Computers, Environment and Urban Systems 54 (2015) 23–31

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/compenvurbsys

A local polycategorical approach to areal interpolation

Jie Lin[a,*], Robert G. Cromley[b,1]

[a] School of Earth Sciences, Zhejiang University, 38 Zheda Road, Hangzhou, Zhejiang 310027, PR China
[b] Department of Geography, University of Connecticut, 215 Glenbrook Road, Storrs, CT 06269, USA

ARTICLE INFO

Article history:
Received 2 December 2014
Received in revised form 28 May 2015
Accepted 29 May 2015

Keywords:
Areal interpolation
Local statistical
Geographically weighted regression
Quantile regression

ABSTRACT

Areal interpolation is a technique used to transfer attribute information from source zones with known values to target zones with unknown values. This paper presents and describes a new polycategorical method that integrates positive aspects of both geographically weighted regression (GWR)-based and quantile regression (QR)-based interpolators for solving areal interpolation problems. Two different types of neighborhoods for selecting observations used to estimate ancillary control densities are presented: one that is spatially based and one that is statistically based. The new polycategorical methods are evaluated against a number of existing methods – areal weighting, pycnophylactic, binary dasymetric, intelligent dasymetric mapping, and GWR using test data from the 2010 census population, the National Land Cover Database 2006 (NLCD2006) and the Topologically Integrated Geographic Encoding and Reference (TIGER) line graph files. The evaluations include several overall error measurement indices as well as maps of the spatial distribution of the error associated with selected methods. Results suggest that with appropriate land cover categories and neighborhoods, the new polycategorical methods provide comparable results to local regression models but with much less computation complexity.

1. Introduction

Often, spatial data are aggregated into areal units for further analysis, even though the data may be collected at the individual level. This aggregation occurs for several reasons (Openshaw & Taylor, 1981): (1) spatial data concerning personal information are restricted by privacy and confidentiality; (2) data in the aggregated form are convenient and require less volume for storage, and also have a computational advantage over data in a disaggregated form; and (3) geography has a long tradition of studying data at the regional level. However, different applications, variations in geographic scale, and the distinct nature of a phenomenon's distribution have worked against the unification of areal units into a single standardized system (Visvalingam, 1991). It is also well known that results of spatial analyses are sensitive to the choice of zoning system associated with the aggregation, which is known as the modifiable areal unit problem (MAUP) (Openshaw & Taylor, 1981).

Related to these issues is the change of support problem in which data are collected for one measurement support but must be transferred to a different support system before analysis is performed. This change of support requires values to be estimated at locations different from those at which the data have been observed (Gelfand, Zhu, & Carlin, 2001). For data collected and reported in areal units, it means that values must be estimated for an alternative zoning system. Areal interpolation (AI) is a procedure for transferring attribute values from one partition of geographic space (a set of source regions) to a different partition (a set of target regions) (Goodchild & Lam, 1980; Lam, 1983). The development of Geographic Information Systems (GIS) has also increased the necessity for AI as a GIS analysis frequently generates new layers of areal units for which non-spatial attribute information must be estimated.

A number of different AI methods have been developed over time to improve the efficiency and accuracy of the transfer procedures. Local models, in which the relationship between an attribute variable and ancillary information is estimated using a selected subset of observations from the full set, are now fairly common in AI procedures as they address the heterogeneity problem with respect to any relationship. By close scrutiny of two statistical interpolators that emphasize local variation, a geographically weighted regression (GWR)-based (Lin, Cromley, & Zhang, 2011) and a quantile regression (QR)-based interpolator (Cromley, Hanink, & Bentley, 2012), this study proposes a new local polycategorical AI procedure that integrates the positive aspects of both aforementioned regressions. QR is a regression with varying parameter estimates like GWR, but these regression models differ in two respects: (1) QR minimizes the sum of

* Corresponding author. Tel.: +86 571 87952453.
E-mail addresses: jielin@zju.edu.cn (J. Lin), robert.cromley@uconn.edu (R.G. Cromley).
[1] Tel.: +1 8604862059.

http://dx.doi.org/10.1016/j.compenvurbsys.2015.05.007
0198-9715/

absolute deviations whereas GWR minimizes the sum of squared deviations, and (2) QR estimates are a function of a position in the statistical distribution (the quantile level) rather than a position in geographic space. The term 'local' used in this paper mainly stresses locally-varying model parameter estimates, which is consistent with Fotheringham's (1997) discussion, but not the selection of a subset of observations used in model calculation.

The rest of this paper is organized as follows: related work is provided in Section 2. Section 3 explains the proposed polycategorical method. The data and comparative methods are presented in Section 4. Section 5 presents and compares the results of different AI methods in conjunction with various control data. Finally, conclusions and future work are discussed in Section 6.

2. Background

AI methods are used to estimate attribute data that are associated with a density surface. The accuracy of the AI method then is a function of how accurately the underlying density surface can be approximated. The assumptions made by the AI method regarding the nature of the density surface within the source and target spatial units have a major impact on model performance. AI methods are sometimes dichotomized as being either *simple* or *intelligent*. Simple AI methods, transfer data from source zones to target zones without using any ancillary data whereas intelligent areal interpolation methods use some form of ancillary data that provide insight to the underlying density surface in order to improve estimation accuracy (Langford, 2006; Langford, 2007; Langford, Maguire, & Unwin, 1991). Ancillary data then are used to infer the internal structure of the density surface within each source zone. For estimating population, two readily available ancillary datasets that are commonly used are remotely sensed land cover data and road networks. In the US land cover data are easily accessible from the National Land Cover Database (NLCD) and road networks can be obtained from the U.S. Census Bureau's 2010 TIGER/Line files. Another ancillary data used more recently in the areal interpolation of population is parcel data. Maantay, Maroko, and Herrmann (2007) developed the so-called cadastral-based expert dasymetric system to redistribute population in urban area, while Tapp (2010) used county address points and parcels to estimate population in rural and transitional areas. Although the assumption in these methods is very straightforward, cadastral data are not as readily available as road networks or land cover data. Moreover, due to a rapid expansion in the use of Web 2.0 applications, numerous forms of volunteered geographic information (VGI) are being produced by general public. These open access geographic data together with other traditional ancillary data can also be used for the areal interpolation of population. Bakillah, Liang, Mobasheri, Jokar Arsanjani, and Zipf (2014) proposed a framework using OpenStreetMap points-of-interest and pre-classified land use land cover categories to infer population at a building level. Lin and Cromley (2015) evaluated geo-located nighttime tweets collected from Twitter.com both as single control data and as an enhancement to other control data for areal interpolation of population as well as different age-specific population groups.

2.1. Areal weighting

Areal weighting, the easiest AI method, assumes that the density surface is uniform within each source polygon (the overall surface is a 3D prism). It calculates the geometric overlay of the source and target zones, and values are estimated for each intersection zone by proportionally weighting the data counts by the area of each intersection polygon with respect to the area of the source polygon that contains it (Goodchild & Lam, 1980). It is the most widely used method among all other methods due to its intuitively simple theory, low data and computation requirement, and it can be easily included in GIS software (Xie, 1995). However, areal weighting usually produces poorer estimates when compared against results of other methods because of its simplistic representation of the density surface and lack of any ancillary data.

2.2. Pycnophylactic interpolation

Tobler's pycnophylactic method (1979), another simple AI method, creates a smooth density surface of raster units from an initial 3D prism surface associated with the source zones. Its original purpose was to construct isopleth maps. Tobler noted that the method could also be used as an interpolation technique for transferring data from one support to another by aggregating the raster values of the surface by each target zone that contains them. The regular grid has also been modeled as an irregular triangular network (TIN) surface in Rase's (2001) pycnophylactic interpolation procedure. Using this method as an AI interpolator in GIS is more difficult because it requires a pycnophylactic surface interpolator, and vector-to-raster and raster-to vector operations as intermediary steps.

2.3. Binary dasymetric method

A direct extension of areal weighting that uses ancillary data and is easy to implement in a GIS is the binary dasymetric method (Fisher & Langford, 1996). For estimating population, the ancillary information is classified into areas containing population and areas that do not. For land cover data, this only requires that land cover categories are reclassified into either a populated or an unpopulated category. For road networks, the length of the road network within each source could replace the area of the source zone if one assumes that the population is located along the road itself (Xie, 1995), or a buffer zone can be created around the road network that would contain the population (Mrozinski & Cromley, 1999). The areas within each source zone that do not contain population are "erased" from the total area of the source zone so that only areas that contain population are retained. The density surface is viewed as a dichotomous prism within each source zone. For binary dasymetric AI, areal weighting is then used to estimate target zone populations from the source zones based on populated areas. Thus binary dasymetric AI can also be termed an areal weighting of populated areas.

2.4. Dasymetric mapping-based polycategorical methods

A finer granularity of the density surface can be estimated by polycategorical AI methods. For example, instead of grouping land cover categories into populated and unpopulated, these categories can be grouped into different levels of population densities. Polycategorical methods either follow the principles of dasymetric mapping (Wright, 1936) or some statistical method. Eicher and Brewer (2001) implemented two polycategorical approaches based on Wright's dasymetric method but they used a totally subjective scheme to estimate the population densities within the different categories. The limiting variable method includes more land cover groups and produces significantly better results than other methods, while the three-class method produces only slightly better results than the binary dasymetric approach.

In contrast, Mennis and Hultgren (2006) developed the intelligent dasymetric mapping (IDM) method to redistribute population between different land cover types in a more objective manner. The method first overlays the layer of source zones against the layer of ancillary zones. Next, the density of each ancillary class is estimated by first associating source zones with a specific