Available online at www.sciencedirect.com

ScienceDirect

Procedia Environmental Sciences 3 (2011) 14–19

Procedia Environmental Sciences

Spatial Statistics 2011 – Mapping Global Change

Multivariate Areal Interpolation for Continuous and Count Data

Konstantin Krivoruchko[a], Alexander Gribov[a], Eric Krause[a]

[a]*Environmental Systems Research Institute, 380 New York St, Redlands, CA, USA, 92373*

Abstract

Geographic information system (GIS) users often need to disaggregate and reaggregate data collected in polygons, but classical kriging models only allow for data collected in points. We discuss our implementation of areal interpolation, a kriging-based disaggregation technique, in the Geostatistical Analyst extension of ArcGIS 10.1 for Gaussian, binomial, and overdispersed Poisson data. All methods allow for surfaces of prediction standard errors. We also allow for the use of a secondary cokriging variable, which can be any of the three above-mentioned distributions. Our areal interpolation model overcomes several computational problems, such as how to handle polygons of vastly different sizes and how to analyze polygons that are overlapping or disjoint.

For Gaussian data averaged over polygons, the output is a surface predicting the value at each individual location. Gaussian polygonal data may arise when continuous point measurements are averaged to polygons in order to protect privacy or reduce overhead, and the original point data is discarded. For polygons containing Poisson counts, the output is a surface predicting the density of counts at each location in the data domain. Our model allows for overdispersed counts and for different observation times between polygons. The output for binomial data is a surface predicting the underlying risk at each location of seeing an individual with a certain trait. Each polygon of the input data must contain a count and a population value. The latter indicates the number of individuals sampled, and the former indicates the number of sampled individuals with a certain trait.

Once a prediction surface has been created, predictions can be aggregated back to a new set of polygons. This allows for the collection of data over one set of polygons and the prediction for a different set of polygons. We discuss diagnostic options for determining how well the data fits a model, and we demonstrate areal interpolation with three case studies.

Keywords: areal interpolation, counts, kriging, Poisson, binomial, negative binomial, overdispersion

doi:10.1016/j.proenv.2010.02.004

1. Gaussian areal kriging model

From the very beginning, geostatistical theory took special attention to data averaging because average values play an important role in both meteorological and geological applications, Gandin and Kagan,1962 [1], Matheron, 1968 [2]. In geostatistical literature, the average value of the variable $\boldsymbol{Z}(\boldsymbol{s})$ in the area $\boldsymbol{A}$ is called the support of $\boldsymbol{Z}(\boldsymbol{A})$ and the averaging statistical model is called block kriging. Changing the support of a variable creates a new variable with different statistical properties. Excellent review on the change of support problem can be found in Gotway and Young, 2002 [3]. These same authors discussed in detail areal Gaussian kriging in Gotway and Young, 2004 [4].

In practice, each measurement is not made in the mathematical point $\boldsymbol{s}$, but in some, usually small, volume v and, therefore, the measurement value $Z(v)$ assigned to the point s is the average of $\boldsymbol{Z}(\boldsymbol{s})$ for all s in volume v.

Spatial correlation between data observed in polygons A_i and A_j and in polygon A_i and point s is estimated using the following covariances:

$$cov\left(Z(A_i), Z\left(A_j\right)\right) = \frac{1}{|A_i||A_j|}\iint_{A_iA_j} cov(Z(s'), Z(s))dsds' \quad (1)$$

$$cov(Z(A_i), \mathrm{Z}(s)) = \frac{1}{|A_i|}\int_{A_i} cov(Z(s'), Z(s))ds' \quad (2)$$

where $|A_i|$ is the area of polygon A_i and $cov(Z(s'), Z(s))$ is the covariance between measurements made in points s and s'. With these covariances, predictions can be made to both polygons and point locations, as described in Cressie, 1993 [5].

A Gaussian areal kriging application which is often overlooked is interpolation of environmental data collected in populated places when actual measurement locations are not provided. For example, the Belarusian catalog of radiocesium Cs-137 soil contamination provides the names of the cities and villages, and the researchers use the centroids of populated places to locate the measurements on the map. However, the size of the populated places can be large enough to use block instead of point kriging. It can be shown that the prediction error is smaller for areal kriging compared to representing the polygons as centroids. This indicates that areal kriging should be preferred when the measurements are collected in relatively large polygons.

Figure 1 on the left shows a subset of the populated places with measured Cs-137 soil contamination values made in Belarus in 1992 [6] and the Gaussian areal kriging prediction map. The semivariogram modeling is presented in the right part of figure 1. The horizontal axis shows the average distance between the polygons (calculated using cells of the overlapping grid). The crosses are the empirical semivariogram values calculated using the available averaged data in the polygons. The line is the estimated point semivariogram model. The bars are the confidence intervals (in figure 1, 90% confidence intervals) calculated assuming that the re-estimated empirical semivariances are normally distributed and uncorrelated.

The estimated point semivariogram in figure 1b is clearly different than the estimated empirical semivariogram values for the polygons. In this case, areal kriging can produce more accurate predictions and prediction standard errors than point kriging with values assigned to the polygons' centroids.

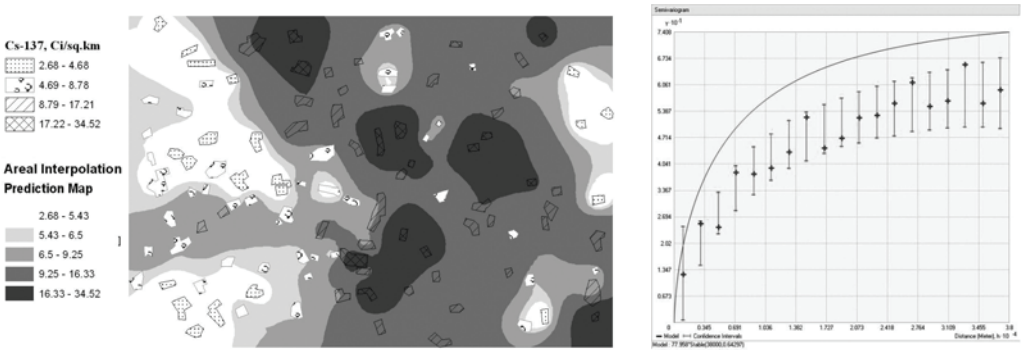

Fig. 1. a) Populated places with measured Cs137 soil contamination values (in Ci/sq.km) and Gaussian areal kriging prediction map. b) Deconvoluted point semivariogram (line) and re-estimated empirical semivariogram values for polygons (crosses) and their 90% confidence intervals (vertical lines).

2. Overdispersed Poisson areal kriging model

Monestiez et al, 2006 [7], developed a modification of classical kriging for modeling events observed in polygons of equal size. This model requires two inputs: the number of counts in an area immediately around location $\boldsymbol{s}$, $\boldsymbol{Z_s}$ and the time spent counting $\boldsymbol{t_s}$ (clearly the longer we observe, the higher the count). The output is a smooth surface that predicts the underlying *density* $\boldsymbol{Y_s}$ of whatever is being counted. It was assumed that $\boldsymbol{Z_s}$ is Poisson distributed so that the conditional expected value and the conditional variance are equal: $\boldsymbol{E[Z_s|Y_s] = Var[Z_s|Y_s] = t_s Y_s}$. This distributional assumption requires modification of the covariance and semivariogram formulas as well as the kriging system. In particular, Monestiez et al [7] estimated the semivariogram of the count density using the following expression:

$$\widehat{\gamma_{\mathrm{Y}}}(h) = \frac{1}{2\sum_\alpha \sum_\beta \frac{t_\alpha t_\beta}{t_\alpha + t_\beta}} \sum_\alpha \sum_\beta \left[\left(\frac{t_\alpha t_\beta}{t_\alpha + t_\beta} \right) \left(\frac{Z_\alpha}{t_\alpha} - \frac{Z_\beta}{t_\beta} \right)^2 - \hat{\mathrm{m}}_{\mathrm{Y}} \right], \tag{3}$$

where $Z_\alpha, \alpha = 1, \ldots, N$ are the measurements of Z_α obtained during time t_α and $\hat{\mathrm{m}}_{\mathrm{Y}}$ is the expected value of Y_s.

Note that the random field Y_s is non-stationary even when the times of observation are the same because the Poisson distribution assumes that the data are not absolutely precise (the mean is equal to the variance), and the data uncertainty varies in the neighboring polygons. For comparison, Gaussian kriging assumes that the measurement error is zero or a constant value. Since the Poisson distribution allows for measurement error, the mass balance property is not satisfied for the Poisson areal kriging model. This is also true for the overdispersed Poisson and the binomial types of areal kriging discussed below.

In practice, the count data variance is usually greater than the mean, and in this paper we use a more general distributional assumption about the relationship between the conditional mean and the conditional variance:

$$E[Z_s|Y_s] = t_s Y_s \text{ and } Var[Z_s|Y_s] = k t_s Y_s + l(t_s Y_s)^2, \tag{4}$$

where $k \geq 0$ and $l \geq 0$ are estimable parameters. Y_s is proportional to the population density (number of counts per square unit) at location s, and k and l are constant for all locations. It is assumed that Y_s is a

positive random field honoring order two stationarity with mean $\hat{m}_Y$ and variance σ_Y^2 for all locations. We assume the conditional independence of $Z_s|Y_s$ for all locations.

Letting $k = 1$ and $l = 0$ gives the first two moments of the Poisson distribution with mean $t_s Y_s$. Similarly, letting $k = 1$ and $l = \frac{1}{r} > 0$ gives the first two moments of the negative binomial distribution with mean $t_s Y_s$ and dispersion parameter r. Although other values of k and l allow for more general types of dispersion, their simultaneous estimation is difficult, if not impossible. Literature suggests that the negative binomial distribution is common in the analysis of count data, and all discussions below are based on that distribution. Complete formulas for covariances and the kriging equations can be found in the full version of this paper.

In the next example we illustrate the usage of overdispersed Poisson areal cokriging. The variable of interest is the number of violent crimes in the 439 census tracts within the city of Houston. The spatial variations of crimes in associations with alcohol distribution and drug-law violations collected at the same administrative units were analyzed in Waller et al, 2007 [8], using the geographically weighted regression and spatially varying regression coefficients models. Since our goal is interpolation rather than regression, we use another explanatory variable which more strongly correlates with the variable of interest – a median value of the estimated road density in the tracts. The violence data was collected during the same time period, and we use a constant value for the time variable in the overdispersed Poisson model given in equation (4). We assume that the road density data is normally distributed.

Figure 2 shows the predicted density of the violent crimes and the associated prediction standard errors. Note that the uncertainty is large in the polygons with large areas. In general, these polygons have relatively low road density.

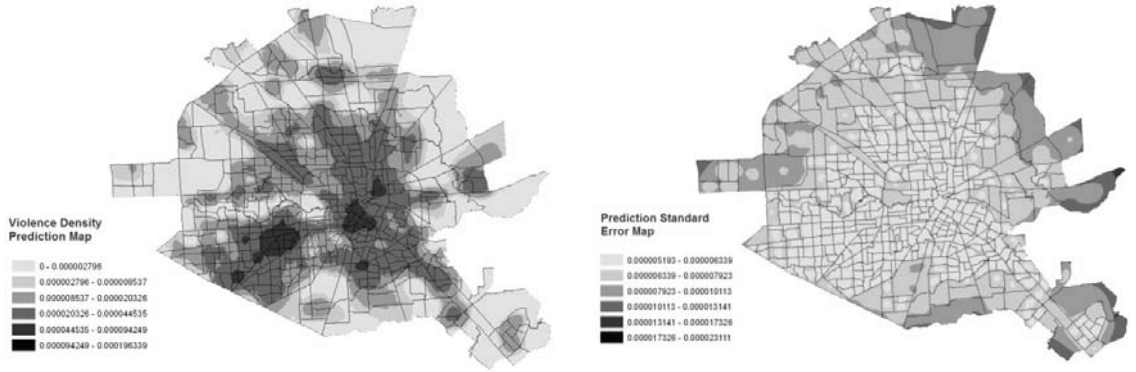

Fig. 2. The predicted density of the violent crimes (a) and the associated prediction standard errors (b).

3. Binomial areal kriging model

Another typical type of count data are samples from known populations, for example, the number of lung cancer cases among the female population of a specific age in a particular region. In this case, counts should be scaled by population size instead of the time of observation.

Formulas for binomial kriging models were derived by McNeill [9] and Lajaunie in 1991 [10]. McNeill's motivation for derivation of binomial kriging was that the measurement error in the observed rates (the total number of observed cases N_i within a fixed time interval in a geographical region divided by the total number of possible cases P_i) is large and varying considerably from one polygon to another. True variability of the rates is considered as a stationary random variable *R(s)* with constant mean and

changing variances, called regionalized risk factor or simply risk by Lajaunie. The average risk R_i in the polygon i is the average of $R(s)$ over polygon A_i:

$$R_i = \frac{1}{|A_i|}\int_{A_i} R(s)ds, \text{ if population in each polygon is homogeneous or} \quad (5)$$

$$R_i = \frac{1}{\int_{A_i} w(s)ds}\int_{A_i} w(s)R(s)ds, \text{ where } w(s) \text{ desribes the variability of } R(s), \text{ if it is not} \quad (6)$$

Weights $w(s)$ can reflect the known population density. In this case, the interpretation of the weights is the proportion of time spent at the location *s* by an individual under the risk. In this paper we assume that this time is unknown and, therefore, equation (5) is used.

The assumptions behind the binomial kriging are the following:

- the observed rates are the sum of true but unknown risk R_i and measurement error ε_i, $Z_i = \frac{N_i}{P_i} = R_i + \varepsilon_i$
- Z_i are independent binomial variables with distribution: $\frac{1}{P_i} Binomial(R_i, P_i)$
- *R(s)* is the only reason for correlation between the rates

We want to map the spatial distribution of unobserved relative risk *R(s)* together with the associated prediction standard errors.

According to McNeill, 1991 [9], and Lajaunie, 1991 [10], the semivariogram is estimated according to the following expression:

$$\gamma_{i,j}^{R} = \gamma_{i,j}^{Z} - \frac{1}{2}\left(\frac{1}{P_i} + \frac{1}{P_j}\right)\mu(1-\mu) + \frac{1}{2}\left(\frac{\sigma_i^2}{P_i} + \frac{\sigma_j^2}{P_j}\right), \quad (7)$$

where μ and σ_i^2 are respectively the mean rate and the rate variance values over the region A_i.

We illustrate the usage of binomial kriging using classical data on the number of lip cancer cases registered during 1975-1980 in 56 districts of Scotland, Kemp et al, 1985 [11]. The percentage of the work force in each district employed in agriculture, fishing, and forestry was used by several authors as an explanatory variable in the regression analysis, see for example, Waller and Gotway, 2004 [4], because the exposure to sunlight is correlated with lip cancer rates. Figure 3 shows cokriging predictions and prediction standard errors assuming that the secondary variable also has a binomial distribution.

Fig. 3. Cokriging predictions of the lip cancer rates (a, binomial data) and prediction standard errors (b, binomial data).

4. Conclusions

Areal cokriging is an inverse problem application and there are many ways for downscaling averaged and aggregated data. Therefore, it is important to have a choice of statistical methods for various cases, and we provide an interactive software environment with a choice of the data distribution, interactive spatial correlation analysis, prediction surface preview, and cross-validation diagnostic graphs and tables.

Areal kriging can be used in addition to or instead of traditional choropleth maps to better represent the data variability and for visualizing “hot spots” that are difficult to recognize when the raw data are displayed (these sorts of maps are sometimes called “heat maps” in GIS literature).

Although other researchers prefer ordinary kriging, we suggest using the simple kriging model because in the case of areal data, both kriging models require specification of the mean value and, therefore, the ordinary kriging constraint to the sum of weights becomes an additional and unnecessary property of the model. In addition, the simple kriging model can be used for simulating new surfaces conditionally to the observed aggregated or averaged data. These surfaces can be useful, for example, in modeling disease outbreaks by allowing the analysis of hypothetical situations.

The models described in this paper can be improved and extended. We suggest a Bayesian generalization of areal kriging because the uncertainty of semivariogram/covariance modeling is higher in areal kriging than in classical point kriging. We also suggest a generalization using a mixed linear model (universal kriging with external trend).

By overcoming the initial hurdles of areal cokriging implementation, we hope to open the door to the analysis of a new class of problems for GIS users. We have developed solutions to many of the problems associated with data disaggregation, but there is much potential to improve and refine the theory and the software implementation. With solutions to the problems listed above, areal cokriging has the potential to become an even more flexible and versatile geostatistical model.

References

[1] Gandin, L.S. and Kagan R.L. 1962. The accuracy of determining the mean depth of snow cover from discrete data. Trudy GGO, 130, 3-10 (In Russian).

[2] Matheron, G. (1968) Osnovy prikladnoi geostatistiki (Principles of Applied Geostatistics). Mir, Moscow, 408 pp. (In Russian)

[3] Gotway CA, Young LJ: Combining incompatible spatial data. Journal of the American Statistical Association 2002, 97:632-648.

[4] Gotway CA, Young LJ: A geostatistical approach to linking geographically-aggregated data from different sources. In Technical report # 2004-012 Department of Statistics, University of Florida; 2004.

[5] Cressie, N.A.C. (1993) Statistics for Spatial Data. Revised ed. John Wiley & Sons, New York.

[6] Krivoruchko K. (2011) Spatial Statistical Analysis for GIS Users. Esri Press, Redlands, California

[7] Monestiez P, Dubroca L, Bonnin E, Durbec JP, Guinet C: Geostatistical modeling of spatial distribution of Balenoptera physalus in the northwestern Mediterranean Sea from sparse count data and heterogeneous observation efforts. Ecological Modelling 2006, 193:615-628.

[8] Waller, L.A., Zhu, L., Gotway, C.A., Gorman, D.M., Gruenewald, P.J. (2007) Quantifying geographic variations in associations between alcohol distribution and violence: A comparison of geographically weighted regression and spatially varying coefficient models. Stochastic Environmental Research and Risk Assessment 21 (5), pp. 573-588.

[9] McNeill L. (1991) Interpolation and smoothing of binomial data for the southern african bird atlas project. South African Statist. J. vol.25, pp. 129-136.

[10] Lajaunie C. (1991) Local risk estimation for a rare non contagious disease based on observed frequencies. Centre de Geostatistique de l’Ecole des Mines de Paris, Fontainebleau, Note N-36/91/G.

[11] Kemp, I., P. Boyle, M. Smans, and C. Muir (1985). *Atlas of Cancer in Scotland, 1975–1980: Incidence and Epidemiologic Perspective*. IARC Scientific Publication 72. Lyon, France: International Agency for Research on Cancer.