SPAM: Spatial Autocorrelation Modelling

Andrew J. R. Gillespie, Syracuse University Research Consulting Services

ABSTRACT

Given (1) a set \(\{x_1, x_2, ... x_n\}\) of \(n\) observations on some spatially distributed variable of interest \(X\) and (2) a rule to determine which pair \((x_i, x_j) : i \neq j\) are considered to be "neighbors", then spatial autocorrelation is defined as a condition where the values of individual observations are dependent upon the values of their respective neighbors. An index of autocorrelation is thus dependent not only upon the observations on \(X\), but also on the hypothesis of structure implied by the rule used to define neighbors. This permits researchers to develop and test hypotheses of relationships among spatially distributed data. Such tests should be made before assumptions of spatial autocorrelation are built into a model.

A program called SPAM, written in SAS Interactive Matrix Language (SAS/IML), facilitates the development and testing of such hypotheses. The modeller may either choose from one of several directionally-based default structures for defining neighbors, or may enter the structure information along with the observations on \(X\). An index of autocorrelation (Moran's \(I\) statistic) is computed along with confidence intervals under two statistical assumptions. Additionally, the modeller may choose to calculate multiple \(I\) values based on differing numbers of "steps" between neighbors and use built-in plotting functions to plot a correlogram showing spatial autocorrelation as a function of the intra-neighbor distance implied by the number of "steps".

INTRODUCTION

When data are collected from a two-dimensional area, for example a forest or a country, researchers are often interested in the spatial distribution of those data: that is, how the values of the observations change as one moves through the study area. One aspect of such questions concerns spatial autocorrelation, which is a measure of the degree to which the value of a random variable at one location is dependent upon the values of observations at neighboring locations. When neighboring observations tend to have similar values, they are said to be positively autocorrelated; if they tend to have very different values, they are negatively autocorrelated. If the values are distributed independently of their neighbors, then they are spatially uncorrelated.

Spatial autocorrelation is relevant in many fields. Geographers have long been interested in the spatial autocorrelation of distributions of human settlements, flows of technology, and the spread of epidemics. Much of the early work in spatial autocorrelation analysis was done by geographers, and is being adopted by researchers in other fields. For example, geologists study spatial autocorrelation to improve their ability to predict oil well locations based on core samples. Ecologists study the spatial autocorrelation of many phenomenon, including biomass distributions and the movement of genotypes through an environment. Cartographers and others who construct maps based on collections of spatial data assume implicitly that their data are spatially correlated; this assumption allows them to draw contours to indicate which areas are similar with respect to some variable of interest.

Although positive or negative spatial autocorrelation is often assumed, the assumption is seldom tested. This may be because the researcher lacks a handy tool for estimating and evaluating the degree of spatial autocorrelation. One of the purposes of this paper is to provide such a tool.

Alternately, some researchers are more interested in looking for spatial trends in field data, or developing and testing hypotheses about how the data are related. For example, an ecologist might hypothesize the existence of a directional gradient in the spatial distribution of some variable. Thus the second purpose of this paper is to illustrate a tool which may be used for exploratory data analysis in the context of seeking trends.

In their comprehensive treatment of spatial autocorrelation, Cliff and Ord (1973) give a thorough discussion of the mathematical theory behind
several estimators of spatial autocorrelation. In particular, they
found through simulation studies that the most generally well-behaved
estimator of spatial autocorrelation is Moran's I statistic. This paper
focuses on application of Moran's I to interval data; the special case
of categorical data is not addressed.

A MEASURE OF SPATIAL AUTOCORRELATION

Measures of spatial autocorrelation depend on both the
vector \( \mathbf{x} = (x_1, x_2, \ldots, x_n) \) of observations on the variable of
interest, and on the rule used to
determine which pairs \( \{(x_i, x_j) : i \neq j\} \) are considered to be neighbors. The
outcome of applying the neighbor rule
to \( \mathbf{x} \) is concisely expressed by a
connection matrix \( \mathbf{W} \), where the
elements \( w_{ij} \) indicate the degree of
dependence of \( x_j \) upon \( x_i \). In the
simplest (binary) case, we have
\( w_{ij} = 1 \) if \( x_i \) is determined by some
criteria to affect \( x_j \) (and thus
constitute a neighbor), and \( w_{ij} = 0 \)
otherwise. Setting \( w_{ij} = w_{ji} \) implies
directional independence between
neighbors, in which case \( \mathbf{W} \) is
symmetric. In other cases, it may be
of interest to make \( w_{ij} \) not
necessarily equal to \( w_{ji} \); for example
in a model of a stream, where
upstream conditions might affect
downstream but not vice versa. The
most complicated model is both
asymmetric and nonbinary, with the
\( w_{ij} \) estimated according to some
hypothesized function of distance
and/or direction. If the vector \( \mathbf{x} \)
contains \( n \) observations, then \( \mathbf{W} \) will
be an \( n \times n \) matrix. The formulae
below are generalized to treat
asymmetric and non-binary connection
matrices.

One estimator of spatial autocorrelation is given by Moran's I
statistic (Moran, 1950):

\[
I = \frac{n \sum w_{ij} z_i z_j}{\sum w_{ij}^2}
\]

(1)

where \( z_i = (x_i - \bar{x}) \), \( \bar{x} \) is the mean of
the vector \( \mathbf{x} \), \( \mathbf{z} = \mathbf{x} - \bar{x} \), a vector of
\( z_i \) values, and \( I \) is an \( n \times n \) matrix
of 1's. It can be shown (Cliff and Ord 1973, 1981) that \( I \) is
asymptotically normally distributed
as \( n \) increases, permitting
construction of confidence intervals

and hypothesis tests based on
standard normal deviates.

The expected value of \( I \) is

\[
E(I) = -\frac{1}{n-1} = -\frac{1}{(n-1)^2}
\]

(3)

The variance of \( I \) may be derived
under two different statistical
assumptions. Under the assumption of
Normality, the variable \( \mathbf{x} \) may be
assumed to be Normally distributed,
with the values \( x_i \) being independent
observations from a Normally
(Gaussian) distributed population. Alternately, under the assumption of
randomization, the distribution of \( X \)
is unspecified, and the value of \( I \) is
considered relative to all \( (n!) \)
possible permutations of the values
\( x_i \) over the \( n \) points. Note that
these two assumptions, Normality and
randomization, concern the
statistical distribution of the
variable \( \mathbf{x} \), not the spatial
distribution; no assumptions are
required about the spatial
distribution, other than that of the
hypothesized structure implied by \( \mathbf{W} \).

Under the normality assumption, the variance of \( I \) is:

\[
V_N(I) = \frac{(n^2 S_1 - n S_2 + 3 \sum w_{ij}^2)}{(\sum w_{ij}^2)^2 (n-1)} - \frac{1}{(n-1)^2}
\]

(4)

Under the randomization assumption, the variance of \( I \) is:

\[
V_R(I) = \left[ \frac{[n(n^2-3n+3) S_1 - n S_2 + 3 \sum w_{ij}^2]}{(n-1)(n-2)(n-3)(\sum w_{ij}^2)^2} \right]
\]

(5)

where \( S_1 = (\sum w_{ij} + w_{ji})^2 / 2 \),
\( S_2 = \sum (w_{ij} + w_{ji})^2 \), and
\( w_{ij} = \frac{1}{n} \sum w_{ij} \), \( w_{ji} = \frac{1}{n} \sum w_{ij} \).

For proofs and derivation, the
interested reader is referred to
Cliff and Ord (1973).
CONSTRUCTION OF THE CONNECTION MATRIX

Consider the hypothetical grid of data in Figure 1 (based on an example in Sokal and Oden 1978). The numbers in the cells correspond to nine observations on some variable X. With such gridded data, one criteria for constructing connection matrices might be orthogonal axes. For example, the Rook's move (from chess) joins each cell to the four cells with which it shares a border (north, east, south, and west). The Bishop's move joins each cell to the four cells adjoining to opposite corners: northeast, northwest, southeast, and southwest. The Queen's move combines Rook and Bishop moves. If one were to calculate I for the data in Figure 1 under each of these moves, one would find that:

1. I is strongly positive for the Bishop's move, since all pairs of neighbors have identical values;
2. I is strongly negative for the Rook's move, since all pairs of neighbors have very different values;
3. I approaches 0 for the Queen's move, since each observation has both similar and different neighbors.

Alternately, a connection matrix may be designed independently of specific directions. Ecologists may be interested in basing connections on geographical closeness, after accounting for some biological model or ecological barrier (Sokal and Oden, 1978a). For example, a Gabriel-connected graph (Gabriel and Sokal, 1969, cited in Sokal and Oden, 1978a) joins two sample points A and B if no other sample point C lies on or in the circle whose diameter is the line AB and whose center is the midpoint of AB. One version of SPAM allows the modeller to enter connection information in the DATA step, with each observation on X accompanied by $n_i = \text{the number of points connected to } x_i$, and, for $j = 1$ to $n_i$, the adjoining point number $k$, observed value $x_k$, and connection weight $w_{ik}$.

SPAM is more flexible than the example above in that it permits the modeller to choose any subset of the eight cardinal directions included in the Queen's move. Users may use one of the structures described above, or may design their own connection matrix based on one or more of the eight directions.

Alternately, a connection matrix may be designed independently of specific directions. Ecologists may be interested in basing connections on geographical closeness, after accounting for some biological model or ecological barrier (Sokal and Oden, 1978a). For example, a Gabriel-connected graph (Gabriel and Sokal, 1969, cited in Sokal and Oden, 1978a) joins two sample points A and B if no other sample point C lies on or in the circle whose diameter is the line AB and whose center is the midpoint of AB. One version of SPAM allows the modeller to enter connection information in the DATA step, with each observation on X accompanied by $n_i = \text{the number of points connected to } x_i$, and, for $j = 1$ to $n_i$, the adjoining point number $k$, observed value $x_k$, and connection weight $w_{ik}$.

COMPUTATION OF I

The matrix versions of formulae given in the above equations are beautifully elegant but generally useless for computation of I, since they depend on $W$. For interesting problems, $n$ may well exceed 200 observations (for example, a 20 x 10 grid); this implies that $W$ would be 200 x 200, which can lead to space limitation problems in IML (or other procedures). Moreover, while large, $W$ is apt to be very sparse, since each observation $x_i$ is likely to be connected to only a small subset of all observations on X. SPAM forgoes the elegant matrix approach in favor of a more efficient iterative vector solution which works with one row of $W$ at a time. This permits analysis of fairly large problems. Results of computations include the estimated and expected value of $I$, as well as estimates of the variance under the two distributional assumptions (normality and randomness) and the t statistic for testing the hypothesis that $I = 0$. These results are reported in the listing file as well as written to an output vector for subsequent plotting in a correlogram.

CORRELOGRAM PLOTTING

Once the I statistic has been calculated, the researcher may be interested in how spatial autocorrelation changes with the distance used to define neighbors. For example, one can calculate I based on pairs of nearest neighbors (the first order I) and compare it to an I value based on points which are separated by 1, 2, or more points in some direction (higher order I's). A plot of I vs. Order is called a
Correlograms are commonly used in econometric studies of time series where autocorrelation coefficients are repeatedly calculated for different lags (time intervals between any pair of measurements). Figure 2 is an example of a correlogram up to the seventh order.

Figure 2. Sample correlogram.

Correlograms may be distance independent, where distances between points are assumed to be constant and uniform; or they may be based on categories of distances between points, with the actual number of points between neighbors being irrelevant. In either case, a connection between any two points is included in no more than one calculation of I. SPAM includes an option to plot distance independent correlograms. Each plot shows the observed values of I for the respective orders, as well as lines denoting 95% confidence limits under both normality and randomness. Distance dependent correlograms might be plotted by using several DATA steps to create several W matrices corresponding to different distances. Sokal and Oden (1978a) discuss biological implications of analysis of spatial autocorrelation and correlograms. Cliff and Ord (1981) contain a more general discussion, with examples based on geographical studies.

**EXAMPLES**

1. **Positive Spatial Autocorrelation**

Consider the following grid of data:

```
   1  5  7  11
   4  8 11 15
   8 12 14 17
  11 14 18 20
```

There is an obvious trend of increasing values from the northwest corner to the southeast corner. This is a typical example of positive spatial autocorrelation: values next to each other are similar to each other, although they change gradually over time. The sample statistics are:

- \( n = 25 \)
- \( E(I) = -0.0417 \)
- \( I = 0.6469 \)
- \( V_R(I) = 0.009934 \)
- \( V_R(I) = 0.0102 \)

The two-tailed \( t \) statistic for the null hypothesis: \( (I = 0) \) is 6.8 and 6.9 under Randomness and Normality respectively. This implies that the spatial autocorrelation statistic is significantly different from 0 at the \( \alpha = 0.05 \) significance level.

2. **Negative Spatial Autocorrelation**

Consider the following grid of data:

```
   1 12 33 8
   33 22 3 15
   4 15 34 23
  36 27 4 12
```

These data are typical of negative spatial autocorrelation, with dissimilar values adjacent to each other. Negative autocorrelation is often associated with clustered data, where the values within clusters are similar but values between clusters are very different. The sample statistics are:

- \( n = 16 \)
- \( E(I) = -0.0667 \)
- \( I = -0.02634 \)
The two tailed statistic for the null hypothesis: \( (I = 0) \) is -1.62 and -1.68 under Randomness and Normality respectively. This implies that the negative autocorrelation coefficient is not statistically significant at \( \alpha = .05 \).

**SUMMARY**

The SAS Interactive Matrix Language (SAS/IML) has been used to evaluate spatial autocorrelation in spatially distributed interval data using Moran's I statistic. This method is suitable for testing assumptions of spatial autocorrelation prior to model building, or for exploring possible trends in spatial data. Correlograms are drawn to show the effect of spatial autocorrelation as a function of distance.

Copies of the SAS/IML code used in SPAM are available upon request from Syracuse University Research Consulting Services, 732 Ostrom Avenue, Syracuse University, Syracuse NY 13244. Bitnet: RESCON@SUVM. I thank Dr. R. Kalinoski of Syracuse University for many interesting discussions on this topic, and for sharing some data.

**LITERATURE CITED**


1Current address: Institute of Tropical Forestry, Box 25000, Rio Piedras, Puerto Rico 00928-2500.