[AM-02-040] Areal Interpolation

Abstract

Areal interpolation is the process of transforming spatial data from source zones with known values or attributes to target zones with unknown attributes. It generates estimates of source zone attributes over target zone areas. It aligns areal spatial data attributes over a single spatial framework (target zones) to overcome differences in areal reporting units due to historical boundary changes of reporting areas, integrating data from domains with different reporting conventions or in situations when spatially detailed information is not available. Fundamentally, it requires assumptions about how the target zone attribute relates to the source zones. Areal interpolation approaches can be grouped into two broad categories: methods that link target and source zones by their spatial properties (area to point, pycnophylactic and areal weighed interpolation) and methods that use ancillary or auxiliary information to control, inform, guide, and constrain the interpolation process (dasymetric, statistical, streetweighted and point-based interpolation). Additionally, there are new opportunities to use novel data sources to inform areal interpolation arising from the many new forms of spatial data supported by ubiquitous web- and GPS-enabled technologies including social media, PoI check-ins, spatial data portals (e.g for crime, house sales, microblogging sites) and collaborative mapping activities (e.g. OpenStreetMap).

Keywords: area weighted, areal interpolation, dasymetric mapping, data integration, geostatistics, historical GIS, interpolation, pychnophylactic, spatial analysis, spatial transformation, zoning

Author & citation

Comber, A. and Zeng, W. (2022). Areal Interpolation. The Geographic Information Science & Technology Body of Knowledge (2nd Quarter 2022 Edition), John P. Wilson (Ed.).
DOI: 10.22224/gistbok/2022.2.2.

Explanation

1. Definitions
2. A Brief Introduction to Areal Interpolation
3. Areal Interpolation Using Zone Spatial Properties
4. Areal Interpolation with Ancillary Information
5. Summary

1. Definitions

Interpolation: the estimation of unknown values from observations with known values.

Areal Interpolation: The process of transforming data from one spatial framework to another.

Source zones: The original areas with known values or attributes before areal interpolation

Target zones: The new areas, for which values will be estimated by the areal interpolation procedure.

Ancillary information: Data that is used to constrain target zone estimates, usually some form of spatial data.

2. A Brief Introduction to Areal Interpolation

In geography, spatial interpolation is the process of re-distributing some observation, measure or count reported over one geographic framework to another. Spatial interpolation can broadly be divided into two methods: point interpolation and areal interpolation. Point interpolation is used to estimate values from point observations at locations where values are unknown but assumed to vary continuously over space e.g. Kernel Density Estimation; and Inverse Distance Weighting, forthcoming).

Areal interpolation is the process of transforming values or attributes reported over spatial data in area or polygon format with known values, referred to source zones, to a new set of areas referred to as target zones with unknown values (Goodchild and Lam, 1980). It is the process of re-distributing data reported over one areal geographic framework to another. There are a number of approaches to areal interpolation and each of them generates estimates of Target Zone attributes over the Source Zones. The attributes are usually counts for which the overall count sum is preserved or rates for which the overall mean is preserved.

Areal interpolation is undertaken for 3 main reasons:

1. As a first step in analyses that require spatial data recorded over different spatial frameworks to be integrated to a common one. It is frequently used to transform data reported at one scale but where the analysis is undertaken at another, for example, to link socio-economic data captured in a population census and reported over small areas to pollution data reported over a 1km grid.
2. To link data when administrative or reporting boundaries have changed, such as periodically occurs with data from different censuses, causing problems for data integration and analyses of historical change. This is a persistent problem. For example, some census areas are commonly defined to include a similar number of people in each area as in the UK, and these have to be adjusted as the underlying population changes.
3. To overcome a lack of reporting or data availability at fine spatial scales. For example, many countries do not report census data over small areas, for reasons related to privacy, a lack of data infrastructures or political will.

There are a number of approaches for areal interpolation and the reallocation of source zone variables to target zones. These are extensively reviewed in Comber and Zeng (2019) with coded illustrations in R, and can be grouped into two broad categories: methods that rely solely on the target and source zone spatial properties and methods that use ancillary (or auxiliary) information to constrain or inform the reallocation.

3. Areal Interpolation Using Zone Spatial Properties

Areal interpolation methods that use only target and source zone spatial properties are usually chosen when information to inform or constrain the allocation to target zones is not available (Comber and Zeng, 2019). There are three commonly used approaches for estimating target zone values: (1) areal weighted interpolation, which uses the intersecting areas of target and source zone to allocate proportionately; (2) pycnophylactic interpolation, which smooths the allocation to minimise discontinuities between adjacent target zones); (3) area-to-point interpolation which treats the target zones as point locations and uses point based, geostatistical methods.

3.1 Areal Weighting

Areal weighting allocates source zone attributes to target zones based on the area of their intersection (Goodchild & Lam, 1980; Lam, 1983). It is implemented using polygon overlay operations and is incorporated into most GIS software packages. It is widely used (Langford, 2006). The disadvantage of this method is that it assumes that the relationship between the source zone attribute being interpolated and the target zone areas is consistent – ie spatially homogenous (Goodchild & Lam, 1980). This assumption can result in inappropriate allocation of values to target zones. For example, source zone population counts are allocated evenly to intersecting target zones even if the target zones include large areas of non-residential land use such as green space or industrial areas. In the absence of ancillary data, areal weighting generates reasonable allocations.

3.2 Pycnophylactic Interpolation

Pycnophylactic interpolation (Tobler, 1979) interpolates the source zone attribute to the target zones in a way that avoids sharp discontinuities between neighbouring target zones. Target zones are frequently regular raster cells (but do not have to be) and the allocated values are iteratively adjusted to generate a smooth target zone surface. Each iteration seeks to improve the smoothness of adjacent target zone values by adjusting the allocation to each target zone, whilst preserving the source zone total (also referred to as mass or volume). It uses the weighted average of target zone neighbours, and the number of neighbours and iterations determines the overall level of smoothing. Pycnophylactic interpolation assumes that no sharp boundaries exist in the distribution of the allocated data, which may not be the case, for example, when target zones are divided by linear features (rivers, railways, roads) or are adjacent to waterbodies. However, it generates intuitively elegant allocations for many urban case studies with many applications (Kounadi, Ristea, Leitner, & Langford, 2018; Comber, Proctor, & Anthony, 2008).

3.3 Area-to-point Interpolation

Area-to-point interpolation is an extension of point interpolation. A control point for each source zone is identified (usually the centroid) and a density value is assigned to that point based on the variable to be re-allocated. The value is interpolated to a regular grid of points using a point interpolation method such as Inverse Distance Weighting (Martin, 1989; Xie 1995) and then converted back to a count value for each source zone. The interpolated values depend on the choice of the control point which affects the resulting surface. For example, in some cases, the source zone geometric centroid may be outside the zone and it may not represent the average distribution of the feature in the source zone as well as a population weighted centroid (Martin, 1989). Point-based interpolators are not volume

preserving and a scaling step needs to be added to ensure that the target zone values match the source zone total.

4. Areal Interpolation with Ancillary Information

Ancillary information that relates to the variable of interest in some way can be used to constrain or guide allocation to target zones. The aim of using ancillary data is to generate source zone estimates that better reflect the actual distribution. They key is to identify spatial data that are related to the feature of interest. For example, ancillary information might be residential areas or the number of houses in target zones that could be used to guide the allocation of source zone population counts. The increased availability and variety of spatial data describing different phenomena has resulted in greater opportunities for interpolation approaches that incorporate ancillary data (Langford, 2013).

There are four commonly used approaches for areal interpolation informed by ancillary data (Comber and Zeng, 2019): (1) Dasymetric mapping in which areal masks are used to guide and constrain allocation; (2) Street-weighting methods, which use road networks to allocate populations proportionately to road lengths in target zones; (3) Statistical approaches, which construct a statistical model of the relationship between ancillary data and the variable, to guide allocation; and (4) Point-based approaches which use point data of some feature as ancillary information.

4.1 Dasymetric mapping

Dasymetric interpolation approaches typically use some kind of spatial mask to guide the redistribution of values to target zones (Langford, 2013). It uses spatial data of features related to the variable being interpolated to identify areas in the target zones to include or exclude from the interpolation. Ancillary data can be areal features or linear and point features with a buffer. The simplest dasymetric approach creates binary masks of areas to include or exclude from the interpolation (Fisher and Langford 1996). Masks created from land use data are commonly be used, for example, exclude non-residential areas in target zones from the interpolation of population data. However, population densities may vary and some dasymetric approaches allocate different proportions to each land use class (Eicher and Brewer 2001). There are a number of considerations for dasymetric approaches. Auxiliary information may not be available and may need to be created from remote sensing data requiring an understanding of multispectral signatures and image classification techniques. However, the availability and easy programmatic access to OpenStreetMap data goes some way to overcoming this problem. Dasymetrtic approaches also assume homogenous population densities in each land use class whether binary or using multiple mask types, which may not be the case.

4.2 Street-weighting

Street-weighting uses vector street network data (Xie, 1995). Several variants exist, the simplest of which uses the network length within the source zone and allocates source zone values proportionately along road segments lying within target zone boundaries. The linear features are intersected with the target zones and the interpolated value estimated by summing each road segment within the target zone boundary. For population related variables, this approach performs better in urban areas with regularly spaced streets and

residences than in rural areas.

4.3 Statistical methods

Statistical methods use ancillary data in conjunction with a regression to model relationships between the spatial distribution of the ancillary data (such as land use types) and the spatial distribution of the variable of interest within the source zone (Goodchild, Anselin, and Deichmann 1993). The model is then applied to the target zones, containing the same ancillary data to predict the target zone values based on the intersecting ancillary data.

4.4 Point-based ancillary information

A final set of approaches use point ancillary data to guide areal interpolation to target zones. Examples of point data include address points, post code points, properties for sale or rent, schools and bus stop locations. What they do is allocate the source zone variable proportionately to the number of points within each target zone. Much point data are available and there are emergent opportunities with the many new forms of data available from 3 general sources:

- Open data initiatives, including data portals and community led open data infrastructures;
- Online service providers such as property sales and rentals data via APIs as well as commercially produced Points-Of-Interest (POI) data;
- Social media, citizen sensing, volunteered geographic information (VGI) activities.

These opportunities relate to the increasing amount of data which are routinely generated as part of our everyday lives with some form of location attached. These new forms of data, user contributed data and VGI have emerged as new sources of point data and although there are potential data quality issues, they provide valuable data sources that can complement official and commercial data (Goodchild, 2007; Bakillah et al., 2014). Zeng and Comber (2020) applied household data from house sales and rental websites to estimate fine scale population for a developed urban area (Leeds, UK) and a rapidly urbanizing one (Qingdao, China).

5. Summary

Areal interpolation seeks to generate estimates of source zones variables over target zones. Typically, the variable describes counts, and although rates can also be estimated where possible these should be converted to counts for the numerator and denominator, the interpolation procedure applied and then the rate calculated.

Each of the approaches for areal interpolation make assumptions about the nature of the relationship between the target zones and the source zone variable. The choice of areal interpolation method should be made in consideration of these assumptions. For example, areal weighted interpolation assumes the source zone variable is homogenously distributed withing the target zones; the statistical method assumes that coefficient estimates constructed over source zones are appropriate to model the variable in target zones; the street weighted for population related variables method assumes that the population are

evenly distributed along each street, etc. The assumptions frequently may not hold and the user is placed in the position of having to make pragmatic decisions to balance the different considerations and assumptions relative to the interpolation task.

Consider the interpolated data in Figures 1 and 2. These show subtly different patterns reflecting these assumptions. Figure 1 shows that estimations of house counts are more evenly distributed under the area weighting approach compared to the pycnophylactic one. This of course relates to their operation and assumptions. Areal weighting simply distributes the target variable based on areas of intersection between target and source zones. Whereas the pycnophylactic approach adjusts adjacent target zone areas estimates to create to smooth rolling surfaces.

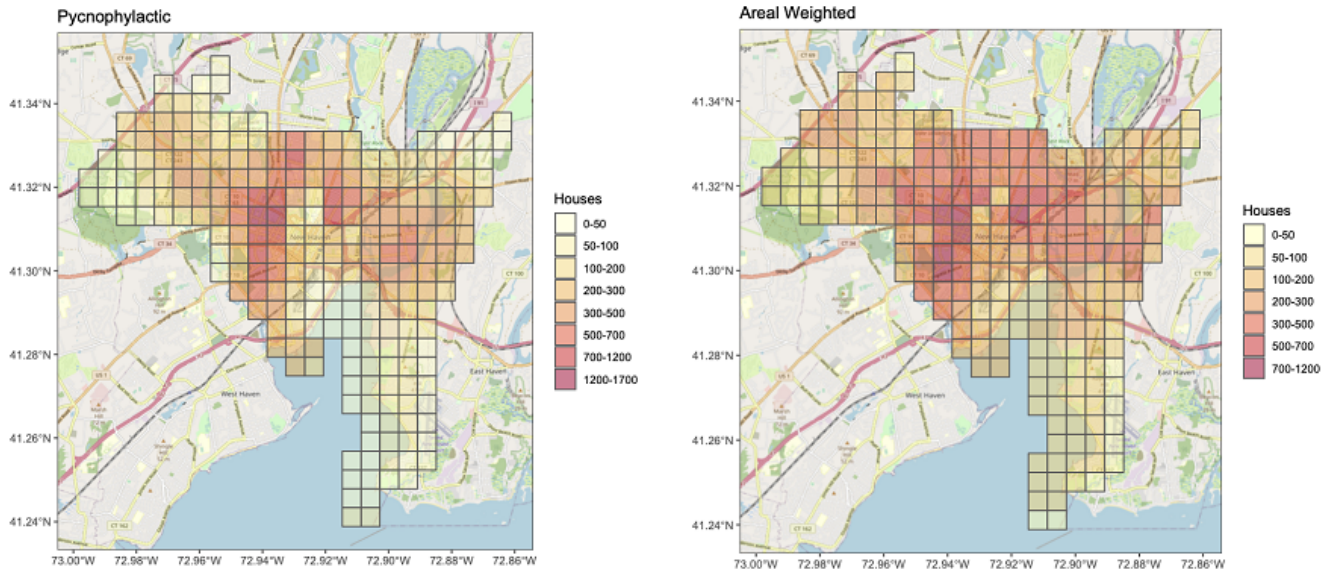


Figure 1. Examples of Pycnophylactic and Areal Weighted Interpolation. Source: author.

Similarly Figure 2 shows are clear differences between a binary dasymetric approach and a street-weighted one. The dasymetric approach results in much lower allocations to areas with few residential properties, such as to the south of the study area around the port, and in areas with large green spaces to the northwest. By contrast both of these areas have roads and streets, and the street weighted method allocates to these areas based on these. As well as some spatial differences in the distribution, these differences in underpinning assumption result in higher allocation densities in the dasymetric approach, precisely because some areas are masked out.

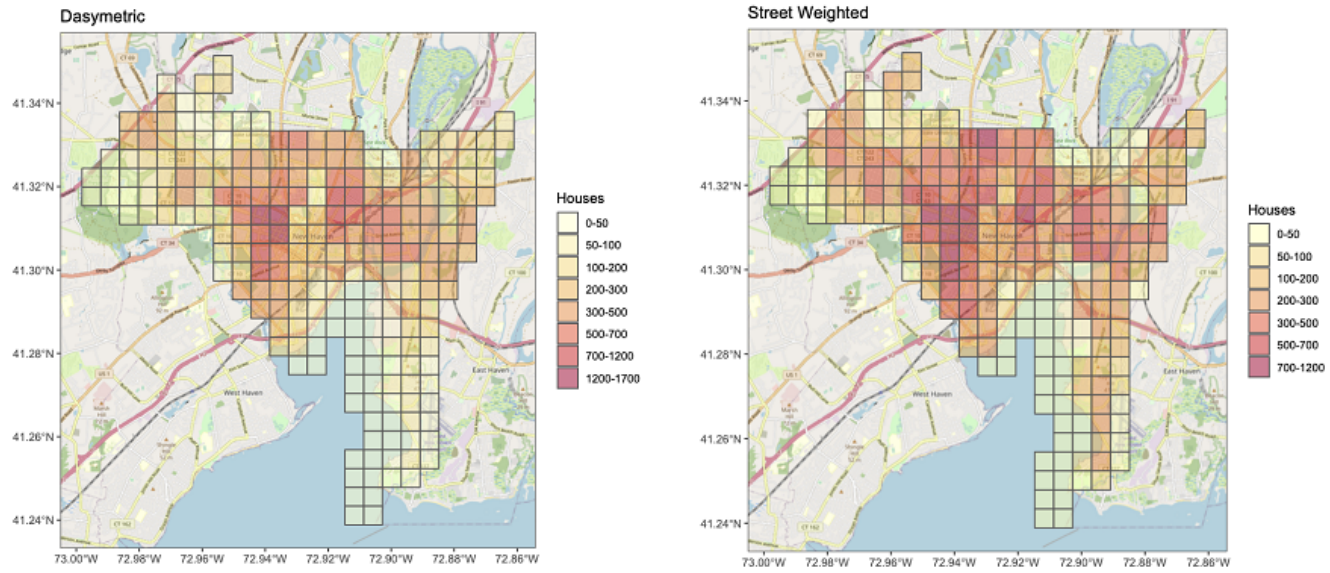


Figure 2. Examples of Dasymetric and Street Weighted Interpolation. Source: author.

This review of areal interpolation is based on Comber and Zeng (2019) which provides a deeper description of the methods and their assumptions, some guidance about implementation and includes coded examples implemented in the R programming language. The interested reader wishing to explore this topic in more depth is advised to explore the paper, the references therein and the associated GitHub site with R code and data (the link is in the Additional Resources section below).

References

Bakillah, M., Liang, S., Mobasheri, A., Jokar Arsanjani, J., & Zipf, A. (2014). Fine-resolution population mapping using OpenStreetMap points-of-interest. International Journal of Geographical Information Science, 28(9), 1940-1963.

Comber, A., & Zeng, W. (2019). Spatial interpolation using areal features: A review of methods and opportunities using new forms of data with coded illustrations. Geography Compass, 13(10), e12465.

Comber, A., Proctor, C., & Anthony, S. (2008). The creation of a national agricultural land use dataset: combining pycnophylactic interpolation with dasymetric mapping techniques. Transactions in GIS, 12(6), 775-791.

Eicher, C. L., & Brewer, C. A. (2001). Dasymetric mapping and areal interpolation: Implementation and evaluation. Cartography and Geographic Information Science, 28(2), 125-138.

Fisher, P. F., & Langford, M. (1996). Modeling sensitivity to accuracy in classified imagery: A study of areal interpolation by dasymetric mapping. The Professional Geographer, 48(3), 299-309.

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography.

GeoJournal, 69(4), 211-221.

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

Goodchild, M. F., Anselin, L., & Deichmann, U. (1993). A Framework for the Areal Interpolation of Socioeconomic Data. Environment and Planning A, 25(3), 383-397.

Kounadi, O., Ristea, A., Leitner, M., & Langford, C. (2018). Population at risk: Using areal interpolation and Twitter messages to create population models for burglaries and robberies. Cartography and Geographic Information Science, 45(3), 205–220.

Lam, N. S. N. (1983). Spatial interpolation methods: a review. The American Cartographer, 10(2), 129-150.

Langford, M. (2006). Obtaining population estimates in non-census reporting zones: An evaluation of the 3-class dasymetric method. Computers, Environment and Urban Systems, 30(2), 161–180.

Langford, M. (2013). An evaluation of small area population estimation techniques using open access ancillary data. Geographical Analysis, 45(3), 324-344.

Martin, D. (1989). Mapping Population Data from Zone Centroid Locations. Transactions of the Institute of British Geographers, 14, 90–97.

Tobler, W. R. (1979). Smooth Pycnophylactic Interpolation for Geographical Regions. Journal of the American Statistical Association, 74(367), 519–530.

Xie, Y. (1995). The overlaid network algorithms for areal interpolation problem. Computers, Environment and Urban Systems, 19(4), 287–306.

Yang, X., Jiang, G-M., Luo, X., Zheng, Z. (2012). Preliminary mapping of high-resolution rural population distribution based on imagery from Google Earth: A case study in the Lake Tai basin, eastern China. Applied Geography 32(2): 221-227.

Zeng, W., & Comber, A. (2020). Using household counts as ancillary information for areal interpolation of population: Comparing formal and informal, online data sources. Computers, Environment and Urban Systems, 80, 101440.