Statistical methods appendix

Spatial autocorrelation (dependency): the covariation of properties within geographic space, where characteristics of nearby locations appear to be correlated. This correlation can be either positive (similar values) or negative (opposite or different values).[1,2]

Moran's I: a measure of spatial autocorrelation that is an extension of standard correlation (frequently expressed as *r* in linear models) to a geographic space. The named *I* can be utilized in hypothesis testing (as a z-score) to define if this arrangement is more or less than expected.[1,2]

Bivariate local Moran's I: an extension of standard bivariate correlations to spatial analysis. Local Moran's I identifies spatial associations that are stronger than expected between one variable (inpatient deaths) in one location and another variable (gunshot deaths) in another location. This yields an analysis that identifies geographic clusters that are outliers with respect to other population means.[1,2]

K-means clustering: this is one of the most commonly used clustering procedures in unsupervised learning. In this technique, a total of n observations gets partitioned into k different sets, each of them named as clusters and each observation belongs to the cluster with the nearest mean.[3]

Hierarchical clustering: this clustering procedure is based on more information than in the K-means clustering, even though the idea is same: attempts to group observations with similar features into clusters. It uses some sort "tree based" procedure in making the decision of creating clusters. Two ways can be done: "top down" and "bottoms up". In both situations, all observations first belong to one cluster; then splitting is performed recursively.[4]

Hierarchical clustering with spatial constraints: this technique falls under the concept of hierarchical clustering, but it implements clustering procedure to incorporate spatial information from the data. This distance-based procedure uses a well-known algorithm called "ward-like" algorithm on two dissimilarity matrices. The first matrix is 'feature space' for example, socio-economic variables and the second matrix is 'constraint space', for example geographical distances. A mixing parameter which takes value between 0 and 1, helps in building relationship between those two matrices to develop underlying clusters in the data set.[5]

Within cluster sums of squares (WCSOS): this is a tool to compare several clustering procedures and helps to choose the best one which explains the data most. The WCSOS is a measure of the variability of the observations within each cluster. In general, a cluster that has a small sum of squares is more compact than a cluster that has a large sum of squares.

References

1. Anselin L, Syabri I, Kho Y. GeoDa: An Introduction to Spatial Data Analysis. *Geogr Anal.* 2006;38(1):5-22.

2. Anselin L, Syabri I, Smirnow O. Visualizing Multivariate Spatial Correlation with Dynamically Linked Windows. In: *New Tools for Spatial Data Analysis: Proceedings of the Specialist Meeting.* University of California, Santa Barbara: Center for Spatially Integrated Social Science (CSISS); 2002.

3. Forgy, EW. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*. 1965;21(3):768-769.

4. Hartigan, JA. *Clustering Algorithms*. Hoboken, NJ: John Wiley and Sons; 1975.

5. Chavent M, Kuentz-Simonet V, Labenne A, et al. ClustGeo: an R package for hierarchical clustering with spatial constraints. *Comput Stat.* 2018;33(4):1799-1822. doi:10.1007/s00180-018-0791-1