

HOT SPOT SPATIAL MODELS FOR DATA REPORTED AS COUNTS OVER GEOGRAPHIC AREAS

Gary Simon
Stern School of Business
New York University
44 West Fourth Street
New York NY - 10012
JUNE 2006

ABSTRACT

Data obtained as counts over geographic regions can be described as Poisson random variables, and this work models the Poisson expected values. There are two major points here. (1) The model parameterizes hot spots in a way that allows the estimation of the hot spot coordinates. (2) Spatial integrals are computed through Green's theorem.

KEY WORDS

Spatial data, spatial integrals, hot spot, Green's theorem

PHONE: 212-998-0451
FAX: 212-995-4003
E-MAIL: gsimon@stern.nyu.edu

1. INTRODUCTION

The problem considered here is that of formulating models to explain count data reported as aggregates over geographical regions. Disease counts, sales data, and crime statistics are generally reported by region. The “unit of analysis” for such data should be individuals; each person either does or does not come down with the disease, does or does not decide to buy a washing machine, does or does not commit a crime. Information however is usually provided as geographical aggregates than at the level of individuals.

This is of course a problem in spatial analysis, and the literature on spatial analysis is enormous. Important resources here are Cliff and Ord (1973, 1981) and Cressie (1991).

Many approaches assess “spatial association” in terms of estimation and hypothesis testing, using either distance or incidence. The distance $d_{i,i'}$ between regions i and i' needs a representative point for each region (typically the center), and these are not always easily available. The incidence is defined as

$$u(i, i') = \begin{cases} 1 & \text{if regions } i \text{ and } i' \text{ share a border} \\ 0 & \text{if regions } i \text{ and } i' \text{ do not share a border} \end{cases}$$

The spatial association notion is that regions which are close to each other (small $d_{i,i'}$ or $u(i, i') = 1$) should present values that are closer to each other than would be expected by chance alone. Cliff and Ord (1981) discuss the I and c statistics for measuring spatial association. Simon (1997) gives an angular interpretation to spatial association.

Spatial random fields used heavily; indeed random fields are the building blocks of the landmark book by Cressie (1991). If o_i is the observed value for region i , we associate this value with point $\begin{pmatrix} x_i \\ y_i \end{pmatrix}$, usually the region’s center. The value o_i is then assumed to

be taken from a distribution with mean μ , standard deviation σ , and with $\text{Cov}(o_i, o_{i'})$

depending on the locations $\begin{pmatrix} x_i \\ y_i \end{pmatrix}$ and $\begin{pmatrix} x_{i'} \\ y_{i'} \end{pmatrix}$. The spatial random field model has

unconditional mean μ at every point; the observed similarities among nearby values are due only to the covariance structure. The special case in which $\text{Cov}(o_i, o_{i'})$ depends only on $d_{i,i'}$ is called isotropic. One can then write $\text{Cov}(d)$, expressing the covariance as a function only of distance. Much analytic work then centers on kriging, the estimation of $\text{Var}(o_i - o_{i'})$, which depends only on $d_{i,i'}$.

This paper departs from these ideas in that a model will be used to specify the mean function $E(o_i)$ though the use of a “spatial force” function. This will lead to a Poisson model for the number of events per region, with the Poisson expected values described

through the “spatial force” function. This function will contain (non-random) parameters with useful interpretations. We will be able to estimate those parameters, which will have interpretations as hot spots and rates of decay from hot spots.

The phrase “hot spot” arises often in scientific and mathematical discussions, but the usage is nearly always casual. For instance, Osterberg (1962) used “hot spot” to speak of a location off the Oregon coast in which krill were affected by nuclear tests (and noted that the spot dissipated over time), Michelson (1964) used the phrase to describe the action of cigarette smoke on the throat, and Salisbury and Hunt (1967) used it in dealing with heat distributions on a moon crater. Wagner (1971) used “hot spot” to describe a location in Hawaii with a plentiful population of a newly-discovered fern species. Harris, MacGregor, and Wright (1980) and also MacGregor and Wong (1980) used “hot spot” with reference to the hottest spot in a chamber, and they noted that this hot spot might move around.

The “hot spot” phrase is used frequently to describe volcanic activity, as in Bonatti (1990), Storey et al (1995), and Kerr (1997).

With regard to spatial methods, Kvam et. al. (2000) and Piegorsch et. al. (1998) used the notion that locations with high values should be used to suggest the next places to collect samples. Hobbs and Maeda (1996) try to close in on radon hot spots and achieve partial success.

The works cited above talk about hot spots but do not endow these hot spots with coordinates that could be estimated. Flather, Knowles, and Kendall (1998) at least provide operational definitions for their selection of various United States counties as hot spots critical to maintaining threatened animal species.

An interesting perspective is that of the paper by Mardia and Gadsden (1977) dealing with a point process in which the points are locations of Hawaiian and mid-Pacific volcanoes. They fit a small circle to the surface of the globe, using an objective function that has the effect of enclosing as many volcanoes as possible within a small circle. The center of their small fitted circle is a hot spot whose coordinates are estimated by their calculations.

In the work developed here, the hot spot is endowed with coordinates, and one end point of the analysis is the identification of the hot spot location.

2. PRELIMINARIES

Suppose that there are I regions in all. The regions may be zip codes or census tracts or voting districts, according to the problem. For the sake of consistent terminology, this paper will identify the regions as counties. Subscript i will always be used for counties. Some counties will have disjoint subsets, and these subsets of counties will be called islands. Each county will consist of one or more islands. The subscript k will always be used for the vertices used to draw maps of the islands, or for the segments that connect those vertices.

Data provides us with $P_i =$ population of county i , and also $o_i =$ observed number of events (disease cases, for example) in county i . The values o_1, o_2, \dots, o_I will be assumed to be independent Poisson random variables. If a statistical model dictates a rate r_i for county i , then $e_i = r_i P_i$ is the expected number of cases. We will compare observed and

expected rates with $G^2 = 2 \sum_{i=1}^I o_i \log \frac{o_i}{e_i}$. It is of course possible to use the

goodness-of-fit chi-squared $\chi^2 = \sum_{i=1}^I \frac{(o_i - e_i)^2}{o_i}$, but that was not done in this paper.

When the model holds, G^2 is asymptotically chi-squared, and the number of degrees of freedom depends on the number of parameters estimated.

For the model of uniformity, $e_i = o_+ \frac{P_i}{P_+}$, where o_+ is the total number of events and P_+ is the total population. The uniformity model is not usually successful in fitting real data.

Let $\mathbf{z} = \begin{pmatrix} x \\ y \end{pmatrix}$ be a point in the plane. Suppose that there is a force related to the probability of coming down with a disease (or buying a washing machine or committing a crime) at point \mathbf{z} and denote the magnitude of the force as $f(\mathbf{z})$, a scalar function of a point.

The expected number of disease cases in \mathbb{B} is approximately

$$E(\mathbb{B}) = \int_{\mathbb{B}} n(\mathbf{q}) d\mathbf{q} \times \frac{\int_{\mathbb{B}} f(\mathbf{q}) d\mathbf{q}}{\int_{\mathbb{B}} d\mathbf{q}} \quad (2.1)$$

The function $n(\mathbf{q})$ describes the density of population over \mathbb{B} , and (2.1) is total population, multiplied by total force, and then divided by area; equivalently, this is total

population multiplied by the average force over \mathbb{B} . (This is approximate because it does not use the locations of the individual people within \mathbb{B} .) In (2.1), the challenging calculation is $\int_{\mathbb{B}} f(\mathbf{q}) d\mathbf{q}$. Here's why:

The value of $\int_{\mathbb{B}} n(\mathbf{q}) d\mathbf{q}$ represents the population of \mathbb{B} , and this information is available from standard sources. Moreover, we are not likely to know the exact details of $n(\mathbf{q})$ anyhow.

The value $\int_{\mathbb{B}} d\mathbf{q}$ is the area of \mathbb{B} , and this is also known from standard sources. (It turns out that the work here will, for rather interesting reasons, still calculate this integral.)

The overall disease force $\int_{\mathbb{B}} f(\mathbf{q}) d\mathbf{q}$ will include parameters. Since these parameters will need to be estimated, there will be repeated requests for new calculations of $\int_{\mathbb{B}} f(\mathbf{q}) d\mathbf{q}$ and thus an efficient computational method will be needed.

The calculation of $\int_{\mathbb{B}} f(\mathbf{q}) d\mathbf{q}$ could in principle be done pixel-by-pixel. This can be time consuming, as county \mathbb{B} may involve thousands of pixels and the classification of pixels as to whether they are inside or outside of \mathbb{B} will be clerically difficult.

This paper instead uses Green's theorem to find $\int_{\mathbb{B}} f(\mathbf{q}) d\mathbf{q}$ through line integrals along the boundary. In geographic data bases, regions are polygons represented by their boundary segments for the purposes of constructing maps. The rectangular-shaped states of Colorado and Wyoming could be represented by as few as four segments, and many hundreds of rectangular counties within the United States have this same property. A complicated geographic region might have several hundred boundary segments. In the first example below, the most complicated county is described by 109 boundary segments. In the second example, the most complicated county has 91 boundary segments. In most of the counties (for both examples) there are fewer than 30 boundary segments.

3. FORCE FUNCTIONS

Let $z = \begin{pmatrix} x \\ y \end{pmatrix}$ denote a general point in two-space. The function $f(z) = c$ is constant and represents a uniform force.

Let now $s = \begin{pmatrix} s_x \\ s_y \end{pmatrix}$ be a single, fixed location. This will be the “hot spot” and the values s_x and s_y will have to be estimated. This might be the location of a potent contaminant or of a dirty air source or of a center of criminal activity.

As usual, $\|z - s\|$ denotes the Euclidean distance between points.

Consider the simple hot-spot force function

$$f(z) = \frac{c \alpha^2}{\alpha^2 + \|z - s\|^2} = \frac{c}{1 + \left(\frac{1}{\alpha}(z - s)\right)' \left(\frac{1}{\alpha}(z - s)\right)} \quad (3.1)$$

The function f will be in density units, such as $\frac{1}{\text{miles}^2}$, reciprocal square miles. The coordinates of z and s will be in miles, and thus α will be in miles. The parameter c will be in units of $\frac{1}{\text{miles}^2}$.

The parameter α has a half-life type interpretation. The maximum force occurs at $z = s$ and has $f(s) = c$. If z_2 is any point for which $\|z_2 - s\| = \alpha$, then $f(z_2) = \frac{c}{2}$, half the maximum.

The force function used for the examples is a modification of the above, namely

$$f(z) = c \left\{ 1 + \frac{\omega \alpha^2}{\alpha^2 + \|z - s\|^2} \right\} \quad (3.2)$$

This is a mix of the uniform force with the simple hot-spot model. The parameter ω assesses the strength of the hot spot relative to uniformity. A negative value for ω would represent a protective effect.

This force function is chosen not only for its modeling credibility, but because $\int f(z) dx$ is solvable as an indefinite integral. (NOTE: The dx is correct.) This will be explained more fully later.

The role of the parameter c is to enforce the condition $\sum_{i=1}^I o_i = \sum_{i=1}^I e_i$.

The examples will be presented next, with the computational issues left for the final sections.

This work treats o_i as a Poisson random variable with mean e_i . The methods will be generalized in future work so that e_i , along with demographic factors, will be a predictor for o_i in a Poisson regression.

4. FIRST EXAMPLE, FLORIDA CANCER RATES

Figure 1 shows cancer rates for the state of Florida.

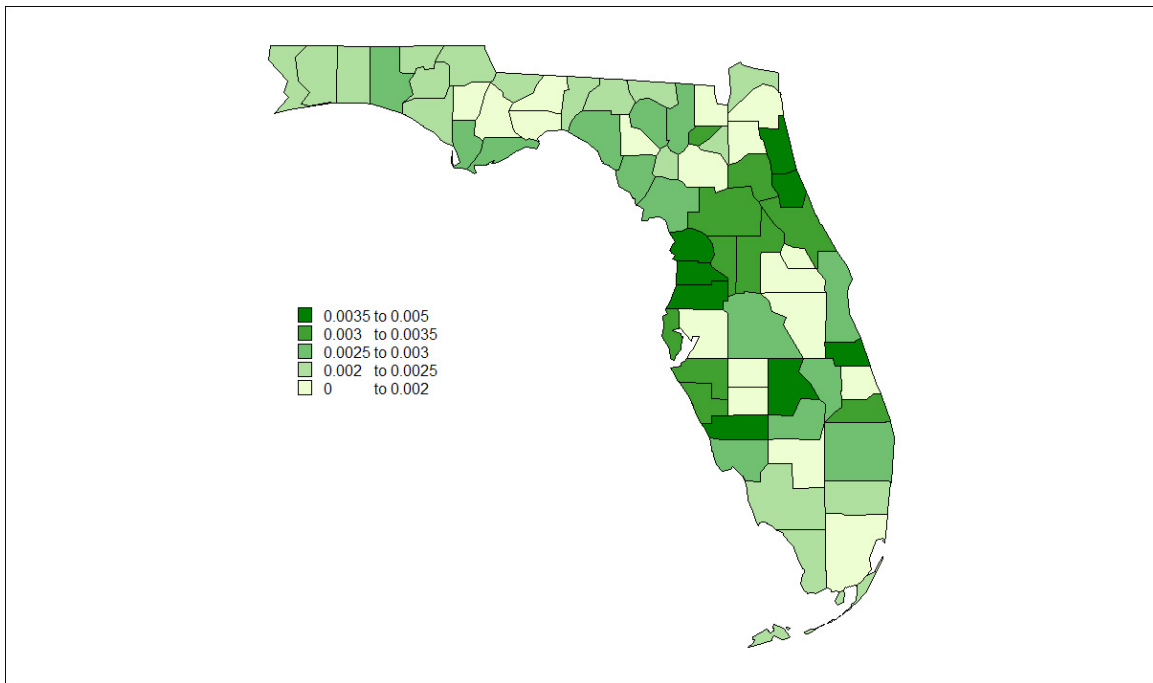


FIGURE 1
Florida Cancer Rates, by County
Rates were computed from average deaths per year
for the period 1998-2002

Source: <http://statecancerprofiles.cancer.gov/cgi-bin/deathrates/deathrates.pl?12&001&00&0&001&1&1&1>

Florida has 67 counties. The test for uniformity over the counties is $G^2 = 2,816.26$, on 66 degrees of freedom.

The map was endowed with an x - y coordinate system with the origin to the southwest of the state. The force function (3.2) was used, producing maximum likelihood estimates $(\hat{s}_x, \hat{s}_y) = (375.8877, 300.6793)$, $\hat{\alpha} = 13.4375$, and $\hat{\omega} = 2.325$.

The hot spot (\hat{s}_x, \hat{s}_y) is (82.56 w long, 28.80 n lat), in Citrus County.

Figure 2 shows the fitted rates.

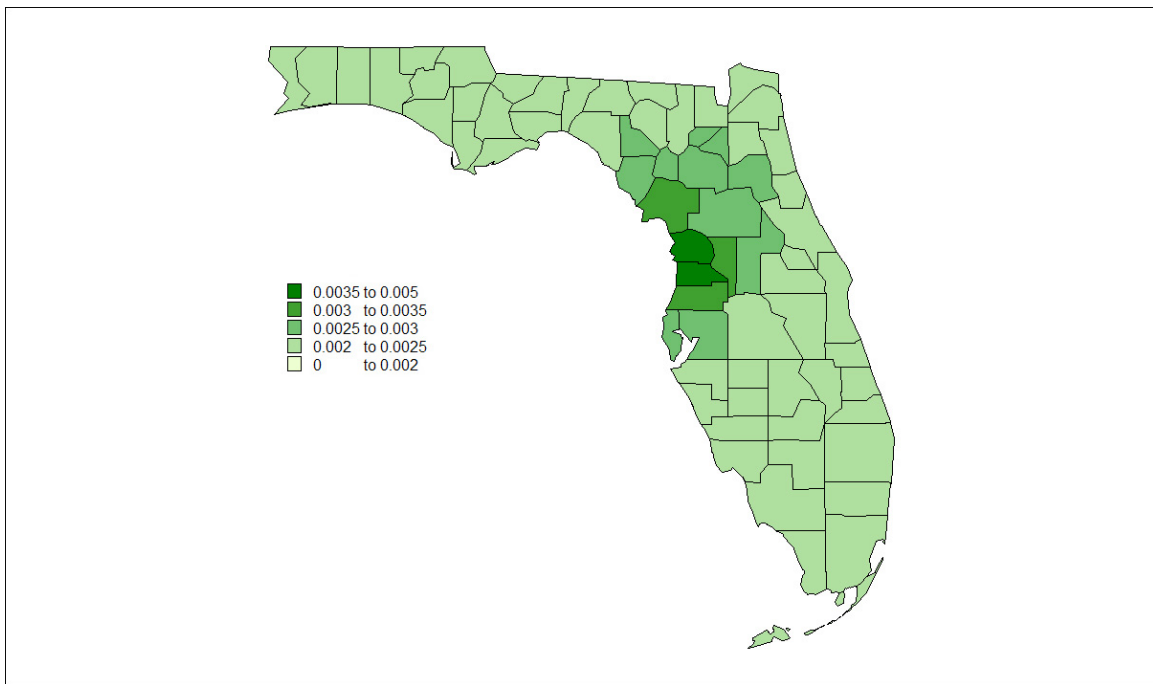


FIGURE 2
Florida *Fitted* Cancer Rates, by County
for the period 1998-2002
using force function (3.2)

The resulting fit is given as $G^2 = 2,246.93$, on $67 - 1 - 4 = 62$ degrees of freedom. This is still not a good fit, however the reduction in G^2 is 569.34, for four degrees of freedom. Moreover, the ratio of the G^2 values is $\frac{2,246.93}{2,816.26} \approx 0.7978$. Thus about 20% of G^2 has been “explained” by the force function.

5. SECOND EXAMPLE, 1916 POLIO RATES

In the summer of 1916, the northeast region of the United States was struck with an epidemic of polio. The details are given in a fascinating article by Trevelyan, Smallman-Raynor, and Cliff, (2005); the authors have generously shared their data. They had usable data on 148 counties, as indicated in Figure 3.

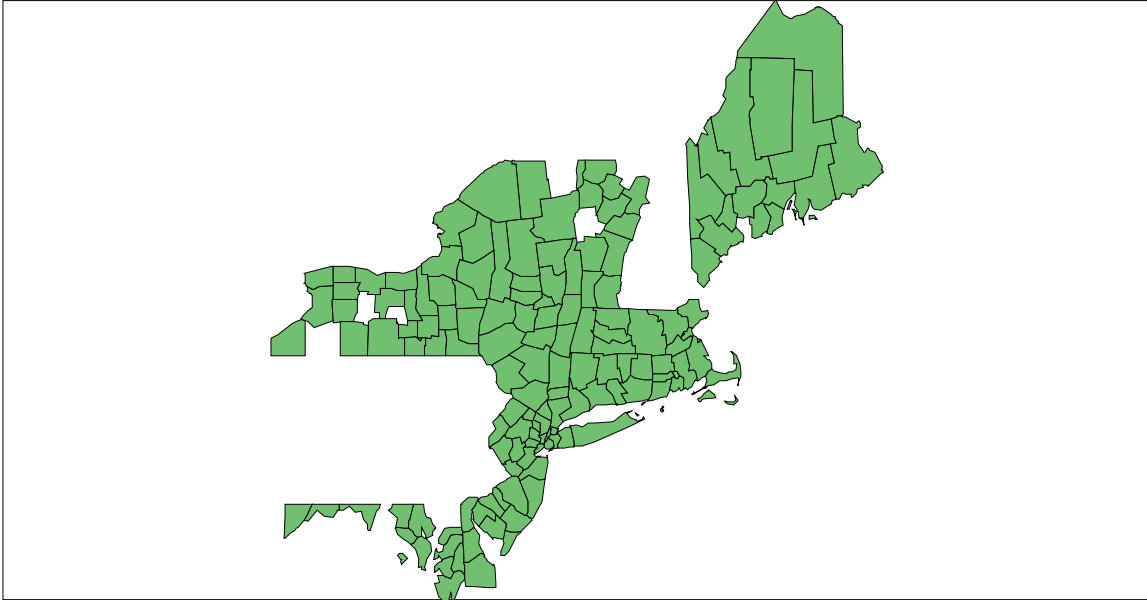


FIGURE 3
Counties with usable data,
Trevelyan, Smallman-Raynor, and Cliff (2005)

In Figure 3, the District of Columbia and each of the five boroughs of New York City are included among the 148 counties. The states of New Hampshire and Pennsylvania did not have usable data.

The analysis here will look only at the rates for the entire year 1916. These are summarized in Figure 4.

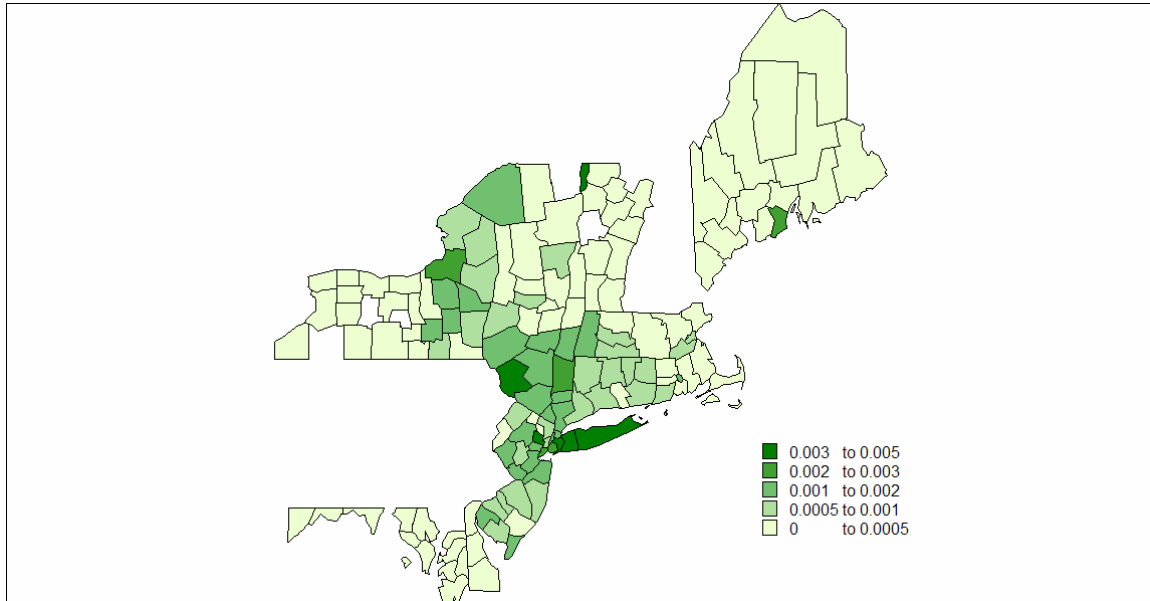


FIGURE 4
1916 Polio Rates
Trevelyan, Smallman-Raynor, and Cliff (2005)

The first reported polio case in this year was in Brooklyn (Kings County, New York), and the rates were highest in the urbanized counties near New York City.

The observed rates are certainly not uniform. Here $G^2 = 16,713.64$, with 147 degrees of freedom.

The map was given an x - y coordinate system with the origin to the southwest of the studied area. The force function (3.2) was again used, producing maximum likelihood estimates $(\hat{s}_x, \hat{s}_y) = (450.78, 135.77)$, $\hat{\alpha} = 56.80$, and $\hat{\omega} = 15.66$

The hot spot (\hat{s}_x, \hat{s}_y) is offshore, east of Ocean County, New Jersey. The force function has radial symmetry, and this particular location for the center provides the best fit. Figure 5 shows the fitted rates.

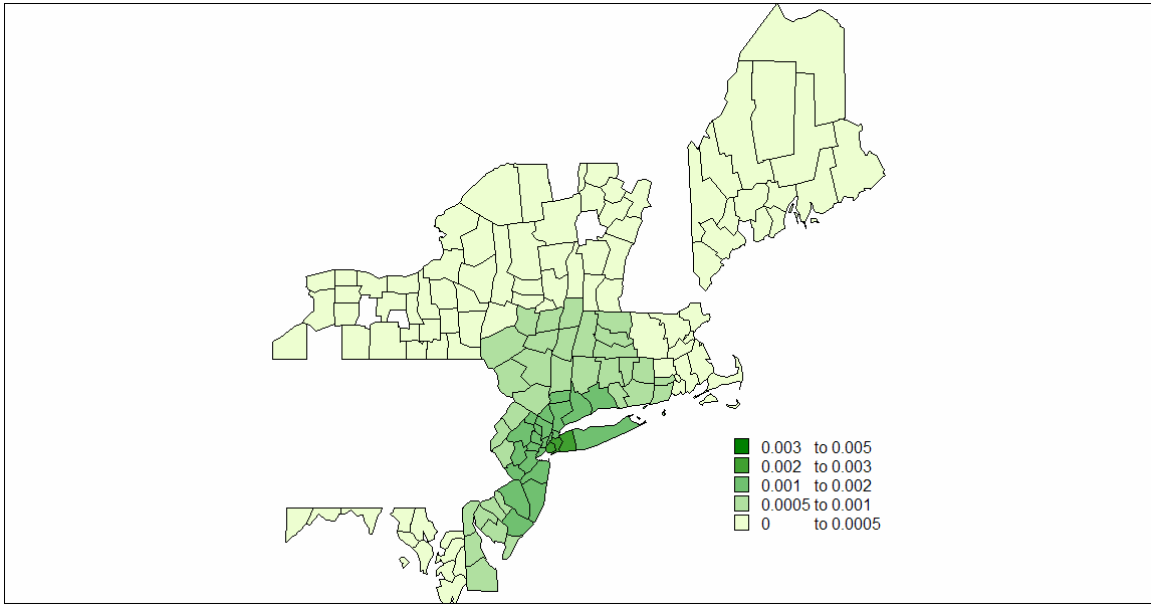


FIGURE 5
1916 *Fitted* Polio Rates
using force function (3.2)

The fit statistic is now $G^2 = 7,045.73$, on $148 - 1 - 4 = 143$ degrees of freedom. The reduction in G^2 is 9,667.91, using four degrees of freedom.

The ratio of G^2 statistics is $\frac{7,045.73}{16,713.64} \approx 0.4216$, so that about 58% of the variability in polio rates has been explained by this model.

6. COMPUTATIONAL ISSUES

The expected counts per county will be found through (2.1). The difficult calculation is

$\int_{\mathbb{B}} f(\mathbf{q}) d\mathbf{q}$, the integral of the force function over a county.

The integral requires a clean parameterization of the county's boundary. Mapping software possesses this information, as it is critical to the computer construction of maps. Figure 6 shows Manhattan (New York County), as produced by MapInfo 7.8.

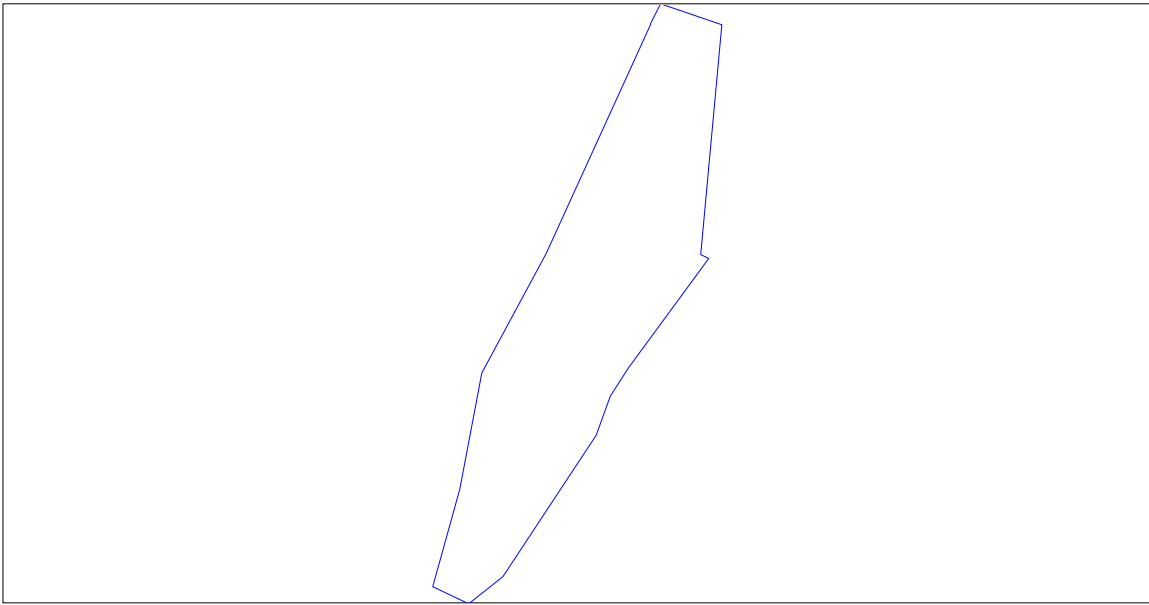


FIGURE 6
New York County (Manhattan), New York
MapInfo 7.8

This program, through the command **Table > Export**, will create a file (*.MIF) listing the longitude and latitude of corner points that it used to create the map. The part of the file useful for the purposes here has this structure:

```
Region 1
16
-73.9072 40.875
-73.9344 40.882
-73.9387 40.8751
.
.
.
-73.9128 40.796
-73.9163 40.7973
-73.9072 40.875
```

The line “Region 1” indicates that Manhattan is constructed as a single connected region. The “16” means that it is drawn as a polygon with 15 corners; the last listed corner is coincident with the first. The remaining lines give the (longitude, latitude) coordinates. Figure 7 shows a plot of those corners.

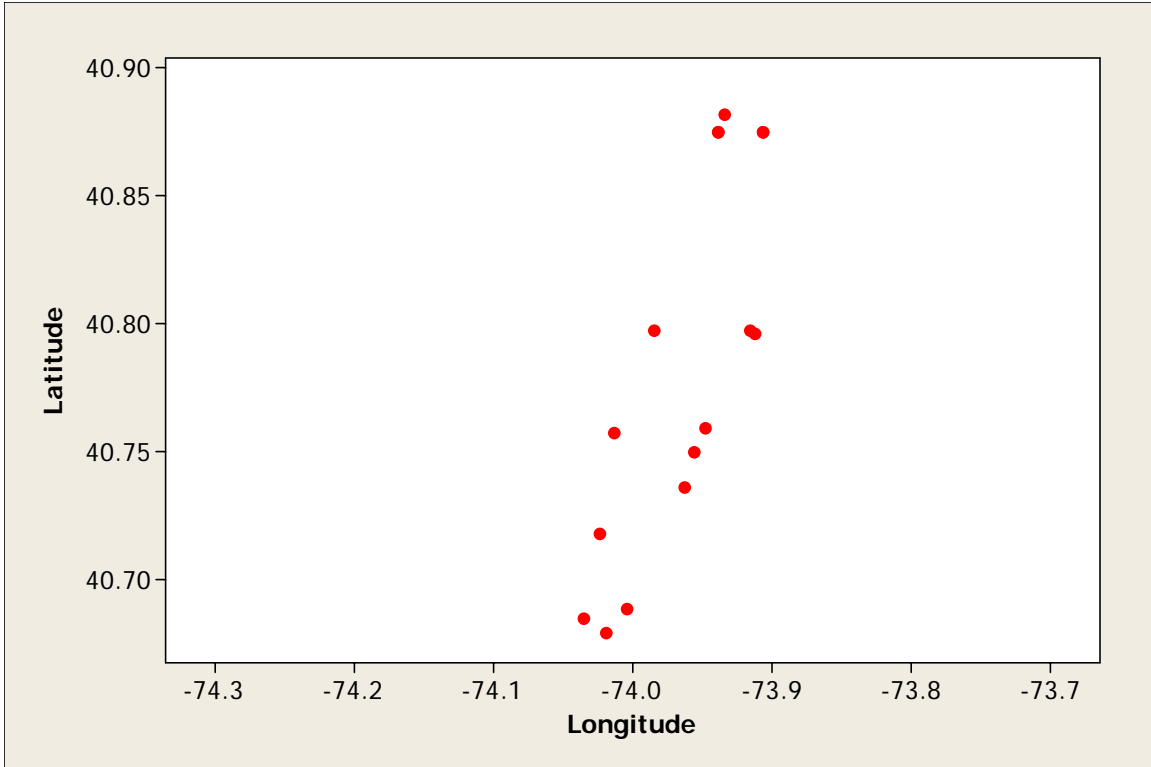


FIGURE 7
New York County (Manhattan), New York
MapInfo 7.8 Plotting Corners

The regions studied in the examples were converted to planar maps. Longitude and latitude values were converted to miles. Each degree of latitude represents $\frac{1}{90}$ of the earth's quadrant, about 69.2 miles. A degree of longitude at the equator is $\frac{1}{360}$ of the earth's circumference, also about 69.2 miles. Away from the equator, a degree of longitude is shorter. At latitude θ , one degree of longitude measures about $\cos(\theta) \times 69.2$ miles. This is a standard result in spherical trigonometry; see Murray (1900). In each example, a typical latitude was used to make the longitude to miles conversion.

Because MapInfo has represented counties as polygons, it becomes reasonable to use Green's theorem to find $\int_{\mathbb{B}} f(\mathbf{q}) d\mathbf{q}$. Green's theorem for connected region \mathbb{A} and for scalar functions P and Q of two variables is

$$\oint_{\partial\mathbb{A}} \{P dx + Q dy\} = \iint_{\mathbb{A}} \left\{ \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right\} dx dy \quad (6.1)$$

This was at one time a standard result in most calculus textbooks.

The integral on the right is a routine double integral, and it's easy enough provided the mechanical integration steps can be performed. The integral on the left is the line integral around the perimeter of \mathbb{A} , done in the counter-clockwise direction.

We will use this for our particular region \mathbb{B} in the form with $P \equiv 0$:

$$\oint_{\partial\mathbb{B}} Q \, dy = \iint_{\mathbb{B}} \frac{\partial Q}{\partial x} \, dx \, dy \quad (6.2)$$

With a modest change in procedure, this can also be done with $Q \equiv 0$ and would give a slightly different, but equivalent, form for the final results.

The function $\frac{\partial Q}{\partial x}$ will play the role of f , the force function. The right side is the total force over area \mathbb{B} , and this will be calculated through the line integral on the left.

Making this work requires finding Q ; that is, the symbolic (indefinite) dx integral of $f = \frac{\partial Q}{\partial x}$ must be calculable. This need motivated the choice for the form of the force function (3.2).

Mapping software can provide $\partial\mathbb{B}$, the boundary of \mathbb{B} , as $m_{\mathbb{B}}$ straight line segments:

Segment 1 goes from $(x_1^{\mathbb{B}}, y_1^{\mathbb{B}})$ to $(x_2^{\mathbb{B}}, y_2^{\mathbb{B}})$.

Segment 2 goes from $(x_2^{\mathbb{B}}, y_2^{\mathbb{B}})$ to $(x_3^{\mathbb{B}}, y_3^{\mathbb{B}})$.

.

.

Segment k goes from $(x_k^{\mathbb{B}}, y_k^{\mathbb{B}})$ to $(x_{k+1}^{\mathbb{B}}, y_{k+1}^{\mathbb{B}})$.

.

.

Segment $m_{\mathbb{B}}$ goes from $(x_{m_{\mathbb{B}}}^{\mathbb{B}}, y_{m_{\mathbb{B}}}^{\mathbb{B}})$ to $(x_1^{\mathbb{B}}, y_1^{\mathbb{B}})$.

This means that

$$\iint_{\mathbb{B}} f(x, y) \, dx \, dy = \oint_{\partial\mathbb{B}} Q \, dy = \pm \sum_{k=1}^{m_{\mathbb{B}}} \int_{\text{Segment } k} Q(x, y) \, dy \quad (6.3)$$

The positive result occurs when the corner points are listed in counterclockwise order, and the negative when in clockwise order.

The segment endpoints $(x_k^{\mathbb{B}}, y_k^{\mathbb{B}})$ used here were obtained from MapInfo and are very close together. For example, the map of Colorado, which could have been described by four corner points, is given with 878 points. (In MapInfo 7.8, using Usa_Maps, select **Query > Select** and choose the condition State = "CO". The list of points can be produced with **Table > Export**. In the MapInfo list, the final listed point will be coincident with the first.)

The parametric form for the k^{th} segment is

$$\begin{cases} x(t) = x_k^{\mathbb{B}} + (x_{k+1}^{\mathbb{B}} - x_k^{\mathbb{B}})t \\ y(t) = y_k^{\mathbb{B}} + (y_{k+1}^{\mathbb{B}} - y_k^{\mathbb{B}})t \end{cases}; 0 \leq t \leq 1 \quad (6.4)$$

Note that $dy = (y_{k+1}^{\mathbb{B}} - y_k^{\mathbb{B}}) dt$. The fact that the points are close together makes it computationally reasonable to use

$$\begin{aligned} \int_{\text{Segment } k} Q(x, y) dy &\approx \int_{\text{Segment } k} \frac{Q(x_k^{\mathbb{B}}, y_k^{\mathbb{B}}) + Q(x_{k+1}^{\mathbb{B}}, y_{k+1}^{\mathbb{B}})}{2} dy \\ &= \int_0^1 \frac{Q(x_k^{\mathbb{B}}, y_k^{\mathbb{B}}) + Q(x_{k+1}^{\mathbb{B}}, y_{k+1}^{\mathbb{B}})}{2} (y_{k+1}^{\mathbb{B}} - y_k^{\mathbb{B}}) dt \\ &= \frac{Q(x_k^{\mathbb{B}}, y_k^{\mathbb{B}}) + Q(x_{k+1}^{\mathbb{B}}, y_{k+1}^{\mathbb{B}})}{2} (y_{k+1}^{\mathbb{B}} - y_k^{\mathbb{B}}) \end{aligned}$$

NOTE: If segment k is judged too long to use this approximation, we can

$$\begin{aligned} \text{consider } \int_{\text{Segment } k} Q(x, y) dy &\approx \\ &\frac{Q(x_k^{\mathbb{B}}, y_k^{\mathbb{B}}) + 4Q\left(\frac{x_k^{\mathbb{B}} + x_{k+1}^{\mathbb{B}}}{2}, \frac{y_k^{\mathbb{B}} + y_{k+1}^{\mathbb{B}}}{2}\right) + Q(x_{k+1}^{\mathbb{B}}, y_{k+1}^{\mathbb{B}})}{6} (y_{k+1}^{\mathbb{B}} - y_k^{\mathbb{B}}) \end{aligned}$$

which is a simple version of Simpson's rule for approximating integrals. Given the scales in which these maps are done, the corner points are only four or five miles apart, and some corners are as close as a few hundred feet. Thus, it's not clear that this pursuit of extra precision will be helpful.

Now find (using the index k modulo $m_{\mathbb{B}}$, so that $m_{\mathbb{B}} + 1 \Leftrightarrow 1$)

$$\begin{aligned} \sum_{k=1}^{m_{\mathbb{B}}} \int_{\text{Segment } k} Q(x, y) dy &\approx \sum_{k=1}^{m_{\mathbb{B}}} \left\{ \frac{Q(x_k^{\mathbb{B}}, y_k^{\mathbb{B}}) + Q(x_{k+1}^{\mathbb{B}}, y_{k+1}^{\mathbb{B}})}{2} (y_{k+1}^{\mathbb{B}} - y_k^{\mathbb{B}}) \right\} \\ &= \sum_{k=1}^{m_{\mathbb{B}}} \frac{Q(x_k^{\mathbb{B}}, y_k^{\mathbb{B}})}{2} (y_{k+1}^{\mathbb{B}} - y_{k-1}^{\mathbb{B}}) \end{aligned} \quad (6.5)$$

(Note the appearance of $y_{k-1}^{\mathbb{B}}$ in the final form.) This final form is easier for computational purposes, since the function Q will change from one optimization step to another while the differences $(y_{k+1}^{\mathbb{B}} - y_{k-1}^{\mathbb{B}})$ will not.

The calculation shown so far has not yet sorted out the \pm issue. Since $f \geq 0$, this should not be a problem. In fact, the calculation of area will be used to resolve the choice between counter-clockwise and clockwise.

A similar trick can be used to find the area of \mathbb{B} , if it is not otherwise provided. The result is the rather interesting formula

$$\text{Area}(\mathbb{B}) = \pm \frac{1}{2} \sum_{k=1}^n (x_k y_{k+1} - x_{k+1} y_k) \quad (6.6)$$

For the examples here, (6.6) was used to calculate the areas and to simultaneously identify which regions were listed in counter-clockwise order and which in clockwise order. The derivation of (6.6) is given in Appendix 1.

With the function $f(z)$ given as (3.2), the solution to $f(z) = \frac{\partial Q}{\partial x}$ is

$$Q = \frac{\omega \alpha^2}{\sqrt{\alpha^2 + (y - s_y)^2}} \tan^{-1} \left(\frac{x - s_x}{\sqrt{\alpha^2 + (y - s_y)^2}} \right)$$

Details are given in Appendix 2. The denominator $\sqrt{\alpha^2 + (y - s_y)^2}$ in the inverse tangent is positive, so that we may take the value of the inverse tangent in the interval $\left[-\frac{\pi}{2}, \frac{\pi}{2} \right]$.

The optimizing is done over s_x , s_y , α , and ω . The parameter c is adjusted so that $\sum_{i=1}^I o_i = \sum_{i=1}^I e_i$. The likelihood function has many local extreme points, so the optimization is done by exhaustive search over (s_x, s_y) starting points and then using derivative-based searches over all parameters.

APPENDIX 1

This appendix shows the derivation of (6.6). The area formula can be derived with Green's theorem with $\frac{\partial Q}{\partial x} \equiv 1$ and $\frac{\partial P}{\partial y} \equiv 0$. Then

$$\begin{aligned} \text{Area}(\mathbb{B}) &= \iint_{\mathbb{B}} \left\{ \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right\} dx dy = \iint_{\mathbb{B}} dx dy \\ &= \oint_{\partial \mathbb{B}} \{ P dx + Q dy \} = \oint_{\partial \mathbb{B}} x dy \end{aligned}$$

Identify Segment k as the line connecting $(x_k^{\mathbb{B}}, y_k^{\mathbb{B}})$ to $(x_{k+1}^{\mathbb{B}}, y_{k+1}^{\mathbb{B}})$. Wrap the subscripts so that point $(m_{\mathbb{B}} + 1)$ is also point 1. Let γ be +1 if the ordering is counter-clockwise, with $\gamma = -1$ for clockwise ordering.

$$\text{Area}(\mathbb{B}) = \oint_{\partial \mathbb{B}} x dy = \gamma \sum_{k=1}^{m_{\mathbb{B}}} \int_{\text{Segment } k} x dy$$

Then

$$\begin{aligned} \int_{\text{Segment } k} x dy &= \int_0^1 \left\{ x_k^{\mathbb{B}} + (x_{k+1}^{\mathbb{B}} - x_k^{\mathbb{B}})t \right\} (y_{k+1}^{\mathbb{B}} - y_k^{\mathbb{B}}) dt \\ &= \int_0^1 x_k^{\mathbb{B}} (y_{k+1}^{\mathbb{B}} - y_k^{\mathbb{B}}) dt + \int_0^1 (x_{k+1}^{\mathbb{B}} - x_k^{\mathbb{B}}) (y_{k+1}^{\mathbb{B}} - y_k^{\mathbb{B}}) t dt \\ &= x_k^{\mathbb{B}} (y_{k+1}^{\mathbb{B}} - y_k^{\mathbb{B}}) + \frac{1}{2} (x_{k+1}^{\mathbb{B}} - x_k^{\mathbb{B}}) (y_{k+1}^{\mathbb{B}} - y_k^{\mathbb{B}}) \\ &= \frac{1}{2} (x_{k+1}^{\mathbb{B}} + x_k^{\mathbb{B}}) (y_{k+1}^{\mathbb{B}} - y_k^{\mathbb{B}}) \end{aligned}$$

In summary,

$$\text{Area}(\mathbb{B}) = \gamma \sum_{k=1}^{m_{\mathbb{B}}} \frac{x_{k+1}^{\mathbb{B}} + x_k^{\mathbb{B}}}{2} (y_{k+1}^{\mathbb{B}} - y_k^{\mathbb{B}}) \quad (\text{A1.1})$$

If the roles of x and y were exchanged, the ordering would be in the other direction, and the result would be

$$\text{Area}(\mathbb{B}) = -\gamma \sum_{k=1}^{m_{\mathbb{B}}} \frac{y_{k+1}^{\mathbb{B}} + y_k^{\mathbb{B}}}{2} (x_{k+1}^{\mathbb{B}} - x_k^{\mathbb{B}}) \quad (\text{A1.2})$$

Since both (A1.1) and (A1.2) give the area, and so must their average. This will produce (6.6) (either with $\gamma = 1$ or $\gamma = -1$), which can be called ‘‘Green’s theorem for polygons.’’ The simple formula (6.6) is undeservedly obscure.

APPENDIX 2

This appendix gives the solution of $f(z) = f\left(\begin{pmatrix} x \\ y \end{pmatrix}\right) = \frac{\omega \alpha^2}{\alpha^2 + \|z - s\|^2}$

$$= \frac{\omega \alpha^2}{\alpha^2 + \left\{ (x - s_x)^2 + (y - s_y)^2 \right\}} = \frac{\partial Q}{\partial x}.$$

Suppose that $\frac{dM}{dx} = \frac{\gamma}{\beta + (x - u)^2}$.

Certainly $M = \gamma \int \frac{1}{\beta + (x - u)^2} dx = \left[\begin{array}{l} x - u \stackrel{\text{let}}{=} \sqrt{\beta} \tan \theta \\ dx \stackrel{\text{let}}{=} \sqrt{\beta} \sec^2 \theta d\theta \end{array} \right]$

$$= \gamma \int \frac{1}{\beta + \beta \tan^2 \theta} \sqrt{\beta} \sec^2 \theta d\theta = \frac{\gamma}{\sqrt{\beta}} \int \frac{1}{1 + \tan^2 \theta} \sec^2 \theta d\theta$$

$$= \frac{\gamma}{\sqrt{\beta}} \int \frac{1}{\sec^2 \theta} \sec^2 \theta d\theta = \frac{\gamma}{\sqrt{\beta}} \theta = \frac{\gamma}{\sqrt{\beta}} \tan^{-1} \left(\frac{x - u}{\sqrt{\beta}} \right)$$

The correspondences are then routine: $\gamma \Leftrightarrow \omega \alpha^2$, $u \Leftrightarrow s_x$, $\beta \Leftrightarrow \alpha^2 + (y - s_y)^2$. Thus

$$Q = \frac{\omega \alpha^2}{\sqrt{\alpha^2 + (y - s_y)^2}} \tan^{-1} \left(\frac{x - s_x}{\sqrt{\alpha^2 + (y - s_y)^2}} \right)$$

REFERENCES:

- Bonatti, E. 1990. Not So Hot “Hot Spots” in the Oceanic Mantle, *Science*, Series 3, Vol. 250, pp 107-111.
- Cliff, A.D., and Ord, J.K., 1973, *Spatial Autocorrelation*, Pion Limited, London.
- Cliff, A.D., and Ord, J.K., 1981, *Spatial Processes: Models and Applications*, Pion Limited, London.
- Cressie, Noel, 1991, *Statistics for Spatial Data*, John Wiley & Sons, New York.
- Flather, C.H., Knowles, M.S., and Kendall, I.A. 1998. Threatened and Endangered Species Geography, *BioScience*, Vol. 48, pp 365-376.
- Haining, Robert, 2003, *Spatial Data Analysis Theory and Practice*, Cambridge University Press, Cambridge.
- Harris, T. J., MacGregor, J. F., and Wright, J. D. (1980) Self-Tuning and Adaptive Controllers: An Application to Catalytic Reactor Control, *Technometrics*, Vol. 22, Number 2, pp 153-154.
- Hobbs, W.E., and Maeda, L.Y., 1996, Identification and Assessment of a Small, Geologically Localized Radon Hot Spot, *Environment International*, vol 22, Supplement 1, pp S809-817.
- Kerr, R. A., 1997. New Way to Hit the hot Spot Hints at a Complex Pacific, *Science*, Series 3, Vol. 276, pp 1198.
- Kvam, Paul H., Tiwari, Ram C., Zalkikar, Jyoti N., 2000, Nonparametric Bayes Estimation of Contamination Levels Using Observations from the Residual Distribution, *Journal of the American Statistical Association*, vol 95, #452, pp 1119-1126.
- MacGregor, J. F., and Wong, A. K. L. (1980) Multivariate Model Identification and Stochastic Control of a Chemical Reactor, *Technometrics*, Vol. 22, Number 4, pp 453-464.
- Mardia, K. V., and Gadsden, R. J. (1977) A Small Circle of Best Fit of Spherical Data and Areas of Vulcanism, *Applied Statistics*, Vol. 26, Number 3, pp 238-245.
- Michelson, Irving. (1964) Cigarettes: Polonium-210, *Science*, New Series, Vol. 143, Number 3609, p 917.
- Murray, Daniel A., 1900, *Spherical Trigonometry*, Longmans, Green, and Company, New York/London/Toronto.

- Osterberg, Charles. (1962) Fallout Radionuclides in Euphausiids, *Science*, New Series, Vol. 138, Number 3539, pp 529-530.
- Piegorsch, Walter W., Smith, Eric P., Edwards, Don, and Smith, Richard L., 1998, Statistical Advances in Environmental Science, *Statistical Science*, vol 13, #2, pp 186-208.
- Salisbury, John W., and Hunt, Graham R. (1967) Infrared Images of Tycho on Dark Moon, *Science*, New Series, Vol. 155, Number 3766, pp 1098-1100.
- Simon, Gary, 1997, An Angular Version of Spatial Correlations, with Exact Significance Tests, *Geographical Analysis*, vol 29, #3, pp 267-278.
- Storey, M. et al 1995. Timing of Hot Spot-Related Volcanism and the Breakup of Madagascar and India, *Science*, Series 3, Vol. 267, pp 852-855.
- Trevelyan, Barry, Smallman-Raynor, Matthew, and Cliff, Andrew D. (2005) The Spatial Structure of Epidemic Emergence: Geographical Aspects of Poliomyelitis in North-eastern USA, July-October 1916, *Journal of the Royal Statistical Society, Series A*, vol 168, part 4, pp 701-722.
- Wagner, W. H., Jr. (1971) *Lindsaea* (Schizoloma) *ensifolia* Swartz in Hawaii, *American Fern Journal*, Vol. 61, Number 2, pp 49-58.