

Appendix: Permutation tests

Measures of Spatial Autocorrelation

The objective is to measure how strong the tendency is for observations from nearby regions to be more (or less) alike than observations from regions farther apart, and then judge whether any apparent tendency is sufficiently strong that it is unlikely to be due to chance alone.

We focus on spatial autocorrelation measures for two types of response variables;
Examples of spatial autocorrelation for binary data:

- *Positive autocorrelation:*

1	1	0	0
1	1	0	0
0	1	1	0
1	1	0	0

- *Negative autocorrelation:*

1	0	1	0
1	1	0	1
1	0	0	1
0	1	1	0

- *No autocorrelation:*

1	1	0	0
0	0	1	0
0	1	0	0
0	0	1	1

1. The General Cross-Product Statistic:

Notation:

- Let Z_i denote the response at the i th location.
- Let Y_{ij} be a measure of how similar or dissimilar the responses are at locations i and j
- Let W_{ij} be a measure of the spatial proximity of locations i and j .

The general cross-product statistics is

$$C = \sum_i \sum_j W_{ij} Y_{ij}$$

e.g., with binary Z_i 's,

1	1	0
1	1	0
0	0	0

for binary data, we define $Y_{ij} = (Z_i - Z_j)^2$ and

$$W_{ij} = \begin{cases} 1 & \text{if locations } i \text{ and } j \text{ are adjacent} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

C is 8 for the table above.

Note:

- C too small \Rightarrow positive spatial autocorrelation.
- C too large \Rightarrow negative spatial autocorrelation.

2. How should the statistical significance of C be judged?

- Comparison to randomization distribution
 - * List all possible arrangements of the observed responses over the locations obtained by permutation of responses.
 - * Compute C for each arrangement, and rank there.
 - * Determine where the data's C -value fits in; P-value for the test is the number of C -values in the randomization distribution as extreme or more extreme than the observed C .
 - * Can do one-sided or two-sided tests.
 - * Example: 9 elements, 4 of "1" type, 5 of type "0". The number of different arrangements is $\frac{9!}{4!5!} = 126$, for instance for the arrangement:

1	1	0
1	1	0
0	0	0

the value of C is 8.

- Monte Carlo approach
 - * Motivated by the fact that complete enumeration of the possible arrangements may be computationally prohibitive even for moderately-sized data sets.
 - * So instead, obtain a random sample from the randomization distribution and follow the same type of procedure.
 - * To implement this random sampling, you merely need to generate n random numbers (one for each data location), rank these random numbers from smallest to largest, then re-arrange the observations in accordance with the ranking of random numbers. C is computed for this arrangement, and the whole process is then repeated m times.

- * The P-value then estimates the proportion of C-values as extreme or more extreme than the observed C , and is given by

$$P = \frac{1 + \text{number of } C - \text{values} \geq \text{observed } C}{1 + m}$$

- * Example presented in class.
- Normal approximation, i.e. $C \sim N(E(C), \text{var}(C))$
- * Define

$$S_0 = \sum_{i \neq j} W_{ij}$$

$$S_1 = \frac{1}{2} \sum_{i \neq j} (W_{ij} + W_{ji})^2$$

$$S_2 = \sum_i (W_{i \cdot} + W_{\cdot i})^2$$

- * Define T_0 , T_1 , and T_2 similarly but for the Y_{ij} 's.
- * Compute $z = \frac{|C - E(C)| - 1}{\sqrt{\text{var}(C)}}$ where $E(C) = \frac{S_0 T_0}{n(n-1)}$ and

$$\text{var}(C) = \frac{S_1 T_1}{2n(n-1)} + \frac{(S_2 - 2S_1)(T_2 - 2T_1)}{4n(n-1)(n-2)} + \frac{(S_0^2 + S_1 - S_2)(T_0^2 + T_1 - T_2)}{n(n-1)(n-2)(n-3)} - [E(C)]^2$$

- * Example (presented in class).

3. Joint-Count Statistics

These are a subclass of general cross-product statistics which are for use with binary data.

Code the data as either 1 (black) or 0 (white).

- Classify the “joins” between contiguous regions as BB, BW, or WW.
- Define $W_{ij} = 1$ if regions i and j share an edge, and 0 otherwise. This definition of contiguity is called the rooks’ definition.
- Count the number of joins of a specified type, e.g. the # of BW joins \equiv BW.

Observe that if we define $Y_{ij} = (Z_i - Z_j)^2$, then

$$C = \sum_i \sum_j W_{ij} Y_{ij} = 2BW$$

i.e. $BW = C/2$.

Likewise, $BB = C^*/2$ where C^* is the value of C obtained by defining $Y_{ij} = Z_i Z_j$.

If the total # of joins in the system is J , then $WW = J - BB - BW$. Some evidence exists that BW is slightly *more sensitive* than the other two, so we will consider BW in further detail.

Notation and results associated with the use of BW

- Let $b = \#$ black regions and $w = \#$ white regions; $b + w = n$.
- Note $E(BW) = \frac{1}{2}E(C)$ and $\text{var}(BW) = \frac{1}{4}\text{var}(C)$.
- Can be shown that

$$T_0 = 2bw,$$

$$T_1 = 2T_0,$$

$$T_2 = 4nbw.$$

- If regions form a rectangular rc lattice, and the rook's contiguity definition is used, then

$$S_0 = 2(2rc - r - c),$$

$$S_1 = 2S_0,$$

$$S_2 = 8(8rc - 7r - 7c + 4).$$

Remarks:

- The same approach can be used for data at irregularly spaced and shaped locations, though the formulas given for S_0 , S_1 and S_2 no longer apply. (The formulas given for T_0 , T_1 and T_2 are still okay though.)
- The test presented in the example is two-sided but a one-sided test, if more appropriate, can be done easily.
- The same approach can be used for BB (and WW) joins. In the case of BB joins,

$$T_0 = b(b - 1),$$

$$T_1 = 2T_0,$$

$$T_2 = 4b(b - 1)^2.$$

- Extensions to polytomous categorical data (i.e. a multi-colored map) are possible.

4. Moran's and Geary's Statistics (Continuous Data)

Moran (1950, *Biometrika*) proposed the following statistics which can be used with continuous data:

$$I = \frac{n \sum_i \sum_j W_{ij} (Z_i - \bar{Z})(Z_j - \bar{Z})}{S_0 \sum_i (Z_i - \bar{Z})^2}$$

where $\bar{Z} = \sum_i Z_i/n$.

Geary (1954, *The Incorporated Statistician*) proposed a similar statistic:

$$c = \frac{n-1}{S_0} \frac{\sum_i \sum_j W_{ij} (Z_i - Z_j)^2}{\sum_i (Z_i - \bar{Z})^2}$$

Remarks:

- Note the superficial resemblance of I to the *ordinary correlation coefficient*.
- Note that $I = \frac{n}{S_0 \sum_i (Z_i - \bar{Z})^2} C$ if we take $Y_{ij} = (Z_i - \bar{Z})(Z_j - \bar{Z})$
- Similarly, c might be related to C by taking $Y_{ij} = (Z_i - Z_j)^2$
- Consequently, I is more sensitive to extreme Z - values whereas c is more sensitive to differences between pairs of Z -values.
- However, I is more popular than c , so we'll only consider I further.
- $E(I) = -\frac{1}{n-1}$ under independence.
- $I < -\frac{1}{n-1} \Rightarrow$ negative autocorrelation.
- $I > -\frac{1}{n-1} \Rightarrow$ positive autocorrelation. (Note: large values of $c \Rightarrow$ negative autocorrelation.)
- NORMAL APPROXIMATION to distribution of I under independence ($n > 25$)
For more information about the Normal Approximation:
(Cliff and Ord, 1981, *Spatial Processes: Models and Applications*, p. 46)
- For smaller sample sizes, can use randomization distribution or Monte Carlo approach to evaluate significance.

5. Generalized cross-product

So far our discussion of join-count has assumed that the W_{ij} 's are binary (0 or 1). In many situation we may be able to measure spatial proximity on a more refined scale (as we do the Y_{ij} 's in going from BW to I or c).

1. Use lengths of common boundary
2. Use actual distance between locations or centroids of locations. This recognizes the fact that interaction between sites does not usually terminate sharply beyond places that share a boundary.
3. Incorporate directionality by allowing $W_{ij} \neq W_{ji}$

A side benefit of using non-binary W_{ij} 's is that the distribution of the test statistic under independence is better approximated by the normal distribution.

6. Spatial Autocorrelation Functions

The statistics considered so far attempt to express information about spatial autocorrelation in a single number. alternatively, we could consider regarding spatial autocorrelation as a function of distance. That is:

- Divide the range of distances into q classes.
- Compute a previously considered spatial autocorrelation measure, e.g. I , once for each of the q distance classes; in other words, we use only those pairs of locations that are within the same distance class.
- plot the statistic, e.g. I_d , versus d . Such a plot is called the *correlogram* corresponding to that statistic.

This last notion is essentially what did with geostatistical data, when we measured spatial dependence via the covariance function or semivariogram.