	(71.34)
Percent Minority	2.76	−4.55	−1.15	26.27	−1.35	3.66	0.71	3.24	5.25	−1.14
	(2.34)	(12.88)	(10.14)	(13.64)	(4.33)	(4.22)	(4.98)	(4.52)	(5.25)	(3.67)
Spatial Lag (r)	0.06**	0.04**	0.03**		0.03**	0.03**	0.03**	0.03**	0.02**	0.03**
	(0.01)	(0.01)	(0.01)		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Spatial Chow Test	6.15	6.88	8.00	9.84	33.86**	84.79**	108.79**	32.87**	27.68**	61.56**
R^2	0.17	0.13	0.14	0.08	0.19	0.33	0.22	0.16	0.14	0.20
N	671	455	657	639	1183	1420	2072	1821	2288	2206

Note: Standard errors in parentheses.
** $p < 0.05$.

TABLE 3 Spatial Lag Models. Dependent Variable: Contribution Amount from Asian Americans at the Zip Code Level (West Only)

	1980	1982	1984	1986	1988	1990	1992	1994	1996	1998
Constant	−269.49	1300.38	1273.46	−5053.18	−5336.42**	−3531.08	−6677.27**	−6443.86**	−3911.88	−9345.54**
	(751.16)	(2339.73)	(1885.05)	(3447.91)	(2213.28)	(2036.34)	(2257.39)	(1930.93)	(2136.92)	(2599.79)
Population	6.38	19.45	19.78	31.31	47.38**	26.87**	33.26	34.32**	56.95**	38.66**
	(6.93)	(14.01)	(12.26)	(21.61)	(15.90)	(13.76)	(17.64)	(13.23)	(15.73)	(17.33)
Percent Asian	40.63**	29.81**	48.12**	42.02**	111.74**	112.45**	214.21**	111.91**	98.17**	167.71**
	(9.71)	(10.59)	(11.29)	(15.61)	(16.98)	(20.68)	(25.57)	(14.62)	(17.49)	(21.41)
Age	−60.94	906.51	1142.12	3532.11**	3267.05**	1461.37	2069.68**	2030.95**	2702.57**	2788.22**
	(294.02)	(708.32)	(683.38)	(1209.70)	(872.03)	(833.55)	(836.89)	(651.67)	(853.53)	(992.76)
Education	119.49	−737.77	−829.18	−805.85	−1181.30**	−354.89	−202.29	226.20	−1142.38**	−38.71
	(180.51)	(543.53)	(433.34)	(751.14)	(524.87)	(494.26)	(510.84)	(449.13)	(538.20)	(617.82)
Income	44.16	141.26	116.14	455.59	914.85**	582.17**	870.44**	421.84**	871.88**	1256.83**
	(90.09)	(168.92)	(150.31)	(268.42)	(205.40)	(206.28)	(244.68)	(168.99)	(239.26)	(257.50)
Percent Minority	−16.79**	−22.42	−33.62**	−18.84	−53.31**	−42.64**	−69.96**	−32.32**	−53.84**	−33.62
	(6.77)	(15.85)	(12.53)	(24.89)	(19.02)	(15.32)	(20.17)	(15.44)	(19.82)	(20.15)
Spatial Lag (r)	0.07**	0.06**	0.06**	0.05**	0.05**	0.05**	0.05**	0.04**	0.05**	0.03**
	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
R^2	0.21	0.11	0.15	0.07	0.24	0.43	0.30	0.20	0.16	0.27
N	291	229	249	273	458	502	698	606	660	752

Note: Standard errors in parentheses.
** $p < 0.05$.

TABLE 4 Spatial Lag Models. Dependent Variable: Contribution Amount from Asian Americans at the Zip Code Level (Northeast and Other Regions)

	1988	1990	1992	1994	1996	1998
Northeast						
Constant	−17164.50**		−8794.43**	−11076.4**	−23848.30**	−1229.86**
	(3999.68)		(2288.65)	(2477.17)	(4253.9)	(3416.58)
Population	39.58**		51.01**	37.51**	97.78**	41.90**
	(18.12)		(12.35)	(12.74)	(22.67)	(18.77)
Percent Asian	219.16**		130.98**	67.95	390.33**	132.32**
	(49.61)		(35.98)	(35.35)	(67.31)	(53.64)
Age	6134.51**		2458.78**	2770.62**	5860.38**	2546.80
	(1596.74)		(916.52)	(989.90)	(1699.22)	(1431.64)
Education	213.88		332.24	1214.24**	1709.21**	1467.45**
	(646.87)		(416.24)	(475.53)	(860.83)	(624.09)
Income	823.68**		485.95**	138.64	715.05**	224.22
	(233.01)		(153.39)	(171.13)	(311.64)	(230.03)
Percent Minority	16.83		−7.10	−5.38	8.66	−1.88
	(20.03)		(11.89)	(12.47)	(23.08)	(17.22)
Spatial Lag (r)	0.00		0.02**			0.02**
	(0.01)		(0.00)			(0.01)
Spatial Error (l)				0.05**	0.03**	
				(0.00)	(0.01)	
R^2	0.17		0.21	0.14	0.17	0.13
N	401		613	516	651	581
Other Regions						
Constant	18.21	−2626.48**	−2903.96	−461.45	−2247.06	569.54
	(1354.16)	(1202.37)	(2307.06)	(1134.60)	(1618.68)	(879.10)
Population	−1.36	26.81**	0.94	−7.71	−3.58	5.81
	(8.88)	(7.49)	15.60	(8.11)	(11.16)	(6.31)
Percent Asian	59.72	104.70**	12.55	29.98	63.31	110.11**
	(43.94)	(25.17)	52.89	(46.32)	(61.39)	(40.53)
Age	201.20	710.69	402.56	314.01	1016.10	283.07
	(426.22)	(446.67)	(795.24)	(389.48)	(568.06)	(300.91)
Education	141.81	41.52	404.67	228.03	−273.28	−266.73
	(280.91)	(245.46)	(518.80)	(258.08)	(354.28)	(196.41)
Income	100.18	332.26**	474.49**	48.15	698.35**	212.99**
	(108.82)	(96.50)	(214.22)	(104.78)	(156.36)	(89.94)
Percent Minority	0.70	3.03	20.74	7.25	13.68	−2.12
	(7.57)	(6.68)	(6.69)	(13.09)	(8.81)	(4.97)
Spatial Lag (r)	0.05	0.01**	0.03	0.06**	0.05**	0.07**
	(0.03)	(0.00)	(0.03)	(0.02)	(0.02)	(0.01)
R^2	0.05	0.11	0.03	0.03	0.07	0.08
N	324	918	760	701	980	874

Note: Standard errors in parentheses.
** $p < 0.05$.

contributions does not appear to be established until the mid-1980s. Once this pattern emerges, the analysis indicates that it persists. Although this relationship may not concur with the bulk of the literature on campaign finance, one must consider that the population bases for these previous studies have been comprised of primarily native-born Americans. Variation on the native-born/foreign-born dimension in those studies was

virtually absent and if present, not a primary concern in the analysis.

These data, in contrast, are rich in variance on the nativity dimension, permitting us to catch a glimpse into the unfolding of the political incorporation dynamic. Hence, while the initial reaction to an insignificant income coefficient is somewhat surprising, some reflection on the context of these data mutes this initial reaction. Despite the oft-heard claim that Asians contribute primarily because of their high income levels (Lew 1987), then, the evidence here implies that the dynamic is considerably more complex. The evidence for a contagion effect is becoming stronger with each passing election, implying that Asian Americans are becoming increasingly sophisticated political actors with more and stronger intragroup ties.[16] As indicated by the significant income coefficient in later years, income levels may partly explain the behavior, but it is certainly not the sole determinant, and not even a significant part of the explanation in the earlier years.

Another strong similarity between the results in the various regions is that the "Percent Asian" variable is significant and generally rising throughout. The only exception is in the "Other Regions," where Asian Americans are also the most sparse. In the rest of the country, however, the main jump in values for this variable, as for the income variable, again occurs in the mid-1980s, where the magnitude of the coefficients makes a clear rise to a new plateau. At this point, activity rises, and the organization of this activity becomes more evident. The basic interpretation of the coefficient on "Percent Asian" is that as the percentage of Asians in a zip code rises, so too do the dollar amounts that are donated by Asians from those areas. Note that this is not simply a function of a larger population base in some areas, since the model controls for differences in population. The effect is over and above the population effect. Hence, Asian Americans are at least as active as others in terms of donating, and far from passive in this form of political behavior. This effect is present even when the income effect is absent, so it is not necessarily related to socioeconomic factors.

On more than one dimension, then, the confluence of a number of factors in the mid-1980s seems to have signaled a silent "new age" for Asian American politics, one that gives credence to the claims that the Asian American group is a "sleeping giant." The giant still appears to be in a state of semi-slumber, but there is evidence that the giant is beginning to stir. Notably, the observed changes in the 1980s coincide with the appearance of the first significant numbers of Asian American candidates for political office.[17] So the period of the 1980s for this group was indeed characterized by change on many political dimensions.

The most notable difference between the results for various regions is that, in all regions except the West, as the percentage of other minority residents rises, there seems to be little to no effect on campaign contributions from Asian Americans. In the West, however, we can see from Table 3 that there is a significant and negative effect. So, on a national scale and in the non-western states, after controlling for other variables, the dollar amounts that flow to candidates neither rises nor falls as the heterogeneity of the minority composition increased. In the West, the dollar amounts from Asian Americans decline in areas that are more ethnically heterogeneous. Hence, while there seems to be some ethnic contribution network at play among Asian Americans, this web of donations does not appear to cast itself more widely to include other ethnic groups on a national scale, and is negatively affected by other ethnic groups in the West. This result accords with much of the ethnic studies literature on political coalitions, namely, that Asian Americans do not generally align themselves with other minority groups to form a broader coalition (Saito 1998; Cho and Cain 2001; Lien 2001). The field of

[16]Note that while the data here are at the zip code level and we are primarily interested in individual behavior, this is not a classic case of the ecological inference problem. The reason is that the dependent variable, the amount in contributions from Asian Americans, we know, is attributable only to the Asian American residents in the zip code areas. We are unsure if the median income of the zip code area is representative of the median income of Asian Americans in the zip code area. However, a zip code area is a relatively small geographic unit, so this should not be a pervasive problem. The issue is how well the variable measures the underlying heterogeneity. Thus, problems that may occur in the analysis and interpretation are all related to the extent that zip codes variables do not adequately capture the underlying heterogeneity. This same caveat applies to the age variable. Other variables like population, percent Asian, and percent minority are less problematic. The percent Asian variable can obviously be attributable only to Asian Americans, and this variable is a basic measure of context. The percent minority variable does not apply only to Asians, but is instead the percentage of the area that is comprised of blacks and Hispanics. However, this variable, like the percent Asian variable, is a measure of context. It gives us an indication of how contribution amounts from Asian Americans vary as a function of context. The population variable allows us to examine how contribution amounts vary as population density changes. Hence, these last three variables are not problematic, as they are measured on a level of interest.

[17]Prior to the 1980s, there were few Asian American candidates. Although there were not a large number of candidates in the 1980s, and many of the candidates who did run were unsuccessful, the rise in numbers was nonetheless significant. During the 1980s, the number of Asian American candidates who ran for office rose to the double digits. This number increased dramatically in the 1990s. See Cho (2002) for a complete list.

campaign finance, where Asian Americans are especially active, does not appear to be an exception.

Interestingly, even after we control for socioeconomic and demographic factors, variables widely recognized to be influential, the spatial lag is still significant. This result holds for every year examined in the entire U.S. (except for 1986) as well as just the western region. The general pattern holds in the other regions as well. In the West, although the patterns among all of the variables is roughly the same, the magnitude of the spatial lag effect is larger and the models generally explain more variance in the data. It is not particularly surprising that the West would exhibit more spatial effects given that it has traditionally hosted and continues to host the bulk of the Asian American populace. The type of ethnic networks that we seek through the analysis pursued here are most likely to occur in locations where Asian Americans have resided the longest. As the length of time increases, there are more opportunities to integrate into the community as well as into the political scene.

Moreover, as the size of the community grows, there are more opportunities and outlets to integrate into the community. Hence, what is surprising is the simple existence of spatial effects that is evident apart from the tried-and-true socioeconomic indicators. Contribution amounts, then, are not generated solely by the nonspatial structural factors that have been identified by earlier research. Given that these effects exist, the relative magnitude of these effects in the West and in the nation align with initial expectations. While some of these spatial lag effects may seem small initially, note that the dependent variable is dollar amounts. In this time span, Asian Americans have contributed millions of dollars. The spatial lag effects, in terms of proportion, has remained largely the same, but the dollar amounts have increased seven-fold, and have regularly exceeded 7 million dollars in the 1990s. Thus, the effect is both substantively and statistically significant.[18] These effects are also likely to be tempered because the Asian American group is considered as a whole here, but ethnic networks are likely to be stronger within the various ethnicities that comprise the larger group.

The autocorrelated geographic patterns that we see in the models with significant spatial lag effects are typical of those patterns we might observe if contagion or diffusion effects were at play. People influencing people, and contributions begetting more contributions. Socioeconomic factors are also at play, but contrary to studies that examine only the atomistic actor outside of the context in which he resides, there is considerable evidence that contextual factors are also at play. The exact manner in which these webs operate is not clear from this analysis. However, we can see the emergence of the idea that candidates can tap into a ethnic contribution network.

For some of the years in the national data, 1988–1998, diagnostics for the spatial model indicate that some spatial error dependence remains. In the data for the western region, there appears to be some remaining spatial error dependence in a few of the years as well (1990, 1992, 1998). The spatial effects in these years are more complex than in the years where the diagnostics indicate no remaining spatial error dependence. That these years are clustered toward the end of the time period examined appears to indicate again that the complexity underlying the contribution dynamic is growing. In earlier years, the spatial lag was sufficient.[19] Because there are remaining spatial effects in these latter years from some unmeasured variable or variables, it is more difficult to expound on the origins of the spatial patterning. There seems to be some effect that can be captured via a spatial lag (i.e. an effect consistent with a diffusion process), but also some effect that may be, perhaps, consistent with a political mobilization or candidate effect story, where the variable (or variables) that measures these effects are not included or perhaps not available. Exploring these additional sources of spatial patterning and the mechanisms that may be lurking beneath these spatial patterns is an obvious extension of the analysis presented here. In general, fit statistics and diagnostics indicated that the spatial lag model was more appropriate, although some spatial error dependence did remain in several of the years. In the Northeast, the spatial error model was more appropriate for two of the years. The evidence, then, seems more consistent with a contagion effect than a mobilization effort,

[18]Moreover, these estimates are conservative. The spatial effects are likely to be much greater. The data here include all donations. If we were to limit the data to donations just to Democrats or just to Republicans, one can see how the diffusion effect would be expected to be larger. Within-party donations are more likely to beget more donations than to beget donations to the other party. Examining all of the donations, then, likely tempers the observed spatial effects.

[19]These models are still not ideal for several reasons. One reason that has already been mentioned is the difficulty of discerning the various types of spatial effects that may exist (i.e., spatial heterogeneity and spatial lag versus spatial error dependence). Another problem is that the model diagnostics show evidence of non-normality. Finally, the inclusion of the spatial dependence parameter did not eliminate the heterogeneity in every case. The data are limited, however, so the difficulty is exacerbated. The data need to be merged into the FEC data at some level of aggregation. The Census provides many socioeconomic variables, but no political variables, which may be useful here.

though there is some evidence of a mobilization effect as well. In any case, these data are complex on the spatial realm, with manifestations of many forms of spatial effects, from lag dependence to error dependence to spatial heterogeneity.

Conclusion

Donation rates across the country vary. Most of the analysis of these rates, however, have focused on the individual actor, acting within his own realm, making decisions based upon his own personal resources. For instance, past research has suggested that socioeconomic factors figure prominently in the decision to contribute (Sorauf 1988; Rosenstone and Hansen 1993; Verba, Scholzman, and Brady 1995; Brown, Powell, and Wilcox 1995; Gierzynski 2000). In the Asian American context specifically, research suggests that socioeconomic factors are at play, but that ethnic cues are also important. In neither the research on minority nor nonminority groups is there a focus on social context and how that might affect the contribution dynamic.

A central question in this article is whether the individual effects remain even after spatial effects are controlled. If the spatial lag is not significant, then we can explain the clustering of donation rates with covariates that measure individual characteristics. That is, "imitation" or diffusion through social networks cannot explain the clustering. Alternatively, if some clustering effect remains even after the covariates related to individual characteristics are controlled, then a diffusion process is likely to be a factor. As we saw from the analysis, spatial effects remain even after individual effects are controlled, implying that some type of diffusion force prominently underlies the Asian American campaign contribution network. In some of the later years, additional spatial error dependence remains, implying perhaps that some elite mobilization effort or candidate effect is fueling some of the spatial patterning. In either case, these spatial effects do not originate solely from the socioeconomic variables included in the model or completely within the realm of the individual-level explanations offered by previous research. Clearly, context plays a role.

By the end of the time span examined here, the patterns that we would expect to occur among the general populous manifest themselves among Asian Americans. Both the spatial lag as well as the socioeconomic indicators become significant predictors of geographical patterns of contributions. Whether these patterns will endure or how these patterns may morph in the future may be questionable, but the roots of the contextual effects have been laid. From the analysis presented thus far, we can see that the story is partly one of the atomistic actor acting alone, but it is also the story of the individual actor within the context of a more broadly defined neighborhood. In this way, the ethnic contribution network is nothing to balk. While less sophisticated political actors act alone, without a clear understanding of the immense benefits that arise from collective action, there is now evidence that Asian Americans are tending away from the less sophisticated individual-level model, and toward a more complex and involved networked model of behavior.

To be sure, Asian Americans bear some unique qualities as only one group of the polity. Nonetheless, the diffusion story likely underlies all of the campaign finance data and has broader implications for political behavior theories.[20] Certainly the social context literature has already advanced these theories, and so empirical studies will not lag much further behind. Here, limiting the analysis makes it feasible. Although broad and varied theories of network and neighborhood effects are posited without matching empirical verification, this analysis begins to move in a new direction and gives gusto to the adage that context matters.

[20]The limitations in computing power and model complexity, at the moment, hinder a larger study of the entire FEC database.

APPENDIX A: LISA Statistics[21]

FIGURE A-1 LISA Statistics for 1980

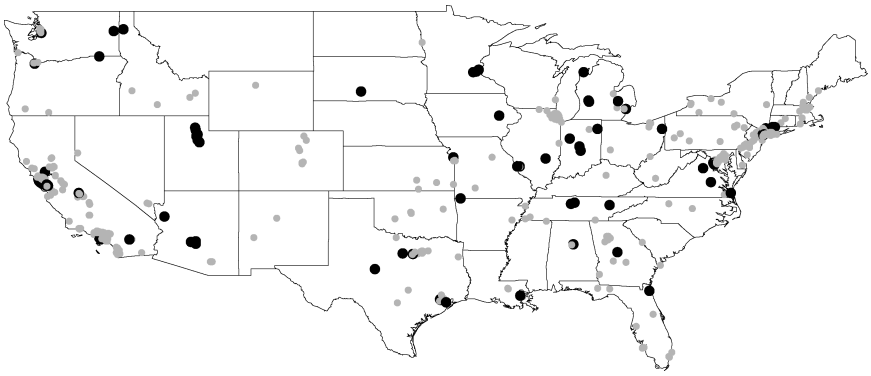
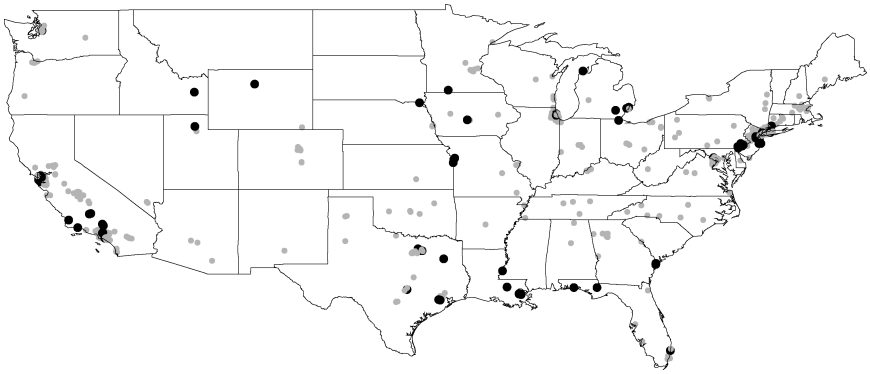


FIGURE A-2 LISA Statistics for 1982

FIGURE A-3 LISA Statistics for 1984



[21]Black dots indicate a zip code with a significant LISA statistic.
Grey dots indicate a zip code with an insignificant LISA statistic.

FIGURE A-4 LISA Statistics for 1986

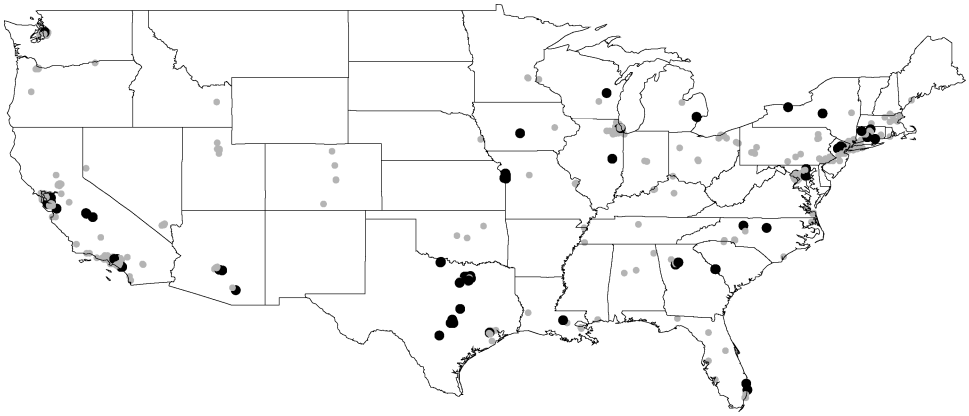


FIGURE A-5 LISA Statistics for 1988

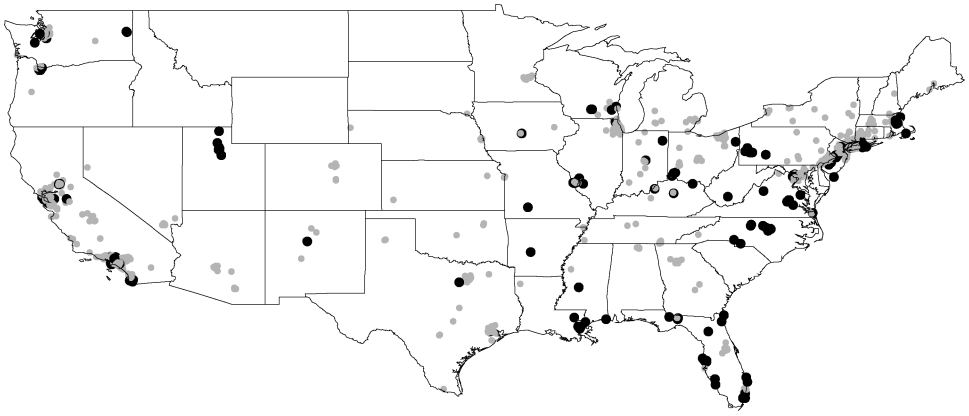


FIGURE A-6 LISA Statistics for 1990

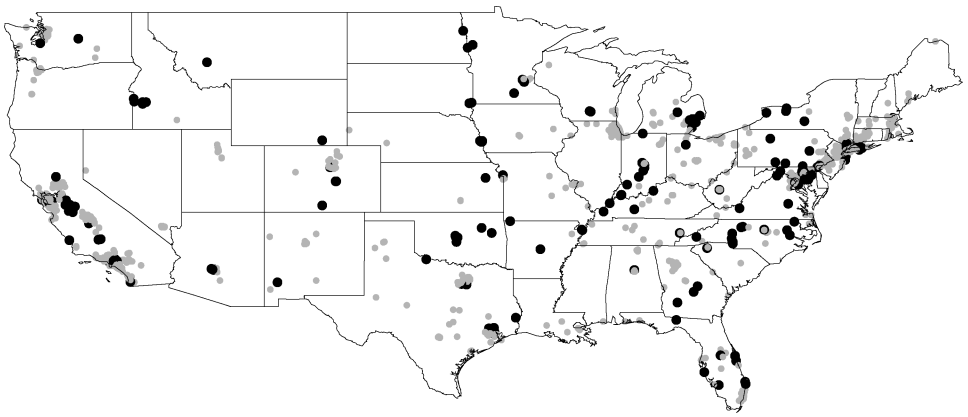


FIGURE A-7 LISA Statistics for 1992

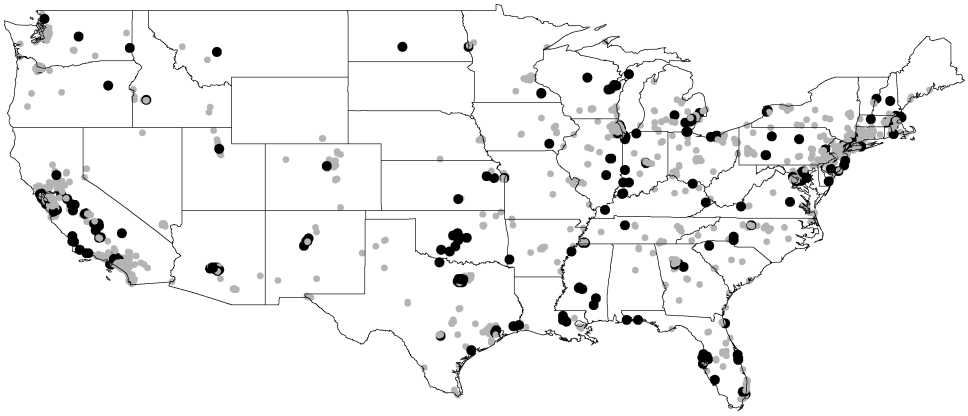


FIGURE A-8 LISA Statistics for 1994

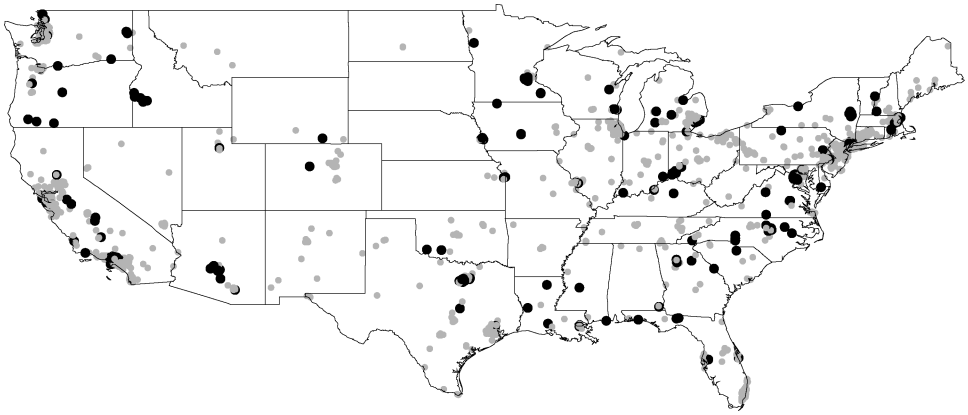


FIGURE A-9 LISA Statistics for 1996

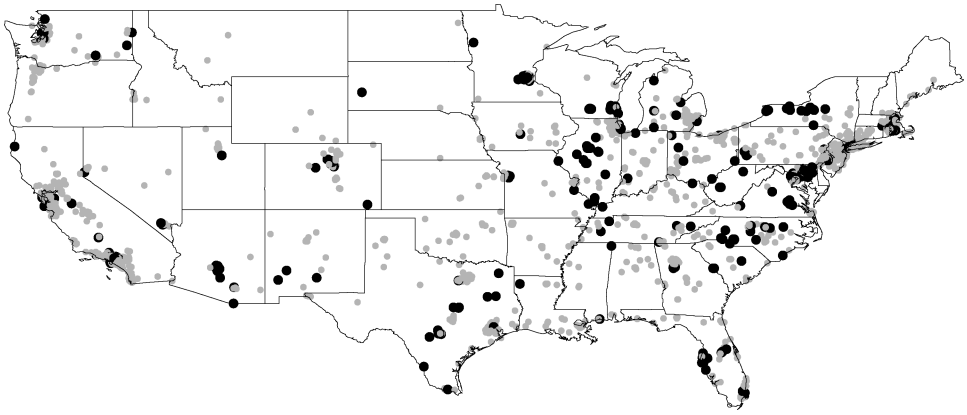
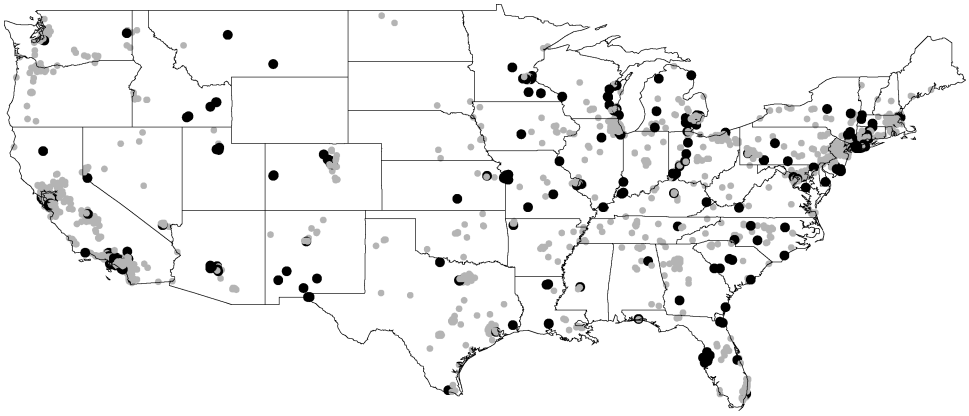


FIGURE A-10 LISA Statistics for 1998



References

Agnew, John A. 1987. *Place and Politics: The Geographical Mediation of State and Society*. Boston: Allen and Unwin.

Anselin, Luc. 1988. *Spatial Econometrics: Methods and Models*. Dordrecht: Kluwer Academic Publishers.

Anselin, Luc. 1990. "What is Special about Spatial Data?" In Daniel A. Griffith, ed., *Spatial Statistics: Past, Present, and Future*. Ann Arbor: Institute for Mathematical Geography, 63–77.

Anselin, Luc. 1995. "Local Indicators of Spatial Association—LISA." *Geographical Analysis* 27(April):93–115.

Anselin, Luc, and Daniel Griffith. 1988. "Do Spatial Effects Really Matter in Regression Analysis?" *Papers, Regional Science Association* 65:11–34.

Anselin, Luc, and Sergio J. Rey. 1991. "Properties of Tests for Spatial Dependence in Linear Regression Models." *Geographical Analysis* 23(April):112–31.

Baybeck, Brady. 2001. "Which Context? Racial Heterogeneity, Geography, and the Individual." Paper Presented at the Annual Meetings of the American Political Science Association, San Francisco, CA, August 30–September 2.

Baybeck, Brady, and Robert Huckfeldt. 2002. "Spatially Dispersed Ties Among Interdependent Citizens: Connecting Individuals and Aggregates." *Political Analysis* 10(Summer):261–75.

Berelson, Bernard F., Paul F. Lazarsfeld, and William N. McPhee. 1954. *Voting: A Study of Opinion Formation in a Presidential Campaign*. Chicago: University of Chicago Press.

Berry, Frances Stokes, and William D. Berry. 1992. "Tax Innovation in the States: Capitalizing on Political Opportunity." *American Journal of Political Science* 36(3):715–42.

Brown, Clifford W., Jr., Lynda W. Powell, and Clyde Wilcox. 1995. *Serious Money: Fundraising and Contributing in Presidential Nomination Campaigns*. New York: Cambridge University Press.

Brustein, William. 1990. "The Political Geography of Fascist Party Membership in Italy and Germany, 1918–1933." In *Social Institutions: Their Emergence, Maintenance and Effects*, ed. Reinhard. Wippler, Karl-Dieter Opp, and Michael Hechter. New York: Aldine de Gruyter, 245–64.

Campbell, Angus, Philip E. Converse, Warren E. Miller, and Donald E. Stokes, 1960. *The American Voter*. Chicago: University of Chicago Press.

Cho, Wendy K. Tam. 1999. "Naturalization, Socialization, Participation: Immigrants and (Non-) Voting." *Journal of Politics* 61(November):1140–55.

Cho, Wendy K. Tam. 2001. "Foreshadowing Strategic Pan-Ethnic Politics: Asian American Campaign Finance Behavior in Varying Multicultural Contexts." *State Politics and Policy Quarterly* 1(September):273–94.

Cho, Wendy K. Tam, and Bruce E. Cain. 2001. "Asian Americans as the Median Voters: An Exploration of Attitudes and Voting Patterns on Ballot Initiatives." In *Asian Americans and Politics: Perspectives, Experiences, Prospects*, ed. Gordon H. Chang. Stanford, CA: Stanford University Press.

Cho, Wendy K. Tam. 2002. "Tapping Motives and Dynamics Behind Campaign Contributions: Insights from the Asian American Case." *American Politics Research* 30(July):347–83.

Cliff, Andrew, and J. Keith Ord (1973). *Spatial Autocorrelation*. London: Pion.

Cliff, Andrew, and J. Keith Ord (1981). *Spatial Processes: Models and Applications*. London: Pion.

Crenson, Matthew A. 1978. "Social Networks and Political Processes in Urban Neighborhoods." *American Journal of Political Science* 22(3):289–95.

Darmofal, David. 2002. "Voter Participation Across Space and Time: Institutions, Contagion Effects, and Turnout, 1828–2000." Dissertation. University of Illinois at Urbana-Champaign.

Espiritu, Yen Le. 1992. *Asian American Panethnicity: Bridging Institutions and Identities*. Philadelphia: Temple University Press.

Eulau, Heinz. 1986. *Politics, Self, and Society*. Cambridge, MA: Harvard University Press.

Gimpel, James G. 1999. *Separate Destinations: Migration, Immigration, and the Politics of Places*. Ann Arbor, MI: University of Michigan Press.

Gierzynski, Anthony. 2000. *Money Rules: Financing Elections in America.* Boulder: CO: Westview Press.

Glazer Nathan, and Daniel Patrick Moynihan. 1972. *Beyond the Melting Pot: The Negroes, Puerto Ricans, Jews, Italians, and Irish of New York City*. Cambridge, MA: MIT Press.

Gray, Virginia. 1973. "Innovation in the States: A Diffusion Study." *American Political Science Review* 67(December): 1174–85.

Huckfeldt, R. Robert. 1979. "Political Participation and the Neighborhood Social Context." *American Journal of Political Science* 23(3):579–92.

Huckfeldt, Robert, and John Sprague. 1987. "Networks in Context: The Social Flow of Political Information." *American Political Science Review* 81(December):1197–216.

Huckfeldt, Robert, and John Sprague. 1992. "Political Parties and Electoral Mobilization: Political Structure, Social Structure, and the Party Canvass." *American Political Science Review* 86(March):70–86.

Huckfeldt, Robert, Eric Plutzer, and John Sprague. 1993. "Alternative Contexts of Political Behavior: Churches, Neighborhoods, and Individuals." *Journal of Politics* 55(2):365–81.

Key, V. O. 1949. *Southern Politics in State and Nation.* Knoxville: University of Tennesee Press.

Kirby, Andrew M., and Michael D. Ward. 1987. "The Spatial Analysis of Peace and War." *Comparative Political Studies* 20(October):293–313.

Kohfeld, Carol W., and John Sprague. 2002. "Race, space, and turnout." *Political Geography* 21(February):175–93.

Johnston, Ron J., and Charles J. Pattie. 1998. "Composition and Context: Region and Voting in Britain Revisited During Labour's 1990s' Revival." *Geoforum* 29(August):309–29.

Johnston, Ron J., and Charles J. Pattie. 2000. "People Who Talk Together Vote Together?: An Exploration of Contextual Effects in Great Britain." *Annals of the Association of American Geographers* 90(March):41–66.

Johnston, Ron J., Charles J. Pattie, Danny F. L. Dorling, Ian MacAllister, Helena Tunstall, David J. Rossiter. 2001. "Social Locations, Spatial Locations and Voting at the 1997 British General Election: Evaluating the Sources of Conservative Support." *Political Geography* 20(January):85–111.

Johnston, Ron J., Charles J. Pattie, Ian MacAllister, David J. Rossiter, Danny F. L. Dorling, and Helena Tunstall. 1997. "Spatial Variations in Voter Choice: Modelling Tactical Voting at the 1997 General Election in Great Britain." *Geographical and Environmental Modelling* 1(May):153–79.

Lien, Pei-te. 2001. *The Making of Asian America through Political Participation.* Philadelpha: Temple University Press.

Lew, Julie. 1987. "Asian Americans More Willingly Stuff Campaign Warchests than Ballot Boxes." *East/West*, 27 August.

Moran, P. A. P. 1948. "The Interpretation of Statistical Maps." *Journal of the Royal Statistical Society. Series B* 10:243–51.

Most, Benjamin A., and Harvey Starr. 1980. "Diffusion, Reinforcement, Geopolitics, and the Spread ofWar." *American Political Science Review* 74(December):932–46.

Most, Benjamin A., and Harvey Starr. 1982. "Case Study, Conceptualizations and Basic Logic in the Study of War." *American Journal of Political Science* 26(4):834–56.

Most, Benjamin A., and Harvey Starr. 1983. "Conceptualizing 'War': Consequences for Theory and Research." *Journal of Conflict Resolution* 27(March):137–59.

Most, Benjamin A., and Harvey Starr. 1984. "International Relations Theory, Foreign Policy Substitutability, and 'Nice' Laws." *World Politics* 36(April):383–406.

O'Loughlin, John. 1987. "Spatial Models of International Conflicts: Extending Current Theories of War Behavior." *Annals of the Association of American Geographers* 76(March):62–80.

O'Loughlin, John. 2002. "The Electoral Geography of Weimar Germany: Exploratory Spatial Data Analyses (ESDA) of Protestant Support for the Nazi Party." *Political Analysis* 10(Summer):217–43.

O'Loughlin, John, Colin Flint, and Luc Anselin. 1994. "The Geography of the Nazi Vote: Context, Confession and Class in the Reichstag Election of 1930." *Annals of the Association of American Geographers* 84(September):351–80.

Ord, J. K. 1975. "Estimation Methods for Models of Spacial Interaction." *Journal of the American Statistical Association* 70:120–26.

Putnam, Robert D. 1966. "Political Attitudes and the Local Community." *American Political Science Review* 60(September):640–54.

Putnam, Robert D. 2000. *Bowling Alone: The Collapse and Revival of American Community*. New York: Simon & Schuster Trade.

Reedy, George E. 1991. *From the Ward to the White House*. New York: Charles Scribner's Sons.

Rom, Mark Carl, Paul E. Peterson, and Kenneth F. Scheve, Jr. 1998. "Interstate Competition and Welfare Policy." *Publius: The Journal of Federalism* 28(Summer):17–37.

Rosenstone, Steven J., and John Mark Hansen. 1993. Mobilization, Participation, and Democracy in America. New York: Macmillan.

Saavedra, Luz Amparo. 2000. "A Model of Welfare Competition with Evidence from AFDC." *Journal of Urban Economics* 47(March):248–79.

Saito, Leland T. 1998. *Race and Politics: Asian Americans, Latinos, and Whites in a Los Angeles Suburb.* Urbana: University of Illinois Press.

Shin, Michael. 2001. "The Politicization of Place in Italy." *Political Geography* 20(March):331–53.

Shin, Michael, and John Agnew. 2002. "The geography of party replacement in Italy, 1987–1996." *Political Geography* 21(February):221–42.

Smirnov, Oleg, and Luc Anselin. 2001. "Fast Maximum Likelihood Estimation of Very Large Spatial Autoregressive Models: A Characteristic Polynomial Approach." *Computational Statistics and Data Analysis* 35(January):301–19.

Sorauf, Frank. 1988. *Money in American Elections*. Glenview: Scott Foresman and Company.

Starr, Harvey. 2001. "Using Geographic Information Systems to Revisit Enduring Rivalries: The Case of Israel." *Geopolitics* 5(Summer):37–56.

Starr, Harvey, and Benjamin A. Most. 1976. "The Substance and Study of Borders in International Relations Research." *International Studies Quarterly* 20(4):581–620.

Starr, Harvey, and Benjamin A. Most. 1978. "A Return Journey: Richardson, 'Frontiers' and Wars in the 1946–1965 Era." *Journal of Conflict Resolution* 22(September):441–67.

Starr, Harvey, and Benjamin A. Most. 1983. "Contagion and Border Effects on Contemporary African Conflict." *Comparative Political Studies* 16(February):92–117.

Sui, Danile, and Peter J. Hugill. 2002. "A GIS-based spatial analysis on neighborhood effects and voter turn-out: a case study in College Station, Texas." *Political Geography* 21(February):159–73.

Tam, Wendy K. 1995. "Asians—A Monolithic Voting Bloc?" *Political Behavior* 17(June):223–49.

Tir, Jaroslav, and Paul F. Diehl. 2002. "Geographic dimensions of enduring rivalries." *Political Geography* 21(February):263–86.

Uhlaner, Carole J., Bruce E. Cain, and D. Roderick Kiewiet. 1989. "Political Participation of Ethnic Minorities in the 1980s." *Political Behavior* 11(June):195–221.

Verba, Sidney, Kay Schlozman, Henry E. Brady. 1995. *Voice and Equality: Civic Voluntarism in American Politics*. Cambridge: Harvard University Press.

Walker, Jack L. 1969. "The Diffusion of Innovations among the American States." *American Political Science Review* 63(September):880–99.

Weatherford, M. Stephen. 1982. "Interpersonal Networks and Political Behavior." *American Journal of Political Science* 26(1):117–43.

Wong, Janelle S. 2000. "The Effects of Age and Political Exposure on the Development of Party Identification Among Asian American and Latino Immigrants in the United States." *Political Behavior* 22(December):341–71.

3

MEASURES OF SPATIAL SEGREGATION

*Sean F. Reardon**
*David O'Sullivan***

The measurement of residential segregation patterns and trends has been limited by a reliance on segregation measures that do not appropriately take into account the spatial patterning of population distributions. In this paper we define a general approach to measuring spatial segregation among multiple population groups. This general approach allows researchers to specify any theoretically based definition of spatial proximity desired in computing segregation measures. Based on this general approach, we develop a general spatial exposure/isolation index ($\tilde{P}^*$), and a set of general multigroup spatial evenness/clustering indices: a spatial information theory index ($\tilde{H}$), a spatial relative diversity index ($\tilde{R}$), and a spatial dissimilarity index ($\tilde{D}$). We review these and previously proposed spatial segregation indices against a set of eight desirable properties of spatial segregation indices. We conclude that the spatial exposure/isolation index $\tilde{P}^*$—which can be interpreted as a

This research was supported by a National Academy of Education Postdoctoral Fellowship to Dr. Reardon, and by the College of Education Research Initiation Grants Program, Pennsylvania State University, and a grant from the Consortium on Children, Youth, and Families, Pennsylvania State University. We thank Glenn Firebaugh, Chad Farrell, Deborah Gorman-Smith, and an anonymous reviewer for helpful comments on earlier drafts. Direct correspondence to Sean F. Reardon, School of Education, 485 Lasuen Mall, Stanford University, Stanford, CA 94305; (650) 736-8517; sreardon@stanford.edu

*Stanford University
**Pennsylvania State University

measure of the average composition of individuals' local spatial environments—and the spatial information theory index $\tilde{H}$—which can be interpreted as a measure of the variation in the diversity of the local spatial environments of each individual—are the most conceptually and mathematically satisfactory of the proposed spatial indices.

1. INTRODUCTION—SEGREGATION AND SPACE

Reliable and meaningful measurement of residential segregation is essential to the study of the causes, patterns, and consequences of racial and socioeconomic segregation. Nonetheless, prior work on residential segregation has been limited by a reliance on methodological tools that do not fully capture the spatial distributions of race and poverty. Scholars have repeatedly pointed out that the most commonly used measures of segregation—such as the dissimilarity index (D), the exposure index (P^*), the variance ratio index (V), and the entropy-based information theory index (H)—are "aspatial," meaning that they do not adequately account for the spatial relationships among residential locations (Grannis 2002; Massey and Denton 1988; Morrill 1991; Reardon and Firebaugh 2002b; Wong 1993; Wong 2002).

In this paper, we take up the challenge of developing measures of spatial segregation that satisfactorily address the problems identified with existing measures of segregation. We begin by arguing for a set of criteria that would be met by a satisfactory spatial segregation measure. We then present a new and general approach to measuring spatial segregation that addresses the key limitations of prior spatial measures. This approach allows researchers to specify theoretically appropriate definitions of how spatial features constrain or enhance the possibility of social interaction. Finally, we review previously proposed measures of spatial segregation and evaluate both these and our new measures against our criteria.

1.1. *Methodological Issues in the Measurement of Spatial Segregation*

Segregation can be thought of as the extent to which individuals of different groups occupy or experience different social environments. A measure of segregation, then, requires that we (1) define the social

environment of each individual, and (2) quantify the extent to which these social environments differ across individuals. Traditional measures of segregation are aspatial, in that they differ from one another only on the second of these criteria, because they implicitly define the social environment as equivalent to some organizational or spatial unit (school, census tract), without regard for the patterning of these units in social space. Much prior discussion of segregation indices, then, has focused only on the matter of the most appropriate mathematical formulation for quantifying differences across social environments (James and Taeuber 1985; Reardon and Firebaugh 2002a; White 1986; Zoloth 1976).

Aspatial segregation measures have been repeatedly criticized in the residential segregation context for their failure to account for the spatial patterning of census tracts (Grannis 2002; Massey and Denton 1988; Morrill 1991; Wong 1993; Wong 2002). In particular, two flaws of aspatial measures are identified: the *checkerboard problem* (Morrill 1991; White 1983) and the *modifiable areal unit problem* (Openshaw and Taylor 1979; Wong 1997). Each of these can be seen as critiques of the definition of the social environment implicit in the traditional segregation measures.

The checkerboard problem stems from the fact that aspatial segregation measures ignore the spatial proximity of neighborhoods and focus instead only on the racial composition of neighborhoods. To visualize the problem, imagine a checkerboard where each square represents an exclusively black or exclusively white neighborhood. If all the black squares were moved to one side of the board, and all white squares to the other, we would expect a measure of segregation to register this change as an increase in segregation, since not only would each neighborhood be racially homogeneous, but most neighborhoods would now be surrounded by similarly homogeneous neighborhoods. Aspatial measures of segregation, however, do not distinguish between the first and second patterns, since in each case the racial compositions of individual neighborhoods are the same (White 1983).

The modifiable areal unit problem (MAUP) arises in residential segregation measurement because residential population data are typically collected, aggregated, and reported for spatial units (such as census tracts) that have no necessary correspondence with meaningful social/spatial divisions. This data collection scheme implicitly assumes

that individuals living near one another (perhaps even across the street from one another) but in separate spatial units are more distant from one another than are two individuals living relatively far from one another but within the same spatial unit. As a result—unless spatial subarea boundaries correspond to meaningful social boundaries—all measures of spatial and aspatial segregation that rely on population counts aggregated within subareas are sensitive to the definitions of the boundaries of these spatial subareas.[1]

Essentially then, the definition of spatial segregation measures requires a redefinition of the social environment implicit in the traditional segregation measures. In fact, the checkerboard problem and the MAUP are both artifacts of a reliance on subarea (e.g., tract) boundaries in the computation of segregation measurement. In principle, a segregation measure that used information on the exact locations of individuals and their proximities to one another in residential space could eliminate the checkerboard problem and MAUP issues entirely.

1.2. *The Dimensions of Spatial Segregation*

Another confusion in the segregation literature also results from relying on census tract boundaries in computing segregation measures. In an often-cited article, Massey and Denton (1988) describe five conceptually distinct "dimensions" of residential segregation: (1) *evenness*, (2) *exposure*, (3) *clustering*, (4) *centralization*, and (5) *concentration*. In their formulation, evenness and exposure are aspatial dimensions

[1]In fact, the MAUP is constituted by two interrelated effects: an *aggregation* (or scale) effect, and a *zoning* effect (Wong 1997). The aggregation effect leads to differences in statistical measures resulting purely from dealing with data that are "less detailed." The difference between a statistic derived from tract data and the same statistic derived for block group data, for example, is an aggregation effect. For segregation measures, greater aggregation usually results in lower measured levels of segregation. The zoning effect refers to the fact that any measure derived from aggregated population data depends on the choice of aggregation zones (i.e., the "modifiable areas"), even if the scale and number of the zones remains fixed. With regard to the census tracts often used in studies of segregation, the effect is *initially* to exaggerate segregation (because tracts are designed to be relatively homogeneous internally). However, over time, if the same zones are retained, measured levels of segregation fall (Massey and Denton 1988).

(allowing that they are nonetheless implicitly spatial because they depend on census tract boundaries), while clustering, concentration, and centralization are explicitly spatial dimensions of segregation, and they require information on the location and size of census tracts to compute.

The distinction between aspatial evenness and spatial clustering, however, is an artifact of the reliance on spatial subareas (e.g., census tracts) at some chosen geographical scale of aggregation. Evenness, in Massey and Denton's formulation, refers to the degree to which members of different groups are over- and underrepresented in different subareas relative to their overall proportions in the population. Clustering refers to the proximity of subareas with similar group proportions to one another. However, evenness at one level of aggregation (say, census tracts), is clearly strongly related to clustering at a lower level of aggregation (say, block groups), since tracts where a minority group is overrepresented will tend to be clusters of block groups where the minority population is overrepresented. Unless subarea boundaries correspond to meaningful social boundaries, the distinction between evenness and clustering is thus arbitrary.

In principle, if we derived a segregation measure from information about the exact locations and spatial environments of individuals and their proximities to one another in residential space, there would be no conceptual distinction at all between evenness and clustering. Any movement of an individual that increased unevenness (by moving a person from a location where his or her group is underrepresented to one where it is overrepresented) would also increase clustering, because it would result in members of the same groups being nearer to one another.

As a result of this insight, we suggest an alternative to the Massey and Denton (1988) dimensions of residential segregation. We argue that there are two primary conceptual dimensions to spatial residential segregation: (1) spatial exposure (or spatial isolation) and (2) spatial evenness (or spatial clustering). *Spatial exposure* refers to the extent that members of one group encounter members of another group (or their own group, in the case of spatial isolation) in their local spatial environments. *Spatial evenness*, or clustering, refers to the extent to which groups are similarly distributed in residential space. Spatial exposure, like aspatial exposure, is a measure of the typical environment experienced by individuals; it depends in part on the

overall racial composition of the population in the region under investigation. Spatial evenness, in contrast, is independent of the population composition.

To see that spatial exposure and evenness are conceptually distinct, consider the four patterns of individual residential locations (not subarea proportions) shown in Figure 1. In the upper half of the diagram are two patterns where black and white households are evenly distributed throughout space. Both of these patterns have low levels of spatial clustering (or high levels of spatial evenness). In the pattern on the upper right, however, there are more black households in the local environment of each white household (and vice versa) than in the pattern on the upper left; this means that the white-black exposure is higher on the right, and the white isolation is higher on the left. In the bottom half of the figure, both patterns

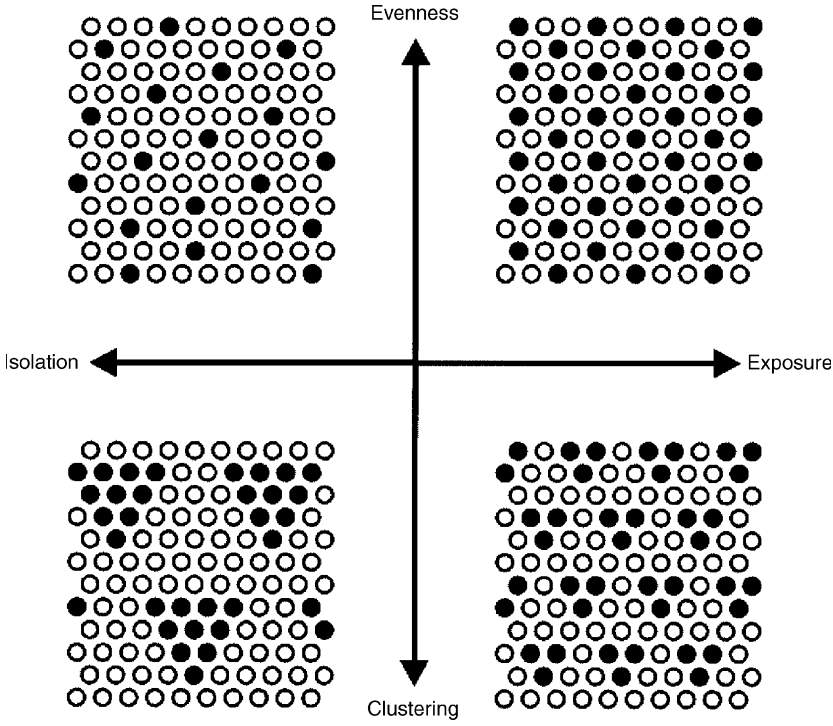


FIGURE 1. Dimensions of spatial segregation.

show greater clustering—but roughly the same levels of exposure—than the corresponding patterns above.

In this framework, Massey and Denton's evenness and clustering dimensions are collapsed into a single dimension. Their exposure dimension remains intact, but it is now conceptualized as explicitly spatial. Their centralization and concentration dimensions can be seen as specific subcategories of spatial unevenness. In some cases, centralization and concentration may be of sufficient theoretical interest to be considered distinct subdimensions; however, we do not consider them further in this paper.

1.3. *Existing Measures of Spatial Segregation*

Many spatial measures have been developed to address the methodological shortcomings identified above (for example, see Frank 2003; Grannis 2002; Jakubs 1981; Massey and Denton 1988; Morgan 1982, 1983a, 1983b; Morrill 1991; Waldorf 1993; White 1983, 1986; Wong 1993; Wong 1998, 1999, 2002), although it is not clear that any of the proposed measures fully solve the problem of measuring spatial segregation. Many of the measures have been developed in a relatively *ad hoc* manner, and none have been evaluated against a conceptually meaningful set of criteria, as has been done for the traditional aspatial measures (James and Taeuber 1985; Reardon and Firebaugh 2002a), so it is unclear whether they reliably produce results consistent with theoretically useful definitions of segregation.

At present, few of the proposed spatial segregation measures have been used in published empirical segregation research. These measures have been ignored in part because they typically are more difficult to compute than the aspatial measures. At present, there is also still little publicly available software to compute spatial segregation measures—Wong's extensions to the Arc/INFO (Wong and Chong 1998) and ArcView GIS software (Wong 2003), and Apparicio's extension to MapInfo GIS (Apparicio 2000) are the only examples that we are aware of. This limitation, however, is likely to become less relevant with the increased availability and ease of use of geographical information system (GIS) software (Longley et al. 2001). However, in the absence of a clear evaluation of the proposed

measures, the development of GIS software is likely to lead to a situation where researchers use a wide variety of different measures, resulting in findings that cannot be easily compared across studies.

2. MEASURES OF SPATIAL SEGREGATION

2.1. *Notation*

Throughout this paper, we use the following notation: consider a spatial region R populated by M mutually exclusive population subgroups (e.g., racial groups), indexed by m. Let p, q, and s index points within the region R; and let r index subareas of the region R (e.g., census tracts). Let τ denote population density and π denote population proportion. In addition, we use a super-positioned tilde ($\sim$) to indicate that a parameter describes the spatial environment of a given point, rather than the point itself. Thus we have

$\tau_p =$ population density at point p,
$\tau_{pm} =$ population density of group m at point p (note that $\sum_m \tau_{pm} = \tau_p$),
$T =$ total population in R (note that $\int_p \tau_p dp = T$),
$\tilde{\tau}_{pm} =$ population density of group m in the local environment of point p,
$\pi_m =$ proportion in group m of total population (e.g., proportion black),
$\pi_{pm} =$ proportion in group m at point p (defined as $\pi_{pm} = \tau_{pm}/\tau_p$),
$\tilde{\pi}_{pm} =$ proportion in group m in the local environment of point p.

Note that the population densities τ_p and τ_{pm} are defined by the population counts per unit area at location p. In practice, these must be estimated from census tract (or other subarea) population counts, most simply by dividing the population count of a tract by its area and assigning the population density this value at each point in the tract. Other density estimation procedures might be used as well, including pycnophylactic ("mass preserving") smoothing and dasymetric mapping (for example, see Dent 1999; Mennis 2003; Tobler 1979). We leave discussion of these estimation methods and of the

sensitivity of segregation measurement to different choices of density estimators, however, for another paper.

2.2. Spatial Proximity and the Local Environment

The measurement of spatial segregation requires that we define the spatial proximity between all pairs of points in a region R. Let $\phi(p, q)$ be a non-negative function that defines the spatial proximity of locations q and p, such that $\phi(p, q) = \phi(q, p)$ and $\phi(q, q) = \phi(p, p)$ for all p, $q \in R$, and with larger values of $\phi(p, q)$ indicating greater proximity. Let $\Phi_p = \int_{q \in R} \phi(p, q) dq$, noting that we do not require $\Phi_p = \Phi_q$ for all p, $q \in R$. We define the population density of the local environment of a point p as the weighted average of the population densities of all other points in R, where points are weighted by their proximity to p:[2]

$$\tilde{\tau}_p = \frac{1}{\Phi_p} \int_{q \in R} \tau_q \phi(p, q) dq. \tag{1}$$

We define $\tilde{\tau}_{pm}$ similarly, substituting τ_{qm} for τ_q in equation (1). Now $\tilde{\tau}_p$ and $\tilde{\tau}_{pm}$ are, respectively, the spatially weighted average population density and the group m population density at point p. For each m, $\tilde{\tau}_{pm}$ describes a spatially smoothed population surface, where the value of $\tilde{\tau}_{pm}$ at location p indicates the group m population density at point p. We define

$$\tilde{\pi}_{pm} = \frac{\tilde{\tau}_{pm}}{\tilde{\tau}_p}. \tag{2}$$

It is trivial to show that, for each location p,

$$\sum_{m=1}^{M} \tilde{\pi}_{pm} = 1. \tag{3}$$

We can think of the $\tilde{\pi}_{pm}$'s as indicating the population composition that a person living at point p would experience in his or her local

[2]Throughout this paper, we use a single integral to denote the summation over all points in a region.

environment, where the local environment is defined by the proximity function ϕ.[3]

The function $\phi(p, q)$ may take on a variety of possible forms, each implying a different definition of the local environment. For example, $\phi(p, q)$ might be a function that declines as the Euclidean distance from p to q increases, which means that the spatial environment of point p is influenced more by the population nearby than by those more distant. The spatial proximity function $\phi(p, q)$ might also incorporate information about physical barriers (such as rivers, railroads, or highways) and patterns of social interaction between locations p and q. Ideally, a spatial proximity function should capture theoretically meaningful patterns of social interaction.

One special case of the spatial proximity function is worth noting. Measures of aspatial segregation implicitly define the local environment of each individual as equivalent to the organizational unit (e.g., census tract, school) containing the individual. Reardon and Firebaugh (2002b) note that this can be seen as a special case of the above definition of the local environment, where spatial proximity is defined such that $\phi(p, q)$ equals some constant c if p and q are both in tract r and $\phi(p, q) = 0$ if p and q are in separate tracts. In this case, equations (1) and (2) yield $\tilde{\pi}_{pm} = \pi_{rm}$ for all m and all $p \in r$, indicating that the group composition of the local environment at each point in r is identical to the group proportions in tract r as a whole, regardless of how population groups are distributed within the tract, or how tracts are arranged in space (Reardon and Firebaugh 2002b). This insight—that the aspatial segregation indices can be seen as spatial indices that depend on a very specific notion of spatial proximity—will prove

[3]Note that we can rewrite equation (2) as

$$\tilde{\pi}_{pm} = \int_{q \in R} \frac{\tau_q \phi(p, q)}{\int_{s \in R} \tau_s \phi(p, s) ds} \pi_{qm} dq.$$

From this, we can see that $\tilde{\pi}_{pm}$ is a density- and proximity-weighted average of the π_{qm}'s for all q in R. In the aspatial case, population density and group proportions are assumed constant within tracts and the spatial proximity of each pair of distinct tracts is zero, so the above yields $\tilde{\pi}_{pm} = \pi_{rm}$, where p is in tract r (see Reardon and Firebaugh 2002b).

useful in our approach to developing spatial segregation measures in this paper.

2.3. *Criteria for Evaluating Spatial Segregation Measures*

Previous methodological work, drawing on the inequality measurement literature (for example, see Schwartz and Winship 1980), has defined a set of criteria for the evaluation of aspatial evenness measures of segregation (James and Taeuber 1985; Reardon and Firebaugh 2002a). Compliance with these criteria implies that a measure will register an appropriate change in segregation levels under specified conditions; conversely, noncompliance implies that it is possible for a measure to respond to changes in population distributions in ways that are inconsistent with conceptually appropriate definitions of segregation. Since the criteria were developed with aspatial measures in mind, Reardon and Firebaugh (2002b) suggest that they may need to be modified in order to apply them to spatial segregation measures. Here we describe a general set of criteria for segregation measures that apply to spatial evenness measures. A subset of these reduce to the Reardon and Firebaugh (2002a) criteria in the special case where a measure is aspatial.[4] In addition to these criteria, we suggest several additional desirable properties that pertain specifically to spatial segregation indices. Five of the criteria—scale interpretability, arbitrary boundary independence, location equivalence, population density invariance, and additive spatial decomposability—apply to measures of spatial exposure as well.

1. *Scale interpretability*: A spatial segregation index should be equal to zero if the group proportions are the same in the local environment of each individual. A segregation index should reach its maximum value (typically normalized to equal 1) if the local environment of each individual is monoracial. An alternate way

[4]In addition to possessing the properties described here, a spatial segregation index should (1) be a continuous function of both the total and group population densities at each point and of the spatial proximity function between all points; (2) allow the computation of segregation among any number of population groups; and (3) correspond to a meaningful (aspatial) segregation measure in the aspatial special case.

of stating this is that a segregation index should reach its maximum value only if the proximity of any two members of different groups is zero. A segregation index should take on a negative value if the population is "hyper-integrated"—if individuals, on average, experience greater diversity in their local environments than the diversity of the population as a whole.[5]

2. *Arbitrary boundary independence*: A spatial segregation measure should be independent of the definitions of census tract (or other subarea) boundaries. In principle, a spatial segregation measure should be computed from information about the exact locations and spatial proximities of individuals in residential space (although in practice, it may be necessary to use tract or other subarea data to estimate local population densities). This will ensure that a measure will not be susceptible to MAUP issues.
3. *Location equivalence*: If the local environments of two points p and q have the same population composition (i.e., if $\tilde{\pi}_{pm} = \tilde{\pi}_{qm}$ for all m) and the same proximity to all other points (i.e., if $\phi(p, s) = \phi(q, s)$ for all $s \neq p, q$),[6] then segregation is unchanged if we treat the two points as one point with a population density equal to the sum of the two original points. While this criterion may seem to have little concrete application, it is a spatial generalization of the aspatial organizational equivalence criterion, which states that if two organizational units (schools, tracts) have the

[5]In the spatial case, unlike the aspatial case, it may be possible—and meaningful—for the average individual to experience greater diversity in his or her local environment than the diversity of the population as a whole. Consider the residential pattern shown in the upper-right corner of Figure 1. If we define the local environment of each household as consisting of itself plus the six households immediately adjacent to it, then each white household will be in a local environment that is 3/7 black, despite the fact that the overall population is only 1/3 black. Likewise, each black household inhabits a local environment that is 6/7 white, despite the fact that the total population is only 2/3 white. In such a case, the population may be said to be hyper-integrated. A segregation index should be negative in this case, indicating that the population is more integrated than expected given the population composition.

[6]In general, this can occur only if the two points have the same population composition ($\pi_{pm} = \pi_{qm}$ for all m) and either (1) the points have the same population density ($\tau_p = \tau_q$), or (2) the points have the same population composition as their local environments ($\tilde{\pi}_{pm} = \tilde{\pi}_{qm} = \pi_{pm} = \pi_{qm}$ for all m).

same composition and are combined into a single unit, segregation is unchanged (James and Taeuber 1985).[7]

4. *Population density invariance*: If the population density τ_{pm} of each group m at each point p is multiplied by a constant factor c, segregation is unchanged. This is a spatial generalization of the aspatial size invariance criterion (James and Taeuber 1985).
5. *Composition invariance*: In general, a measure of spatial evenness should be independent of the population composition and should depend only on the distribution of groups in space. More formally, the composition invariance criterion states that if the proportions of groups change in the population while the relative distribution of groups in space remains the same, then segregation is unchanged. The key to operationalizing this seemingly intuitive concept is to define what it means for the relative distribution of groups to remain the same. As Coleman, Hoffer, and Kilgore (1982) point out, determining whether an index is composition invariant always "depends on a specific definition of what it means to say that the distribution [of individuals among organizational units] is 'kept the same' while the [group] proportion changes" (p. 177).[8]

   The literature on segregation measurement provides several different definitions of composition invariance. James and Taeuber (1985) say that the distribution of individuals in space is the same if the population density of group m at each point is multiplied by a constant c and the population density of all other groups at each point is unchanged.[9] Coleman, Hoffer, and Kilgore (1982), however,

[7]Note that this criterion implies that if $\phi(p, q)$ is defined so that $\phi(p, q) = c$ for all points p and q in tract r and $\phi(p, s) = \phi(q, s)$ for all points p and q in tract r and all points s not in tract r, then a segregation measure that satisfies the locational equivalence criterion will be unchanged if we consider the entire population of the tract to be located at a single point within the tract (e.g., the centroid).

[8]We would like to thank a thoughtful anonymous reviewer who suggested the importance of clarifying the meaning of the composition invariance criterion.

[9]Another proposed definition of composition invariance is given by Gorard and Taylor (2002), who argue that the distribution of individuals in space is the same if the *proportion* of group m in each location is multiplied by a constant (while the total enrollment in each school remains the same). This definition, however, is not internally consistent, since it is not symmetric—multiplying group m's proportion by c implies multiplying each other group's composition in each location p by $d_p = (1 - c\pi_{pm})/(1 - \pi_{pm})$. Unless π_{pm} is constant for all p, then d_p varies across p, so composition invariance under this definition is dependent on which group is considered.

argue that this is an arbitrary definition of composition invariance; they imply that any segregation measure can be considered composition invariant under an appropriate definition of segregation. For example, if we define segregation as the ratio of actual to potential pair relations between members of different groups (as Coleman and colleagues do), then any change in the population composition that leaves this ratio unchanged should leave a composition invariant segregation index unchanged. The variance ratio index (V) can be defined as simply (one minus) the ratio of actual to potential pair relations between members of different groups, which means that V is composition invariant under this definition of segregation (Coleman, Hoffer, and Kilgore 1982).

This definition of composition invariance is somewhat tautological, of course, since it suggests that if we believe that an index appropriately measures what we take to be segregation (specifically, spatial evenness/clustering) in some meaningful sense, then that measure will be composition invariant by an appropriate definition of the criterion—a change in the population composition that leaves what is measured by a particular index unchanged will necessarily leave segregation unchanged, as measured by that index. As a result, we take the position that the traditional composition invariance criterion espoused by James and Taeuber (1985) is less important than is ensuring that a measure of segregation has a sound conceptual basis. If a segregation index measures exactly that quantity that we believe defines spatial segregation, then the index will be composition invariant by definition. That said, we nonetheless evaluate the measures discussed in the paper against the traditional (James and Taeuber) composition invariance criterion, in order to preserve continuity with prior research.

6. *Transfers and exchanges*: A key criterion for a segregation measure is a definition of how segregation should change in response to the movement of individuals in social space. Transfers and exchanges, as we define them here, are specific types of such movement. We suggest here spatial extensions of the Reardon and Firebaugh (2002a) multigroup transfer and exchange criteria; in addition, we suggest an additional exchange criterion.

   - *Transfers*: If an individual of group m is transferred from point p to q, and if the proportion of group m in the local

environments of all points closer to p than q is greater than the proportion of group m in the local environments of all points closer to q than p, segregation is reduced. In the aspatial case, this reduces to the usual transfer criterion (James and Taeuber 1985; Reardon and Firebaugh 2002a).

- *Exchanges (Type 1)*: If an individual of group m from point p is exchanged with an individual of group n from point q, and if the proportion of group m in the local environments of all points closer to p than q is greater than the proportion of group m in the local environments of all points closer to q than p, and if the proportion of group n in the local environments of all points closer to q than p is greater than the proportion of group n in the local environments of all points closer to p than q, segregation is reduced. In simpler terms, if an exchange moves two individuals of different groups to locations where they are less likely to encounter members of their own group (and hence, more likely to encounter members of other groups), then segregation should be reduced. In the aspatial case, this reduces to the usual exchange criterion (James and Taeuber 1985; Reardon and Firebaugh 2002a).
- *Exchanges (Type 2)*: If an individual of group m from point p is exchanged with an individual of group n from point q, and if the proportion of group m is greater than the proportion of group n in the local environments of all points closer to p than q, and if the proportion of group n is greater than the proportion of group m in the local environments of all points closer to q than p, segregation is reduced. Although this formulation of the exchange criterion does not reduce to the familiar exchange criterion, it has a compelling logic: if two individuals of groups m and n change places in a way that makes the proportions of groups m and n more similar in the local environments of at least some places (and leaves them unchanged in all others), while leaving the proportions of all other groups unchanged everywhere, then segregation should be reduced.[10]

[10]Note that in the two-group case, the type 2 exchange criterion is a special case of the type 1 criterion; in the multigroup case, however, they are distinct criteria—the conditions of a type 2 exchange can be met without meeting those of a type 1 exchange, and vice versa.

7. *Additive spatial decomposability*: If X spatial subareas are aggregated into Y larger spatial areas, then a segregation measure should be decomposable into a sum of within- and between-area components.
8. *Additive grouping decomposability*: If M groups are clustered in N supergroups, then a segregation measure should be decomposable into a sum of independent within- and between-supergroup components.

3. A GENERAL APPROACH TO MEASURING SPATIAL SEGREGATION

We now turn to developing and evaluating new and proposed measures of spatial segregation. We begin by describing a new approach to measuring spatial segregation and use this approach to develop several measures of spatial exposure and spatial evenness. Conceptually, we measure spatial exposure and spatial evenness as follows. We begin by computing the spatially weighted group composition of the local environment of each location (or person) in the region of interest. Typically, we will weight this measure so that locations near another location contribute more to its local spatial environment than do more distant locations (a "distance-decay" effect).

To measure spatial exposure, we compute the average composition of the local environments of members of each group. To measure spatial evenness, we examine how much variation there is among the racial compositions of the local environments of everyone living in the region of interest. If each person's spatial environment is relatively similar in racial composition, there is little spatial unevenness; conversely, if there is considerable variation across persons in the racial composition of their spatial environments, there is high spatial segregation (unevenness).

Our approach in this paper provides a general framework for measuring spatial segregation among multiple population groups. This approach encompasses, as special cases, traditional aspatial measures, both two-group and multigroup. Our approach here assumes complete data about the residential locations of individuals (though these data may be estimated from tract or other aggregated data, of course). Our approach does not, however, assume any specific functional form defining spatial proximities between locations. In fact, we deliberately do not specify a functional form for the spatial proximity function, as we wish to call attention to the fact that many meaningful definitions are possible. The flexibility of our approach allows researchers to

specify a definition of local social environments derived from theoretical considerations of patterns of social interaction.

3.1. *A General Spatial Exposure Segregation Index*

Equation (2) above defines a surface $\tilde{\pi}_{pm}$, which gives, at each point p in R, the proportion of the population in the local neighborhood who are members of group m. This can be interpreted as the exposure to group m for a person residing at location p. These $\tilde{\pi}_{pm}$ surfaces are the basis of the class of spatial segregation measures we develop here.

We define the spatial exposure of group m to group n as the average percentage of group n in the local environments of each member of group m.

$$ {}_m\tilde{P}^*_n = \int_{q \in R} \frac{\tau_{qm}}{T_m} \tilde{\pi}_{qn} dq. \tag{4} $$

We likewise define the spatial isolation of group m as simply the spatial exposure of group m to itself:

$$ {}_m\tilde{P}^*_m = \int_{q \in R} \frac{\tau_{qm}}{T_m} \tilde{\pi}_{qm} dq. \tag{5} $$

In the aspatial case, equations (4) and (5) are equivalent to the usual exposure and isolation indices (Bell 1954; Lieberson and Carter 1982a, 1982b). Although formulated slightly differently, Morgan's (1983b) distance-decay interaction index, ${}_mPC_n$, can be seen as a special case of equation (4), where the spatial proximity function used to compute $\tilde{\pi}_{qm}$ is defined based on estimated contact rates between each tract and its surrounding areas.[11]

[11]Schnell and Yoav (2001) develop sociospatial isolation measures using a related approach. Their approach differs from ours, however, in that they construct $\tilde{\pi}_{pm}$ as a sociospatially weighted average of the π_{qm}'s (see footnote 3) without weighting for population density. In addition, they average population compositions in the logistic scale, a technique that makes their measure difficult to interpret. Finally, they construct sociospatial isolation measures for individuals, rather than populations, though it would be a simple matter to average their individual isolation measures over all individuals to construct population-average exposure measures as we do in equations (4) and (5).

3.2. *A General Approach to Measuring Spatial Evenness*

Now recall that we define the evenness dimension of spatial segregation as the extent to which individuals of different groups occupy or experience different social environments. Given the population density distribution and the $\tilde{\pi}_{pm}$ exposure surfaces, we know the population density at each location and the group proportions in the local environment of each location; these are all we will need to construct a set of spatial segregation measures.

Knowing the population density (τ_p) at each location and the group proportions (the $\tilde{\pi}_{pm}$'s) in the local environment of each location, we can construct a variety of potentially useful multigroup spatial segregation measures. By substituting the $\tilde{\pi}_{pm}$'s for the π_{pm}'s in Reardon and Firebaugh (2002a, table 2), we can derive spatial generalizations of all their aspatial multigroup segregation measures (D, G, H, C, P, R). Because the aspatial measures are special cases of the spatial measures, and because the aspatial criteria described by Reardon and Firebaugh (2002a) are special cases of the spatial criteria described above, spatial measures derived this way cannot, in general, meet any of the spatial criteria that are not met by their aspatial analogs. We focus here, therefore, on deriving and describing a spatial version of the entropy-based information theory segregation index (H), since the aspatial multigroup H has been shown to be preferable to other aspatial measures on the basis of these criteria (Reardon and Firebaugh 2002a).

In addition, we describe and evaluate two additional measures—spatial versions of the dissimilarity index (D) and the relative diversity index (R). We evaluate the spatial dissimilarity index because the aspatial dissimilarity index has been used so commonly in segregation research. We evaluate the spatial relative diversity index because the aspatial R meets most criteria for an aspatial index (Reardon and Firebaugh 2002a), suggesting that it may make a useful spatial index as well.

3.3. *The Spatial Information Theory Segregation Index*

Following Theil (1972), we compute the spatially weighted entropy—a measure of population diversity (see Pielou 1977; White 1986)—at each point p as

$$\tilde{E}_p = -\sum_{m=1}^{M} (\tilde{\pi}_{pm}) \log_M (\tilde{\pi}_{pm}). \tag{6}$$

This is the entropy of the local environment of p. It is analogous to the entropy of an individual tract, E_r, that is used in the computation of the aspatial segregation index H (and in fact, if we define the local environment of p to be tract r, then $\tilde{E}_p = E_r$), except that $\tilde{E}_p$ incorporates spatially weighted information on the racial composition at all points in R, not only the racial composition of the tract where p is located.

Now we define the spatial information theory index, denoted $\tilde{H}$:

$$\tilde{H} = 1 - \frac{1}{TE} \int_{p \in R} \tau_p \tilde{E}_p dp, \tag{7}$$

where E is the overall regional entropy of the total population given by

$$E = -\sum_{m=1}^{M} (\pi_m) \log_M (\pi_m). \tag{8}$$

The spatial information theory index, $\tilde{H}$, is a measure of how much less diverse individuals' local environments are, on average, than is the total population of region R. It will be equal to 1—indicating maximum segregation—only when each individual's local environment is monoracial. If each individual's local environment has the same racial composition as the total population, then $\tilde{E}_p = E$ for all p, and $\tilde{H}$ will be zero—indicating complete integration.

3.4. *Additional Spatial Segregation Indices*

We define a spatial relative diversity index $\tilde{R}$ as

$$\tilde{R} = 1 - \int_{p \in R} \frac{\tau_p \tilde{I}_p}{TI} dp, \tag{9}$$

where I is the interaction index, a measure of population diversity (Lieberson 1969; White 1986):

$$I = \sum_{m=1}^{M} (\pi_m)(1 - \pi_m), \tag{10}$$

and where $\tilde{I}_p$ is the spatially-weighted interaction index at point p:

$$\tilde{I}_p = \sum_{m=1}^{M} (\tilde{\pi}_{pm})(1 - \tilde{\pi}_{pm}). \tag{11}$$

Like $\tilde{H}$, $\tilde{R}$ is a measure of how much less diverse individuals' local environments are, on average, than is the total population of region R.[12]

Finally, we define a spatial dissimilarity index as

$$\tilde{D} = \sum_{m=1}^{M} \int_{p \in R} \frac{\tau_p}{2TI} |\tilde{\pi}_{pm} - \pi_m| dp. \tag{12}$$

Unlike its aspatial analog, the spatial dissimilarity index cannot be interpreted as the proportion of the population who would have to

[12]Unlike in the aspatial case, $\tilde{R}$ is not easily related to the $\tilde{P}^*$ spatial exposure indices. In the aspatial case, in a two-group population, we have (Reardon and Firebaugh 2002a):

$$1 - R = \frac{{}_mP_n^*}{\pi_n} = \frac{{}_nP_m^*}{\pi_m}.$$

In general, the spatial version of these equalities does not hold, since the spatial versions of the quantities above are given by

$$1 - \tilde{R} = \int_{p \in R} \frac{\tau_p \tilde{\pi}_{pm} \tilde{\pi}_{pn}}{T \pi_m \pi_n} dp$$

$$\frac{{}_m\tilde{P}_n^*}{\pi_n} = \int_{p \in R} \frac{\tau_p \pi_{pm} \tilde{\pi}_{pn}}{T \pi_m \pi_n} dp$$

$$\frac{{}_n\tilde{P}_m^*}{\pi_m} = \int_{p \in R} \frac{\tau_p \tilde{\pi}_{pm} \pi_{pn}}{T \pi_m \pi_n} dp.$$

These are equal only if $\tilde{\pi}_{pm} = \pi_{pm}$ holds for all p and m (see footnote 3).

move to achieve complete integration. However, it can be interpreted as a measure of how different the composition of individuals' local environments are, on average, from the composition of the population as a whole.

3.5. *Prior Proposed Measures of Spatial Segregation*

As we noted above, we are not the first to propose measures of spatial segregation. Table 1 summarizes proposed spatial segregation measures. Those that rely explicitly on tract boundaries and contiguity patterns are noted in column 3; these measures will each be necessarily susceptible to MAUP issues. The other indices are computed, in principle, from more general functions of spatial or social distance, although tract boundaries and contiguity are generally used to approximate spatial distance.

Among the proposed measures of spatial evenness, most are modifications of the aspatial dissimilarity index D (Jakubs 1981; Morgan 1982, 1983a; Morrill 1991; O'Sullivan and Wong 2004; Waldorf 1993; Wong 1993; Wong 1998); these generally incorporate some spatial contiguity weight into the computation of D, or characterize the distance between tracts in terms of "relocation efforts." As each of these measures is a generalization of D, they will necessarily fail to meet any of the criteria that the aspatial D fails to meet (Reardon and Firebaugh 2002a). In particular, they fail to meet the exchange criterion and the decomposition criteria. Moreover, most of these are based explicitly on tract boundaries, and so are susceptible to MAUP issues. Because of these shortcomings, we do not consider these measures further here.

Morgan (1983b:215) defines a symmetric spatial segregation index IC_2 that is a spatial analog to the variance ratio index or the standardized exposure index. However, IC_2 is well-defined only for spatial proximity functions where $\tilde{\pi}_{pm} = \pi_{pm}$ holds for all p and m, since otherwise the standardized versions of the exposure indices are not, in general, equal (see footnote 9). When IC_2 is well-defined, it can be seen as a special case of our relative diversity index $\tilde{R}$; thus we do not consider IC_2 further here.

Among the other proposed measures of spatial evenness, the remainder (save our new measures) do not correspond to any known

TABLE 1
Proposed Spatial Segregation Measures

Measure	Aspatial Analog	Tract-based?	Original Citation	Brief Description
Spatial Evenness Measures				
D(adj)	*D*	Y	(Morrill 1991)	*D* adjusted for tract contiguity
D(w)	*D*	Y	(Wong 1993)	*D(adj)* adjusted for contiguous tract boundary lengths
D(s)	*D*	Y	(Wong 1993)	*D(w)* adjusted for tract perimeter/area ratio
SD(m)	*D*	Y	(Wong 1998)	Multigroup *D* computed using composite population counts
DBI	*D*	Y	(Jakubs 1981; Morgan 1982)	*D* adjusted for relocation efforts needed to achieve integration
MDBI	*D*	Y	(Morgan 1983a)	*DBI* with alternate definition of complete segregation
RDI	*D*	N	(Waldorf 1993)	*D* adjusted for relocation efforts
S	*D*	N	(O'Sullivan and Wong 2004)	*D* based on population density surfaces
IC_2	V	N	(Morgan 1983b)	Standardized distance-decay exposure
S	—	N	(Wong 1999)	Intersection of deviational ellipses
SP	—	N	(White 1983)	Ratio of mean within-group proximity to mean total proximity
SP	—	N	(Grannis 2002)	Multigroup version of White's *SP*
I	—	Y	(Frank 2003)	Moran's *I*–spatial autocorrelation among adjacent tracts
$\tilde{H}$	*H*	N	This paper	*H* based on local environment group proportions
$\tilde{R}$	R	N	This paper	*R* based on local environment group proportions
$\tilde{D}$	*D*	N	This paper	*D* based on local environment group proportions
Spatial Exposure Measures				
*PC**	*P**	N	(Morgan 1983b)	*P** based on distance-decay exposure, special case of $\tilde{P}$*
$\tilde{P}$*	*P**	N	This paper	*P** based on local environment group proportions

aspatial measure. White (1983) proposed a spatial proximity index to measure spatial segregation; Grannis (2002) proposed a multigroup version of this index, which we will denote as *SP*. The index is a measure of the average spatial proximity between two members of the same group divided by the average proximity between two members of the population. In principle, this measure does not depend on tract boundaries (White uses tract boundaries in estimating proximities; the measure would, however, be independent of tract boundaries if we had information on individuals' exact locations). Moreover, it has an intuitive appeal as a measure of spatial segregation. We consider *SP* a potentially useful measure of spatial segregation, and evaluate it alongside our new measures later in this paper. In our notation, the White/Grannis spatial proximity index is defined as

$$SP = \sum_{m=1}^{M} \frac{\pi_m P_{mm}}{P_{tt}}, \tag{13}$$

where P_{mm}, the average proximity between two members of group m, is defined as

$$\begin{aligned} P_{mm} &= \frac{1}{T_m^2} \int_p \int_q \tau_{pm} \tau_{qm} \phi(p, q) dq dp \\ &= \frac{1}{T_m^2} \int_p \tau_{pm} \tilde{\tau}_{pm} \Phi_p dp. \end{aligned} \tag{14}$$

P_{tt} is defined similarly. White suggests using a decreasing function of the distance between p and q for the proximity function ϕ here, though in principle, any desired proximity function could be used in equation (14) (Grannis 2002; ; White 1983).

Scholars have suggested several additional spatial evenness measures. Wong's deviational ellipse (1999) introduces the novel idea of comparing the overall spatial distribution of different population subgroups. However the measure is problematic because the deviational ellipse provides only a very generalized approximation of subgroup spatial distributions. In more recent work O'Sullivan and Wong (2004) use a density estimation method to approximate and compare the spatial distributions of two population subgroups.

However, because the resulting measure is a spatial generalization of *D*, this approach will fail to meet a number of the criteria under consideration. Finally, Frank (2003) suggests a segregation measure based on the spatial autocorrelation of tract compositions. Like all other measures that depend explicitly on tract boundary definitions, however, it is susceptible to MAUP issues. Thus, of the proposed spatial evenness measures, the White/Grannis spatial proximity index *SP* appears the most promising candidate for satisfying the spatial segregation criteria above.

There are far fewer candidates for a spatial exposure index. As we noted above, Morgan's *PC** is a special case of our more general proposed spatial exposure index, so we will not evaluate it separately here (Morgan 1983b). We are not aware of any other proposed spatial exposure indices.

4. EVALUATION OF THE SPATIAL SEGREGATION INDICES

We now turn to evaluating the indices against the criteria articulated above. We evaluate here four measures of spatial evenness ($\tilde{H}$, $\tilde{D}$, $\tilde{R}$, and *SP*), and one measure of spatial exposure ($\tilde{P}^*$).

Scale Interpretability. Each of the three evenness measures we derive—$\tilde{H}$, $\tilde{D}$, and $\tilde{R}$—meets the scale interpretability criterion. Each has a maximum of 1, obtained under complete segregation, is equal to zero if each local environment has a composition equal to that of the whole population, and is negative in the case of hyperintegration. The spatial proximity index has no theoretical maximum and is equal to 1 under perfect evenness, with values less than one indicating hyperintegration (White 1983). The lack of a theoretical maximum makes comparative studies using *SP* potentially difficult. The spatial exposure index is, by definition, bounded between 0 and 1.

Arbitrary Boundary Independence. Each of the five indices is computed based on population density information at each point; as a result, each of the indices is free of MAUP issues in principle, though the estimation of population density information from aggregated (tract) data may still lead to some MAUP issues, but these are due

to data collection methods rather than segregation computation methods.

Population Density and Location Equivalence. Both of these criteria are easily assessed using simple algebra. Like their aspatial counterparts, all five indices satisfy the population density invariance criteria. Each of the measures except *SP* meets the location equivalence criterion.

Composition Invariance. As we noted above, the composition invariance criterion as stated by James and Taueber (1985) may be inappropriate, since it rests on a particular definition of what it means for the distribution of individuals in space to remain the same as the population composition changes. Nonetheless, it is useful to evaluate the indices against the James and Taueber formulation of composition invariance for the sake of continuity with prior work. When we do so, we find that, since D, H, and R do not satisfy the criterion in the aspatial multigroup case, their spatial analogs likewise do not meet it (though $\tilde{D}$ is composition invariant in the two-group case). Likewise, simple algebra shows that the spatial proximity index is not composition invariant either. The criterion does not apply to the spatial exposure index $\tilde{P}^*$.

Although none of $\tilde{H}$, $\tilde{D}$, $\tilde{R}$, and *SP* are composition invariant (by the James and Taeuber definition) in the general spatial multigroup case, we have conducted a preliminary set of exploratory simulation analyses in order to determine whether one or more of the indices approximates composition invariance. In general, the behavior of each of the indices with respect to a change in the population composition of the type suggested by James and Taeuber is complex, particularly when one or more groups makes up a small portion of the population and/or when the group whose size is changing is highly segregated from some groups and not highly segregated from others. Nonetheless, a series of exploratory simulations analogous to those conducted by James and Taeuber (1985:16–18) show that $\tilde{H}$ and $\tilde{D}$ are less sensitive to changes in the population composition, in general, than are $\tilde{R}$ and *SP*.

It is worth noting that a failure to meet the James and Taeuber composition invariance criterion does not imply that a measure may not be composition invariant by some other definition. For example,

if the population composition changes but if the ratio of the average diversity (as measured by the entropy function E) of individuals' local environments to the total diversity of the population remains constant, then $\tilde{H}$ will be composition invariant under a corresponding definition of invariance. Finally, failure to meet a composition invariance criterion indicates that a segregation measure is affected by the population composition; this does not, however, imply that the measure is a measure of spatial exposure, rather than of spatial evenness. A measure of exposure should increase when the proportion of the group that others are exposed to grows in the population. But, as James and Taueber (1985) show, the aspatial H and V (as well as the spatial $\tilde{H}$ and $\tilde{R}$) both may increase or decline in response to increases in one group's share of the population, indicating that their responsiveness to population composition is not due to a confounding of exposure and evenness measurement.

Transfers and Exchanges.[13]

- *Transfers*: None of the spatial segregation measures we describe here meet the transfer criterion in the general spatial case.
- *Exchanges*: In the most general case, none of the evenness measures meet the type 1 exchange criterion, and only $\tilde{H}$ meets the type 2 exchange criterion. Several of the measures, however, meet the exchange criteria if the region R is symmetric under ϕ. We say that points p and q are symmetric if there is a one-one mapping between the set of points closer to p than q and those closer to q than p, and if corresponding points and their local environments have similar population density ratios. This condition is unlikely to be met completely, but may be approximated in real residential patterns. (See Section A.1 in the Appendix for a more precise definition and related discussion.)

  Both $\tilde{H}$ and $\tilde{R}$ meet the type 1 exchange criterion under conditions of spatial symmetry. Moreover, while only $\tilde{H}$ meets the type 2 exchange criterion in the most general case, $\tilde{R}$ also meets the criterion when the region is symmetric under ϕ.

  Under conditions of spatial symmetry, the spatial dissimilarity index $\tilde{D}$, like its aspatial counterpart, satisfies only a weak version of

[13]See Sections A.2 and A.3 in the Appendix for all proofs.

each of the exchange criteria. An exchange that moves a group m member away from locations with higher proportions of group m and nearer to points with lower proportions of group m will never result in an increase in $\tilde{D}$. In most cases, however—as long as there is set of symmetric points (see Section A.1 in the Appendix) s and s' in R such that $\tilde{\pi}_{sm} > \pi_m > \tilde{\pi}_{s'm}$ or $\tilde{\pi}_{sn} < \pi_n < \tilde{\pi}_{s'n}$—a type 1 exchange will register an appropriate decrease in $\tilde{D}$. Likewise, as long as there is some point s closer to p than q such that $\tilde{\pi}_{sm} > \tilde{\pi}_{sn}$ or some point s' closer to q than p such that $\tilde{\pi}_{s'n} > \tilde{\pi}_{s'm}$, then a type 2 exchange will register an appropriate decrease in $\tilde{D}$.

The spatial proximity index SP fails to meet either of the exchange criteria, even under conditions of spatial symmetry.

Additive Spatial Decomposability. Suppose a spatial region R is subdivided into K subregions. In the aspatial case, any rearrangement of individuals within subregion k will not affect segregation within any other subregion; nor will it affect the between-subregion segregation level. In this case, Reardon and Firebaugh (2002a) define a segregation measure as organizationally decomposable if it can be written as a sum of $K+1$ independent components—a between-subregion component and K within-subregion components, with each of the K within-subregion components indicating the amount by which total segregation would be reduced if segregation within subregion k were eliminated by rearranging individuals within k while leaving the location of all other individuals outside k unchanged.

An additive spatial decomposition is not so neatly defined, since rearranging individuals within subregion k may change the spatial environments of individuals in other subregions. As a result, the between- and within-subregion components of segregation are not necessarily independent. Nonetheless, we can define meaningful spatial decompositions of both $\tilde{R}$ and $\tilde{H}$ that incorporate a spatial interaction term that accounts for the spatial interaction between locations in different subregions.

To describe the spatial decomposition of $\tilde{H}$, we first require a refinement to our earlier notation. For some region S, define $\tilde{E}_{p|S}$ as the spatial entropy at point p as defined in equation (6), except that the $\tilde{\pi}_{pm}$'s are computed from equations (1) and (2) using only the points in region S.

In this notation, $\tilde{E}_p$ as defined in equation (6) would be written $\tilde{E}_{p|R}$, since all points in region R might contribute to the spatial environment of point p. Now we can write $\tilde{H}$ as the sum of three components:

$$\tilde{H} = \sum_{k \in R} \frac{t_k}{TE}(E - E_k) + \sum_{k \in R} \left[\int_{p \in k} \frac{\tau_p}{TE} \left(\tilde{E}_{p|k} - \tilde{E}_{p|R} \right) dp \right] + \sum_{k \in R} \frac{t_k E_k}{TE} \int_{p \in k} \frac{\tau_p}{t_k E_k} \left(E_k - \tilde{E}_{p|k} \right) dp \tag{15}$$

The first term on the right-hand side of equation (15) is simply the aspatial segregation between the K subregions (see Reardon and Firebaugh 2002a). The integral in the third term on the right-hand side is the spatial segregation within subregion k, ignoring points outside of k. We can rewrite equation (15) as

$$\tilde{H} = H_K + \sum_{k \in R} \left[\int_{p \in k} \frac{\tau_p}{TE} \left(\tilde{E}_{p|k} - \tilde{E}_{p|R} \right) dp \right] + \sum_{k \in R} \frac{t_k E_k}{TE} \tilde{H}_k. \tag{16}$$

The middle term is an interaction term that reflects the contribution to spatial segregation that results from the spatial proximity of points within different subareas. In general, for a given subregion k, the integral will be positive if, on average, subregions outside of k decrease the diversity of the local environments of individuals within subregion k, and negative if they increase it.

It is useful to consider a few special cases of equation (16). First, suppose that each population group is evenly distributed within each subregion k—this would be the case if, for example, the subregions were census tracts and we assumed that the population of each tract were evenly distributed throughout the tract. In this case, $\tilde{H}_k = 0$ and $\tilde{E}_{p|k} = E_k$ for each k. Then $\tilde{H}$ is simply the sum of the aspatial segregation among the tracts and a between-tract spatial segregation term:

$$\tilde{H} = H_K + \sum_{k \in R} \left[\int_{p \in k} \frac{\tau_p}{TE} \left(E_k - \tilde{E}_{p|R} \right) dp \right]. \tag{17}$$

Second, consider a special case where the spatial proximity of points in separate subregions is zero. This is the case, of course, for aspatial measures, but it also might be the case if, for example, a city were divided into distinct subareas through natural or manmade barriers—rivers, major highways, and the like—across which there were no spatial interaction. In this case, the interaction term would be zero, and $\tilde{H}$ can be written as the sum of a between-subarea aspatial segregation component and K within-subarea spatial segregation measures:

$$\tilde{H} = H_K + \sum_{k \in R} \frac{t_k E_k}{TE} \tilde{H}_k. \tag{18}$$

Finally, suppose that the K subregions are large compared to the size of the local spatial environments of individuals. Then, for most individuals—except for those located near the boundaries between subregions—we will have $\tilde{E}_{p|k} \approx \tilde{E}_{p|R}$. As a result, the spatial interaction term will be relatively small, and the decomposition in equation (18) will hold approximately.

The spatial relative diversity index $\tilde{R}$ can be decomposed in the same way as $\tilde{H}$, by substituting $\tilde{I}$ for $\tilde{E}$ in equations (15) through (18). $\tilde{D}$, however, like its aspatial counterpart, cannot be meaningfully decomposed into between- and within-subregion components. Nor are we able to construct a decomposition of SP.

The spatial exposure index, however, does have a useful spatial decomposition. Using similar notation as in equation (16), we can write

$${}_m\tilde{P}^*_n = \sum_{k \in R} \frac{t_{km}}{T_m} \left[{}_m\tilde{P}^*_n \right]_k + \sum_{k \in R} \left[\int_{p \in k} \frac{\tau_{pm}}{T_m} \left(\tilde{\pi}_{pn|R} - \tilde{\pi}_{pn|k} \right) dp \right]. \tag{19}$$

The first term in the right-hand side of the equation is a weighted average of the spatial exposure within each subarea. The second term is an interaction term that reflects the contribution to spatial exposure that results from the spatial proximity of points within different subareas.

Additive Grouping Decomposability. Following Reardon and Firebaugh (2002a), a spatial segregation index $\tilde{S}$ meets the grouping decomposability criterion if we can write

$$\tilde{S} = \tilde{S}_N + \sum_{n=1}^{N} g(\tilde{S}_n), \tag{20}$$

where $\tilde{S}_N$ is the segregation calculated among the N supergroups, $\tilde{S}_n$ is the segregation among the groups making up supergroup n, and g is a strictly increasing function on the interval [0,1] with $g(0) = 0$. As in the aspatial case, only $\tilde{H}$ yields a meaningful grouping decomposition. Because the between-supergroup decomposition of H depends only on the decomposition of E, and because the $\tilde{\pi}_{pm}$'s sum to 1 for all p; the decomposition of $\tilde{H}$ into between- and within-supergroup components has the same form as the aspatial H:

$$\tilde{H} = \tilde{H}_N + \sum_{n=1}^{N} \frac{t_n E_n}{TE} \tilde{H}_n. \tag{21}$$

Table 2 summarizes the compliance of the spatial segregation measures with the criteria we describe. Of the four spatial evenness measures, the spatial proximity index SP is clearly the least satisfactory, as it fails to meet almost every criterion. The spatial information theory index $\tilde{H}$ appears the most satisfactory, as it satisfies the exchange criteria in the widest range of cases and is also the only index that has both a meaningful spatial and grouping decomposition. Of the remaining two, $\tilde{D}$ is arguably less satisfactory than $\tilde{R}$, as $\tilde{R}$ meets the exchange criteria in a wider range of cases and can also be spatially decomposed.

5. DISCUSSION AND CONCLUSION

Despite the existence of a number of proposed measures of spatial segregation, such measures have not been widely used in residential segregation research. In fact, a reading of the existing spatial segregation literature provides little guidance about which of the many proposed measures are most useful. In this paper, we have—at the risk of further cluttering an already cluttered field—developed several new measures of spatial segregation, based on a new spatial proximity approach. Several key features of our approach are notable. First, our approach avoids, in principle, MAUP issues by using point-to-point

TABLE 2
Properties of Spatial Segregation Indices

	Information Theory ($\tilde{H}$)	Relative Diversity ($\tilde{R}$)	Dissimilarity ($\tilde{D}$)	Spatial Proximity (SP)	Spatial Exposure ($\tilde{P}^*$)
Scale interpretability	✓	✓	✓	**X**	✓
MAUP-free	✓	✓	✓	✓	✓
Location equivalence					
Aspatial	✓	✓	✓	**X**	✓
Spatial	✓	✓	✓	**X**	✓
Population density invariance					
Aspatial	✓	✓	✓	✓	✓
Spatial	✓	✓	✓	✓	✓
Compositional invariance					
Aspatial 2-Group	**X**	**X**	✓	**X**	—
Aspatial *M*-Group	**X**	**X**	**X**	**X**	—
Spatial 2-Group	**X**	**X**	✓	**X**	—
Spatial *M*-Group	**X**	**X**	**X**	X	—
Transfers					
Aspatial 2-Group	✓	✓	**X**[a]	**X**	—
Aspatial *M*-Group	✓	**X**	**X**	**X**	—
Spatial 2-Group	**X**	**X**	**X**	**X**	—
Spatial *M*-Group	**X**	**X**	**X**	**X**	—

continued

TABLE 2
Continued

	Information Theory ($\tilde{H}$)	Relative Diversity ($\tilde{R}$)	Dissimilarity ($\tilde{D}$)	Spatial Proximity (SP)	Spatial Exposure ($\tilde{P}^*$)
Exchanges (type 1)					
Aspatial 2-Group	✓	✓	X[a]	X	—
Aspatial M-Group	✓	✓	X[a]	X	—
Spatial 2-Group	✓[b]	✓[b]	X[a,b]	X	—
Spatial M-Group	✓[b]	✓[b]	X[a,b]	X	—
Exchanges (type 2)					
Aspatial 2-Group	✓	✓	X[a]	X	—
Aspatial M-Group	✓	✓	X[a]	X	—
Spatial 2-Group	✓	✓[b]	X[a,b]	X	—
Spatial M-Group	✓	✓[b]	X[a,b]	X	—
Additive spatial decomposability					
Aspatial 2-Group	✓	✓	X	X	✓
Aspatial M-Group	✓	✓	X	X	—
Spatial 2-Group	✓	✓	X	X	✓
Spatial M-Group	✓	✓	X	X	—
Additive grouping decomposability					
Aspatial	✓	X	X	X	—
Spatial	✓	X	X	X	—

[a] *The dissimilarity index satisfies only a weak form of the principles of transfers and exchanges in these cases: transfers and exchanges that move individuals away from local environments with higher group proportions and nearer to those with lower group proportions will never result in an increase in* $\tilde{D}$, *though they may result in no change in* $\tilde{D}$.

[b] *The indices do not meet the criterion in general, though they do meet it if the region* R *is symmetric under* ϕ.

proximity functions rather than tract contiguity matrices. Second, our approach is nonspecific regarding the choice of a spatial proximity function. This enables (requires, actually) researchers to specify their underlying assumptions about socio-spatial proximity, and it facilitates research that compares segregation levels based on different theoretical bases for defining spatial proximity. Further, our approach yields, as special cases, traditional aspatial segregation measures (both two-group and multigroup) and makes clear the assumptions about spatial proximity inherent in such measures. Finally, our approach yields measures of both spatial exposure/isolation and spatial evenness/clustering.

In addition to developing a new set of spatial segregation measures, we review and evaluate all previously proposed measures of spatial evenness and exposure as well as our new measures. Here we conclude that the spatial information theory index $\tilde{H}$ is the best of the spatial evenness measures, when judged against the criteria we have outlined. Likewise, we conclude that the spatial exposure/isolation index $\tilde{P}^*$—which is a spatial generalization of the familiar P^* exposure and isolation index—is a satisfactory measure of spatial exposure. We suggest that researchers rely on these measures in future research in order to ensure comparability across studies.

We do not, however, specify or recommend a particular proximity function for use in computing the measures. In fact, it seems likely that research that compares segregation levels of $\tilde{H}$ based on different proximity functions could be useful in understanding the processes that organize residential space. For example, parallel studies of a number of cities might reveal that using a simple fast (short) distance decay formulation for the proximity function results in a different rank-ordering of cities by segregation levels than does using a proximity function with a slower distance decay characteristic. Interpretation of such results would indicate something about the geographical scale at which segregation occurs in the cities in question.

While $\tilde{H}$ and $\tilde{P}^*$ are, in principle, very satisfactory spatial segregation indices, several important issues remain in operationalizing and computing these measures. The first is that although our approach relies on complete data about individual residential locations, such data are rarely available, though they can be estimated from readily available tract data using a variety of methods. The simplest method would be to assume an even population density

within each tract, though this will result in sharp discontinuities in density at tract edges. Alternatively, spatial smoothing of population can be performed using various methods: kernel density estimation (Silverman 1986); group-specific pycnophylactic smoothing, which redistributes each population group within tracts such that tract totals are honored but population groups are moved toward neighboring tracts with large populations of the same group (Tobler 1979); or dasymetric mapping, which uses street networks or zoning patterns to estimate population densities (Mennis 2003). Although all of these methods are computationally intensive, they can be readily automated within typical GIS software packages. We are currently developing a set of tools that will allow researchers to estimate smooth population density surfaces using these methods and to use these smooth density surfaces in the computation of spatial segregation measures.

A second practical issue in our approach is that it requires the numerical evaluation of integrals over the study region R. In practice, this means dividing R into small cells for computational purposes, but it is unclear how sensitive the resulting measures of segregation will be to the choice of cell size. A third issue arises at edges of a study region R. Omitting data from outside the study region (say, a city, or metropolitan area), may be convenient (or necessary, if data are not available), but this may affect the estimation of the population density and racial composition for points near the edge of the study region, which will in turn affect the measured segregation.

The scale of the chosen proximity function relative to the scale of cells, tracts, and the study region is likely to be the critical factor determining how sensitive computed segregation measures are to variation, respectively, in the density estimation method, the choice of cell size, and the approach to treating edge conditions. Future work should determine how sensitive $\tilde{H}$ and $\tilde{P}^*$ are to choices regarding these issues.

An additional issue pertains to the potential need to use complex spatial proximity functions. It may be relatively simple to use a "bounded Gaussian" distance-decay spatial proximity function—a proximity function that is strictly a decreasing function of the Euclidean distance between two points and that goes to zero at some defined distance; such a function is computationally efficient, because it is defined identically at each point and because the cut-off distance removes the necessity to perform numerical integration over the entire study region R. Such a function, regardless of its precise mathematical

form (Gaussian, negative exponential, negative power law, etc.) has a certain intuitive appeal but nonetheless has only weak theoretical support and no supporting empirical evidence. It may be thought of as accounting for the varied behavior of individuals by aggregating many individual life-spaces into an overall average.

A more realistic spatial proximity function might take into account obstacles and discontinuities in the spatial fabric, such as highways or rivers. These will disrupt the neat mathematics of a bounded-Gaussian proximity function, however, and call for special treatment, with the proximity of locations on opposite sides of boundaries being set much lower than their simple Euclidean separation distance would dictate. Conversely "promoters" or channels for sociospatial interaction, such as street networks and public transportation services, would be treated in the opposite sense, increasing the proximity of locations connected by them (Grannis 2002).

While our approach to measuring spatial segregation enables us to account for complicated patterns of spatial proximity, such patterns do complicate the implementation of the measures, since special programming in GIS software is necessary to incorporate them. Moreover, any proximity function that is not the same at all locations requires the representation of the function by a (very large) interaction matrix that records the proximity between every pair of locations in the study region. Calculation and manipulation of such a matrix will impose significant computational burdens on any implementation of the proposed measure. Given the complexity of programming complex, user-defined spatial proximity functions for a number of locations, we expect that most users of these measures will initially prefer to use some simple distance-decay function until software tools are available to automate the use of more complex spatial proximity functions.

Finally, the measures we have developed here apply most obviously to the case of spatial residential racial segregation. In principle, however, we can extend this approach to measure segregation according to any population characteristic. For example, we could generate measures of spatial income segregation simply by computing some income variation statistic (such as the variance) within each local environment and then computing a measure of the variation in this statistic across all points in the region. In addition, we can extend this approach to measure other types of segregation, simply by defining an

appropriate proximity function. For example, we can measure the segregation of social networks by defining some social proximity function that indicates how near to one another any two individuals are within a social network (see Reardon and Firebaugh 2002b). Because of the generality of the measures with regard to the proximity function, our approach here may yield useful measures of social segregation in any domain, so long as an appropriate social proximity function is specified.

APPENDIX A: COMPLIANCE OF THE MEASURES WITH THE TRANSFER AND EXCHANGE CRITERIA

A.1. *The Spatial Symmetry Condition*

To evaluate the conditions under which a spatial segregation index meets the exchange criterion, we first provide a definition of spatial symmetry. Given a spatial proximity function $\phi(p, q)$ that is defined for all points $p, q \in R$, we say that R is symmetric under ϕ if for each pair of distinct points $p, q \in R$, we can divide R into three subregions, R_p, R_q, and R_0, where $\phi(s, p) > \phi(s, q)$ for all $s \in R_p$; $\phi(s, q) > \phi(s, p)$ for all $s \in R_q$; and $\phi(s, p) = \phi(s, q)$ for all $s \in R_0$, and such that for each point $s \in R_p$ there exists a unique corresponding point $s' \in R_q$ such that $\phi(s, p) - \phi(s, q) = \phi(s', q) - \phi(s', p)$ and $\frac{\tau_s}{\tilde{\tau}_s} = \frac{\tau_{s'}}{\tilde{\tau}_{s'}}$. If we denote

$$\Delta_s(p, q) = \frac{\tau_s}{\tilde{\tau}_s}[\phi(s, p) - \phi(s, q)], \qquad (A-1)$$

then we have

$$\Delta_s(p, q) = -\Delta_{s'}(p, q) \qquad (A-2)$$

for all symmetric points s and s' in R.

Several examples of spatial symmetry are notable. First, in the usual aspatial case, the region R is divided into distinct subregions (tracts), and $\phi(p, q)$ is defined such that $\phi(p, q) = c$ if p and q are in the same subregion and $\phi(p, q) = 0$ otherwise. It is simple to show that R is symmetric under ϕ in this case. To see this, consider the case where p and q are in different tracts. Then R_p consists of the tract containing point p, R_q consists of the tract containing point q, and R_0 is the remainder of R. Now if we assume the populations in both R_p

and R_q are both located at a single point (this will not change segregation as measured by any index satisfying the locational equivalence criterion—see footnote 4), then the conditions for symmetry are met.

A second example of spatial symmetry results if region R extends infinitely in all directions, with constant population density (in which case $\tau_s = \tilde{\tau}_s$ for all s), and if $\phi(p, q)$ depends only on the Euclidean distance between points p and q. Finally, if R is large compared to the scale of local environments defined by ϕ, and if the population density changes relatively little over distances comparable to the scale of local environments, then $\tau_s \approx \tilde{\tau}_s$ for all s in R and R is approximately spatially symmetric under ϕ.

A.2. *Evaluation of the Exchange Criteria*

We can evaluate each index's compliance with the principles of transfers and exchanges by taking the derivative of the index with respect to a transfer or exchange x. Moreover, because an exchange consists of a pair of complementary transfers, failure to satisfy the type 1 exchange criterion implies that a measure does not satisfy the transfer criterion. Likewise, a measure that meets the transfer criterion will necessarily meet the type 1 exchange criterion.

We first evaluate the behavior of the indices with respect to an exchange. When x involves the exchange of a member of group m at point p with a member of group n at point q, then the derivative of $\tilde{H}$ with respect to x is

$$\frac{d\tilde{H}}{dx} = \frac{1}{TE} \int_{s \in R} \Delta_s(p, q) \ln \frac{\tilde{\pi}_{sn}}{\tilde{\pi}_{sm}} ds. \qquad (A-3)$$

Now we divide the region R into three subregions, R_p, R_q, R_0, such that $\phi(s, p) > \phi(s, q)$ for all $s \in R_p$; $\phi(s', q) > \phi(s', p)$ for all $s' \in R_q$; and $\phi(r, p) = \phi(r, q)$ for all $r \in R_0$. Now

$$\frac{d\tilde{H}}{dx} = \frac{1}{TE} \left[\int_{s \in R_p} \Delta_s(p, q) \ln \frac{\tilde{\pi}_{sn}}{\tilde{\pi}_{sm}} ds + \int_{s' \in R_q} \Delta_{s'}(p, q) \ln \frac{\tilde{\pi}_{s'n}}{\tilde{\pi}_{s'm}} ds' \right]. \qquad (A-4)$$

Now suppose that $\tilde{\pi}_{sm} > \tilde{\pi}_{sn}$ and $\tilde{\pi}_{s'm} < \tilde{\pi}_{s'n}$ for all $s \in R_p$ and all $s' \in R_q$. In this case, equation (A-4) yields $\frac{d\tilde{H}}{dx} < 0$, so $\tilde{H}$ satisfies the type 2 exchange criterion.

In general, $\tilde{H}$ does not satisfy the type 1 exchange criterion, since equation (A-4) can be negative under conditions of a type 1 exchange. If R is symmetric under ϕ, however, then we can exploit the one-to-one mapping of points in R_p and R_q to write equation (A-4) as

$$\frac{d\tilde{H}}{dx} = \frac{1}{TE} \int_{s \in R_p} \Delta_s(p,q) \left(\ln \frac{\tilde{\pi}_{sn}\tilde{\pi}_{s'm}}{\tilde{\pi}_{s'n}\tilde{\pi}_{sm}} \right) ds, \qquad (A-5)$$

where s' is the point in R_q corresponding to the point s in R_p. For every point $s \in R_p$, $\phi(s, p) - \phi(s, q) > 0$. When $\tilde{\pi}_{sm} > \tilde{\pi}_{s'm}$ and $\tilde{\pi}_{sn} < \tilde{\pi}_{s'n}$, then equation (A-5) yields $\frac{d\tilde{H}}{dx} < 0$, so $\tilde{H}$ satisfies the type 1 exchange criterion if R is symmetric under ϕ.

The derivative of $\tilde{R}$ with respect to an exchange x is

$$\frac{d\tilde{R}}{dx} = \frac{2}{TI} \int_{s \in R} \Delta_s(p,q)[(\pi_m - \tilde{\pi}_{sm}) - (\pi_n - \tilde{\pi}_{sn})] ds \qquad (A-6)$$

Note that, unlike $\tilde{H}$, the condition that $\tilde{\pi}_{sm} > \tilde{\pi}_{sn}$ and $\tilde{\pi}_{s'm} < \tilde{\pi}_{s'n}$ for all $s \in R_p$ and all $s' \in R_q$ is not sufficient to ensure that $\tilde{R} < 0$, so $\tilde{R}$ does not, in general, satisfy the type 2 exchange criterion. However, under the condition of spatial symmetry, we can write equation (A-6) as

$$\frac{d\tilde{R}}{dx} = \frac{2}{TI} \int_{s \in R_p} \Delta_s(p,q)[(\tilde{\pi}_{s'm} - \tilde{\pi}_{sm}) + (\tilde{\pi}_{sn} - \tilde{\pi}_{s'n})] ds. \qquad (A-7)$$

When either $\tilde{\pi}_{sm} > \tilde{\pi}_{s'm}$ and $\tilde{\pi}_{sn} < \tilde{\pi}_{s'n}$ or $\tilde{\pi}_{sm} > \tilde{\pi}_{sn}$ and $\tilde{\pi}_{s'm} < \tilde{\pi}_{s'n}$, then equation (A-7) yields $\frac{d\tilde{R}}{dx} < 0$, so $\tilde{R}$ satisfies both exchange criteria if R is symmetric under ϕ.

The derivative of $\tilde{D}$ with respect to an exchange x is

$$\frac{d\tilde{D}}{dx} = \frac{1}{2TI} \int_{s \in R} \Delta_s(p,q)(z_{sn} - z_{sm}) ds, \qquad (A-8)$$

where

$$z_{sk} = \begin{cases} 1 & \text{if } \tilde{\pi}_{sk} > \pi_k \\ -1 & \text{if } \tilde{\pi}_{sk} < \pi_k \\ 0 & \text{if } \tilde{\pi}_{sk} = \pi_k. \end{cases}$$

In the aspatial case, D satisfies only a weak form of the type 1 exchange criterion; the specified exchange may not reduce segregation, but will never increase it (Reardon and Firebaugh 2002a). In the spatial case, however, $\tilde{D}$ does not satisfy even this weak form of the type 1 exchange criterion, as the expression in equation (A-8) may be positive in some cases. Under the spatial symmetry condition, however, equation (A-8) can be written

$$\frac{d\tilde{D}}{dx} = \frac{1}{2TI} \int_{s \in R_p} \Delta_s(p,q)[(z_{sn} - z_{sm}) - (z_{s'n} - z_{s'm})]ds. \qquad (A-9)$$

When either (1) $\tilde{\pi}_{sm} > \tilde{\pi}_{s'm}$ and $\tilde{\pi}_{sn} < \tilde{\pi}_{s'n}$, or (2) $\tilde{\pi}_{sm} > \tilde{\pi}_{sn}$ and $\tilde{\pi}_{s'm} < \tilde{\pi}_{s'n}$ is true, then equation (A-9) yields $\frac{d\tilde{D}}{dx} \leq 0$, so $\tilde{D}$ satisfies a weak form of both exchange criteria if R is symmetric under ϕ.

If we assume that $\Phi_p = \Phi_q = \Phi$ for all $p, q \in R$, then the derivative of SP with respect to an exchange x is

$$\frac{dSP}{dx} = \frac{2\Phi}{T_m T_n P_{tt}} \left[\pi_n\left(\tilde{\tau}_q \tilde{\pi}_{qm} - \tilde{\tau}_p \tilde{\pi}_{pm}\right) + \pi_m\left(\tilde{\tau}_p \tilde{\pi}_{pn} - \tilde{\tau}_q \tilde{\pi}_{qn}\right)\right] \qquad (A-10)$$

In general, this quantity can be positive under the conditions of either exchange criterion. This is true for both the spatial and aspatial cases and for the two-group and multigroup versions of the index, so SP does not satisfy either of the exchange criteria in any case.

A.3. *Evaluation of the Transfer Criterion*

We next examine the behavior of the indices with respect to a transfer x of a person of group m from point p to q. Because $\tilde{H}$ and $\tilde{R}$ meet the first exchange criterion only when R is symmetric under ϕ, we need only evaluate $\tilde{H}$ and $\tilde{R}$ with respect to the transfer criterion in the case when R is symmetric under ϕ. In this case, we have

$$\frac{d\tilde{H}}{dx} = \frac{1}{TE} \int_{s \in R_p} \Delta_s(p,q) \left[\left(\tilde{E}_{s'} - \tilde{E}_s \right) + \ln \frac{\tilde{\pi}_{s'm}}{\tilde{\pi}_{sm}} \right] ds + \frac{1}{TE} \left(\tilde{E}_p - \tilde{E}_q \right) \tag{A-11}$$

$$\frac{d\tilde{R}}{dx} = \frac{2}{TI} \int_{s \in R_p} \Delta_s(p,q) \left[\left(\tilde{I}_{s'} - \tilde{I}_s \right) + \left(\tilde{\pi}_{s'm} - \tilde{\pi}_{sm} \right) \right] ds + \frac{1}{TI} \left(\tilde{I}_p - \tilde{I}_q \right) \tag{A-12}$$

Both of these quantities may be positive when $\tilde{\pi}_{sm} > \tilde{\pi}_{s'm}$ for all $s \in R_p$ and all $s' \in R_q$, so neither $\tilde{H}$ nor $\tilde{R}$ meets the transfer criterion.

Because the aspatial D satisfies the transfer criterion only in the two-group case (Reardon and Firebaugh 2002a) and the spatial $\tilde{D}$ satisfies only a weak form of the type 1 exchange criterion, and then only under conditions of spatial symmetry, we need only evaluate $\tilde{D}$ with regard to the transfer criterion in the two-group case under conditions of spatial symmetry. The derivative of $\tilde{D}$ with respect to a transfer x in this case is

$$\begin{aligned} \frac{d\tilde{D}}{dx} = \frac{-1}{TI} \Bigg[& \int_{s \in R_p} \Delta_s(p,q) [(1 - \tilde{\pi}_{sm}) z_{sm} - (1 - \tilde{\pi}_{s'm}) z_{s'm}] ds \\ & + \left(\tilde{\pi}_{pm} - \pi_m \right) z_{pm} - \left(\tilde{\pi}_{qm} - \pi_m \right) z_{qm} \Big], \end{aligned} \tag{A-13}$$

where z_{pm} and z_{qm} are as in equation (A-8). This expression can be either positive or negative under the transfer criterion conditions, so $\tilde{D}$ does not meet the transfer criterion.

Because the spatial proximity index SP does not meet the exchange criterion in any case, we know it will not meet the transfer criterion.

REFERENCES

Apparicio, Phillipe. 2000. "Les indices de ségregation résidentielle: un outil intégré dans un système d'information géographique." *Cybergéo, revue européenne de géographie* 134.

Bell, W. 1954. "A Probability Model for the Measurement of Ecological Segregation." *Social Forces* 43:357–364.

Coleman, James, Thomas Hoffer, and Sally Kilgore. 1982. "Achievement and segregation in secondary schools: A further look at public and private school differences." *Sociology of Education* 55:162–82.

Dent, Borden D. 1999. *Cartography: Thematic Map Design*. Boston, MA: McGraw-Hill.

Frank, Andrea I. 2003. "Using Measures of Spatial Autocorrelation to Describe Socio-economic and Racial Residential Patterns in US Urban Areas." Pp. 147–62 in *Socio-economic Applications of Geographic Information Science, Innovations in GIS*, edited by D. Kidner, G. Higgs, and S. White. London: Taylor and Francis.

Gorard, Stephen, and Chris Taylor. 2002. "A Comparison of Measures in Terms of 'Strong' and 'Weak' Compositional Invariance." *Sociology* 36:875–95.

Grannis, Rick. 2002. "Discussion: Segregation Indices and Their Functional Inputs." Pp. 69–84 in *Sociological Methodology*, Vol. 32, edited by Ross M. Stolzenberg. Boston, MA: Blackwell Publishing.

Jakubs, John F. 1981. "A Distance-Based Segregation Index." *Journal of Socio-economic Planning Sciences* 15:129–36.

James, David R. and Karl E. Taeuber. 1985. "Measures of Segregation." Pp. 1–32 in *Sociological Methodology*, Vol. 14, edited by Nancy Brandon Tuma. San Francisco, CA: Jossey-Bass.

Lieberson, Stanley. 1969. "Measuring Population Diversity." *American Sociological Review* 34:850–62.

Lieberson, Stanley, and Donna K. Carter. 1982a. "A Model for Inferring the Voluntary and Involuntary Causes of Residential Segregation." *Demography* 19:511–26.

———. 1982b. "Temporal Changes and Urban Differences in Residential Segregation: A Reconsideration." *American Journal of Sociology* 88:296–310.

Longley, P., M. F. Goodchild, D. J. Maguire, and D. W. Rhind. 2001. *Geographic Information Systems and Science*. New York: Wiley.

Massey, Douglas S., and Nancy A. Denton. 1988. "The Dimensions of Residential Segregation." *Social Forces* 67:281–315.

Mennis, J. 2003. "Generating Surface Models of Population Using Dasymetric Mapping." *Professional Geographer* 55:31–42.

Morgan, Barrie S. 1982. "An Assessment of Some Technical Problems in the Comparative Study of Residential Segregation." *Transactions–Institute of British Geographers* 7:227–32.

———. 1983a. "An Alternate Approach to the Development of a Distance-Based Measure of Racial Segregation." *American Journal of Sociology* 88:1237–49.

———. 1983b. "A Distance-Decay Interaction Index to Measure Residential Segregation." *Area* 15:211–16.

Morrill, R. L. 1991. "On the Measure of Spatial Segregation." *Geography Research Forum* 11:25–36.

Openshaw, S., and P. Taylor. 1979. "A Million or So Correlation Coefficients: Three Experiments on the Modifiable Area Unit Problem." Pp. 127–44 in *Statistical Applications in the Spatial Sciences*, edited by N. Wrigley. London: Pion.

O'Sullivan, David, and David W. S. Wong. 2004. "A Density Surface-Based Approach to Measuring Spatial Segregation." Presented at the Annual Meeting of the Association of American Geographers, March 14–19, Philadelphia, PA.

Pielou, E. C. 1977. *Mathematical Ecology*. New York: Wiley.

Reardon, Sean F., and Glenn Firebaugh. 2002a. "Measures of Multigroup Segregation." Pp. 33–67 in *Sociological Methodology*, Vol. 32, edited by Ross M. Stolzenberg. Boston, MA: Blackwell Publishing.

———. 2002b. "Response: Segregation and Social Distance—A Generalized Approach to Segregation Measurement." Pp. 85–101 in *Sociological Methodology*, Vol. 32, edited by Ross M. Stolzenberg. Boston, MA: Blackwell Publishing.

Schnell, Izhak, and Benjamini Yoav. 2001. "The Sociospatial Isolation of Agents in Everyday Life Spaces as an Aspect of Segregation." *Annals of the Association of American Geographers* 91:622–36.

Schwartz, Joseph, and Christopher Winship. 1980. "The Welfare Approach to Measuring Inequality." Pp. 1–36 in *Sociological Methodology*, Vol. 9, edited by Karl F. Schvessler. San Francisco, CA: Jossey-Bass

Silverman, B. W. 1986. *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

Theil, Henri. 1972. *Statistical Decomposition Analysis*, Vol. 14, edited by H. Theil. Amsterdam: North-Holland Publishing Company.

Tobler, Waldo R. 1979. "Smooth Pycnophylactic Interpolation for Geographical Regions." *Journal of the American Statistical Association* 74:519–29.

Waldorf, Brigitte S. 1993. "Segregation in Urban Space: A New Measurement Approach." *Urban Studies* 30:1151–64.

White, Michael J. 1983. "The Measurement of Spatial Segregation." *American Journal of Sociology* 88:1008–18.

———. 1986. "Segregation and Diversity Measures in Population Distribution." *Population Index* 52:198–221.

Wong, David S. 1993. "Spatial Indices of Segregation." *Urban Studies* 30:559–72.

Wong, David W. S. 1997. "Spatial Dependency of Segregation Indices." *Canadian Geographer* 41:128–36.

———. 1998. "Measuring Multiethnic Spatial Segregation." *Urban Geography* 19:77–87.

———. 1999. "Geostatistics as Measures of Spatial Segregation." *Urban Geography* 20:635–47.

———. 2002. "Spatial Measures of Segregation and GIS." *Urban Geography* 23:85–92.

———. 2003. "Implementing Measures of Spatial Segregation in GIS." *Computers, Environment and Urban Systems* 27:53–70.

Wong, David W. S., and Wing K. Chong. 1998. "Using Spatial Segregation Measures in GIS and Statistical Modeling Packages." *Urban Geography* 19:477–85.

Zoloth, Barbara S. 1976. "Alternative Measures of School Segregation." *Land Economics* 52:278–98.

Hypersegregation in U.S. Metropolitan Areas: Black and Hispanic Segregation along Five Dimensions
Author(s): Douglas S. Massey and Nancy A. Denton
Source: *Demography*, Vol. 26, No. 3 (Aug., 1989), pp. 373-391
Published by: Springer on behalf of the Population Association of America
Stable URL: http://www.jstor.org/stable/2061599
Accessed: 30/05/2011 21:06

Demography, Vol. 26, No. 3, August 1989

Hypersegregation in U.S. Metropolitan Areas: Black and Hispanic Segregation Along Five Dimensions

Douglas S. Massey and Nancy A. Denton
Population Research Center, 1155 E. 60th Street,
NORC/University of Chicago, Chicago,
Illinois 60637

Residential segregation has traditionally been measured by using the index of dissimilarity and, more recently, the P^* exposure index. These indices, however, measure only two of five potential dimensions of segregation and, by themselves, understate the degree of black segregation in U.S. society. Compared with Hispanics, not only are blacks more segregated on any single dimension of residential segregation, they are also likely to be segregated on all five dimensions simultaneously, which never occurs for Hispanics. Moreover, in a significant subset of large urban areas, blacks experience extreme segregation on all dimensions, a pattern we call hypersegregation. This finding is upheld and reinforced by a multivariate analysis. We conclude that blacks occupy a unique and distinctly disadvantaged position in the U.S. urban environment.

Since Duncan and Duncan's (1955) seminal paper, ecologists have relied primarily on the index of dissimilarity to measure residential segregation. Over the years, however, and especially since the critique of Cortese, Falk, and Cohen (1976), many other measures of segregation have been proposed (see James and Taeuber, 1985, and White, 1986, for reviews). In a recent paper, we identified 20 such measures and undertook a detailed conceptual and statistical analysis of their properties and interrelationships (Massey and Denton, 1988a). On both theoretical and empirical grounds, we concluded that segregation is a global construct that subsumes five distinct dimensions of spatial variation.

These five dimensions are evenness, exposure, clustering, centralization, and concentration. *Evenness* is the degree to which the percentage of minority members within residential areas equals the citywide minority percentage; as areas depart from the ideal of evenness, segregation increases. *Exposure* is the degree of potential contact between minority and majority members; it reflects the extent to which groups are exposed to one another by virtue of sharing neighborhoods in common. *Clustering* is the extent to which minority areas adjoin one another in space; it is maximized when minority neighborhoods form one large, contiguous ghetto and minimized when they are scattered widely in space. *Centralization* is the degree to which minority members are settled in and around the center of an urban area, usually defined as the central business district. Finally, *concentration* is the relative amount of physical space occupied by a minority group; as segregation rises, minority members are increasingly concentrated within a small, geographically compact area.

A high level of segregation on any one of these dimensions is problematic because it isolates a minority group from amenities, opportunities, and resources that affect social and economic well-being (cf. Logan, 1978; Massey, Condran, and Denton, 1987; Schneider and Logan, 1982, 1985). As high levels of segregation accumulate across dimensions, the deleterious effects of segregation multiply because isolation intensifies. Indices of evenness and exposure, by themselves, cannot capture this multidimensional layering of segregation and, therefore, misrepresent the nature of black segregation and understate its severity. Not

only are blacks more segregated than other groups on any single dimension of segregation, they are also more segregated across all dimensions simultaneously. In an important subset of urban areas, blacks are extremely segregated on each dimension, a pattern we call hypersegregation. The purpose of this article is to show that blacks occupy a unique and distinctly disadvantaged position in U.S. urban society by comparing their pattern of segregation with that of another disadvantaged minority group, Hispanics, and to demonstrate that the pattern of hypersegregation holds after the application of statistical controls for major confounding variables.

Data and Measures

Data are taken from the 1980 Summary Tape Files (STF4) of the U.S. Bureau of the Census (1980) and cover the 50 largest standard metropolitan statistical areas (SMSAs) plus 10 others that contain large numbers of Hispanics. The units of analysis are census tracts. Hispanics are defined by using the Spanish-origin item, and whites and blacks are identified from the census question on race, both 100 percent items (U.S. Bureau of the Census, 1982). The cross-classification of race and Spanish origin permits the definition of the mutually exclusive ethnic/racial categories (black Hispanics, white Hispanics, non-Hispanic blacks, and non-Hispanic whites) that are employed in this analysis. These groups were created by subtracting white and black Hispanics from the respective total counts of whites and blacks. For convenience, we refer to non-Hispanic whites as Anglos, though we are well aware that the terms "Anglo" and "Hispanic" mask considerable diversity in national origins and characteristics (Bean and Tienda, 1987; Greeley, 1974). A more detailed description of the data set is found in Massey and Denton (1987).

In our earlier methodological article (Massey and Denton, 1988a), we described in detail the choice of an index for each of the five dimensions of segregation, so only a brief review of their computational formulas is provided here. Evenness is measured with the traditional index of dissimilarity, which varies between 0 and 1.0, and represents the proportion of minority members that would have to change tracts to achieve an even distribution (Jakubs, 1977, 1979, 1981). The dissimilarity index may be defined as

$$D = \sum_{i=1}^{n} \frac{t_i|p_i - P|}{2TP(1 - P)}, \tag{1}$$

where t_i and p_i are the total population and minority proportion of areal unit i and T and P are the population size and minority proportion of the whole city, which is subdivided into n areal units.

Exposure is measured with the P^* measure, which has two basic forms. The first is the interaction index (${}_xP_y^*$), which measures the probability that members of minority group X share a tract with members of majority group Y. The other is the isolation index (${}_xP_x^*$), which measures the probability that group X members share a tract with each other. Both measures vary between 0 and 1.0 and in the two-group case sum to unity. Since higher values on the isolation index signify greater segregation, we chose it as our indicator of exposure. It is computed as the minority-weighted average of each unit's minority proportion (Lieberson, 1980, 1981):

$${}_xP_x^* = \sum_{i=1}^{n} \left[\frac{x_i}{X}\right]\left[\frac{x_i}{t_i}\right], \tag{2}$$

where x_i and t_i are the numbers of X members and the total population of tract i and X represents the number of X members citywide.

Clustering is the extent to which tracts inhabited by minority members adjoin one another, or cluster, in space. A high degree of clustering implies a residential structure in which minority areas are contiguous and closely packed, creating one large ethnic or racial enclave, whereas a low level of clustering means that minority areal units are widely scattered around the urban environment. The index of clustering we selected is White's (1983) index of spatial proximity, *SP*. It takes the average proximity between members of the same group and the average proximity between members of different groups and then computes a weighted average of these quantities. The average proximity between group X members is

$$P_{xx} = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{x_i x_j c_{ij}}{X^2}, \tag{3}$$

and the average proximity between members of X and Y is

$$P_{xy} = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{x_i y_j c_{ij}}{XY}, \tag{4}$$

where Y is the number of Y members citywide, x_i and y_j are the numbers of X and Y members in units i and j, and c_{ij} is a distance function between these two areas, defined here as a negative exponential: $c^{ij} = \exp(-d^{ij})$. The term d^{ij} indicates the linear distance between the centroids of units i and j, and d^{ii} is estimated as $0.6a_i \times 5$, where a_i is the area of the tract. The negative exponential assumes that the likelihood of intragroup interaction drops off rapidly with distance (White, 1983).

Average proximities may also be calculated among Y members (P_{yy}) and among all members of the population (P_{tt}) by analogy with equation (3). White's index is the average of intragroup proximities, P_{xx}/P_{tt} and P_{yy}/P_{tt}, weighted by the fraction of each group in the population:

$$SP = \frac{XP_{xx} + YP_{yy}}{TP_{tt}}, \tag{5}$$

producing a ratio that equals 1.0 when there is no differential clustering between X and Y and a ratio that is greater than 1.0 when members of X live nearer to each other than to members of Y. The ratio would be less than 1.0 in the unusual circumstance that members of X resided closer to members of Y than to other X members. In our data, all *SP* indices varied between 1 and 2, so we subtracted 1.0 from each index to produce a measure that varied between 0 and 1.

The fourth dimension of segregation is centralization, which is the degree to which a group is located near the center of an urban area. During the 1960s and 1970s, blacks were increasingly isolated in central cities, away from suburban areas where whites congregated increasingly (Farley et al., 1978; Massey and Denton, 1988b). Centralization is measured by an index that reflects the extent to which a group is spatially distributed close to, or far away from, the central business district (CBD). It compares a group's distribution by distance from the CBD to the distribution of land area around the CBD by using a formula adapted from Duncan (1957), Duncan, Cuzzort, and Duncan (1961), and Glaster (1984):

$$CE = \left(\sum_{i=1}^{n} X_{i-1} A_i\right) - \left(\sum_{i=1}^{n} X_i A_{i-1}\right), \tag{6}$$

where the n areal units are ordered by increasing distance from the central business district and X_i and A_i are the respective cumulative proportions of X's population and land area

in tract i. This index varies between $+1$ and -1, with positive values indicating a tendency for group X members to reside close to the city center and negative values indicating a tendency to live in outlying areas. A score of 0 means that the group has a uniform distribution throughout the metropolitan area. The index therefore gives the proportion of X members required to change residence to achieve a uniform distribution of population around the central business district.

The last dimension of segregation that we consider is concentration, which is the relative amount of physical space occupied by a minority group in the urban environment. Concentration is a relevant dimension of segregation because discrimination often restricts minorities to a small number of neighborhoods that together comprise a small share of the urban environment (Hirsch, 1983; Kain and Quigley, 1975; Spear, 1967). It is measured by computing the average amount of physical space occupied by group X relative to group Y and comparing this quantity with the ratio that would be achieved if group X were maximally concentrated and group Y were maximally dispersed. This relative concentration index is computed as follows:

$$CO = \frac{\left[\sum_{i=1}^{n} \frac{x_i a_i}{X} \Big/ \sum_{i=1}^{n} \frac{y_i a_i}{Y}\right] - 1}{\left[\sum_{i=1}^{n_1} \frac{t_i a_i}{T_1} \Big/ \sum_{i=n_2}^{n} \frac{t_i a_i}{T_2}\right] - 1}, \tag{7}$$

where areal units are ordered by geographic size from smallest to largest, a_i is the land area of unit i, and the two numbers n_1 and n_2 refer to different points in the rank ordering of areal units from smallest to largest: n_1 is the rank of the tract where the cumulative total population of areal units equals the total minority population of the city, summing from the smallest unit up; n_2 is the rank of the tract where the cumulative total population of units equals the minority population totalling from the largest unit down. T_1 equals the total population of tracts from 1 to n_1, and T_2 equals the total population of tracts from n_2 to n. As before, t_i refers to the total population of area i and X is the number of group X members in the city.

The numerator of this index divides the average land area of units inhabited by group X members by the average area of units inhabited by Y members, and the denominator takes the average that would be obtained if X members lived in the smallest space possible and divides it by the average that would be obtained if Y members fit into the largest possible area. The quotient is then standardized to vary between -1.0 and $+1.0$. A score of 0 means that the two groups are equally concentrated in urban space. A score of -1.0 means that Y's concentration exceeds X's to the maximum extent possible, and a score of 1.0 means the converse.

Spatial Segregation of Blacks

These five indices were computed for blacks in 60 metropolitan areas and are reported in Table 1. Since measures of evenness and isolation were analyzed in detail elsewhere (Massey and Denton, 1987, 1988b), we focus on the remaining three dimensions of segregation. Intercorrelations between the measures are shown at the bottom of the table. They range from 0.105 to 0.877 and average 0.525. Although the five dimensions overlap empirically, no index perfectly replicates another. Two indices share at most 77 percent common variance and at the least only 1 percent. In general, the evenness, exposure, and clustering indices are more highly intercorrelated than the centralization and concentration measures. The interrelationships among the indices were discussed in detail in our earlier article (Massey and Denton, 1988a).

Five key metropolitan areas with large minority populations are highlighted at the top of the table, and regional and national averages are reported at the bottom. Measures of black clustering are shown in the *SP* columns of Table 1. In general *SP* indices above 0.600 are very high and imply the existence of a large enclave of contiguous tracts containing most blacks. Indices between 0.400 and 0.600 are still high but indicate the presence of scattered black neighborhoods away from the principal ghetto. *SP* values between 0.100 and 0.400 are moderate and correspond to a pattern of scattered black and racially mixed neighborhoods. Finally, indices under 0.100 are very low, indicating a spatial configuration dominated by racially mixed neighborhoods that are widely scattered about the city (see Massey and Denton, 1988a; White, 1983, 1986).

In most cities, the clustering of blacks is moderate or low. The average *SP* index for all 60 SMSAs is only 0.292, and 14 metropolitan areas have indices in the lower range (under 0.100). Another 29 display indices that are in the moderate range (under 0.400). Clustering is notably lower in Western SMSAs, with a regional average of only 0.141, as well as in the South, where the average is 0.259. In short, blacks in the vast majority of metropolitan areas do not live in a spatially distinct ghetto of contiguous minority tracts.

Despite the scant evidence of clustering in most metropolitan areas, spatial agglomeration is pronounced in SMSAs with large black populations. The lowest clustering indices are generally observed in metropolitan areas with very few black residents, such as Albany, Albuquerque, Bakersfield, Minneapolis, Sacramento, and Tucson. Although SMSAs with clustering indices in the high or very high range are few in number, they generally include areas with the largest urban black populations in the United States. Nine SMSAs have *SP* indices of 0.600 or more, including Chicago, Los Angeles, Baltimore, Cleveland, Detroit, Newark, Philadelphia, and Milwaukee. Metropolitan areas with indices in the 0.400 to 0.600 range include New York, Atlanta, Gary, Kansas City, Memphis, Washington, Buffalo, Boston, and Indianapolis.

As is obvious from this list, the clustering of black neighborhoods is especially prevalent in older industrial areas of the Northeast and Midwest. The regional average was 0.474 among SMSAs in North Central states and 0.368 among those in the Northeast. In these areas, blacks segregated on one dimension also tend to be segregated on others. Among the nine SMSAs with clustering indices above 0.600, seven had dissimilarity indices of 0.800 or more and all were greater than 0.750. All of these SMSAs had P^* isolation indices in excess of 0.600, and six of the nine areas displayed indices greater than 0.700.

The *CE* column of Table 1 contains indices of black centralization, which measure the extent to which blacks are distributed closely round the central business district. We found in earlier work that blacks have little access to the suburbs of U.S. cities (Massey and Denton 1987, 1988b), so it is not surprising to find that most SMSAs display very high levels of black centralization. In general, a *CE* index above 0.800 is very high, indicating that 80 percent of the black population would have to move to be uniformly distributed in the urban environment. More than two-thirds of the metropolitan areas (43 of 60) display centralization indices of 0.800 or more; and this list contains all SMSAs with high or very high clustering indices. Only eight metropolitan areas have centralization indices below 0.600: Miami, Anaheim, Ft. Lauderdale, Greensboro, Jersey City, Salt Lake City, Tampa, and San Antonio.

The *CO* column in Table 1 displays black concentration indices, which indicate the extent to which blacks occupy a small amount of urban space relative to Anglos. In this context, an index value of 0.700 or greater indicates a high level of concentration, with black residents being packed into a limited number of geographically small census tracts. Blacks in 28 of the SMSAs—nearly half—experience a high level of spatial concentration. This list includes 14 of the 17 SMSAs we have already identified as being highly or very highly segregated on the dimensions of clustering and centralization, including Chicago,

Table 1. Five Indices of Black Residential Segregation in 60 U.S. SMSAs in 1980

Metropolitan area	Segregation index				
	D	${}_bP_b^*$	*SP*	*CE*	*CO*
Key SMSAs					
Chicago	0.878	0.828	0.793	0.872	0.887
Los Angeles–Long Beach	0.811	0.604	0.765	0.859	0.695
Miami	0.778	0.642	0.344	0.463	0.565
New York	0.819	0.627	0.468	0.795	0.892
San Francisco–Oakland	0.717	0.511	0.282	0.836	0.687
Other SMSAs					
Albany–Schenectady–Troy	0.622	0.276	0.088	0.848	0.748
Albuquerque	0.390	0.050	0.008	0.795	0.371
Anaheim–Santa Ana–Garden Grove	0.458	0.038	0.018	0.576	−0.442
Atlanta	0.762	0.714	0.398	0.827	0.686
Austin	0.608	0.349	0.123	0.778	0.567
Bakersfield	0.644	0.346	0.101	0.827	0.652
Baltimore	0.747	0.723	0.622	0.857	0.763
Birmingham	0.419	0.496	0.059	0.830	0.775
Boston	0.774	0.550	0.491	0.871	0.799
Buffalo	0.794	0.635	0.443	0.884	0.882
Cincinnati	0.723	0.543	0.158	0.883	0.669
Cleveland	0.875	0.804	0.743	0.898	0.927
Columbus	0.724	0.571	0.321	0.933	0.854
Corpus Christi	0.717	0.267	0.130	0.910	0.793
Dallas–Fort Worth	0.771	0.645	0.334	0.749	0.693
Dayton	0.780	0.650	0.336	0.861	0.600
Denver–Boulder	0.685	0.410	0.211	0.719	0.385
Detroit	0.867	0.773	0.846	0.924	0.842
El Paso	0.347	0.050	0.013	0.687	0.382
Fort Lauderdale	0.816	0.702	0.237	0.593	0.784
Fresno	0.624	0.377	0.159	0.968	0.598
Gary–Hammond–East Chicago	0.906	0.773	0.561	0.887	0.869
Greensboro–Winston-Salem	0.564	0.496	0.053	0.601	0.613
Houston	0.695	0.593	0.238	0.840	0.569
Indianapolis	0.762	0.623	0.411	0.942	0.804
Jersey City	0.765	0.604	0.335	0.560	0.555
Kansas City	0.789	0.689	0.461	0.921	0.857
Louisville	0.718	0.628	0.249	0.894	0.699
Memphis	0.695	0.737	0.440	0.817	0.550
Milwaukee	0.839	0.695	0.689	0.951	0.944
Minneapolis–St. Paul	0.693	0.306	0.102	0.944	0.890
Nashville–Davidson	0.647	0.551	0.244	0.744	0.628
Nassau–Suffolk	0.755	0.469	0.179	0.643	0.194
New Orleans	0.683	0.688	0.327	0.906	0.584
Newark	0.816	0.692	0.755	0.859	0.919

(Table continues)

Table 1. Continued

Metropolitan area	Segregation index				
	D	${}_bP_b^*$	*SP*	*CE*	*CO*
Norfolk–Virginia Beach	0.628	0.625	0.199	0.712	0.559
Oklahoma City	0.710	0.560	0.250	0.886	0.546
Paterson–Clifton–Passaic	0.815	0.489	0.277	0.876	0.929
Philadelphia	0.788	0.696	0.673	0.855	0.757
Phoenix	0.594	0.225	0.041	0.945	0.548
Pittsburgh	0.727	0.541	0.272	0.812	0.821
Portland	0.685	0.316	0.168	0.956	0.826
Providence–Warwick–Pawtucket	0.731	0.253	0.120	0.818	0.803
Riverside–San Bernadino	0.488	0.160	0.048	0.896	0.212
Rochester	0.679	0.437	0.321	0.874	0.792
Sacramento	0.559	0.209	0.096	0.900	0.509
St. Louis	0.814	0.729	0.264	0.931	0.893
Salt Lake City–Ogden	0.533	0.041	0.006	0.443	0.384
San Antonio	0.641	0.358	0.229	0.523	0.544
San Diego	0.643	0.263	0.171	0.902	0.537
San Jose	0.487	0.066	0.032	0.795	0.177
Seattle–Everett	0.682	0.294	0.137	0.952	0.791
Tampa–St. Petersburg	0.735	0.507	0.246	0.581	0.493
Tucson	0.466	0.088	0.014	0.910	0.253
Washington, D.C.	0.693	0.672	0.450	0.850	0.441
Averages					
Total	0.693	0.488	0.292	0.816	0.642
Northeast	0.757	0.522	0.368	0.808	0.757
North Central	0.804	0.665	0.474	0.912	0.836
South	0.669	0.550	0.259	0.752	0.612
West	0.592	0.250	0.141	0.830	0.449
Intercorrelations					
D	1.000	0.795	0.856	0.169	0.702
*P**	0.795	1.000	0.877	0.105	0.528
SP	0.856	0.877	1.000	0.175	0.575
CE	0.169	0.105	0.175	1.000	0.466
CO	0.702	0.528	0.575	0.466	1.000

New York, Los Angeles, Detroit, Cleveland, Gary, Newark, Philadelphia, and Baltimore—in other words, the largest black settlements in the Northern states.

We have thus identified a significant core of large metropolitan areas in which blacks are highly segregated on multiple dimensions. This conclusion is supported visually by the three panels of Figure 1, which plot indices for each of the three spatial dimensions against the index of dissimilarity. SMSAs on the plot are indicated by two-letter codes, which are paired with the metropolitan areas in Table 2. Eight SMSAs with very high *SP* values are circumscribed by an oval, and those in the moderately high range are enclosed by a rectangle.

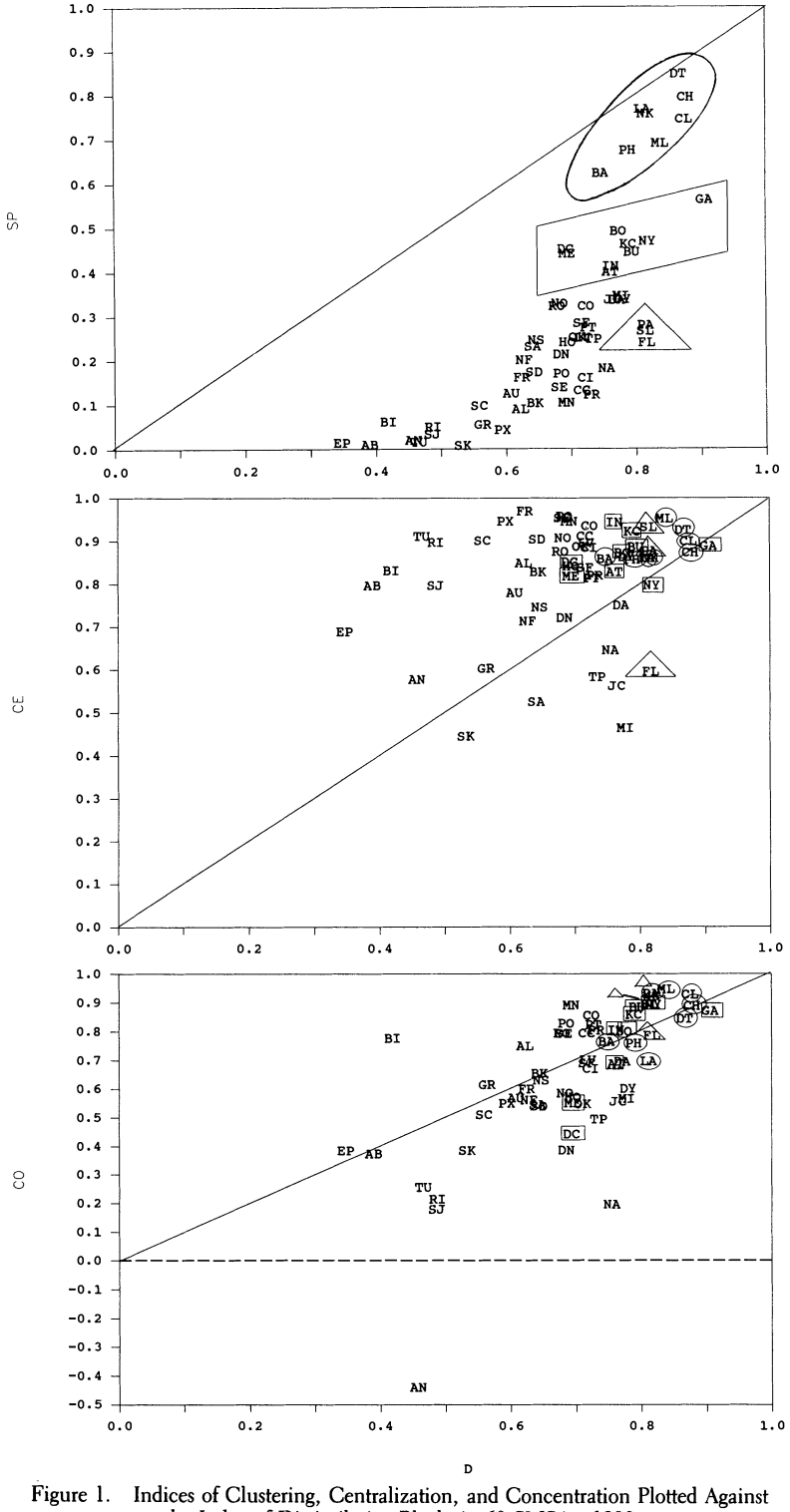


Figure 1. Indices of Clustering, Centralization, and Concentration Plotted Against the Index of Dissimilarity: Blacks in 60 SMSAs, 1980

The figure demonstrates that SMSAs with high *SP* indices generally also have high levels of dissimilarity. Three SMSAs, however, have high dissimilarity measures but only moderate clustering indices—Paterson, St. Louis, and Ft. Lauderdale—and these are enclosed by a triangle.

SMSAs that were enclosed as groups by an oval, a rectangle, or a triangle in the top panel of the figure are enclosed individually by those symbols in the middle and lower panels, so the relative positions of the three sets of SMSAs can be compared and contrasted across panels. The middle panel plots indices of centralization and shows that the same SMSAs that were highly clustered are also highly centralized. They are packed tightly in the upper right sector of the plot, just above the diagonal. The bottom panel shows that the same SMSAs have high levels of geographic concentration, again being packed tightly just above the diagonal in the upper right sector of the scatterplot. Although the plots are not shown, the same SMSAs have high isolation P^*s and display a similar pattern when plotted against D.

We can thus identify an important subset of major urban areas in which blacks are very highly segregated on all five dimensions of residential segregation. If we let "high segregation" mean a dissimilarity index of 0.600 or more, an isolation P^* of 0.700 or more,

Table 2. Names and Abbreviations of SMSAs Used in Study of Spatial Segregation

Metropolitan area	Abbreviation	Metropolitan area	Abbreviation
Albany	AL	Memphis	Me
Albuquerque	AB	Miami	MI
Anaheim	AN	Milwaukee	ML
Atlanta	AT	Minneapolis	MN
Austin	AU	Nashville	NS
Bakersfield	BK	Nassau–Suffolk	NA
Baltimore	BA	New Orleans	NO
Birmingham	BI	New York	NY
Boston	BO	Newark	NK
Buffalo	BU	Norfolk	NF
Chicago	CH	Oklahoma City	OK
Cincinnati	CI	Paterson	PA
Cleveland	CL	Philadelphia	PH
Columbus	CO	Phoenix	PX
Corpus Christi	CO	Pittsburgh	PT
Dallas	DA	Portland	PO
Dayton	DY	Providence	PR
Denver	DN	Riverside	RI
Detroit	DT	Rochester	RO
El Paso	EP	Sacramento	SC
Fort Lauderdale	FL	St. Louis	SL
Fresno	FR	Salt Lake City	SK
Gary	GA	San Antonio	SA
Greensboro	GR	San Diego	SD
Houston	Ho	San Francisco	SF
Indianapolis	IN	San Jose	SJ
Jersey City	JC	Seattle	SE
Kansas City	KC	Tampa	TP
Los Angeles	LA	Tucson	TU
Louisville	LV	Washington, D.C.	DC

an *SP* index of 0.600 or greater, a centralization score of 0.800 or higher, and a concentration index in excess of 0.700, then blacks in six SMSAs are highly segregated on all five dimensions (Baltimore, Chicago, Cleveland, Detroit, Milwaukee, and Philadelphia), and in another four SMSAs they are segregated on four dimensions (Gary, Los Angeles, Newark, and St. Louis). Together these 10 SMSAs contain 29 percent of metropolitan blacks and 23 percent of all blacks in the United States.

In short, roughly one-quarter of the American black population lives in an urban environment that is hypersegregated. Blacks in these cities are very unevenly distributed among tracts and live in small, densely settled, monoracial neighborhoods that are part of large agglomerations of contiguous tracts clustered tightly around the city center. Residents of such an environment would be very unlikely to come into regular contact with a member of Anglo society, except through participation in the labor force, an option that is denied to the quarter of central-city blacks who are under- or unemployed (Lichter, 1988). Blacks without jobs would rarely meet, and would be extremely unlikely to know, an Anglo resident of the same metropolis.

On the other hand, if we establish very liberal criteria for defining a "low" level of black segregation (e.g., $D < 0.600$, $P^* < 0.500$, $SP < 0.600$, $CE < 0.800$, and $CO < 0.700$), then blacks in nine SMSAs experience low segregation on at least four of the five dimensions: Albuquerque, Anaheim, El Paso, Greensboro, Salt Lake City, San Jose, Phoenix, Riverside, and Tucson. But together these SMSAs contain only 2 percent of metropolitan blacks and only 1.5 percent of all blacks in the United States, so very few blacks experience a residential pattern that might be called "integrated."

Spatial Segregation of Hispanics

The distinctiveness of the residential situation faced by blacks is emphasized by the data in Table 3, which presents indices of dissimilarity, isolation, clustering, centralization, and concentration for Hispanics in the 60 metropolitan areas. No Hispanic clustering index was high ($SP > 0.600$) or even moderately high ($SP > 0.400$) by black standards. The Hispanic *SP* indices generally fell into the moderate range, with 4 SMSAs lying between 0.300 and 0.400 (San Antonio, Los Angeles, Chicago, and Fresno) and 10 located between 0.150 and 0.300 (New York, Newark, Miami, El Paso, Corpus Christi, Bakersfield, Philadelphia, Paterson, San Diego, and Albuquerque). At the same time, relatively few SMSAs displayed high Hispanic centralization or concentration indices. Whereas black centralization exceeded 0.800 in 43 cases, only 19 Hispanic indices did so; and only 9 SMSAs evinced Hispanic concentration indices of 0.700 or more, compared with 28 for blacks.

In general, low to moderate levels of segregation were observed for Hispanics on all dimensions. The average level of dissimilarity was 0.436 (compared with 0.693 for blacks), with average indices of isolation, clustering, centralization, and concentration of 0.201, 0.090, 0.713, and 0.398, respectively (compared with indices of 0.488, 0.292, 0.816, and 0.642 for blacks). Even in SMSAs with very large Spanish population, such as Los Angeles, San Antonio, Miami, New York, and Chicago, there was little evidence of high segregation on multiple dimensions. For example, the largest concentration of Hispanics in the United States is in Los Angeles, where people of Spanish origin number more than 2 million and represent 28 percent of the metropolitan population. The respective indices of segregation for Hispanic Angelinos were, in the same order as before, 0.570, 0.501, 0.333, 0.772, and 0.619. None of these values would be considered high by black standards.

In general, then, high levels of Hispanic segregation do not appear to correlate strongly across dimensions, and in no SMSA do Hispanics experience the multidimensional hypersegregation of blacks. The relative independence of the indices is also evident in Figure 2, which plots Hispanic clustering, centralization, and concentration indices against the index

of dissimilarity. In the top panel, the two highest sets of clustering indices are enclosed by an oval and a rectangle; and in subsequent panels, SMSAs from these groups are identified individually by these symbols. In the middle panel, centralization indices are obviously much more dispersed than was true for blacks. The ovals and rectangles are scattered widely rather than concentrated in the upper right sector of the graph. Concentration indices in the bottom panel are even more scattered, with no detectable grouping of ovals or rectangles.

If we adopt the same criteria used to define high segregation for blacks ($D > 0.600$, $P^* > 0.700$, $SP > 0.600$, $CE > 0.800$, and $CO > 0.700$) and consider the multidimensional structure of Hispanic segregation, the contrast between the two groups stands out clearly. In no metropolitan area are Hispanics highly segregated on five or even four dimensions, and in only four areas are they segregated on as many as three dimensions—Chicago, New York, Newark, and Paterson. Three of these areas are dominated by Puerto Rican populations, which display high levels of segregation compared with other Hispanic groups, a pattern that has been attributed to the Afro-American ancestry of this group (Massey and Bitterman, 1985). Moreover, several of the largest Hispanic concentrations in the United States are not highly segregated on *any* dimension at all, including Los Angeles, San Antonio, Miami, and San Diego. Indeed, a lack of high segregation on any dimension is the most common pattern for Hispanics; among the 60 SMSAs in the data set, 37 were not highly segregated on any of the five dimensions. Thus not only are Hispanics less segregated than blacks on any single dimension, they are very unlikely to accumulate high levels of segregation across multiple dimensions simultaneously.

The Hispanic–Black Differential in Multivariate Perspective

It thus appears that blacks occupy a unique position in the American urban landscape. They are more segregated than Hispanics on every dimension of segregation, and in an important core of metropolitan areas—primarily older industrial areas of the Northeast and Midwest—they are extremely segregated on all five dimensions simultaneously, an unusual condition we have called hypersegregation. This condition is not replicated anywhere by Hispanics or by any other group we have examined (see Langberg and Farley, 1985, and Massey and Denton, 1987, 1988b, for data on Asian segregation).

We hesitate, however, to make a strong statement about the relative segregation of blacks and Hispanics in U.S. metropolitan areas, since the two groups differ on many variables that directly influence patterns of residential location. Differences in regional concentration, relative population size, nativity composition, socioeconomic status, and local economic conditions could account for all or part of the black–Hispanic differential in the SMSAs under study, and it would be wrong to infer that black segregation is exceptional from a descriptive study of the indices alone.

To test how robust the apparent contrast between blacks and Hispanics is, we estimated multivariate models of segregation that directly compare the two groups, controlling for possible confounding variables. Table 4 presents regression equations measuring the impact of selected explanatory factors on each of the five dimensions of segregation. We pool black and Hispanic indices for each dimension and then regress them on a set of factors that are theoretically expected to influence the level of minority segregation. For each of the five regressions, a dummy variable under the heading "Minority group" indicates whether the segregation index pertains to blacks (Blacks = 1) or Hispanics (Blacks = 0). If the black–Hispanic differential is explained by variables in the model, then the coefficient for this dummy variable should be statistically insignificant.

Four of the five indices of segregation were transformed into logits before undertaking the regression analyses, since their limited range (0–1) would bias ordinary least squares (OLS) estimates. For any limited-range variable P, the logit transformation—logit(P) =

Table 3. Five Indices of Hispanic Residential Segregation for 60 SMSAs in 1980

Metropolitan area	Segregation index				
	D	${}_hP_h^*$	*SP*	*CE*	*CO*
Key SMSAs					
Chicago	0.635	0.380	0.317	0.813	0.746
Los Angeles–Long Beach	0.570	0.501	0.333	0.772	0.619
Miami	0.519	0.583	0.240	0.542	0.360
New York	0.657	0.399	0.263	0.841	0.878
San Francisco–Oakland	0.402	0.193	0.083	0.628	0.340
Other SMSAs					
Albany–Schenectady–Troy	0.367	0.036	0.006	0.499	0.358
Albuquerque	0.429	0.505	0.149	0.768	0.470
Anaheim–Santa Ana–Garden Grove	0.416	0.310	0.115	0.635	0.449
Atlanta	0.337	0.021	0.003	0.696	0.349
Austin	0.449	0.336	0.100	0.639	0.454
Bakersfield	0.545	0.421	0.197	0.761	0.401
Baltimore	0.381	0.015	0.007	0.657	0.306
Birmingham	0.226	0.009	0.001	0.625	0.461
Boston	0.579	0.129	0.083	0.788	0.705
Buffalo	0.491	0.077	0.028	0.808	0.590
Cincinnati	0.303	0.010	0.001	0.704	0.236
Cleveland	0.554	0.082	0.047	0.842	0.704
Columbus	0.350	0.013	0.003	0.789	0.414
Corpus Christi	0.516	0.636	0.225	0.644	0.712
Dallas–Fort Worth	0.478	0.240	0.085	0.732	0.511
Dayton	0.328	0.010	0.002	0.702	0.282
Denver–Boulder	0.475	0.274	0.104	0.778	0.494
Detroit	0.451	0.065	0.062	0.746	0.374
El Paso	0.512	0.741	0.223	0.737	0.145
Fort Lauderdale	0.255	0.053	0.008	0.307	−0.065
Fresno	0.454	0.446	0.286	0.800	−0.221
Gary–Hammond–East Chicago	0.562	0.237	0.105	0.835	0.694
Greensboro–Winston-Salem	0.321	0.010	0.001	0.495	0.318
Houston	0.464	0.328	0.119	0.818	0.532
Indianapolis	0.332	0.012	0.004	0.777	0.392
Jersey City	0.488	0.465	0.108	0.129	0.606
Kansas City	0.422	0.104	0.031	0.854	0.508
Louisville	0.281	0.009	0.001	0.645	0.219
Memphis	0.406	0.013	0.004	0.707	0.206
Milwaukee	0.562	0.162	0.078	0.848	0.718
Minneapolis–St. Paul	0.418	0.046	0.013	0.860	0.498
Nashville–Davidson	0.371	0.012	0.003	0.559	0.158
Nassau–Suffolk	0.362	0.096	0.027	0.606	0.222
New Orleans	0.251	0.063	0.029	0.809	0.250
Newark	0.656	0.263	0.255	0.807	0.796

(Table continues)

Table 3. Continued

Metropolitan area	Segregation index				
	D	${}_hP_h^*$	*SP*	*CE*	*CO*
Norfolk–Virginia Beach	0.284	0.020	0.003	0.721	−0.026
Oklahoma City	0.316	0.054	0.011	0.804	0.338
Paterson–Clifton–Passaic	0.722	0.375	0.190	0.802	0.900
Philadelphia	0.629	0.216	0.193	0.780	0.549
Phoenix	0.494	0.321	0.077	0.925	0.348
Pittsburgh	0.419	0.013	0.003	0.642	0.211
Portland	0.250	0.028	0.005	0.782	−0.081
Providence–Warwick–Pawtuckett	0.567	0.085	0.038	0.718	0.716
Riverside–San Bernadino	0.364	0.316	0.101	0.829	0.338
Rochester	0.580	0.116	0.081	0.808	0.688
Sacramento	0.364	0.165	0.054	0.756	−0.030
St. Louis	0.340	0.019	0.003	0.754	0.468
Salt Lake City–Ogden	0.308	0.090	0.013	0.216	0.206
San Antonio	0.569	0.665	0.384	0.532	0.660
San Diego	0.421	0.269	0.185	0.793	0.140
San Jose	0.445	0.317	0.118	0.729	0.109
Seattle–Everett	0.213	0.026	0.003	0.846	0.222
Tampa–St. Petersburg	0.489	0.175	0.071	0.701	0.211
Tucson	0.519	0.431	0.122	0.866	0.219
Washington, D.C.	0.307	0.054	0.017	0.758	0.517
Averages					
Total	0.436	0.201	0.090	0.713	0.398
Northeast	0.543	0.189	0.106	0.686	0.602
North Central	0.438	0.095	0.055	0.794	0.503
South	0.387	0.202	0.077	0.656	0.331
West	0.417	0.288	0.122	0.743	0.251
Intercorrelations					
D	1.000	0.699	0.786	0.324	0.656
*P**	0.699	1.000	0.951	0.185	0.321
SP	0.786	0.951	1.000	0.243	0.400
CD	0.324	0.185	0.243	1.000	0.306
CO	0.656	0.321	0.400	0.306	1.000

$\ln[P/(1 - P)]$—creates a new variable ranging from negative to positive infinity, thus enabling the use of OLS estimation procedures. The concentration index was not transformed because it included a few negative values, reflecting its theoretical range from -1 to $+1$.

The explanatory variables fall into one of five categories: indicators of acculturation, socioeconomic status, population composition, regional location, metropolitan context, and sample selectivity. Specific variables were selected by following a line of empirical research and theoretical reasoning developed in our earlier papers (Massey and Denton, 1987, 1988b). Initial tests revealed few problems with nonlinearity or multicolinearity among variables in the equations, and controls for compositional diversity among Hispanics proved to be in-

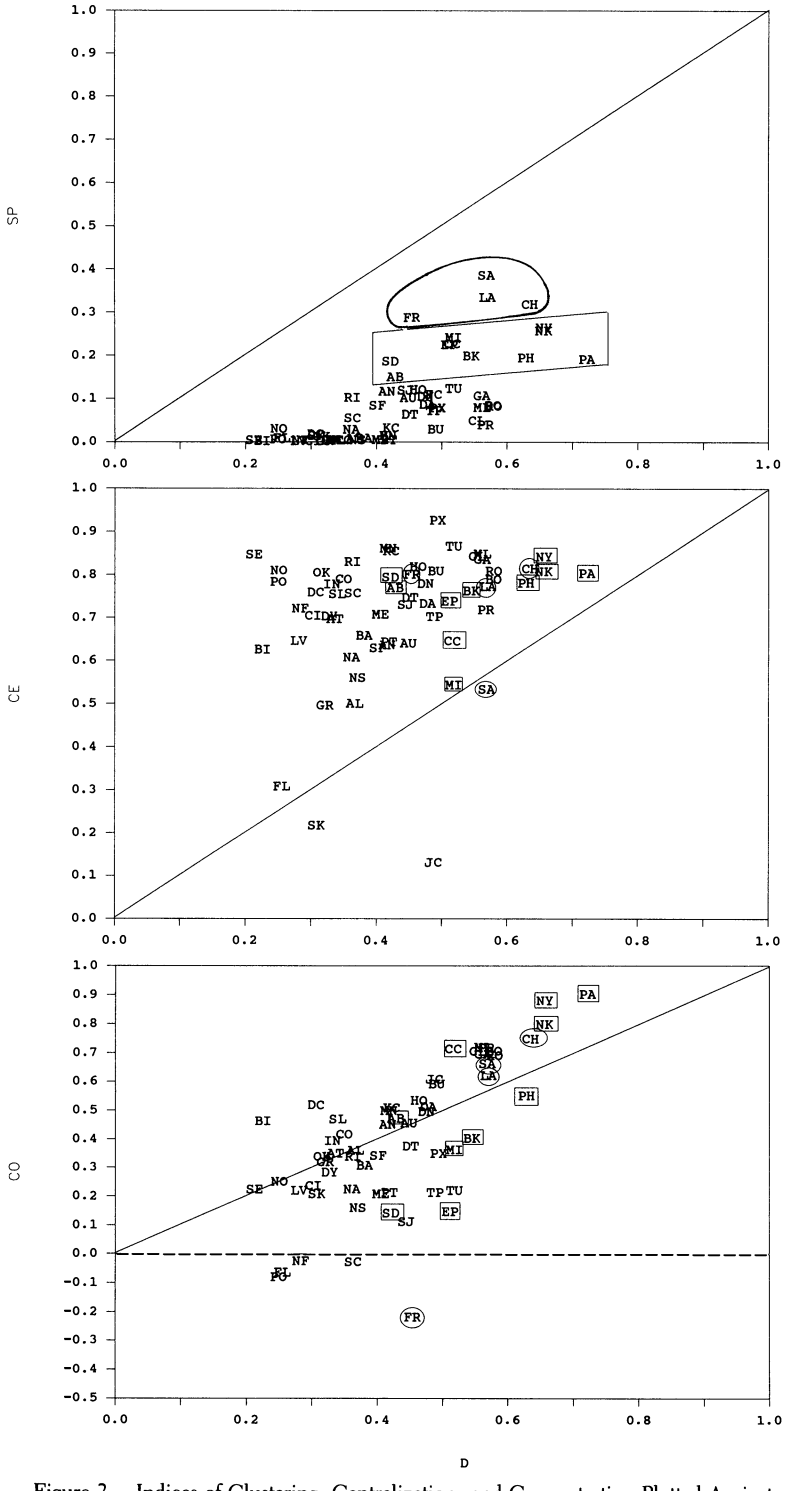


Figure 2. Indices of Clustering, Centralization, and Concentration Plotted Against the Index of Dissimilarity: Hispanics in 60 SMSAs, 1980

Table 4. Regression of Five Segregation Indices on Selected Explanatory Variables: 60 SMSAs in 1980

Explanatory variables	Dependent variables									
	D		*P**		*SP*		*CE*		*CO*	
	B	SE	B	SE	B	SE	B	SE	B	SE
Acculturation										
Proportion native born	0.055	0.446	−0.897	0.740	−0.066	1.170	0.868	0.900	0.047	0.214
Socioeconomic status										
Occupational status	−0.096[a]	0.018	−0.222[a]	0.030	−0.248[a]	0.048	−0.015	0.037	−0.040[a]	0.009
Family income	−0.022	0.018	−0.021	0.030	−0.028	0.047	−0.078[a]	0.036	−0.020[a]	0.009
Population										
Composition										
Minority proportion	1.429[a]	0.414	7.138[a]	0.687	6.528[a]	1.085	−0.279	0.835	0.010	0.198
Proportion black Hispanic	0.295	0.844	−4.647[a]	1.402	−6.675[a]	2.215	−0.627	1.705	−0.112	0.405
Minority group										
Blacks	1.114[a]	0.177	1.654[a]	0.294	1.430[a]	0.465	0.271	0.357	0.193[a]	0.085
Region										
Northeast	0.495[a]	0.230	0.265	0.381	0.754	0.602	−0.149	0.464	0.032	0.110
North Central	0.573[a]	0.167	0.272	0.277	0.641	0.437	0.512	0.337	0.210[a]	0.080
South	0.066	0.130	0.003	0.217	0.163	0.342	−0.558[a]	0.263	0.092	0.080
Metropolitan context										
Housing inflation	1.085	1.770	0.573	2.939	6.239	4.645	−0.675	3.574	−0.040	0.849
Employment growth	−0.162	3.356	−5.349	5.572	−7.245	8.806	2.096	6.776	−0.330	1.610
Minority immigration	0.251	0.524	0.526	0.871	−0.304	1.376	0.004	1.059	0.145	0.252
Growth differential	−1.027	2.001	−7.283[a]	3.321	−2.855	5.249	−2.065	4.039	−0.853	0.960
Age of housing	0.002	0.011	−0.006	0.019	0.003	0.029	−0.016	0.023	0.011[a]	0.005
Selectivity										
$P - 1$	0.749[a]	0.175	1.268[a]	0.290	2.355[a]	0.459	0.672[a]	0.354	0.212[a]	0.084
Intercept	2.647[a]	0.887	6.036[a]	1.473	4.722[a]	2.328	2.771	1.792	1.655[a]	0.426
R^2 (adjusted)	0.803[a]		0.881[a]		0.787[a]		0.312[a]		0.607[a]	
n	120		120		120		120		120	

Note: B = regression coefficient; SE = standard error.

[a] $p < 0.05$.

significant and were eliminated. A full description of the explanatory variables, the theoretical model, and the selectivity correction is given in Massey and Denton (1987).

The estimates of Table 4 generally confirm the exceptional nature of black segregation in U.S. metropolitan areas. The large discrepancy between black and Hispanic segregation indices observed in the earlier tables cannot be accounted for by the explanatory factors that we have identified. The coefficient for black minority status is large and highly significant in four of the five regression equations. Black race is particularly significant in the equations for dissimilarity and spatial isolation, where the coefficient exceeds its standard error by a factor of about six. All four of the equations fit the data well, accounting for between 61 and 88 percent of the intermetropolitan variance in segregation. The only dimension on which black race was not significant was centralization, the equation that most poorly fits the data, explaining only 31 percent of the variance in the *CE* index.

In other words, controlling for a variety of possible confounding factors, blacks are significantly more segregated than Hispanics on four of five dimensions of residential segregation. Apart from black minority status, segregation was strongly reduced by rising socioeconomic status; and on three dimensions (dissimilarity, isolation, and clustering), it was strongly increased by a high proportion of minority members. A relatively large number of black Hispanics reduced the level of Hispanic isolation and clustering, probably by promoting intergroup contact with blacks. Regional location in the Northeast and North Central states increased the level of dissimilarity, and a North Central location increased the level of concentration. In all equations, the selectivity coefficient was highly significant, indicating that the large metropolitan areas chosen for study are considerably more segregated than the smaller ones we left out, creating a selection bias that was corrected by using the technique of Olsen (1980).

Conclusion

Many earlier studies have documented the persistent and high degree of black residential segregation in U.S. metropolitan areas (Duncan and Duncan, 1957; Farley, 1977; Massey, 1979; Massey and Denton, 1987; Sorensen, Taeuber, and Hollingsworth, 1975; Taeuber and Taeuber, 1965). This investigation not only confirms these earlier studies but suggests that black segregation is even more extreme than previously imagined. By focusing on the index of dissimilarity, and more recently on measures of exposure, earlier work has understated the unique situation of blacks in American urban areas and has not appreciated the full extent of their segregation in U.S. society. Alone among U.S. minority groups, blacks often face conditions of hypersegregation.

Being black not only greatly accentuates the level of segregation on any single dimension but also increases markedly the dimensionality of segregation, generating an accumulation of segregation across multiple dimensions simultaneously. From our descriptive analyses, we identified a significant core of 10 large metropolitan areas within which blacks are very highly segregated on at least four dimensions of residential segregation. These areas contain 29 percent of all urban blacks in the United States. They include Baltimore, Chicago, Cleveland, Detroit, Milwaukee, and Philadelphia—which are highly segregated on all five dimensions—as well as Gary, Los Angeles, New York, and St. Louis—which are highly segregated on four dimensions.

In no SMSA were Hispanics highly segregated on more than three dimensions simultaneously, and in 37 of the 60 SMSAs, they were not highly segregated on any dimension at all. Even in large Hispanic settlements such as Los Angeles, Miami, San Antonio, San Francisco, and San Diego, segregation was low or moderate on all dimensions. In other words, not only is the average level of Hispanic segregation lower on any given dimension,

but there is a striking absence of the multidimensional layering of high segregation across dimensions. To be sure, layering does occur in a few cities; but it is always at a moderate level. In SMSAs such as San Antonio, Miami, and Corpus Christi, Hispanics are moderately but consistently segregated across all five dimensions, implying a more restricted social environment than if they displayed low segregation on some dimensions. Hispanics never, however, display both multidimensional layering and high segregation.

Blacks are thus unique in experiencing multidimensional hypersegregation. The contrast between them and Hispanics is not easily explained by different socioeconomic characteristics, varying population sizes, different regional locations, or contrasting metropolitan conditions. Although our models cannot eliminate the view that some unmeasured objective factor accounts for the discrepancy between blacks and Hispanics, the models lend credence to the view that blacks remain the object of significantly higher levels of Anglo prejudice than Hispanics. Two decades after the 1968 Civil Rights Act, blacks still have not achieved the freedom to live where they want.

These results underscore the complexity of urban segregation patterns and the extent to which they have been oversimplified in the past by using one or two indices. Groups differ not only in the degree of their segregation but also in the dimensional structure of their segregation. A minority that is highly segregated on only one dimension is "less segregated," in a very real sense, than one highly segregated on five. Likewise, a group that is moderately segregated on five dimensions is "more segregated" than one displaying low levels on four and a moderate level on the fifth. Recognizing five distinct dimensions of segregation yields considerably more information than using one dimension by itself.

An appreciation of the multidimensional structure of segregation is especially important in the case of blacks. Segregation becomes more profound as it accumulates across dimensions, and hypersegregation across five dimensions simultaneously implies a level of spatial isolation that is much greater than heretofore recognized. From studies based on the index of dissimilarity, it has been known for some time that blacks are unevenly distributed in many metropolitan areas, meaning that most tracts where blacks live contain a disproportionate number of black residents. Our results, however, paint a more extreme picture. Not only are blacks in our largest cities disproportionately likely to share tracts with other blacks, they are very unlikely to share a tract with any whites at all. Moreover, if they go to the adjacent neighborhood, or to the neighborhood adjacent to that, they are still unlikely to encounter a white resident. These agglomerations of monoracial tracts are densely settled and geographically restricted, comprising a small portion of the urban environment closely packed around the city center, a zone known for poverty and social disorganization long before blacks arrived there (Park and Burgess, 1925).

This extreme level of residential segregation across multiple dimensions is important because of the social isolation it implies. For blacks in large ghettos of the north, this isolation must be extreme. Unless a resident of these ghettos works in the Anglo-dominated economy, he or she is unlikely to come into contact with anyone other than another black ghetto-dweller. Indicators of the accompanying social isolation are not hard to find. Over the past decade, black ghetto speech has grown progressively more distant from the standard English spoken by most non-Hispanic whites (cf. Labov, 1972; Labov and Harris, 1986), and black marriage, fertility, and family patterns have diverged more sharply from the mainstream (Espenshade, 1985; Farley, 1984; Farley and Allen, 1987; Pratt et al., 1984). Over the same period, poverty, labor force withdrawal, and unemployment have come to be increasingly concentrated in inner-city black neighborhoods (Wilson, 1987), particularly for young men (Lichter, 1988). Our results suggest that the extremity of black residential segregation and its unique multidimensional character may help explain the growing social and economic gap between the black underclass and the rest of American society.

Acknowledgments

The research reported in this article was supported by National Institute for Child Health and Human Development Grants HD-18594 and HD-22992. We thank Mitchell Eggers and two anonymous reviewers for their helpful comments. We are also grateful to John M. Goering for suggesting the term hypersegregation to us.

References

Bean, F. D., and M. Tienda. 1987. *The Hispanic Population of the United States*. New York: Russell Sage.

Cortese, C. F., R. F. Falk, and J. C. Cohen. 1976. Further considerations on the methodological analysis of segregation indices. *American Sociological Review* 41:630–637.

Duncan, O. D. 1957. The measurement of population distribution. *Population Studies* 11:27–45.

Duncan, O. D., R. P. Cuzzort, and B. Duncan. 1961. *Statistical Geography: Problems in Analyzing Area Data*. Glencoe, Ill.: Free Press.

Duncan, O. D., and B. Duncan. 1955. A methodological analysis of segregation indices. *American Sociological Review* 20:210–217.

———. 1957. *The Negro Population of Chicago*. Chicago: University of Chicago Press.

Espenshade, T. J. 1985. Marriage trends in America: Estimates, implications, and causes. *Population and Development Review* 11:193–246.

Farley, R. 1977. Residential segregation in urbanized areas of the United States in 1970: An analysis of social class and racial differences. *Demography* 14:497–529.

———. 1984. *Blacks and Whites: Narrowing the Gap?* Cambridge, Mass.: Harvard University Press.

Farley, R., and W. R. Allen. 1987. *The Color Line and the Quality of Life in America*. New York: Russell Sage.

Farley, R., H. Schuman, S. Bianchi, D. Colasanto, and S. Hatchett. 1978. "Chocolate city, vanilla suburbs:" Will the trend toward racially separate communities continue? *Social Science Research* 7:319–344.

Glaster, G. C. 1984. On the measurement of metropolitan decentralization of blacks and whites. *Urban Studies* 21:465–470.

Greeley, A. M. 1974. *Ethnicity in the United States: A Preliminary Reconnaissance*. New York: Wiley.

Hirsch, A. R. 1983. *Making the Second Ghetto: Race and Housing in Chicago 1940–1960*. New York: Cambridge University Press.

Jakubs, J. F. 1977. Residential segregation: The Taeuber index reconsidered. *Journal of Regional Science* 17:281–303.

———. 1979. A consistent conceptual definition of the index of dissimilarity. *Geographical Analysis* 11:315–321.

———. 1981. A distance-based segregation index. *Journal of Socio-Economic Planning Sciences* 15:129–136.

James, D. R., and K. E. Taeuber. 1985. Measures of segregation. Pp. 1–32 in N. Tuma (ed.), *Sociological Methodology 1985*. San Francisco: Jossey-Bass.

Kain, J. F., and J. M. Quigley. 1975. *Housing Markets and Racial Discrimination: A Microeconomic Analysis*. New York: National Bureau of Economic Research.

Labov, W. 1972. *Language in the Inner City*. Philadelphia: University of Pennsylvania Press.

Labov, W., and W. A. Harris, 1986. De facto segregation of black and white vernaculars. Pp. 1–24 in D. Sankoff (ed.), *Current Issues in Linguistic Theory: Diversity and Diachrony*. Philadelphia: John Benjamins.

Langberg, M., and R. Farley. 1985. Residential segregation of Asian Americans in 1980. *Sociology and Social Research* 69:51–61.

Lichter, D. T. 1988. Racial differences in underemployment in American cities. *American Journal of Sociology* 93:771–792.

Lieberson, S. 1980. *A Piece of the Pie: Blacks and White Immigrants Since 1880*. Berkeley: University of California Press.

———. 1981. An asymmetrical approach to segregation. Pp. 61–82 in C. Peach, V. Robinson, and S. Smith (eds.), *Ethnic Segregation in Cities*. London: Croom Helm.

Logan, J. R. 1978. Growth, politics, and the stratification of places. *American Journal of Sociology* 84:404–416.

Massey, D. S. 1979. Residential segregation of Spanish Americans in United States urbanized areas. *Demography* 16:653–664.

Massey, D. S., and B. Bitterman. 1985. Explaining the paradox of Puerto Rican segregation. *Social Forces* 64:306–331.

Massey, D. S., G. A. Condran, and N. A. Denton. 1987. The effect of residential segregation on black social and economic well-being. *Social Forces* 66:29–56.

Massey, D. S., and N. A. Denton. 1987. Trends in the residential segregation of blacks, Hispanics, and Asians. *American Sociological Review* 52:802–825.

———. 1988a. The dimensions of residential segregation. *Social Forces* 67:281–315.

———. 1988b. Suburbanization and segregation in U.S. metropolitan areas. *American Journal of Sociology* 94:592–626.

Olsen, R. J. 1980. A least squares correction for selectivity bias. *Econometrica* 48:1815–1820.
Park, R. E., and E. W. Burgess. 1925. *The City*. Chicago: University of Chicago Press.
Pratt, W. F., W. D. Mosher, C. A. Bachrach, and M. C. Horn. 1984. Understanding U.S. fertility: Findings from the National Survey of Family Growth, Cycle III. *Population Bulletin* 39(5).
Schneider, M., and J. R. Logan. 1982. Suburban racial segregation and black access to local public resources. *Social Science Quarterly* 63:762–770.
———. 1985. Suburban municipalities: The changing system of intergovernmental relations in the mid-1970s. *Urban Affairs Quarterly* 21:87–105.
Sorensen, A., K. E. Taeuber, and L. J. Hollingsworth, Jr. 1975. Indexes of racial residential segregation for 109 cities in the United States, 1940–1970. *Sociological Focus* 8:125–142.
Spear, A. H. 1967. *Black Chicago: The Making of a Negro Ghetto, 1890–1920*. Chicago: University of Chicago Press.
Taeuber, K. E., and A. F. Taeuber. 1965. *Negroes in Cities: Residential Segregation and Neighborhood Change*. Chicago: Aldine.
U.S. Bureau of the Census. 1980. *Census of Population and Housing 1980, Summary Tape File 4A* [Machine-readable data file]. Washington, D.C.: U.S . Bureau of the Census (producer). Ithaca, N.Y.: National Planning Data Corporation (distributor).
———. 1982. *1980 Census of Population and Housing, PHC80-R1-A, Users' Guide Part A. Text*. Washington, D.C.: U.S. Government Printing Office.
White, M. J. 1983. The measurement of spatial segregation. *American Journal of Sociology* 88:1008–1019.
———. 1986. Segregation and diversity: Measures in population distribution. *Population Index* 52:198–221.
Wilson, W. J. 1987. *The Truly Disadvantaged*. Chicago: University of Chicago Press.