Geostatistical Alternatives for Incorporating Covariates in Areal Interpolation

P. C. Kyriakidis[1], M. F. Goodchild[2]

[1]Department of Geography, University of California Santa Barbara.,
Ellison Hall 5710, University of California, Santa Barbara, CA 93106, U.S.A.
Telephone: +001 (805) 893-2266
Fax: +001 (805) 893-3146
Email: phaedon@geog.ucsb.edu

[2]Department of Geography, University of California Santa Barbara,
Ellison Hall 5707, Santa Barbara, CA 93106, U.S.A.
Telephone: +001 (805) 893-8049
Fax: +001 (805) 893-3146
Email: good@geog.ucsb.edu

1. Introduction

A major problem in spatial data interoperability is the integration of different measurements obtained over different spatial units (zones) and pertaining to different spatial attributes. Areal interpolation methods, such as proportional areal weighting and dasymetric mapping, have been long ago introduced for coping with the problem of different measurement units, and are collectively known as cartographic areal interpolation methods (Haining 2003). Such methods are typically based on geometrical characteristics of source and target zones, and in particular the area of zones defined by their intersection, and are applicable in a single attribute setting, i.e., for the spatial prediction of unknown attribute values at a set of target zones from known measurements of the same attribute obtained at a set of source zones.

Areal interpolation methods accounting for auxiliary variables or covariates fall in the realm of statistical methods, as they typically involve a regression model linking data of the attribute of interest to data of covariates available at the source zones (Haining 2003). Although traditionally such statistical areal interpolation methods have been applied in a global setting; that is, using all available data to infer the parameters of a regression model, interest has been shifted more recently to making these regression parameters spatially variable within the study region, under what is termed Geographically Weighted Regression (GWR); see, for example, Fotheringham et al. (2000).

The common application of both cartographic and statistical areal interpolation methods, however, fails to account for the differences in measurement units pertaining to source data and target values. In other words, all data are assumed to inform zones of the same shape and size, and in particular data are often associated with so called representative points, such as polygon centroids, within their respective zones. In addition, no unifying framework exists to date that can accommodate both statistical and cartographic areal interpolation methods.

In this paper, we propose a comprehensive framework for areal interpolation accounting for auxiliary variables based on geostatistics. More precisely, we extend our prior work (Kyriakidis and Goodchild 2007) on geostatistical areal interpolation involving data of a single attribute to account for covariate information. In doing so, we

provide an analysis of scale effects on linear regression coefficients with spatially correlated errors, and illustrate the assumptions and limitations of such an approach to data integration. We then offer an extension of coKriging to account for data of different spatial units, and prove that regression-based approaches can be seen as particular cases of this extended coKriging approach under specific assumptions and conditions. Last, we illustrate the implementation of a Matlab-based toolbox for geostatistical areal interpolation, which can handle both cartographic and statistical approached in a unified framework.

2. Approach

Our framework is appropriate when areal data are defined as integrated measurements of point attribute values within regular or irregular shaped spatial units, such as pixels or polygons. More precisely, we conceptualize underlying (latent and unobserved) point fields of different attributes, whose values are then aggregated within arbitrary units to yield observed source data and unknown target values. Note that this conceptualization also includes point source data and target values as particular cases of infinitesimally small sampling units. Based on pre-defined models of spatial auto- and cross-correlation between these latent fields, we derive linear regression coefficients in a generalized least squares (GLS) setting for the unobserved point data, as well as for the observed data on the dependent variable and its predictors at the source zones. In doing so, we illustrate that what is generally conceived as the Modifiable Area Unit Problem (MAUP) is actually predictable and thus an effect, not a problem. We then illustrate under which conditions such regression coefficients inferred at the source zones can be extrapolated to the target zones for prediction; these conditions apply both in a global prediction setting as well in the local setting of GWR.

The above conditions indicate that traditional GLS-based regression models are only appropriate when point values of the dependent attribute and its predictors are upscaled using the same scheme, i.e., when both aggregation and zoning aspects of the MAUE act in the same way for both the response and its predictors. To go beyond this rather uncommon situation, we propose an extension of coKriging that accounts for scale differences between source data and target values. We illustrate that such a coKriging approach is also appropriate when one wants to include lagged variables in the regression model. Last, we demonstrate that particular models of auto- and cross-covariances between the latent fields can yield identical regression coefficients as those obtained using traditional regression models.

The geostatistical data integration approaches proposed in this paper have been implemented in the form of a Matlab toolbox, whose capabilities and functionalities are briefly presented, along with some directions for future research and code improvement.

3. Acknowledgements

The authors would like to acknowledge funding from the National Geospatial Intelligence Agency under the project "*Strategic Enhancement of NGA's Geographic Information Science Infrastructure*".

4. References

Fotheringham AS, Brunsdon C and Charlton ME, 2000, *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*, John Wiley & Sons, Chichester, UK.

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

Kyriakidis PC and Goodchild MF, 2007: A geostatistical perspective on areal interpolation: The case of cartographic methods, *International Journal of Geographical Information Science* (under revision).