Marco Ruocco
G274 Introduction to Geographical Data Analysis
Final Take-Home exam

Review of the paper:
Webster, R., Oliver, M.A., Muir, K.R., and Man, J.R., (1994) "*Kriging the Local Risk of a Rare Disease from a Register of Diagnoses*", Geographical Analysis, Vol.26,p. 168-185

The basic idea of this study is that from local information based on electoral ward area about frequency of disease and total number of children it is possible to derive a map of risk over the entire region, using geostatistical methods and in particular the cokriging prediction scheme. Every area is considered as a point (its centroid), so that the geostatistical problem is one of predicting points on a uniform grid using irregularly distributed data points. The specific problem is that the characteristics of the frequency distribution are different from those of the risk, and in particular it is needed to make predictions on one variable (risk) using correlation information concerning another variable (frequency). A second problem, related to the database used, concerns the adoption of electoral wards as spatial location of healthy children and as reference areas for the analysis, while the disease occurrence data is available with the precise address for each child. The first problem is solved statistically, the second with a sub-optimal solution.

Geostatistics has been originally applied in mining problems, such as in the prediction of ore distributions, but now is applied to spatial data in general, as this case confirms. The general initial condition for a geostatistical analysis is to have sparsely sampled point with a known value, belonging to a variable that can be described by a variogram, a graph depicting the correlation of a variable at a range of distances. Such variogram can be used for predicting estimates of the variable in specific points, which are unbiased weighted averages of the data with minimum variance.

The study of the distribution of human diseases is in this case devoted to the rare non-contagious cancer among children. The principle is to identify the pattern of the disease and to seek the factors in the environment that are responsible for it, with the objective of relating the risk of developing the disease to environmental "triggers". The assumption is that if the risk is dependent on environmental factors, then it might show a degree of autocorrelation that can be analyzed by geostatistical methods. Such methods are modified versions of those used in the earth sciences: in fact, the values of the variables are not volumes of materials but the number of individual cases in a span of time, and the actual location of the population is approximated to some known reference coordinates (i.e. the electoral wards, see below). Both issues require special care in the application of geostatistical methods.

Cancer amongst children is a rare disease (about 120 per million) but serious, and a study of its spatial distribution can inform us about the ways to limit the problem. Previous analyses have found concentrations of child leukemia around nuclear installations, and investigations have been carried out to discover the causal links of this pattern. One solution is to verify if there are alternative causes, and this requires a careful mapping of the disease that is in the end to be compared to the distribution of environmental factors. Research has concentrated in the identification of clusters in the spatial pattern, with two main approaches: the first is based on cell counts and includes Poisson probability mapping, with results depending on cell size and independent of spatial correlation; the second is based on the distance between cases and on nearest neighbor analysis. Both approaches are considered weak by the authors when applied to this kind of data, supposedly because the former in particular does not consider geostatistical analysis of correlation as a way to extract more information from the data.

In the hypothetical case that there are no point sources of the disease, the tendency of the disease to cluster might not be supported by any evidence. One of the basic assumptions of the paper is that it is better to map the risk of developing a disease than the actual incidence of the disease. In other words, when we try to verify alternative causes of the disease other than point sources (e.g. nuclear installations), we need to consider that such point sources might not exist, thus invalidating any clustering hypothesis. Also we need to shift our attention to risk rather than to a study of disease occurrence pattern, because risk might show more interesting patterns than disease occurrence alone (and the two are related, see below).

The study was based on the extraction of the records of disease incidence among children for the period 1980-1984 in the West Midland Area of England, for a total of 595 cases. For each child diagnosed there is an address with postcode, from which a pair of coordinates can be obtained (the usual problem of imprecise rural areas post codes applies here). The resulting map of incidence shows a concentration of cases in urban areas, mainly because it is where children live. The objective here was to explore the data for spatial autocorrelation , and then estimate the risk, taking in account the varying density of population, and mapping the results to identify meaningful patterns.

The paper presents a brief review of geostatistical theory comprising the basic model of a random variable expressed in terms of mean and random variation. Its variance defines the semivariogram which is a function that relates the semivariance to a specific lag h. A function for covariance at a specific lag is also obtained using the value of the variable at (x) and (x+h), and its mean.

All children are exposed to the risk of developing cancer, and this risk is a regionalized random variable R(x) that can vary from place to place. The overall risk is the number of children who do develop cancer divided by the total number of children in the Region. The estimation of local risk is based on the local rate of incidence, frequency, which is, in this case, the number of cases in the electoral ward divided by the number of children in the ward. The data are indexed according to the centroids of the electoral ward. There is no better information available about the location of healthy children than the subdivision by electoral ward, so that the precise information about sick children cannot be used in the analysis. Preliminary data analysis shows that a Poisson distribution does not represent the counts of the occurrences of the disease, even if the disease is rare, mainly because the populations in the wards are too variable.

The calculation of the variogram of frequency was based on an isotropic type with a discrete lag interval of 8 km. From the variogram of frequencies we need to obtain the variogram of the underlying risk. Basically we want to extend the electoral ward-based geostatistical information to a measure of local risk across the region that is valid beyond the pre-defined centroids of the wards. We know how frequencies vary according to an electoral ward subdivision, but we need to know how the composite entity of risk is distributed. The risk would be obtained from frequency, which is a realization of a random variable that depends both on the underlying risk and on the number of children exposed to that risk. In other words we need an equation that related the variogram of frequency to the variogram of risk taking in account he fact that frequency is drawn from a

binomial distribution. In fact the number of cases is a binomial variable that depends on local risk and on the number of children in the ward. The procedure used is as follows:

1) The conditional expectation, variance and product of frequency are expressed in terms of local risk and number of children per ward.
2) From the relations expressed in the equations of (1) the conditional squared differences of the frequencies are obtained.
3) The expected value of the square difference is expressed as a function of the variogram of risk, mean risk, number of children in the wards centered at (x) and (x+h), and variance of underlying risk. An important point to consider at this stage is that while the variances are variable from ward to ward and therefore the previous variograms cannot be considered in the strict sense, on the other hand the averages of the variances can be used in the estimation of the empirical variogram.
4) The final step of this series of equations relates the estimated variogram of risk to the variogram of frequency: the risk variogram is equal to the frequency variogram minus a term that is a function of estimates of frequency mean, risk variance and an average of an arithmetic combination of population numbers of all pairs of wards.
5) In the estimation of the variogram of risk there is no variance of risk value available for substitution in the equation. An iterative process, by which the variance is first omitted, then estimated from the sill of the variogram, and finally re-input in the model, is repeated until the sill of the variogram and the variance converge to a single value.

A model variogram is fitted to the empirical variogram, using Whittle's elementary correlation function, which approaches the sill asymptotically. The sill is therefore defined at 50 km range, or 95% of sill as a limit. The variogram of the frequency is much larger than the one of the risk, and presents a considerable nugget variance. The interpretation of the nugget is that it represents the error in estimating the risk per ward by the observed frequency. From the forms of variograms of risk and frequency it is possible to infer that the risk of a child developing cancer has a coarse patchy distribution, with patches 50 km across and with parts of the Region with considerable higher values than others.

The next task is to use the Kriging method to predict the risk at all areas in the region and to map it. In fact while we have information about risk from the frequency per ward, but that way we don't include the information concerning correlated neighbors. A fine grid is set over the area and all intersections are

estimated through kriging and then contoured. The Kriging method here used is co-kriging, because both the correlations of frequency and risk variables are used in the equations. In particular we need to estimate the risk, of which we have no values, using known frequency values:

1) Risk is expressed in terms of weighted prediction based on frequency, and its variance is the expected value of the squared difference between weighted frequency and risk.
2) The co-kriging equations set to minimize the variance are based on the covariance of frequency and on the covariances between frequency and risk.
3) Since the cross variogram of the frequencies and the risk is equal to the variogram of the risk, also covariance of frequency is equal to the covariance of risk (except in the autocovariance case).
4) The maximum number of observations per estimate is set to 100 for accounting the minimal weight of the more distant values.
5) The estimates, originally for points, are to be considered as for areas equal to the electoral wards. The final map has a 2 km resolution and displays unbiased estimates with least variance.

The resulting map has a relatively low variance, lowest in cities where the data points are most dense, therefore the risk map is reliable. The risk map has a patchy distribution as indicated previously by the variogram, with large risk in rural areas and small risk in urban areas. There is therefore a consistent difference between the risk map and the map of cases, which respectively show high risk in western and southwestern parts of the region and a concentration in the center-east.

According to the risk variogram, risk of childhood cancer is strongly autocorrelated, indicating that wards with large risk occur near others with similarly large risk, in a patchy configuration with patches of 50 km (the range of the variogram). The regional pattern shows a concentration in rural areas away from urban centers, arising from the fact that children in rural areas are exposed to common diseases at an older age than their urban counterparts.

The research can be improved by extending the data sets for gaining greater confidence in the variograms, and by computing variograms over a variable support (in fact, electoral wards do not cover the same area), in order to obtain a final average variogram. The problem of georeferencing the healthy children can also be solved by using the 1991 census with archived addresses information.

This study is an example of how geostatistical analysis, point patterns and area analyses converge towards the same application. While the core of the study is geostatistical, based on co-kriging (even if it is a particular case, where the cross covariance is equalized to the autocovariance by analytical methods), the subdivision in wards is clearly a problem for lattice analysis, which in turn is solved as a point pattern problem with attributes, using the centroids as a summarizing spatial information for each area. The geostatistical core requires better data sets and a more sophisticated approach to analysis, as indicated above, but in essence prediction of risk based on covariance extends the set of solutions available for extracting information from data.

The paper does not offer great insight into the interpretation of the results, suggesting that the aim was only one of introducing the technique in a context that has never been exposed to that approach. A question might be what kind of implications the adoption of kriging as a prediction scheme of risk generates in identifying spatial patterns and environmental causes for the disease. In other words, can kriging predict a point sourced cluster of disease risk around a nuclear installation, or would it rather consider only the covariance over the random field without detecting localized features? While the quest for a clustered pattern might be using a weak approach to data analysis, the additional adoption of geostatistical methods might cause the assumption of the existence of clusters to fail in favor of a covariance distribution based approach, which cannot resolve any suspected features. This might have an impact on the political decisions about claiming for example a nuclear installation to be responsible of child leukemia.