

About Spatial Statistics

From Fotheringham et al., *Geographically Weighted Regression*; Bailey & Gatrell, *Interactive Spatial Data Analysis*; & Mitchell, *The ESRI Guide to GIS Analysis*, vol. 2.

- *Aspatial* data contain only attribute (i.e. feature) information, while *spatial* data contain both attribute and locational information.
 - 'Local attribute space' vs. 'local geographic space' (see Fotheringham et al., 3-6).
 - *Stationarity*: a relationship is locationally uniform across space (i.e. is 'global'); differences in values may depend on the *relative location* of the measurements (i.e. their distance and direction between two points) but not on the *absolute location* of the measurements.
 - *Global* statistics are valid for stationary relationships because the values of the observations are independent of each other.
 - *Non-stationarity*: a relationship varies by *absolute location* across space (i.e. has 'local' features).
 - *Local* statistics are valid for non-stationary relationships because the values of the observations are spatially dependent on each other (i.e. *Tobler's First Law of Geography*: nearby objects are more alike than are objects that are farther away—*spatial autocorrelation*).
- Global trend is also called *first order*. Local trend is also called *second order*,
- *Non-stationarity* can be anticipated for several reasons.
 - Spatially non-random human and/or software sources of error (i.e. non-sampling error that is spatial).
 - Spatially non-random sampling error.
 - Model misspecification that has spatial effects.
 - Intrinsic spatial variation in a relationship.
- There are basic problems in trying to significant detect local patterns within global trends:
 - The global (i.e. first order) trend itself may be an artifact of the scale of geographic aggregation ('modifiable areal unit problem').
 - Testing for significance in local patterns is vulnerable to Type I error in multiple hypothesis testing.
- *Directional patterns of relationships*:
 - *Isotropic*: the relationship varies only with distance between points.
 - *Anisotropic*: the relationship varies by distance and direction between points.
- *Three fundamental questions*:
 - What are the spatial patterns? These must be visualized.
 - What are their dimensions? The patterns must be quantified.
 - How can we explain them? The patterns must be modeled.
- "A *model* is a simplification of, and approximation to, some aspect of the world" (King et al., *Designing Social Inquiry*, 49).

- Models can be assessed as more or less plausible in the ways they *abstract* features of 'reality' as we conceptualize it.
- *Practical problems in analyzing spatial data*
 - *Geographic scale of analysis*: patterned variation at one scale may be mere random variation at another scale.
 - *There is no natural ordering in space* (i.e. nothing like time intervals); i.e. much spatial data involves irregularly distributed sets of sites or spatial dependence that extends in various directions.
 - *How to conceptualize and measure 'boundary' or 'edge' phenomena*: because unless it is, say, a coastline, a phenomenon is likely to occur on the other side, on which we may not have collected data, meaning that the sample is spatially biased.
 - *Results of data analysis may depend on the data's level of aggregation of zonal measurements*: 'the modifiable areal unit problem'.
 - *Complexity of topography*: 'distances,' 'boundaries,' 'edge effects,' and the like are artefacts of the simplifying nature of models.
- *Four types of spatial data*
 - *Point patterns*: e.g., 'unbounded' location counts of population, homicides, illnesses, voting, wealthy households, poor households.
 - *Spatially continuous*: e.g., climate, air pollution, groundwater levels (see 'geostatistics' and 'kriging').
 - *Area*: e.g., population, socioeconomic, voting, or crime data by district, city, or country.
 - *Spatial interaction*: flows between origins and destinations such as investment, trade, transportation, migration, trips.
- *Common approach to spatial statistics*: conceptualize a spatial variable as having two components (Bailey & Gatrell, 33):
 - A first-order component: global, uniform spatial variation (linear or non-linear) in a variable's mean and variance.
 - A second-order component of deviations from the mean whose structure may include local effects.

Fundamental questions about data uncertainty and bias

- *Fundamental questions about data uncertainty and bias*:
 - Basic questions need to be asked about the design of the data collection (e.g., sampling design), the conceptualization and measurement of the variables, sources of non-sampling error (e.g., GPS imprecision, misregistration of aerial photos), and the spatial structure of error.
 - Additional questions about uncertainty inherent in GIS presentation of data: e.g., map projections, map precision (i.e. number of significant digits and rounding), lengths and areas, fuzzy classifications.
 - *Measure of average error (statistical deviation)*: root mean square error—the square root of the average squared error (see Freeman et al., *Statistics*; and Longley et al., *Geographic Information Systems and Science*).
 - Basic questions need to be asked about who's conducting and sponsoring the research, for what purpose, and under what spatio-

temporal political, social, and cultural conditions—GIS and the social construction of reality.

- "... many problems of error propagation in GIS are not amenable to analysis" (Longley et al., 338).
- It's important to take a critical perspective on data, explicitly recognize the various levels of data uncertainty and bias, compare multiple data sources whenever possible, and report the uncertainties and biases.
 - See Longley et al. (342) on using 'sensitivity analysis' to check for errors/bias in raster data.
- *Hypothesis testing*: the use of GIS data for *inferential* statistics is valid only if the data can be reasonably defended as a random and independent sample of a specific population. Otherwise the data are useful for *descriptive* statistics only.

Geographically weighted regression

- See Anselin ('Spatial data analysis with GIS') about 'spatially lagged variables' as "An easy way to compare the value at a location to that of its 'neighbors'": a *spatial lag variable* is a weighted average of the values in neighboring locations.
 - Spatial lag: measures how characteristics of nearby spatial units or populations affect a local area.
- *Geographically weighted regression* (see Fotheringham et al.): tests for significance of spatial vs. aspatial effects; can combine spatial and aspatial variables; can test the relative plausibility of models; can map various pertinent statistics (e.g., y-intercept, slope coefficients, standard errors, t-values, residuals, diagnostic test results); provides the possibility for mapping variation *within* district sub-units of a larger space; controls for lurking variables (as does aspatial regression).
 - Use aspatial and spatial programs to graphically and numerically examine distributional features of variables (pronounced skewness, outliers), and to make appropriate modifications.
 - *Problems*: modifiable areal unit problem and choosing search kernel type and bandwidth; other problems intrinsic to regression analysis (including multiple hypothesis testing and Type I error); early stage in the software's development.
 - *GWR software program*

Geostatistics

- *Geostatistics* is a part of the more general field of spatial statistics. Using continuous, surface data, it quantifies the effect of location on sample variability.
 - *Spatial autocorrelation*: correlation between spatial random variables depends on the distance and/or direction between locations.
 - *Intrinsic hypothesis (intrinsic stationarity)*: the most basic premise of geostatistics—the distribution of the difference between pairs of sample points is the same across the study area, which means that the distribution depends on relative location but not on absolute location

(i.e. the differences are consistent but not constant across the study space).

- *Kriging* is a geostatistical estimation technique that uses a linear combination of surrounding sampled values to minimize errors in making predictions about a surface in unsampled areas.
 - Derive weights to apply to each zone of sampled data; and model the xy spatial autocorrelation structure via a *variogram*.
 - *Variogram* (or *semi-variogram*, which is half of the variogram): a function of the distance and direction between two locations, which quantifies their spatial autocorrelation.
 - It is the variance of the difference between two variables at two locations: the distance between paired locations (x-axis) and the squared difference of the values of all linked pairs of locations (y-axis).
 - The *variogram model* assumes that an xy spatial relationship has:
 - A *trend*: a first-order, global component of a constant mean value:
 - *Forms*: linear, spherical, exponential, or Gaussian.
 - Use the *histogram* in *Geostatistical Analysis* to check for data normality and outliers, and use the associated tool to transform the data as necessary (e.g., toward normality).
 - Use the *Trend Analysis* tool to detect a trend.
 - A random, spatially correlated, second-order component, which may be uniform or may have direction (isotropic or anisotropic).
 - A *residual error* term.
 - *Covariance*: a statistical measure of the correlation between two variables.
 - *Lag*: a distance-class interval (of approximately equal distances and directions between any two locations) that is used for variogram computation.
 - Lag is the vector that separates any two locations. As a vector, it has both distance and direction.
 - *Bin*: a classification of lags according to similar distance and direction.
 - Lag intervals are chosen via trial and error, trying to maximize detail at small lags without capturing random error alone.
 - Usually 50-80% of the maximum pair distance is the highest lag distance used.
 - A rule of thumb: multiply the lag size times the number of lags, which should be about half of the largest distance among all points. Also, if the range of the fitted semivariogram model is very small relative to the extent of the empirical semivariogram, then you can decrease the lag size. Conversely, if the range of the fitted

semivariogram model is large relative to the extent of the empirical semivariogram, you can increase the lag size.

- *Nugget*: the magnitude a variogram's discontinuity (i.e. random error) at the origin—the values of a variable at locations that are very close together; such discontinuity (i.e. the existence of a nugget) is based on amounts of measurement error and/or spatial discontinuity; it is the y-intercept.
- *Range*: the distance to where the variogram begins to flatten out.
 - The larger the range of a variogram (i.e. the distance at which it becomes a constant), the smoother the surface.
- *Sill*: the value that the variogram attains at the range (i.e. the value on the y-intercept).
 - The sill equals the variance of the random variable.
 - The higher the *sill of a variogram* (i.e. the upper limit of any variogram model; its value for distances beyond its range), the higher the prediction variances.
 - *Partial sill*: the value of the sill minus the nugget.
 - Zonal anisotropy exists when there are different sills in different directions (as detected by *anisotropic variograms*), but there may be both geometric and zonal anisotropies.
- *Search neighborhood*: an elliptical area centered on a point or block being kriged.
- *Cross-validation*: checks for severe model-fit problems by specifying a variogram model and a search neighborhood, kriging values at each sampled location (leaving a particular value out), then comparing the kriged and sampled values. The difference between them is the *cross-validated residual*.
- *Simple Kriging*: assumes the local means are constant and equal to the population mean, which is known.
- *Ordinary Kriging*: assumes that the local means aren't necessarily closely related to the population mean and so only uses the samples in the local neighborhood of the estimate.
- *Universal Kriging* is appropriate if there is a gradual trend in the data so that the mean is no longer spatially constant and the variogram is no longer appropriate to model the xy spatial autocorrelation structure.
 - We assume that we know the shape of the trend (either first order ['linear drift'] or second order ['quadratic drift'] and kriged the residuals. See the *intrinsic hypothesis* for a weaker assumption.
- *Cokriging*: a more abundantly sampled variable is combined with a less abundantly sampled variable to make the surface predictions.
 - We must assess the cross-relationships of the two sampled variables via a *cross-variogram*.
- To kriged a surface, in sum, obtain data at sampled, discrete locations; make predictions for the values at unsampled locations; and obtain the prediction errors.
 - E.g., using continuous, surface data, in *Geostatistical Analyst* use *Explore/Histogram* to check for data normality/ outliers (click on low

end to highlight close points/close values and upward to high end to highlight far points/far values), if necessary, use the tool to transform toward normality.

- Does *Explore/Trend Analysis* tool reveal a trend (i.e. first-order effect)?
 - If so, detrend the data to measure the second-order effect via the semivariogram (which plots on the x-axis difference in the location of two points, and on the y-axis the squared difference of the values at the two points).
 - But keep the model as simple as possible: only detrend the data if doing so significantly decreases the prediction errors.
- Use *interpolation* procedures (e.g., appropriate types of Kriging) to predict the surface values at unsampled points and to obtain the prediction errors, and compare the results for best model fit/lowest prediction errors.
- *Semivariogram*: see above on choosing *lag size*; choose *model* form, *anisotropy* (if appropriate), *nugget*, *search direction*.
- *Searching Neighborhood*: choose *Method*, *Neighbors to Include*, *Include at Least*, *Shape Type*. Dots highlighted red will exert the most influence (due to their close location), while dots not colored will exert no influence. The circle displays the area of influence as determined by the semivariogram's range.
- *Cross-validation of model*: predicts a value at each observed location deleting that observed location from the data, thus providing a measure of the model's fit. Examine the 'Predicted' chart (by clicking the tab). The lower the *root-mean-square*, the better the fit.
- Create appropriate kinds of maps and evaluate them; right-click a map layer & click *Compare*: lower root-mean square is better.
- Clip the interpolated map if necessary to the original map:
 - After checking that the correct features are selected in the toolbar dropdowns, *Data Frame Properties/Data Frame/click 'Enable'/ click 'Specify Shape'*.
 - In table of contents, right-click the map's layer name, click *Properties/Extent/dropdown arrow/ set the extent* as required.