Areal Interpolation and Dasymetric Modeling

Fang Qiu¹, Robert Cromley²

¹University of Texas at Dallas, Richardson, Texas, USA, ²Department of Geography, University of Connecticut, Storrs, Connecticut, USA

Frequently in spatial analysis, data are collected using one measurement system while analyses are conducted using a different measurement system. In these two systems, data regarding individual objects often are aggregated into areal units because (1) data concerning personal information are restricted by privacy and confidentiality regulations; (2) aggregated data require less storage and have a computational advantage over data in disaggregated form; and (3) geographers traditionally study data at a regional level (Openshaw and Taylor 1981). Areal interpolation refers to the procedures for transferring attribute data from one partitioning of geographic space (a set of source units) to another (a set of target units) (Goodchild and Lam 1980). Areal interpolation is needed to estimate attribute information for different geographic partitionings at the same scale (an alternative geography problem), for spatial partitionings at a finer resolution (a small area problem), for reconciling boundary changes in spatial units over time (a temporal mismatch problem), or for incomplete coverage (a missing data problem).

Geographic information systems (GISs) have increased the need to change measurement systems because a GIS integrates source data from different systems into a common database, and GIS analyses also create new data layers with different spatial units from source layers, such as in an overlay operation.

As areal interpolation methods for solving these aforementioned estimation problems developed, several critical components to the solution methodology emerged. First, different statistical methods were adapted for areal interpolation, ranging from simple proportions to more sophisticated procedures such as expectation maximization (EM) algorithms and various forms of regression analysis. Second, many of these methods began to incorporate ancillary data that act as control units on the geographic distribution of the attributes being estimated. In doing so, source data are used to estimate a density surface for control units, such as in dasymetric mapping, followed by this density surface being aggregated by target zones to create final estimates. Consequently, researchers have focused not only on better analytical methods but also on the incorporation of better data that are available for source zones and control zones.

The purpose of this special issue is to examine the status of these developments in areal interpolation. The first three articles involve expanding, improving, and redeveloping a popular areal interpolation approach, the EM algorithm. Schroeder and Van Riper (2013) present a geographically weighted expectation-maximization (GWEM) dasymetric interpolation algorithm. Their approach expands existing EM algorithms by introducing the benefits of geographically weighted regression (GWR), which allows the densities of different control units to have a different ratio among all source units. The GWEM algorithm incorporates the best of both EM and GWR, while overcoming the limitations of each individual technique. Schroeder and Van Riper use their GWEM algorithm to interpolate historical 1980 census data backward to 1970 data with land use and land cover data as control units. It achieves better accuracy than that
reported in earlier studies. Additional improvements are made by incorporating target-density weight as extra ancillary information.

Sridharan and Qiu (2013) present a spatially disaggregated EM algorithm using light detection and ranging (LiDAR)-derived residential building volume as the control variable. Their model specification improves upon pixel-based approaches using a least squares approximation to the EM algorithm in which areal interpolation is treated as missing data. LiDAR-derived building volume accounts for the vertical distribution of population, which one-dimensional length data such as roadways and two-dimensional area data are unable to measure. These authors propose new approaches for the initialization of, iterative adjustment in, and stopping criterion for their EM model that are more appropriate for varying-sized control units, replacing the equal-sized pixel control units used in earlier EM applications. A case study compares the performance of various designs and then evaluates these designs against earlier areal interpolation models in terms of accuracy and precision.

Griffith (2013) applies the EM method to the problem of missing data. His research builds on his earlier work using the EM method to impute missing values. Here, the problem is to apply the EM algorithm to a georeferenced Poisson random variable of counts. A mixed model is implemented and contrasted with a Poisson-gamma mixture (i.e., negative binomial) model. Several experiments show that temporal covariates decay in usefulness over time and that temporal covariates (a 10-year lag) produce better results than a purely spatial contemporaneous covariate of the structure of reported values. He also explores the use of regional total constraints and of spatially structured random effects.

The next three articles explore nontraditional data for source, target, and control units, with improved algorithms specifically designed to work with these data. Leyk, Nagle, and Buttenfield (2013) summarize research that is different from most other areal interpolation applications in this special issue, where values at area source units often are spatially disaggregated into smaller units, such as households (Bentley, Cromley, and Atkinson-Palombo 2013; Sridharan and Qiu 2013), before reaggregating into target units. The source data in their application are household records from public use microdata sample (PUMS) files, which contain rich attribute information. The uniqueness of these source data lies in not only that they constitute a small 5% sample but also that their spatial location is unknown. A maximum entropy model spatially allocates records from the PUMS files to census tracts based on a set of related variables and a limiting variable (i.e., a control unit) in the form of land cover data. The sampling weights imputed using the maximum entropy model also provide household-level uncertainty for the dasymetric modeling process.

Bentley, Cromley, and Atkinson-Palombo (2013) extend a network interpolation method proposed by Maantay, Maroko, and Herrmann (2007) by coupling areal control units in the form of cadastral lots with the already-used linear control units of street segments. They implement their extension of the method by incorporating Tobler’s (1979) pycnophylactic method to smooth interpolated values so that the population density on one side of a street segment is similar to that on the other side of that segment, while maintaining the pycnophylactic property. The extended method is successfully applied to a flow modeling problem. The distinctiveness of their application is that the target units are dynamically defined service areas during the flow modeling process. Their new network interpolation methods exhibit better performance than an area-based interpolation approach to flow modeling.

Finally, Langford (2013) explores the accuracy of areal interpolation in providing small area population estimates, where “small areas” are defined as target zones much smaller than those of
the finest geographic resolution census division. He also examines the importance of ancillary data sources used in intelligent areal interpolation, particularly high-quality, open access data available at no cost. If areal interpolation is to have widespread usage by the general public, users need control data that not only are easy to access, to understand, and to use, but also provide accurate results. Dasymetric mapping methods using released open access data for the United Kingdom perform particularly well when estimating counts for small area estimates in Langford’s case study.

In conclusion, the six articles appearing in this special issue reflect some of the recent advancements in the research of areal interpolation, particularly those based on dasymetric modeling. These advancements are illustrated by the various novel approaches to areal interpolation introduced in these articles, with a special emphasis on the improvement upon EM-based methods. Contributions to the literature also lie in this special issue’s summaries of explorations of some newly available data sources for improving areal interpolation accuracy (e.g., three-dimensional LiDAR data, household sample data, and open access data), along with the algorithms particularly developed to take advantage of these data. We are grateful for the opportunity that Dr. Daniel Griffith, the current editor of Geographical Analysis, has provided to publish this special issue here. We also appreciate the time and effort devoted by many reviewers to provide constructive comments and insightful suggestions that improved each article.

References


