CO-282

POPULAR CARTOGRAPHIC AREAL INTERPOLATION METHODS VIEWED FROM A GEOSTATISTICAL PERSPECTIVE

KYRIAKIDIS P.
University of California Santa Barbara, MYTILENE, GREECE

ABSTRACT

Cartographic areal interpolation methods are viewed in this work within a general geostatistical framework for spatial prediction that accounts explicitly for spatial resolution differences between source and target data and for spatial auto-correlation in attribute values. More precisely, both source data and target values are assumed to pertain to the same spatial attribute, and are defined as aggregates of an underlying (unobserved or latent) attribute surface parameterized by a known (or estimated) mean component and a point-level spatial correlation model. It is demonstrated that popular cartographic areal interpolation methods, e.g., proportional areal weighting, can be viewed as particular cases of the geostatistical framework under very restrictive assumptions on the point correlation model. Last, the geostatistical framework provides measures of reliability (error variances) for the target predictions, and can thus be used to provide uncertainty statements regarding interpolated values obtained from traditional cartographic areal interpolation methods, typically cast in a deterministic setting.

BACKGROUND AND OBJECTIVES

In many geographical applications, attribute values from one existing spatial partition must be transformed to another (Haining, 2003); see, for example, Figure 1. In public health applications, e.g., Waller and Gotway (2004), it is often required to integrate socio-economic, environmental and health data, originally available over census tracts, prediction grid nodes of an air pollution model, and administrative reporting zones, respectively. Areal interpolation (Goodchild and Lam, 1980), has been developed for this spatial data transformation, and amounts to predicting the unknown (target) attribute values at the required partition (target zones) from a set of known (source) attribute data available on a different partition (source zones).

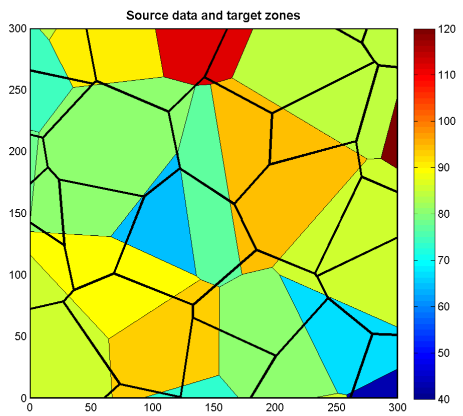

Figure 1: A schematic representation of areal interpolation. Source data (filled polygons) are used to predict target values at target zones (superimposed empty polygons); the color legend to the right corresponds to the level of the source data.

Several areal interpolation methods can be found in the literature, and can be distinguished in cartographic and statistical approaches (Haining, 2003). In the first family (Goodchild and Lam, 1980; Lam, 1983; Fisher and Langford, 1995; Mrozinsky and Cromley, 1999), spatial prediction is typically based on geometrical characteristics of the source and target zones, such as their areas of overlap. In the second family, e.g., Flowerdew and Green (1994), spatial prediction is based on ancillary data typically linked to source data via a regression model. In cartographic areal interpolation methods, the reliability of the resulting target predictions, i.e., the prediction error variance, is not reported since such methods are cast in a deterministic setting. On the contrary, statistical areal interpolation methods provide a measure of reliability regarding the target predictions, since these methods are formulated in a probabilistic framework.

The objective of this paper is to present key concepts of geostatistical areal interpolation and to highlight its close connection with popular cartographic methods for areal interpolation, and in particular, proportional area weighting.

APPROACH AND METHODS

Areal interpolation is formulated in this work within a general geostatistical framework of change of support (Journel and Hujbregts, 1978; Cressie, 1993; Goovaerts, 1997), with the term support being equivalent to that of a zone in two dimensions. This framework accounts explicitly for resolution differences between source and target data and for spatial auto-correlation in attribute values. More precisely, both source and target data (here pertaining to the same spatial attribute) are assumed to be functionally linked via an aggregation mechanism to an underlying (unobserved or latent) point attribute surface or field. In practice, this surface is approximated by a set of discretization points, whose locations are specified by the user. In Figure 2, for example, a regular grid of (100x100) nodes is used to discretize the source and target zones of Figure 1. Random discretization points are also an option, since a regular grid must often be constructed too fine to adequately cover a study region partitioned into zones with complex geometry. It should be stressed here that the actual attribute values of this surface are unknown (hence the term latent or underlying surface); it is only the locations of these values (discretization points) that are required by geostatistical areal interpolation.

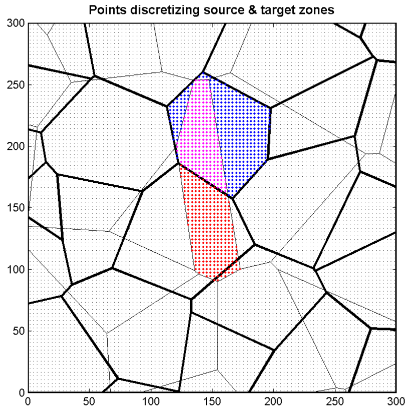

Figure 2: Regular grid of points (black dots) discretizing a set of source zones (thin lines in the background) and target zones (thicker lines in the foreground). Dots colored red belong to a source zone, blue to a target zone, and magenta to the zone of intersection.

The latent attribute surface is parameterized by a known or estimated mean (here assumed constant or stationary throughout the study region) and a spatial correlation model; the latter dictates the smoothness of the point surface. Since the only available data are the source data of Figure 1, the above statistical characteristics of the point field must be estimated from these data. Figure 3 furnishes a graphical representation of three parametric distance-decay models of spatial correlation; namely, a spherical, a Gaussian and an exponential correlation function. The parameters, e.g., range (distance at which spatial correlation reaches zero), of these models must be estimated from the available source data; this is an under-determined inverse problem, which must be solved iteratively by trial-and-error procedures and prior knowledge regarding the spatial distribution of the latent point surface. Kyriakidis (2004) and Nagle, Kyriakidis and Sweeney (2011) provide further details regarding this challenging parameter estimation problem.

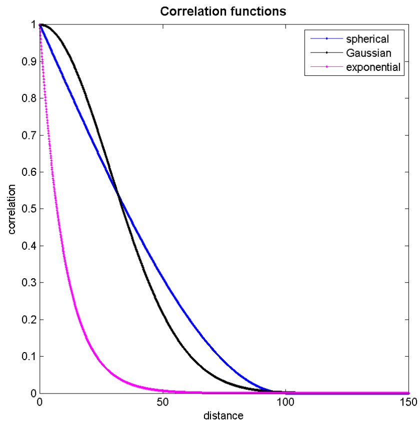

Figure 3: Distance-decay correlation functions: spherical and Gaussian model with range approximately 100 distance units and exponential model with range approximately 50 distance units.

To illustrate the spatial patterns implied by a spatial correlation function, Figure 4 presents two latent surfaces whose smoothness is dictated by the spherical correlation model of Figure 3, without taking into account the actual levels and locations of the source data (unconditional case). In other words, the surfaces of Figure 4 do not reproduce, when their point values are aggregated within their respective source zones, the corresponding source data; this translates to patterns of high and low attribute values appearing in differ positions from one surface to the other. Figure 5 presents another set of two latent surfaces whose smoothness is dictated by the Gaussian correlation function of Figure 3. One can easily appreciate the much smoother spatial characteristics of these latter set of surfaces (Figure 5) as opposed to the former (Figure 4). All surfaces shown hereafter constitute discrete approximations of continuous surfaces on a (300x300) regular grid.

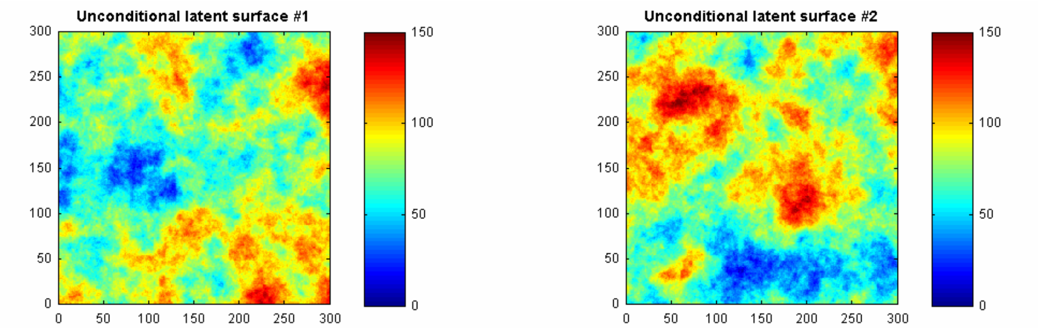

Figure 4: Two unconditional latent surfaces whose smoothness is dictated by the spherical correlation model.

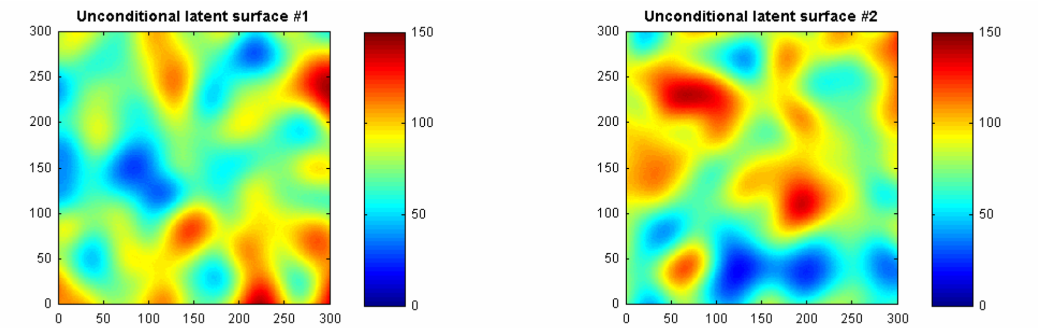

Figure 5: Two unconditional latent surfaces whose smoothness is dictated by the Gaussian correlation model.

When the magnitude and location of the source data is accounted for (conditional case), these latent surfaces are transformed to the surfaces shown in Figures 6 and 7. The point attribute values comprising these surfaces reproduce exactly the corresponding source data, when the former aggregated within their respective source zones. In other words, these latent surfaces have the smoothness characteristics imposed by their respective correlation models, within the degrees of freedom dictated by the measured area-level source data. Note that spatial patterns of high and low attribute values tend to appear at nearby locations within the study region from one realization to the other, since their occurrence is controlled by the area-level source data.

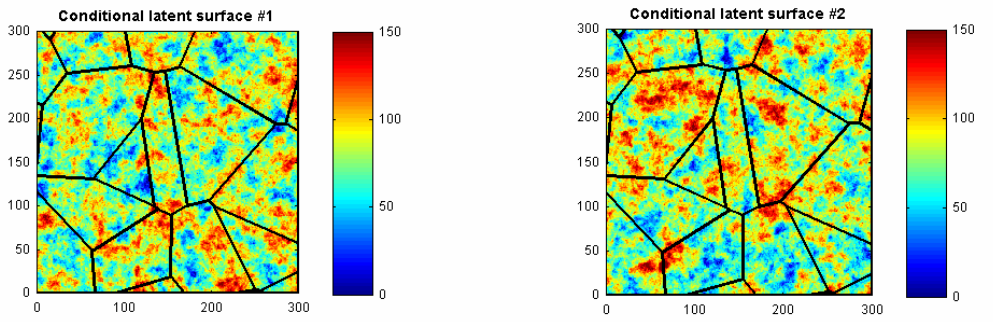

Figure 6: Two latent surfaces with smoothness imposed by the spherical correlation model, which reproduce when aggregated within source zones the corresponding area-level source data.

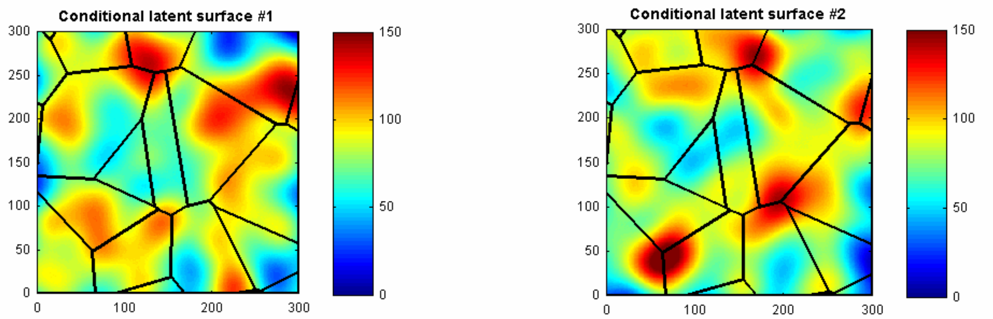

Figure 7: Two latent surfaces with smoothness imposed by the Gaussian correlation model, which reproduce when aggregated within source zones the corresponding area-level source data.

Figure 8 provides two ultimately rough latent attribute surfaces, whose smoothness is dictated by the pure nugget effect correlation model, also termed white noise model. Such a function (not shown) could be plotted in Figure 3 as a set of points with zero ordinate, and one point at unit ordinate. Spatial patterns of high and low values in the surfaces of Figure 8 are completely random (salt and pepper texture), yet their occurrence is still constrained by the area-level source data. In other words, the point attribute values comprising these surfaces also reproduce, when aggregated within their respective source zones, the corresponding area-level source data. Such surfaces, however, correspond to the extremely particular (yet unrealistic) case of lack of point-level spatial correlation, which surprisingly underlies the choropleth map (Kyriakidis, 2004) and the proportional area weighting method of areal interpolation (Kyriakidis and Goodchild, 2011).

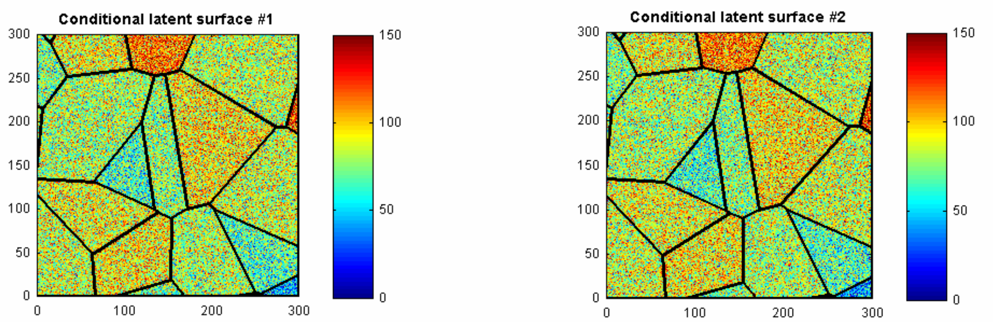

Figure 8: Two completely random latent surfaces, which reproduce when aggregated within source zones the corresponding area-level source data.

To further elucidate this issue, Figure 9 provides the conditionally expected attribute surface, given the source data, computed by averaging an ensemble of possible latent surfaces with the smoothness of the spherical (left) and nugget (right) correlation models. Both these surfaces reproduce, when aggregated within their respective source zones, the corresponding area-level source data, yet provide two significantly different representations of the expected spatial patterns of the point-level attribute values. The conditionally expected surfaces of Figure 9 should be differentiated from the expected surfaces (not shown) of the unconditional case corresponding to Figures 4 and 5. In the latter case, the expected latent surfaces (not shown) are flat and equal to the mean of the source data, since the source data magnitude and location is not accounted for. Note that conditionally expected surface for the random or nugget correlation model corresponds to the choropleth map (Figure 9, right).

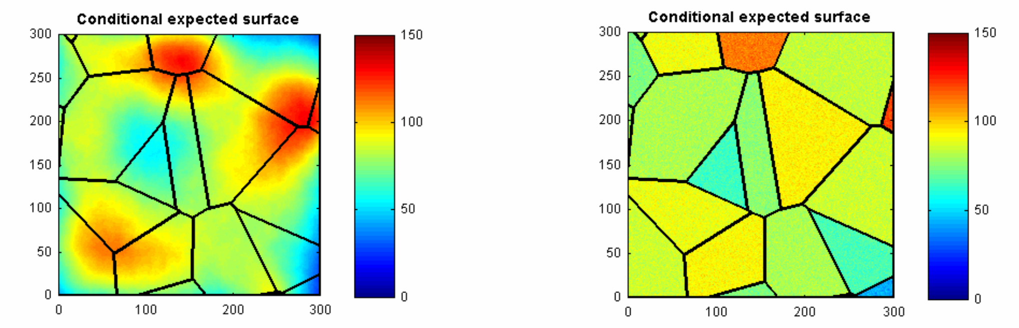

Figure 9: Conditionally expected attribute surfaces, given the area-level source data, for the spherical (left) and nugget or random (right) correlation models.

At this point, one could opt to predict the unknown target attribute values at the target zones of Figure 1. To do so, one simply needs to aggregate the point level attribute values of the expected surfaces of Figure 9 within their respective target zones. The resulting target predictions would be different for each case, since the expected latent surfaces are different, or equivalently the correlation models used to construct these expected surfaces are different. One can demonstrate that the target predictions resulting from such a procedure are identical to those obtained by proportional area weighting (Kyriakidis and Goodchild, 2011). To compute, however, the uncertainty in those target predictions, no matter the postulated point-level correlation model, one would need to: (a) aggregate the attribute values comprising each conditional latent surface, such as those shown in Figures 6 to 8, within their respective target zones, and (b) compute the variance of the resulting target predictions. Such a measure of uncertainty regarding the target predictions is an all-important piece of information, typically ignored by traditional areal interpolation methods.

An alternative route can be pursued to directly predict the unknown target attribute values, instead of choosing to construct a conditionally expected latent point surface and subsequently aggregate attribute values within target zones. Geostatistical areal interpolation provides this alternative route, which calls for the correlation between any two source zones, as well as the correlation between any two source and target zones. These quantities are computed from the postulated correlation function (such as those depicted in Figure 3) for the point surface by aggregating correlation values between points discretizing any two zones (Figure 2). More precisely, the correlation between a source and target zone is evaluated by: (a) computing distances between all possible pairs consisting of points within the source zone (red and magenta colored points in Figure 2) and points within the target zone (blue and magenta colored points in Figure 2), (b) transforming these distances into correlation values via the distance-decay, point-level, correlation model (one of those depicted in Figure 3), and (c) computing the average of these correlation values. Such average zone-to-zone correlation values are computed for all pairs of source-to-source zones, as well as for all pairs of source-to-target zones. Subsequently, a Kriging system is solved to compute the weights that should be assigned to the source data for predicting any unknown target value (Chilès and Delfiner, 1999). The benefit of this procedure is that the error variance for the target predictions can be readily computed without resorting to ensemble averaging as in the procedure outlined in the previous paragraph.

Figure 10 provides the target predictions derived via geostatistical areal interpolation using the spherical correlation model of Figure 3, along with the original source data for comparison. Figure 11 provides the corresponding target predictions using the Gaussian correlation model of Figure 3 and the random or nugget model (not shown). Note again, that the case of geostatistical areal interpolation with a nugget point-level correlation model corresponds to areal interpolation with proportional area weighting. In this case, the resulting Kriging weights are only a function of the area of overlap between the source and target zones. From Figures 10 and 11, one can detect that, although target predictions have similar magnitude across all correlation models, the case of the nugget model yields smoother (with smaller variance) predictions. In other words, proportional area weighting yields less extreme predictions than the other variants of geostatistical areal interpolation in this example. This finding may have important implications in practice, given the critical nature of extreme values in a plethora of spatial analysis operations and physical process simulations.

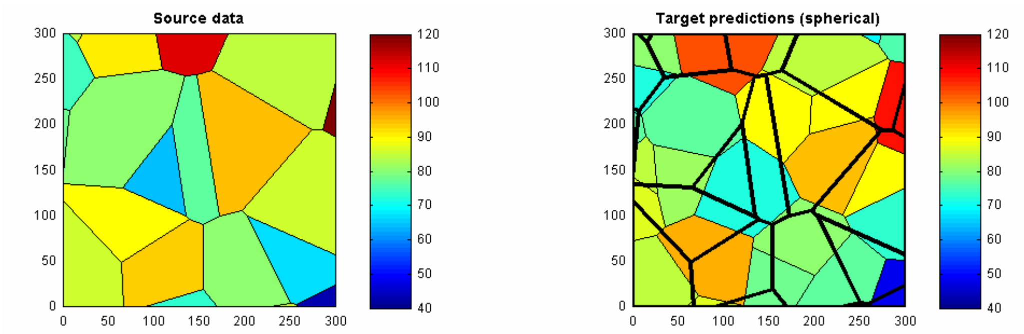

Figure 10: Source data (left) and target predictions (right) obtained by geostatistical areal interpolation using a spherical correlation model. Source zone boundaries (depicted with thick lines) are superimposed on target zones (depicted with thin lines) on the right for comparison.

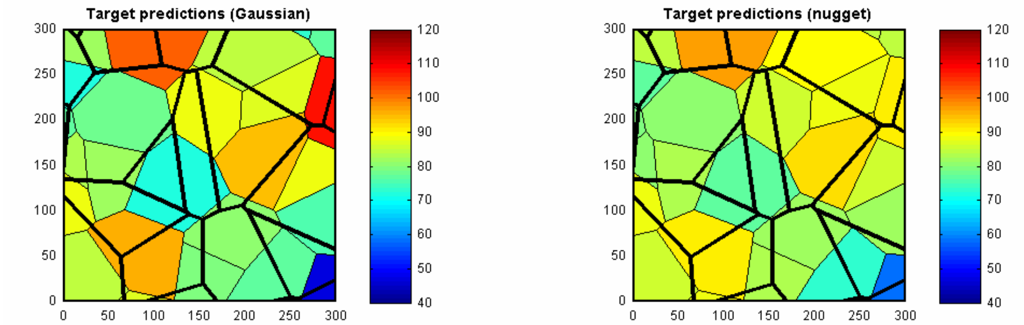

Figure 11: Target predictions obtained by geostatistical areal interpolation using a Gaussian (left) and a random or nugget (right) correlation model. Source zone boundaries (depicted with thick lines) are superimposed on target zones (depicted with thin lines) for comparison.

Last, the uncertainty regarding target predictions is quantified by the Kriging-derived prediction error variance, i.e., the variance of the error committed when predicting the unknown target values (Goovaerts, 1997). This error variance is a function of the point-level spatial correlation model, as well as the spatial configuration of the source and target zones, not of the actual magnitude of the source data. Generally, and for a given point-level spatial correlation model, error variances are expected to be lower the more a target zone is contained within or surrounded by source zones. In addition, the redundancy between neighboring source zones also contributes to the reduction of the error variance. Figures 12 and 13 provide prediction error standard deviations at the target zones for four different point-level spatial correlation models. It can be seen that, across all correlation models, prediction error standard deviations are higher for smaller target zones located near the boundary of the study region and at the same time surrounded by fewer and larger source zones. Point-level correlation models with small range (Figure 13) lead to smaller overall error variances, as compared to correlation models with larger range (Figure 12). At the limit, a pure nugget effect correlation model leads to very small error variances (Figure 13, right), since the only control on the Kriging weights (hence on the Kriging variance) is the area of overlap between source and target zones. In this case, prediction error standard deviations are lowest for target zones engulfed in small source zones, such as those in the central part of the study region.

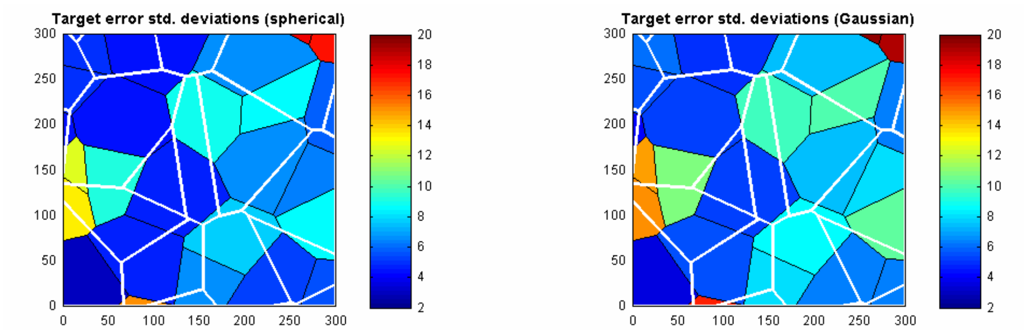

Figure 12: Prediction error standard deviations at target zones derived from geostatistical areal interpolation using a spherical (left) and Gaussian (right) point-level correlation model; source zone boundaries are superimposed with white lines.

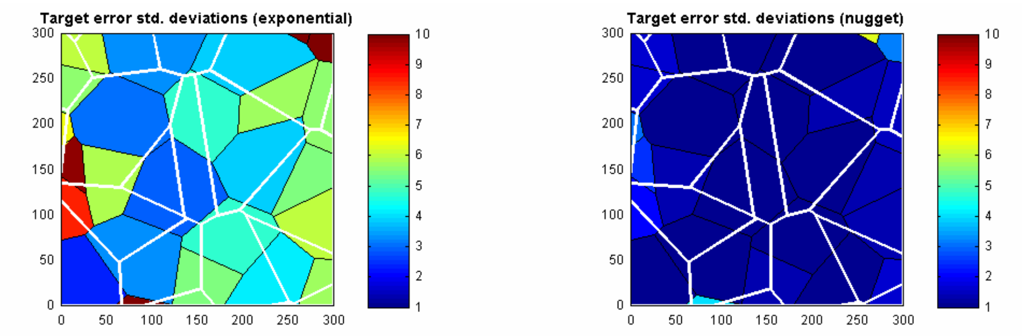

Figure 13: Prediction error standard deviations at target zones derived from geostatistical areal interpolation using an exponential (left) and a pure nugget (right) point-level correlation model; source zone boundaries are superimposed with white lines.

RESULTS

A real-world application of geostatistical areal interpolation is presented in this section, involving population density data from Northern California in the U.S. More precisely, the source data consist of population densities at the county level (shown with thin red lines in Figure 14), and the objective is to predict target population densities at 3-digit zip-code regions (shown with thick black lines in Figure 14). The study region is discretized by a regular grid of (200x200) points, and geostatistical areal interpolation is performed to obtain the target predictions and associated error variances. In the first experiment, a pure nugget effect correlation model is postulated for the latent attribute surface, entailing complete absence of spatial correlation at the point level. The resulting weights, and consequently the target predictions, are identical to those obtained via proportional area weighting (see Figure 14, left). In the second experiment, a (isotropic) point spherical correlation model with a range of 50km is postulated for the latent surface, based on assumed knowledge of the spatial distribution of population density in the region. The resulting target predictions differ considerably (exhibit more variability) from the proportional area weighting case, thus highlighting the flexibility of the geostatistical areal interpolation framework.

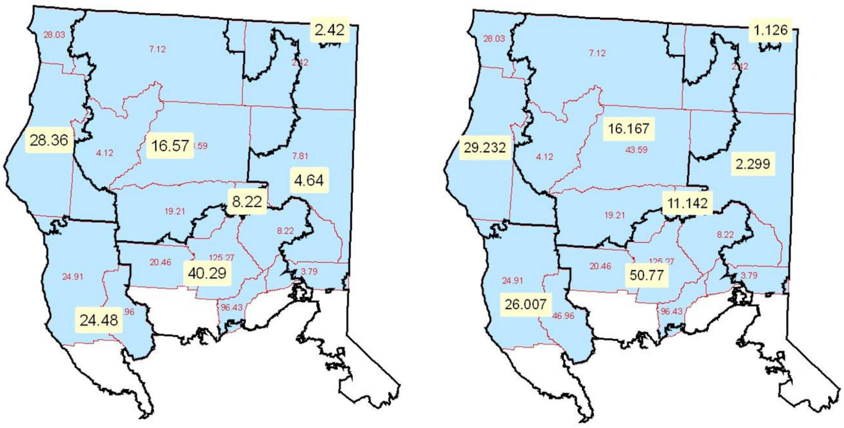

Figure 14: Geostatistical areal interpolation of population densities in Northern California. Thin red lines delineate county polygons (source zones), whereas thick black lines delineate 3-digit zip-code regions (target zones). Source data are shown with red numbers, whereas target predictions are shown in text boxes.

CONCLUSION AND FUTURE PLANS

This paper outlined the basic concepts of geostatistical areal interpolation and its connections to popular cartographic methods; namely, proportional area weighting. Of particular interest is the fact that such methods can be regarded as special cases of the geostatistical areal interpolation framework, albeit under very restrictive assumptions on the point correlation model or equivalently the smoothness of the latent attribute surface. In addition, the geostatistical framework provides measures of reliability (error variances) for the target predictions, and can thus be used to furnish such confidence statements for cartographic areal interpolation methods traditionally cast in a deterministic setting.

Future theoretical research should prove that geostatistics-derived target predictions are mass preserving, i.e., satisfy the pycnophylactic constraint (Mrozinsky and Cromley, 1999), no matter the correlation model adopted for the underlying point attribute field; a first theoretical attempt can be found in Kyriakidis and Goodchild (2011). In that work, it is also demonstrated that geostatistical areal interpolation yields as a particular case (again adopting the random point-level correlation model) the dasymetric solution of Wright (1936). In terms of computation, the processing time required for geostatistical areal interpolation, in particular the evaluation of average correlation values between pairs of irregular zones, should be reduced; parallel processing appears to be a promising avenue to reduce this computational cost (Guan, Kyriakidis, and Goodchild, 2011). Last, in terms of dissemination, easy-to-use GIS-based computer programs are required for rendering the developed analytical framework accessible to a wider pool of geospatial analysts; to this respect, the Python scripting language offers an effective means for porting existing computer code to a GIS environment.

REFERENCES

Chilès, J.P., and Delfiner, P. (1999): *Geostatistics: Modeling Spatial Uncertainty*, New York, NY: John Wiley & Sons.

Cressie, N.A. (1993): *Statistics for Spatial Data* (Revised Edition), New York, NY: John Wiley & Sons.

Fisher, P., and Langford, M. (1995): Modeling the errors in areal interpolation between zonal systems by Monte Carlo simulation, *Environment and Planning A*, 27, 211-224.

Flowerdew, R., and Green, M. (1994): Areal interpolation and types of data, in: S. Fotheringham, and P. Rogerson (Eds.), *Spatial Analysis and GIS*, 121-145, London: Taylor & Francis.

Goodchild, M.F., and Lam, N.S. (1980): Areal interpolation: A variant of the traditional spatial problem, *Geo-Processing*, 1, 297-312.

Goovaerts, P. (1997): *Geostatistics for Natural Resources Evaluation*, New York, NY: Oxford University Press.

Guan, Q., Kyriakidis, P.C., and Goodchild, M.F. (2011): A parallel computing approach to fast geostatistical areal interpolation, *International Journal of Geographical Information Science* (in press).

Haining, R. (2003): *Spatial Data Analysis: Theory and Practice*, Cambridge, UK: Cambridge University Press.

Journel, A.G., and Huijbregts, Ch.J. (1978): *Mining Geostatistics*, New York, NY: Academic Press.

Kyriakidis, P.C. (2004): A geostatistical framework for area-to-point spatial interpolation, *Geographical Analysis*, 36, 259-289.

Kyriakidis, P.C., and Goodchild, M.F. (2011): A geostatistical perspective on cartographic methods of areal interpolation, *International Journal of Geographical Information Science* (in press).

Lam, N.S. (1983): Spatial interpolation methods: A review, *The American Cartographer*, 10, 129-149.

Mrozinski, R.D., and Cromley, R.G. (1999): Singly- and doubly-constrained methods of areal interpolation for vector-based GIS, *Transactions in GIS*, 3, 285-301.

Nagle, N., Kyriakidis, P.C., and Sweeney, S. (2011): A geostatistical linear regression model for small-area data, *Geographical Analysis*, 43, 38-60.

Waller, L.A., and Gotway, C.A. (2004): *Applied Spatial Statistics for Public Health Data*, Hoboken, NJ: John Wiley & Sons.

Wright, J.K. (1936): A method of mapping densities of population with Cape Cod as an example, *Geographical Review*, 26, 103-110.