ResearchGate

See discussions, stats, and author profiles for this publication at:
https://www.researchgate.net/publication/257340941

An Evaluation of Small Area Population Estimation Techniques Using Open Access Ancillary Data

Article *in* Geographical Analysis · July 2013

DOI: 10.1111/gean.12012

CITATIONS
20

READS
248

1 author:

Mitchel Langford
University of South Wales
65 PUBLICATIONS **1,315** CITATIONS

SEE PROFILE

All content following this page was uploaded by Mitchel Langford on 28 October 2016.

The user has requested enhancement of the downloaded file.

geographical analysis

Geographical Analysis (2013) **45**, 324–344

An Evaluation of Small Area Population Estimation Techniques Using Open Access Ancillary Data

Mitchel Langford

GIS Research Unit, Faculty of Advanced Technology, University of Glamorgan, Pontypridd, Wales, U.K.

National census data represent the "gold standard" for authoritatively portraying a country's residential population distribution, but their aggregated counts for fixed administrative areas present problems for many geographic information system (GIS) analyses. Intelligent areal interpolation algorithms assist by transferring data from one zonal system to another using ancillary data to improve accuracy. All areal interpolation methods make assumptions and generate errors, and performance varies with both specific location and the data inputs used. This study adds to our understanding of the relative merits of alternative methods by comparing dasymetric, street network, and surface-based models interpolating across two spatial resolutions. It examines the importance of the ancillary data source used to drive the process, particularly the efficacy of open access products. Results from an empirical study show that interpolation accuracy is influenced by the choice of ancillary data input as well as the methodological approach adopted. The strongest overall performance is delivered by dasymetric mapping combined with open access data identifying the locations of buildings. Open access data sets offer considerable potential for widening the use of intelligent population interpolation tools, especially if plug-in tools to execute these algorithms can be made available for commonly used GIS software packages.

Introduction

Small area estimates of population counts and other demographic variables are essential for the effective integration and analysis of disparately sourced data sets in geographic information system (GIS)-based modeling. The demography expressed via a national census typically represents the "gold standard" in terms of accurately and authoritatively portraying the magnitude, characteristics, and spatial distribution of a country's residential population. However, to adhere to statutory obligations regarding confidentiality and nondisclosure requirements, and to help

Correspondence: Mitchel Langford, GIS Research Unit, Faculty of Advanced Technology, University of Glamorgan, Pontypridd, Wales CF37 1DL, U.K.
e-mail: mitchel.langford@southwales.ac.uk

Submitted: January 17, 2012. Revised version accepted: December 24, 2012.

doi: 10.1111/gean.12012
© 2013 The Ohio State University

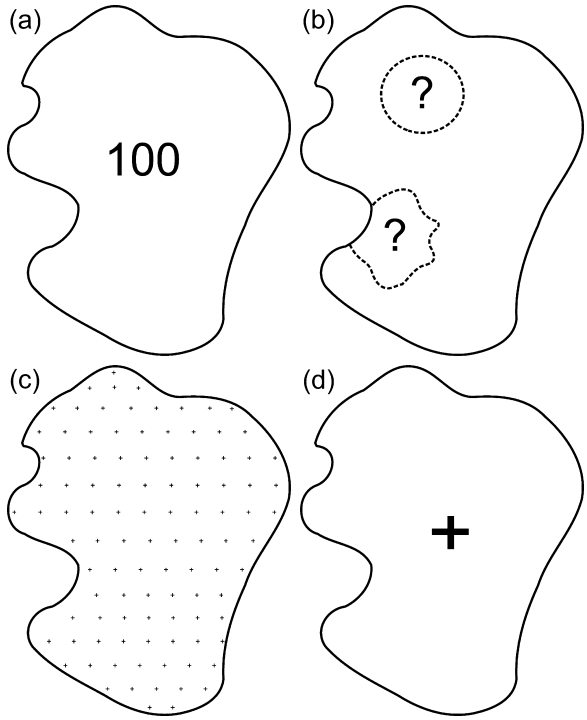

Figure 1. Simple areal interpolation. Aggregated data are available for source zone (a) but required for target zones (b). Areal weighting distributes population uniformly across source zone (c), while a centroid places all population at a representative point (d). Either distribution model may be used to provide estimates for target zones.

constrain potential data volumes, individual census records typically are not released. Instead, aggregated counts are reported for a specified set of fixed areas. This poststratification can be problematic for GIS analysts attempting to undertake spatial modeling because population counts often are required for alternative and incongruent areal units to compute an incidence ratio or to enable the integration of geographic information derived from disparate sources. This situation arises frequently in GIS analyses because the boundaries of natural phenomena (e.g., watersheds, land cover parcels), those of socioeconomic data sets constructed using alternative discrete geographies (e.g., postal delivery zones, police beat areas), or analytical zones created directly by a GIS (through operations such as spatial buffering or the computation of network distance travel time catchments) typically overlap census zones in complex ways.

Enabling the transfer of attribute data from one zonal system to another requires some form of areal interpolation (Flowerdew and Green 1994). In this context, census units with a known population count (Fig. 1a) are termed source units, whereas those requiring an estimated count (Fig. 1b) are termed target units (Goodchild, Anselin, and Deichmann 1993). Only if source units nest completely and without overlap inside target units is areal interpolation unnecessary, because simple summation on the basis of inclusion suffices. More typically, target units partially overlap multiple source units, occupy a subspace within a source unit, or consist of a mixture of source units lying completely within a target unit while others partially intersect it. As the proportion of partially intersecting source units increases or as the size of a target unit lying completely within

a source unit diminishes, the interpolation task becomes progressively more demanding and prone to estimation error (Sadahiro 2000a).

A wide variety of algorithms have been developed to perform areal interpolation (e.g., Lam 1983; Flowerdew and Green 1989; Harvey 2000; Cai et al. 2006; Lo 2008). The simplest of these assign source population counts to a representative point location, then estimate target counts by summing all point counts that fall inside their boundary or utilize only the respective geometries of the intersecting areal units (Goodchild and Lam 1980). Over the last two decades, the increasing sophistication of GIS software and growing availability of digital spatial data sets have led to the development of many "intelligent" areal interpolation algorithms (e.g., Flowerdew and Green 1992; Fisher and Langford 1995; Xie 1995). These algorithms seek to provide more accurate target zone estimates by utilizing additional information residing within a GIS database to provide guidance about the probable internal distribution of source zone population. Ultimately, every areal interpolation algorithm is based upon underlying assumptions regarding the likely distribution of population within source units. Consequently, the accuracy of provided estimates is always a reflection of the validity of these assumptions and of the appropriateness of the ancillary data sets employed.

To test interpolation performance and to compare alternative areal interpolation algorithms, estimates must be generated for a set of target units whose true values are also known. This necessity has led to the widespread adoption of a testing framework in which researchers build a population distribution model using one level in the hierarchy of census spatial units (e.g., *block groups* in the United States or *census wards* in the United Kingdom) and examine its performance by interpolating to a lower level within the same hierarchy (e.g., *blocks* in the United States or *output areas* in the United Kingdom). Table 1 illustrates the popularity of this approach within the current literature. This general strategy renders several implications: first, our understanding of the relative performance of competing areal interpolation algorithms is largely based on models constructed using census data that are not the most spatially detailed because this resolution is reserved for testing; second, although constructing and using models based on the finest level of census data is possible, and we might reasonably assume that these are the most accurate data, we seldom are able to evaluate formally their performance at the finest spatial resolution; and third, whatever level of census hierarchy is used to construct a population distribution model, we cannot be sure of its performance when estimating counts for target units considerably smaller than those of the finest census zone division.

The purpose of this article is threefold. First, it aims to contribute to the growing set of studies comparing performance among alternative intelligent areal interpolation methods (e.g., Hawley and Moellering 2005; Brinegar and Popick 2010; Tapp 2010; Zandbergen and Ignizio 2010). Despite such testing, Zandbergen and Ignizio (2010) acknowledge that all methods have assumptions, flaws, and errors that their performance may vary with location and data conditions, and that no single "best method" has yet been established. This study includes recent additions to the methodology literature among the subset of techniques that are formally evaluated and compared. Second, by utilizing known population counts for U.K. unit postcodes (UPCs), this article explores the accuracy of the tested techniques in providing small area estimates, where *small areas* are defined as geographic target units demonstrably smaller than those of the finest census division. Finally, and perhaps most significantly, this article focuses attention on the importance of the ancillary data sources used to drive intelligent interpolation processes. This factor often is critical in determining interpolation performance, yet it often has been overlooked in discussions concerning the relative merits and underlying assumptions of competing

Table 1 Selection of Previous Literature Reporting Areal Interpolation Techniques and the Testing Framework Adopted

Authors/year	Synopsis	Study area location	Source units used to construct model	Target units used to test model
Fisher and Langford 1995	Modeling areal interpolation errors	Leicestershire County, United Kingdom	U.K. wards	Randomized aggregations of U.K. enumeration districts (EDs)
Yuan, Smith, and Limp 1997	Modeling census population with Landsat imagery	Central Arkansas, United States	U.S. block group	No testing undertaken
Eicher and Brewer 2001	Areal interpolation using dasymetric mapping	Various U.S. states	U.S. county	U.S. block group
Mennis 2003	Dasymetric mapping and population modeling	Southeastern Pennsylvania, United States	U.S. census tract	U.S. block group
Holt, Lo, and Hodler 2004	Dasymetric-based estimation of population density	Metropolitan Atlanta, Georgia, United States	U.S. census tract dated 1980, 1990, and 2000	U.S. census tract dated 1990
Langford 2006	A three-class dasymetric model	Leicestershire County, United Kingdom	U.K. ward	U.K. enumeration district
Mennis and Hultgren 2006	Dasymetric mapping for areal interpolation	Colorado, United States	U.S. census tract	U.S. block group
Langford 2007	Dasymetric mapping using raster pixel maps	Leicestershire County, United Kingdom	U.K. ward	U.K. OA
Merwin, Cromley, and Civco 2009	Areal interpolation using a neural network	Connecticut, United States	U.S. town, U.S. town, U.S. census tract	U.S. census tract, U.S. block group, U.S. block group
Su et al. 2010	Multilayer dasymetric mapping	Metropolitan Taipei, Taiwan	County	Chinese li
Zandbergen and Ignizio 2010	Comparing small area estimation techniques	Various U.S. states	U.S. census tract	U.S. block group
Brinegar and Popick 2010	Comparing small area estimation techniques	Sample areas drawn from 24 U.S. states	U.S block group	U.S. block
Tapp 2010	Areal interpolation in rural areas	North Carolina, United States	U.S. county	U.S. block group
Zhang and Qiu 2011	Surface-based interpolation from point data sets	Collin County, Texas, United States	U.S. census tract	Zip code tabulation areas

methodologies. In particular, research summarized in this article scrutinizes the efficacy of high-quality open access (i.e., "no cost") data sets for use as ancillary inputs to intelligent areal interpolation models. I am unaware of any published examination to date of the effectiveness of several of the specific open access data sources that are addressed here.

Theoretical background and previous methods

If census counts and their corresponding zones are the only information available to a GIS analyst, choices for areal interpolation are limited. In this situation, *areal weighting* may be used (Goodchild and Lam 1980; Lam 1983), which makes the assumption that source zone population is evenly spatially distributed within the zone boundary (Fig. 1c). Intersection zones are created by the overlay of source and target zones, and each receives a fraction of the source zone count based upon its area size relative to the overlapping source zone area. Intersection zones and their interpolated counts are aggregated to form the target zone estimates. This process can be expressed algebraically as

$$\hat{P}_t = \sum \frac{A_{ts}}{A_s} \cdot P_s \tag{1}$$

where $\hat{P}_t$ is the estimated population of target zone t, P_s is the known population of source zone s, A_s is the area of source zone s, and A_{ts} is the area of intersection between target zone t and source zone s.

Alternatively, a source zone centroid may be available (Fig. 1d)—either a computed geometric centroid or, preferably, a population-weighted centroid typically provided by the data supplier and deemed to represent the center of gravity of the contained population. Then, the population can be allocated on the basis of the inclusion of such points within a target zone's boundary. Fig. 1b and 1d indicate that neither target zone in this example would receive any population count using this method. The allocation of counts to target zones based on this all-or-nothing rule is prone to considerable error unless source zones are small relative to target zones (Sadahiro 2000b). Likewise, estimates based on simple areal weighting may contain significant error because the assumption of uniform population distribution within a source zone is seldom an accurate reflection of reality.

Increasingly, census counts and source and target zone boundaries are not the only information available when conducting areal interpolation using a GIS. In recognition of this change in data availability, numerous intelligent areal interpolation algorithms have been developed over the last two decades. They aim to improve interpolation accuracy by utilizing additional information residing in a GIS to provide guidance about the internal distribution of source zone population. Fig. 2a illustrates a typical scenario where a range of data sets is available in the GIS—in this example, polygons depicting building locations, lines portraying a road network, and points identifying bus stops. These data sets (as well as any number of possible alternatives) have the potential to inform *where* population might reasonably be expected to be located within source zone boundaries. The following sections describe in greater detail some specific intelligent areal interpolation methodologies that utilize this general concept.

Interpolation by dasymetric mapping

Dasymetric mapping can be defined as a technique in which attribute data collected within an arbitrary areal unit is more accurately distributed within that unit by the overlay of additional

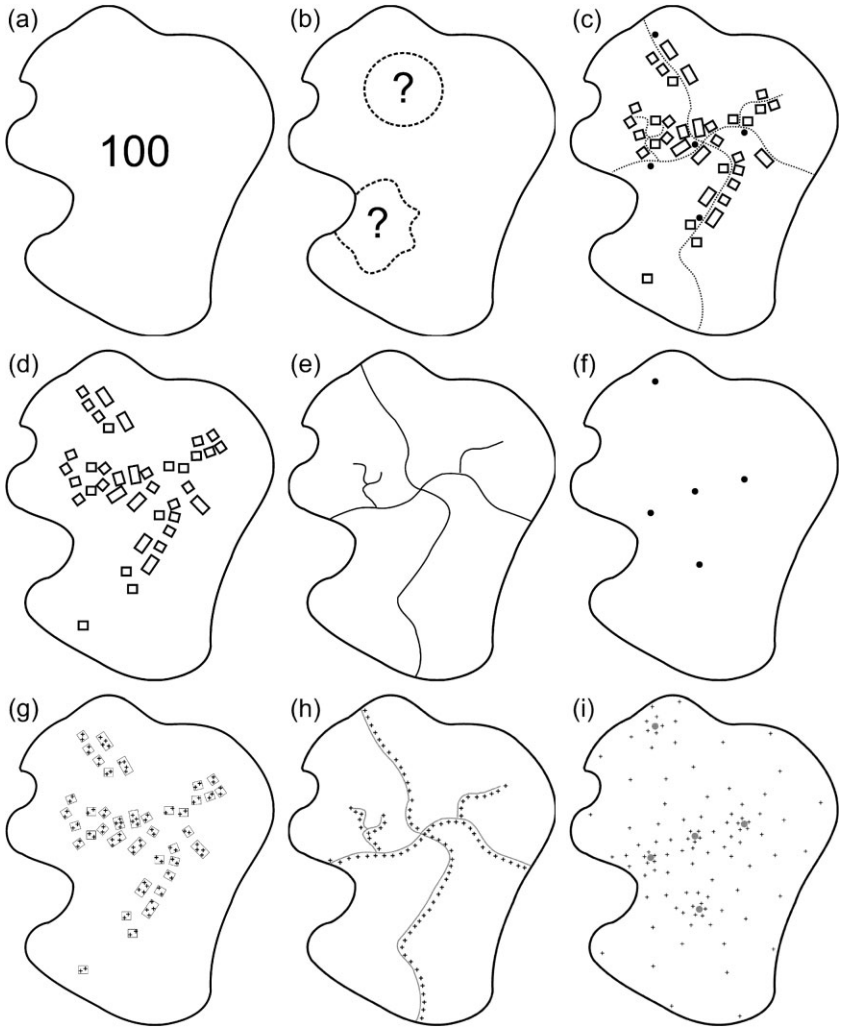

Figure 2. Schematic example of intelligent areal interpolation. Population aggregated for source zone (a) are required for target zones (b). Ancillary data residing in a GIS (c) can be used to model the internal distribution of source zone population. For example, the spatial distribution of buildings (d), of roads (e), or bus stops (f) can be used to yield intelligent estimates of source zone population distribution, shown in (g), (h), and (i), respectively.

geographic boundaries derived from ancillary data sources. These boundaries help to exclude, restrict, or confine the attribute in question in order to generate internal subdivisions that possess increased homogeneity and that better represent the actual underlying geographic distribution. For example, population counts in census zones may be more accurately distributed by the overlay of water bodies, vacant land, and other land use parcels within which people are not expected to live. Dasymetric mapping was first developed in the early 20th century as a cartographic technique aimed at addressing some of the issues associated with choropleth mapping (Wright 1936). Fisher and Langford (1995) demonstrate how dasymetric mapping principles also can be used to enhance areal interpolation algorithms. This insight has led to a rejuvenated

interest in dasymetric mapping; in 2008, Petrov (2008) reported that over 60% of all indexed journal articles explicitly using the term have been published post-2004.

The definition of dasymetric mapping allows a number of methodological variants to be identified. This article restricts itself to the simplest form whereby source zones are subdivided into populated and unpopulated subzones (the *binary dasymetric method*). The rationale behind this approach is that people live in houses, or more generally in residential areas, and if their location within source zone boundaries can be identified, then the population count can be evenly distributed only within this subspace. Fig. 2d and 2g illustrate this idea; the population of the source zone is distributed only within the areas identified as housing. This redistribution can be expressed algebraically as

$$\hat{P}_t = \sum \frac{A_{tsp}}{A_{sp}} \cdot P_s \tag{2}$$

where A_{tsp} is the area of intersection between target zone t and source zone s having land cover identified as populated, and A_{sp} is the area of source zone s having land cover identified as populated.

Although the binary variant has been the most widely used to date, multiclass dasymetric mapping also is possible using category weightings derived through empirical sampling (Mennis 2003) or statistical regression (Langford 2006) approaches. In addition, recently sophisticated dasymetric-based methods have been proposed using, for example, cadastral-based data inputs (Maantay, Maroko, and Herrmann 2007) or other assorted data such as topography, land use zoning, and transportation layers progressively applied in a multilayer, multiclass dasymetric model (Su et al. 2010). However, the implementation of such solutions imposes greater demands in terms of ancillary data requirements and overall computational complexity, and actual improvements in performance over simple binary dasymetric interpolation often have been shown to be relatively modest (e.g., Cromley, Hanink, and Bentley 2012). For this reason and to retain a focus on the influence of ancillary data inputs rather than the complexities of competing dasymetric methodologies, such models are not addressed in this article.

Most studies to date have employed land cover derived from classified satellite imagery as the ancillary data input. In the early 1990s, when the dasymetric technique was first proposed, the availability of suitable vector data sets in the public domain was at best uncommon or, as in the United Kingdom, largely nonexistent. Satellite imagery does have some inherent advantages such as its universal coverage, but it also demands an understanding of multispectral signatures and image classification techniques that may be outside many GIS analysts' skill sets. Reibel and Bufalino (2005) note that perhaps for this reason the use of this method has been largely restricted to computational experiments conducted by academics. To address this issue, other researchers have sought to find alternative ancillary data sources. For example, Reibel and Agrawal (2007) use preclassified national land cover data, whereas Moon and Farmer (2001) employ manually digitized map data, and Langford (2007) experiments with information derived from raster pixel maps. A recent development in the United Kingdom has been the opening up of access to high-quality national mapping agency vector data sets for cost-free and unrestricted public use as part of the Making Public Data Public initiative. This development offers a new opportunity to implement binary dasymetric interpolation in the United Kingdom using an accessible, freely available, and highly consistent ancillary data resource. I believe that the usefulness of this new

resource for areal interpolation is tested for the first time in the experimental study summarized here.

Interpolation by street weighting

The street-weighting method was first proposed by Xie (1995) and utilizes vector street network data. Several variants of the methodology exist, the simplest of which is the network length version, which takes source zone population and uniformly distributes it along all road segments lying within its boundaries. Next, these linear features are intersected with a target zone and an estimated count derived by summing the population contained along each component vector within the target zone's boundary. Fig. 2e and 2h illustrate this idea, which can be expressed algebraically as

$$\hat{P}_t = \sum \frac{L_{ts}}{L_s} \cdot P_s \tag{3}$$

where L_{ts} is the length of each street vector in the intersection zone between source zone s and target zones t, and L_s is the total length of street vectors found in source zone s. The only significant difference between dasymetric interpolation and street weighting is in the dimensions of the ancillary data set used. In the former, the extent of an area object within the intersection zone drives the process; in the latter, the length of a linear object accomplishes it.

The rationale behind this approach is as follows: in modern society, people tend to live in close proximity to transport links and specifically to roads. The use of a vector data set avoids some of the problems previously noted with respect to satellite imagery, but other potential problems do exist. For example, not all roads have residential housing alongside them, and even when they do, the density of occupancy may not be uniform along all roads within each source zone. Despite these concerns, Hawley and Moellering (2005) suggest that this technique performs better than the binary dasymetric method, although other studies dispute this viewpoint and report the opposite finding (Tapp 2010; Zandbergen and Ignizio 2010). All such outlooks are based on empirical studies conducted exclusively within the United States, and a tendency exists to judge the relative merits of these alternative methodologies without considering the influence that the ancillary data set might have in determining their respective performances.

In the United States, vector road network data have been publicly available for a considerable time via TIGER line files (U.S. Census Bureau 1993); but such open access provision often has not been mirrored elsewhere, including in the United Kingdom. Once again, the recent U.K. policy change related to open access to national mapping agency vector data sets has transformed this situation. Road network data derived from Ordnance Survey's (OS's) VectorMap® District (Ordnance Survey, Southampton, Hampshire, U.K.) products are tested for use in street-weighted interpolation in the experiments summarized in this article. I believe that this is the first study to do so. An open access alternative also now exists in the form of volunteered geographic information. The OpenStreetMap geographic database (OpenStreetMap 2012) is freely downloadable, too, although its testing is not summarized here.

Interpolation by surface volume integration

As its name implies, surface-based interpolation requires a statistical surface to be fitted to source zone data. The volume beneath this surface represents the population count, and once such a

surface has been constructed, the volume can be measured for any desired target zone boundary. An early example of this methodology is Tobler's pycnophylactic interpolation (Tobler 1979), which, like areal weighting, requires no ancillary data. It redistributes source zone population internally such that total volume is preserved while sharp transitions in value across adjacent zone boundaries are eliminated. In the United Kingdom, Bracken and Martin (1989) use a distance-weighted kernel estimator to create a discrete surface by distributing counts initially placed at population-weighted centroids into proximal cells residing in an overlaid grid. The same methodology was adopted by Harris and Longley (2000), although they replaced the single centroid for each source zone with multiple initial population points based on address code records.

Zhang and Qiu (2011) propose an addition to the methodological literature about surface-based interpolation. This technique exploits a set of ancillary points that are believed to provide a reasonable proxy for centers of population density; their original article employs schools to demonstrate the approach. A surface based on a linear distance decay function around these points is constructed. Within the confines of each source zone, the volume beneath the surface is scaled to represent the known population count (i.e., the pycnophylactic property), while its slope is adjusted such that the surface touches zero at the maximal zonal distance from a control point. Another possibility is to invert the decay profile and to utilize points associated with an absence of population (e.g., the location of waste disposal plants). The model is described algebraically as

$$\begin{aligned} \hat{D}_{si} &= a_s \cdot W_{si} \left(W_{si} \in [0,1]\right) \\ W_{si} &= \left(1 - \frac{\lambda_{si}}{\lambda_{s\max}}\right)^q \quad (q \geq 1) \end{aligned} \tag{4}$$

where $\hat{D}_{si}$ is the estimated population density assigned to cell i in source zone s, a_s is a constant of proportionality for source zone s, W_{si} is a weighting for cell i in source zone s, λ_{si} is the distance from cell i in source zone s to the nearest control point, $\lambda_{s\max}$ is the maximum value of λ_{si} in source zone s, and q is a distance decay weighting factor. Zhang and Qiu (2011) report good results when comparing estimates from equation (4) with both dasymetric and street-weighting methods in an empirical case study based in Texas in the United States. The chosen study area, Collin County, located in the Dallas/Forth Worth metroplex, was characterized by particularly rapid urban growth between 2000 and 2007, experiencing a population increase of almost 50%. This surface model is included in the experimental study summarized here to test further the generality of Zhang and Qiu's findings.

Data and methodology

A study area was defined consisting of the unitary authority boundary of the city of Cardiff in South Wales, United Kingdom. This is a predominantly urban region containing a total population in 2001 of 305,353. The U.K. 2001 Census is published with a variety of geographical resolutions, the finest of which is the output area (OA). With a target size of 125 households, or approximately 300 people, this unit is roughly comparable to a U.S. census block. OAs are aggregated to form lower super output areas (LSOAs), creating the next division in the U.K. census hierarchy. With a target population of 1,500 people, these zones roughly compare to a U.S. block group. The study area contains 203 LSOAs with an average population of 1,504 and 991 OAs with an average population of 308. These two nested levels of the U.K. census hierarchy support evaluation of the performance of areal interpolation models using the framework dis-

cussed previously, that is, constructing a model using LSOA data and then testing its performance by comparing estimated OA values with their known census counts.

The boundaries of the OAs used in the 2001 Census were generated automatically via a process described in detail by Martin (2002). In essence, this procedure consisted of an initial tessellation of individual postal address points to create Thiessen polygons, which were subsequently aggregated to achieve a target population count and finally clipped to follow a variety of road, railway, river, and administrative boundary layers to yield the final OA polygons. The same postal address points also are aggregated into UPCs, which are the finest hierarchical unit in a discrete geographic referencing system designed to aid postal delivery. Each UPC typically identifies a set of about 15 houses, and the U.K. Office for National Statistics (ONS) has released a population count for each UPC derived from 2001 Census returns. OS, the U.K. national mapping agency, provides a product called Code-Point® (Ordnance Survey) with polygons, in which the boundary extent of every UPC is mapped. These boundaries are constructed in a similar fashion to those for OAs, with the basic requirement that they surround all address points sharing the same postcode. A postcode can relate to a single building, whereas tower blocks and similar buildings can be assigned multiple postcodes (although Cardiff has relatively few high-rise buildings compared to a typical U.S. city). Such instances are represented in the Code-Point data set as small square polygons identified as "vertical streets." Such cases were excluded from the set of postcode target zones employed in this study because it is unreasonable to expect any of the interpolation models tested to predict population counts for individual buildings. This action resulted in a subset of 5,525 UPC polygons in the Cardiff study area with a reported mean population of 53 persons. The exclusion of these very-high-density population points has some implications in terms of the results presented later because the reported performance of all tested models deteriorates somewhat when these high-density points are included.

UPC polygons are considerably smaller than OAs (an average area of 16,000 m^2 compared to approximately 140,000 m^2) and thus provide an opportunity to evaluate areal interpolation performance for target zones significantly smaller than the United Kingdom's finest census division. Population distribution models constructed with OA-level census data can be used to estimate UPC population and compared to the ONS-supplied counts. Ensuring that the UPC boundaries used in the study were closely dated to the census enumeration date was essential because they are subject to update and modification on a three-month cycle.

Using the preceding data sets, areal interpolation was applied across two spatial resolutions in this study, from LSOA to OA and from OA to UPC. A variety of alternative methods and ancillary data sources were used to perform the interpolations.

In the first experiment, areal weighting was included because it defines the lowest common denominator in terms of methodological sophistication and acts as a useful benchmark against which other techniques may be measured. Intelligent methods are anticipated to outperform this technique. Nevertheless, it remains a useful reference point for assessing the degree of improvement achieved, particularly in relation to the additional data requirements and operational complexities of intelligent methodologies.

Estimates based on binary dasymetric interpolation were generated using two alternative ancillary data sets. First, a binary mask depicting the location of residential land cover was created from a classified Landsat ETM+ image dated July 24, 1999 (Fig. 3c illustrates a subset of these data). Many previous studies used classified satellite imagery to provide intelligence about the probable population placement within source zone boundaries. The open availability of imagery from sites such as the Earth Science Data Interface (Global Land Cover Facility 2011)

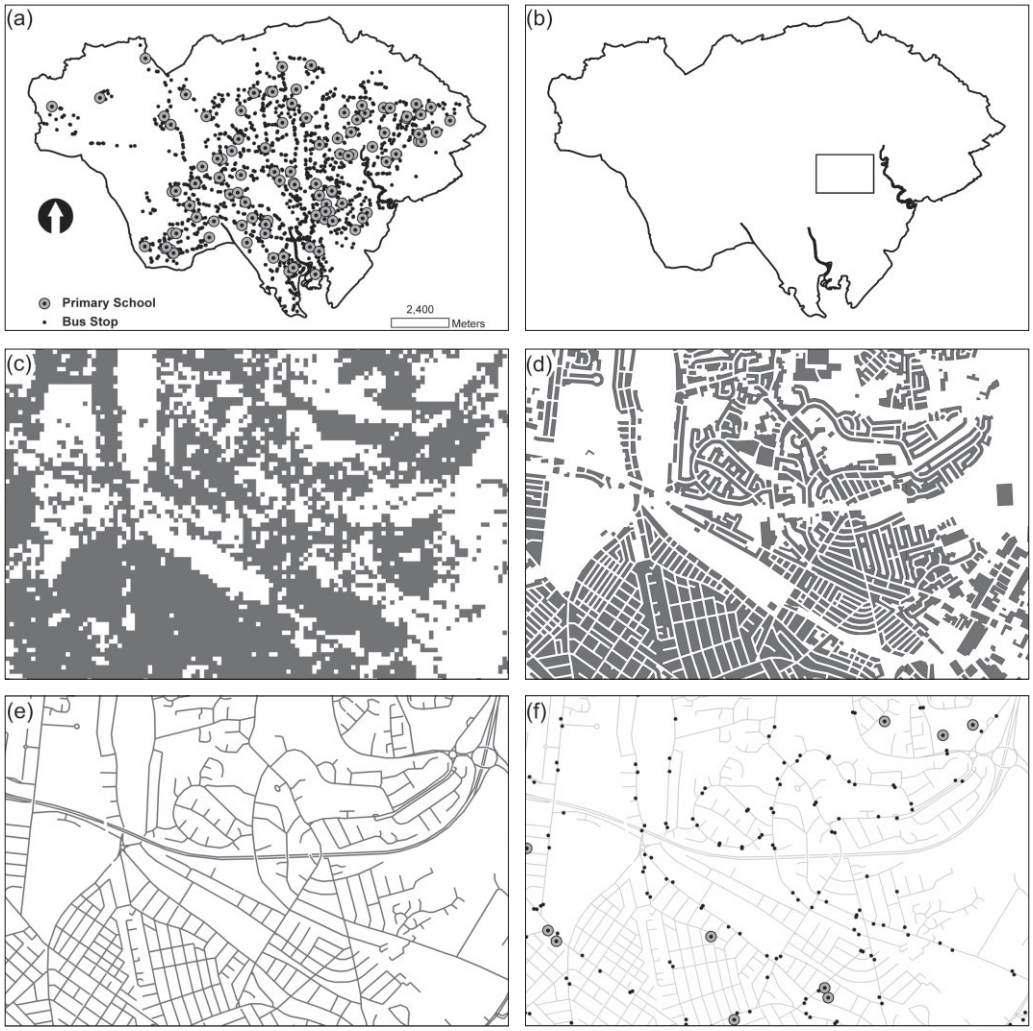

Figure 3. Ancillary data used in the intelligent areal interpolation models: (a) primary school and bus stop points within the study area; (b) location of the detail maps; (c) residential land cover from satellite imagery; (d) building polygons from OS VectorMap District; (e) street vectors from OS VectorMap District; (f) bus stops and primary schools.

ensures this option remains attractive, but issues are associated with this choice. In addition to the need for specialized skills, further potential limitations are imposed by the relatively coarse 30-m spatial resolution of Landsat imagery. Also, some degree of error is inherent in any multispectral classification, and the precise outcome of the classification process varies somewhat from one analyst to another because it is based on unique decisions regard training site selection and the generation of spectral signatures. Finally, in the particular study summarized here, the image predates the census data by two years, raising another potential source of error. The problem of ancillary information being temporally noncoterminous with census data is not uncommon in intelligent areal interpolation, and it affected, to varying degrees, all the data sets used in the study's interpolation experiments.

Binary dasymetric interpolation also was conducted using an open access vector map product, OS VectorMap® District (Ordnance Survey 2011a). This is one of several mapping and geographic information products recently made freely available as part of a U.K. government initiative. The government's aim is to deliver greater access to geographic information in the United Kingdom with the intention of creating new economic and social value. VectorMap® District data are described by OS as "a brand new mid-scale vector dataset specifically designed to display information on the web" (Ordnance Survey 2011b). The term *District* in the name relates to its scale, nominally reported as 1:25,000, and differentiates it from the more detailed (and not open access) VectorMap® Local product. The data set includes a number of layers, such as railway tracks and stations, woodland parcels, and surface water, and is available in both vector and raster formats. It has been downloadable via an unrestricted web-based interface since May 2010.

To deploy this resource in the study, the buildings' polygon layer was utilized for dasymetric mapping (Fig. 3d illustrates a subset of the data). Unfortunately, no attribute field is available to differentiate building type or usage, such as residential versus commercial or industrial; this is perhaps the biggest potential source of error with this ancillary input. Although an analyst with local knowledge may be able to resolve this shortcoming to some extent and then exclude those polygons that are not residential, any such manipulation is always subjective and not easily repeatable, and hence this action was not undertaken in this study. In comparison to the dasymetric mask derived from satellite imagery, general patterns of residential distribution correspond well here, although local knowledge reveals that some buildings associated with commercial and industrial activity are included in this data set but not in the classified ETM+ data. However, OS VectorMap District provides a much cleaner and more precise spatial definition of the building outlines, which may be advantageous when interpolating to small UPC target zones. The exact time stamp of this data set is uncertain, but it is regularly updated by the OS and thus can be assumed to be much more current than the 2001 census data. Areal interpolation using VectorMap District was undertaken in both vector and raster modes, with the latter using a raster grid cell resolution of 5 m.

Interpolation via street weighting also employed VectorMap District data, using the arcs contained within the roads layer (Fig. 3e illustrates a subset of these data). Here, a classification field is available to differentiate road type so some roads (e.g., those labeled as motorway, the equivalent of U.S. freeways) could be excluded because they are unlikely to be directly associated with residential addresses. However, this action was not undertaken in the study so that the source could be evaluated "as presented" in an unmodified state. As noted, the data are more current than the 2001 census data, creating a potential source of error. The original algorithm by Xie (1995) is described in terms of vector processing. Essentially, the same actions may be performed via raster processing without conceptual difference, although some distortion might be introduced due to pixilation effects. The use of a fine-grained 5-m raster grid resolution will limit such problems. Interpolation using both vector and raster modes in the experiments allow differences to be identified and discussed. In raster mode, each source zone population count is uniformly distributed across all designated road cells within its boundary, with these values being summed across target zones to generate population estimates. The "width" of rasterized roads was undifferentiated in the study; both motorways and residential roads, for example, were effectively represented by a single 5-m cell through which the corresponding vector line passes.

As discussed earlier, OA and UPC boundaries have a similar providence in that both are partly derived from an initial tessellation of postal address points subsequently amalgamated and

then potentially clipped to various other geographic features, including roads. As a consequence, the boundaries of both zone sets often, but not always, follow road center lines (although compared to a U.S. census block, an OA is still likely to contain many more internal road vectors). This alignment can give rise to a specific source of error for street-weighted interpolation: the likelihood that all population allocated to a street vector becomes assigned to just one or the other adjacent target polygon when it might be better shared between them, especially because people logically reside in houses offset from the roads. One solution is to apply a spatial buffer to the street network and then to allocate population to these area features (Cromley and McLafferty 2002). Although presented as a modification of street weighting, this action essentially converts the method into a dasymetric approach as described by equation (2). Results for this variant methodology, calculated in vector mode using a 15-m buffer, are included in the following discussion.

The model described by Zhang and Qiu (2011) is adopted here to test a surface-based interpolation algorithm that requires a point data set to provide a proxy for concentrations of population. Retaining a focus on open access data, the study utilizes two alternatives: the sites of primary schools, which are widely available via local government websites, and the location of bus stops, which may be downloaded from the U.K. National Public Transport Access Node database (Department for Transport 2012). Zhang and Qiu suggest that schools are an appropriate choice because they typically are located close to population centers to minimize travel distance for students and to act as community centers. Bus stops also are closely associated with population distribution because they are used predominantly by pedestrians. Their placement often is designed to ensure access to public transport within a 400-m walking distance (O'Sullivan and Morrall 1996; Murray 2001), described by Atash (1994) as the distance an average American will walk rather than drive. A total of 112 schools and 1,618 bus stops are located within the Cardiff study area (Fig. 3a and 3f).

To demonstrate the modeling process, a selected LSOA source zone is displayed along with its internal OA boundaries in Fig. 4a. An indication of the nature of the population densities arising from the various models is shown by distributing the LSOA population using raster cells. Both areal weighting and the surface model distribute population throughout the LSOA and subsequently yield the lowest cell densities. However, areal weighting distributes population uniformly (Fig. 4b), whereas the surface model assigns higher densities to cells lying closest to a control point (i.e., a school or bus stop), with progressively lower values recorded as proximity diminishes (Fig. 4c). The dasymetric model assigns a uniform density to the subset of cells that are deemed to be occupied and zero to all others. The spatially more precise OS VectorMap District data generate higher populated densities (Fig. 4e) than the classified ETM+ data (Fig. 4d). Street weighting distributes population uniformly along the road network. Presented in raster mode, the designated road cells occupy the least space inside the LSOA boundary and generate notably higher density values than any other method (Fig. 4f).

Results and discussion

Performance measures for the various experiments undertaken are summarized in Table 2 for interpolation from LSOA to OA and in Table 3 for interpolation from OA to UPC. Accuracy is measured using the root mean squared error (RMSE) metric described by Fisher and Langford (1995) and used in many succeeding articles (e.g., Eicher and Brewer 2001; Hawley and Moellering 2005; Reibel and Bufalino 2005; Tapp 2010; Zandbergen and Ignizio 2010). RMSE

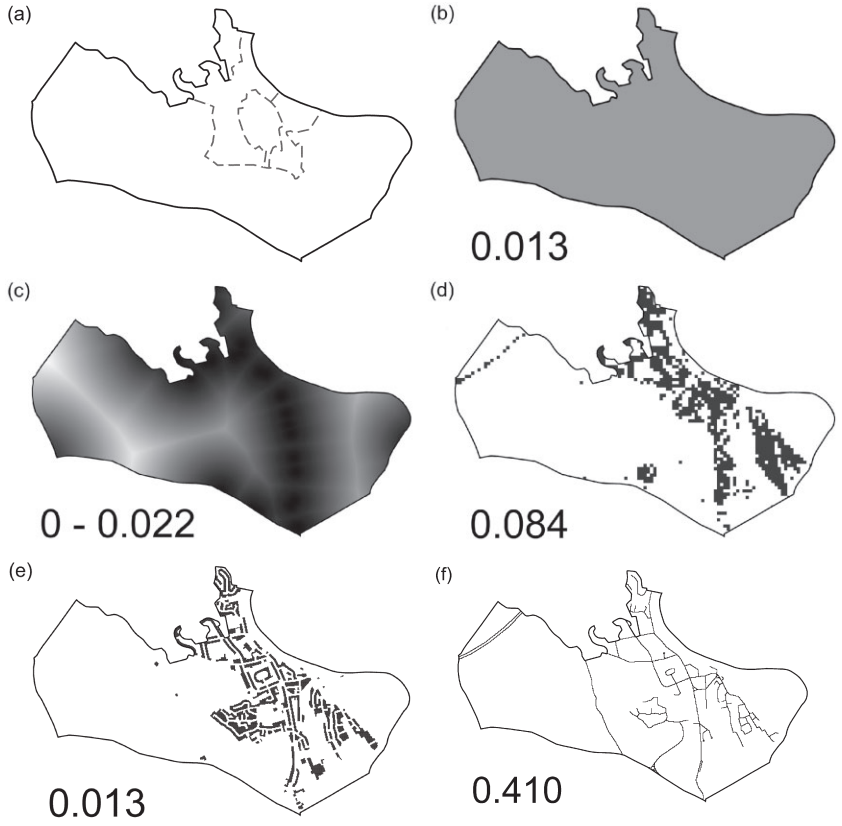

Figure 4. Population distribution modeling in action: (a) a selected LSOA zone and its internal OA boundaries; (b) areal weighting assigns a single low density of 0.013 to every cell; (c) the surface model assigns varying densities, 0–0.022, among all cells, and binary dasymetric models assign a single density but to populated cells only, (d) 0.084 for classified ETM+ data and (e) 0.130 for OS VectorMap District data; (f) street weighting assigns a high density of 0.410 to cells defining the road network.

allows a direct comparison between alternative methods applied to a common set of source and target units, and thus affords a useful indication of the magnitude of errors encountered within each table. RMSE is influenced by the absolute size of estimated values, making it less useful for comparing between different sets of source and target units, particularly where resolution change is involved (e.g., UPC counts are inherently smaller than OA counts, resulting in lower RMSE values from this factor alone). The coefficient of variance (CoV), computed as the RMSE score divided by the average known target zone population, provides a relative error metric that is better suited for cross-resolution comparisons.

Considering first the LSOA to OA results in order of increasing accuracy, as expected, areal weighting performs least well, with an RMSE score of 225 and a CoV of 0.731, setting a benchmark against which the intelligent methods may be judged. Interpolation using the surface model returns only modest levels of improvement. With primary schools used as the ancillary data input, an RMSE value of 220 is achieved, although this improves to 217 when the schools data set is replaced with the bus stops data set. The street-weighting algorithm results rank fourth,

Table 2 Interpolation Results from LSOA to OA Zones

Interpolation method	RMSE	CoV
Areal weighting using zone boundaries only	225.2	0.731
Population density surface using primary schools	220.4	0.715
Population density surface using bus stops	217.5	0.706
Street weighting using OS VectorMap District roads (vector-mode processing)	180.9	0.587
Street weighting using OS VectorMap District roads (raster-mode processing)	178.6	0.580
Street weighting using OS VectorMap District roads with 15-m buffer (vector-mode processing)	170.5	0.553
Binary dasymetric using OS VectorMap District buildings (vector-mode processing)	152.9	0.496
Binary dasymetric using OS VectorMap District buildings (raster-mode processing)	152.9	0.496
Binary dasymetric using classified land cover from Landsat ETM+	144.9	0.471

Note: Mean population of target units is 308.

Table 3 Interpolation Results from OA to UPC Zones

Interpolation method	RMSE	CoV
Street weighting using OS VectorMap District roads (vector-mode processing)	39.3	0.745
Street weighting using OS VectorMap District roads (raster-mode processing)	35.5	0.674
Population density surface using primary schools	35.4	0.672
Population density surface using bus stops	33.8	0.642
Areal weighting using zonal boundaries only	30.3	0.575
Street weighting using OS VectorMap District roads with 15-m buffer (vector-mode processing)	30.3	0.574
Binary dasymetric using classified land cover from Landsat ETM+	23.8	0.451
Binary dasymetric using OS VectorMap District buildings (vector-mode processing)	21.6	0.410
Binary dasymetric using OS VectorMap District buildings (raster-mode processing)	21.6	0.410

Note: Mean population of target units is 53.

with the result generated by raster-mode processing showing a marginal improvement over that generated by a vector-mode implementation; both display a considerable gain in accuracy compared to the surface models. A clue as to the most likely explanation for the benefits of raster processing can be seen in the modified algorithm, which employs buffered roads and returns a better RMSE value (170). First, this outcome offers strong evidence that the issue of roads coinciding with zone boundaries is a factor in the results and further suggests that one of the

effects of raster-mode processing is for discrete raster cells to act somewhat like a small spatial buffer, helping to ameliorate the road alignment problem to some extent. However, the binary dasymetric methods provide the best estimates among the solutions tested at this resolution of interpolation. Using OS VectorMap District as the ancillary data source yields an RMSE score of 153, with no difference whatsoever between raster- and vector-mode implementations. Despite the visual appeal of the VectorMap masks, which appear to offer greater spatial precision in the depiction of building locations compared to the classified ETM+ data set, the latter produces the lowest recorded RMSE value (145). This success can be attributed to the ability of multispectral satellite imagery to discriminate, to a degree, among the probable uses of a built-up area. Within this study area, the housing stock reflects very different roof construction methods and materials from those typically associated with, for example, factories, commercial buildings situated on industrial parks, and shopping centers. Although problems may arise with classification error and a relatively coarse spatial resolution, they appear to be more than compensated for by an ability to differentiate between residential and nonresidential buildings.

At this resolution of interpolation, all intelligent methods provide an improvement over areal weighting, with the best showing a considerable gain in accuracy. Nevertheless, their effectiveness varies substantially across the specific methodologies tested, and even within a given methodology, the choice of ancillary data makes a notable difference.

The OA to UPC results presented in Table 3 have RMSE values that are universally lower, which is expected because target size population is less. CoV values, which are appropriate for making cross-resolution comparisons, show a broadly similar range between the best- and worst-performing models as those for the LSOA to OA results. This finding implies that population distribution models constructed using the finest U.K. census division data are able to interpolate down to target units as small as UPCs with a potentially useful degree of precision, although given the range of outcomes reported here, a careful consideration of the methodology is important. However, very-high-density high-rise housing blocks have been excluded from this study, implying that interpolation performance could deteriorate if these high-density blocks were included.

The most striking feature in Table 3 is that areal weighting is no longer the worst performer. The street-weighting algorithms using unbuffered roads yield RMSE values of 39 and 35 for vector- and raster-mode processing, respectively, compared to a score of 30 for simple areal weighting. The interpolations based on the construction of a surface model also return weak results, with both underperforming the simpler areal weighting. Street weighting using buffered roads matches the areal-weighting performance but is still a disappointing outcome given the extra data and processing involved in its implementation. Despite the overall lower performance of these models, they show patterns that are similar to the LSOA to OA results. Specifically, a raster-mode implementation of street weighting outperforms the vector-mode implementation, both of which are improved by using buffered road features. Furthermore, the use of bus stops as ancillary data in the surface model again provides a modest improvement over the model constructed using primary school locations.

These poor results are unexpected, and in the case of street weighting, the issue of zone boundaries coinciding with road center lines may partly explain the outcome. Because UPCs are specifically constructed around residential address points they often may have buildings contained within them but possibly very few internal road segments, which potentially amplifies the problem. Perhaps the conceptual model begins to fail at this scale of implementation as well. People do not actually reside on the street center lines but rather in buildings that are offset from

such features. Furthermore, population is not uniformly distributed along street segments; if building features are directly modeled to contain people (as in the dasymetric approach), we might more accurately capture the situation where some sections of a street vector have residential housing lying adjacent to them, whereas other sections do not.

The poor performance of the surface models at both resolutions of interpolation also demands some consideration. These results contradict the success reported by Zhang and Qiu (2011), who describe a model deploying school locations as control points in Collin County, Texas, as prevailing over several dasymetric methods and achieving comparable accuracy with a street-weighting approach. This outcome raises the important question of why this method appears to work well in Collin County but not in the city of Cardiff. Ultimately, the answer must lie either in the choice of ancillary data sets used or in the underlying assumptions of this method: that a carefully selected point data set can represent the focus of population clusters in space, and that a mathematical distance decay model is effective in describing the pattern of density distribution around such points. In the Texas study, schools were selected as control points because they are reasonable predictors of household locations (i.e., they are positioned to minimize student travel distances). Presumably, the rapid rate of urban development and the specific nature of planning practices in Collin County lead to a situation where residential population shows a strong propensity to be concentrated around schools, with a clear distance decay effect. Apparently, this is not the case in Cardiff, and one can only speculate as to possible reasons: for example, in a mature city such as Cardiff, the locations of schools may have been established many decades ago, whereas subsequent urban regeneration, new housing developments, and the migration of population from the inner city to suburban locations might lead to a situation where the same spatial association is no longer well preserved. In this study, a model based on school locations actually performs marginally less well than one using bus stops as the ancillary input. Not all bus stops are associated with the collection of passengers from their homes; some are associated with important destinations such as shopping centers and places of employment. Nevertheless, the majority of bus stops are located with the aim of transporting people between home and work. Furthermore, bus stops are a much more dynamic feature than school sites and can be readily introduced or discontinued in response to changing residential patterns. Finally, 1,618 bus stops act as ancillary information inputs into the surface model, compared to only 112 school sites. Together, these factors may help to explain why bus stops prove to be a superior choice of ancillary data in this study.

The inherent information content of the ancillary data sets also could help explain why the dasymetric models produce the best overall outcomes across both resolutions of interpolation. Over 21,800 building polygons are in the OS VectorMap District data set, and 1.68 million pixels are classified as residential in the ETM+ image, all of which contribute to the spatial allocation of population. The information content of the ancillary data used in the surface models seems quite modest in comparison, and this potential shortfall must be made up for by a reliance on the validity of the distance decay model of density distribution around the control points.

Once again, no difference exists in the scores between raster- and vector-mode implementations of binary dasymetric mapping using the OS VectorMap District data; both returned an RMSE value of 22. At this resolution, the map-based ancillary source proves marginally superior to the classified ETM+ input, which returns an RMSE value of 24. When working with target zones as small as these inner-city U.K. postcodes, the need for spatial precision in the population distribution model begins to outweigh any advantages the ETM+ data may hold in terms of their ability to discriminate among residential and nonresidential land cover. The

satellite image used here is broadly concurrent with the census enumeration data, while the OS VectorMap District considerably postdates it. If this discrepancy could be eliminated at a future date (e.g., by using the U.K. 2011 Census statistics), a still greater margin of superiority might manifest itself.

Conclusions

This article shows that the performance of intelligent areal interpolation can be significantly influenced by the qualities of the specific ancillary data used to drive a population distribution model and that the choice of ancillary data and of methodology requires careful consideration. The underlying rationale and theoretical assumptions on which any particular technique is based must be carefully considered because models that perform well in one environment may prove to be much less successful in another. A surface-based model using school locations as control points performed well in a previous empirical study based in Texas but produces relatively poor results in Cardiff despite the use of similar information. The street-weighting algorithm also has been shown to perform well in previous studies (e.g., Hawley and Moellering 2005), yet in this study, it was unable to match the strength of binary dasymetric interpolation using OS VectorMap District data despite the use of information from essentially the same ancillary resource. Land cover information derived from Landsat ETM+ imagery appears to remain a strong contender for driving intelligent interpolation models, although the broadly comparable performance of open access map data depicting building polygons now makes this a preferred option, at least in the U.K. context, given its simpler demands in terms of data preparation and its ability to be applied in either vector- or raster-modes of operation. Frequent discussion occurs concerning the relative merits, differences in theoretical principles, and validity of underlying assumptions among competing intelligent interpolation techniques; but even though methodology remains important, the significance of ancillary data inputs for the ultimate performance of intelligent areal interpolation methods should not be overlooked.

Once again, paraphrasing Zandbergen and Ignizio's assertion (2010), all areal interpolation methods have their errors, and their performance will inevitably vary with specific conditions. It might be added that performance also can vary according to the choices made in terms of ancillary data. Results presented here relate only to one specific study conducted in a predominantly urban environment. Although this geographical context includes natural green spaces such as parks and school grounds, and exhibits the typical patterns of residential and commercial/industrial zoning seen in most European cities, sparsely populated rural areas generally offer a more challenging environment (Tapp 2010) in which the relative merits of specific intelligent areal interpolation techniques may change substantially. Results summarized in this article also relate only to interpolation across resolution and not to interpolation between alternative geographies at the same resolution. Interpolation performance between incongruent spatial units of broadly similar size remains an important issue, and such testing could form the basis of further research.

This article contributes to a growing literature about comparative testing of alternative areal interpolation algorithms for population estimation. This activity remains necessary because conflicting results appear, and no one methodology has yet proven itself to outperform all others universally. Thus, only by increasing the evidence base through replications using as wide a range of geographical environments and circumstances as possible can researchers better understand the conditions under which any particular methodology might be expected to perform.

The increasing availability of open access data sets offers considerable potential for widening the adoption of intelligent population interpolation tools. In addition to the inherent attractiveness of a no-cost option, many such data sets are available as points, lines, and polygons stored in vector format, eliminating the need for multispectral image-processing skills associated with the use of satellite imagery or for familiarity with raster GIS-processing techniques. Perhaps the most convincing argument for encouraging the widespread use of open access data sets in population interpolation tasks is the accuracy achieved in the experiments conducted here. Dasymetric mapping with building polygons obtained from an open access product yields consistently good results across different resolutions and performs particularly well when estimating counts for highly challenging small area urban estimates.

Wider adoption of intelligent areal interpolation could be further encouraged if methodologies such as those addressed in this article are implementable through the use of simple plug-in tools for the most commonly used GIS software packages. Although the computations and algorithmic steps needed to implement dasymetric mapping, street weighting, or surface-based interpolations are not particularly complex, they still discourage ready usage in many circumstances. As geographical data and GIS capabilities continue to migrate from specialist software and skilled professionals to the public domain and the nonexpert user via web-based interfaces, the ability to use these techniques through simple point-and-click and wizard-style interfaces is increasingly necessary to prevent them from becoming confined to the experimental research domain. A final advantage of such easy accessibility would undoubtedly be the generation of comparative results from a diverse range of case studies, which in turn would provide a much broader base of evidence to support any claims of methodological or ancillary data source superiority.

References

Atash, F. (1994). "Redesigning Suburbia for Walking and Transit: Emerging Concepts." *Journal of Urban Planning and Development* 120, 48–57.

Bracken, I., and D. Martin. (1989). "The Generation of Spatial Population Distributions from Census Centroid Data." *Environment and Planning A* 21, 537–43.

Brinegar, S. J., and S. J. Popick. (2010). "A Comparative Analysis of Small Area Population Estimation Methods." *Cartography and Geographic Information Science* 37, 273–84.

Cai, Q., G. Rushton, B. Bhaduri, E. Bright, and P. Coleman. (2006). "Estimating Small-area Populations by Age and Sex Using Spatial Interpolation and Statistical Inference Methods." *Transactions in GIS* 10, 577–98.

Cromley, E. K., and S. L. McLafferty. (2002). *GIS and Public Health*. New York: Guilford.

Cromley, R., D. Hanink, and G. Bentley. (2012). "A Quantile Regression Approach to Areal Interpolation." *Annals of the Association of American Geographers* 102, 763–77.

Department for Transport. (2012). "NaPTAN: The National Public Transport Access Node Database." Available at http://www.dft.gov.uk/naptan/ (accessed 15 December 2011).

Eicher, C., and C. Brewer. (2001). "Dasymetric Mapping and Areal Interpolation: Implementation and Evaluation." *Cartography and Geographic Information Science* 28, 125–38.

Fisher, P. F., and M. Langford. (1995). "Modeling the Errors in Areal Interpolation between Zonal Systems by Monte Carlo Simulation." *Environment and Planning A* 27, 211–24.

Flowerdew, R., and M. Green. (1989). "Statistical Methods for Inference between Incompatible Zonal Systems." In *Handling Geographical Information: Methodology and Potential Applications*, 239–47, edited by M. Goodchild and S. Gopal. New York: Longman.

Flowerdew, R., and M. Green. (1992). "Developments in Areal Interpolation Methods and GIS." *Annals of Regional Science* 26, 67–78.

Flowerdew, R., and M. Green. (1994). "Areal Interpolation and Types of Data." In *Spatial Analysis and GIS*, 121–45, edited by S. Fotheringham and P. Rogerson. London: Taylor and Francis.

Global Land Cover Facility. (2011). "Earth Science Data Interface." Available at http://glcfapp.glcf.umd.edu:8080/esdi/index.jsp (accessed 15 December 2011).

Goodchild, M. F., and N. S. Lam. (1980). "Areal Interpolation: A Variant of the Traditional Spatial Problem." *Geo-Processing* 1, 297–312.

Goodchild, M. F., L. Anselin, and U. Deichmann. (1993). "A Framework for the Areal Interpolation of Socioeconomic Data." *Environment and Planning A* 25, 383–97.

Harris, R. J., and P. A. Longley. (2000). "New Data and Approaches for Urban Analysis: Modelling Residential Densities." *Transactions in GIS* 4, 217–34.

Harvey, J. (2000). "Small Area Population Estimation Using Satellite Imagery." *Statistics in Transition* 4, 611–33.

Hawley, K., and H. Moellering. (2005). "A Comparative Analysis of Areal Interpolation Methods." *Cartography and Geographic Information Science* 32, 411–23.

Holt, J., C. P. Lo, and T. Hodler. (2004). "Dasymetric Estimation of Population Density and Areal Interpolation of Census Data." *Cartography and Geographic Information Science* 31, 103–21.

Lam, N. S. (1983). "Spatial Interpolation Methods: A Review." *American Cartographer* 10, 129–49.

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, 161–80.

Langford, M. (2007). "Rapid Facilitation of Dasymetric-based Population Interpolation by Means of Raster Pixel Maps." *Computers, Environment and Urban Systems* 31, 19–32.

Lo, C. P. (2008). "Population Estimation Using Geographically Weighted Regression." *GIScience and Remote Sensing* 45, 131–48.

Maantay, J. A., A. R. Maroko, and X. Herrmann. (2007). "Mapping Population Density in the Urban Environment: The Cadastral-based Expert Dasymetric System (CEDS)." *Cartography and Geographic Information Science* 34, 77–102.

Martin, D. (2002). "Geography for the 2001 Census in England and Wales." *Population Trends* 108, 7–15.

Mennis, J. (2003). "Generating Surface Models of Population Using Dasymetric Mapping." *The Professional Geographer* 55, 31–42.

Mennis, J., and T. Hultgren. (2006). "Intelligent Dasymetric Mapping and its Application to Areal Interpolation." *Cartography and Geographic Information Science* 33, 179–94.

Merwin, D., R. Cromley, and D. Civco. (2009). "A Neural Network-based Method for Solving 'Nested Hierarchy' Areal Interpolation Problems." *Cartography and Geographic Information Science* 36, 347–65.

Moon, Z. K., and F. I. Farmer. (2001). "Population Density Surface: A New Approach to an Old Problem." *Society and Natural Resources* 14, 39–49.

Murray, A. (2001). "Strategic Analysis of Public Transport Coverage." *Socio-Economic Planning Sciences* 35, 175–88.

O'Sullivan, S., and J. Morrall. (1996). "Walking Distances to and from Light-Rail Transit Stations." *Transportation Research Record* 1538, 19–26.

OpenStreetMap. (2012). "OpenStreetMap: The Free Wiki World Map." Available at http://www.openstreetmap.org (accessed 15 September 2012).

Ordnance Survey. (2011a). "OS VectorMap District." Available at http://www.ordnancesurvey.co.uk/oswebsite/products/os-vectormap-district/index.html (accessed 15 December 2011).

Ordnance Survey. (2011b). "Ordnance Survey Launches OS OpenData in Groundbreaking National Initiative." Available at http://www.ordnancesurvey.co.uk/oswebsite/media/news/2010/April/OpenData.html (accessed 15 December 2011).

Petrov, A. N. (2008). "Setting the Record Straight: On the Russian Origins of Dasymetric Mapping." *Cartographica* 43, 133–36.

Reibel, M., and A. Agrawal. (2007). "Areal Interpolation of Population Counts Using Pre-classified Land Cover Data." *Population Research and Policy Review* 26, 619–33.

Reibel, M., and M. E. Bufalino. (2005). "Street Weighted Interpolation Techniques of Demographic Count Estimation in Incompatible Zone Systems." *Environment and Planning A* 37, 127–29.

Sadahiro, Y. (2000a). "Accuracy of Count Data Transferred through the Areal Weighting Interpolation Method." *International Journal of Geographical Information Science* 14, 25–50.
Sadahiro, Y. (2000b). "Accuracy of Count Data Estimated by the Point-in-Polygon Method." *Geographical Analysis* 32, 64–89.
Su, M., M. Lin, H. Hsieh, B. Tsai, and C. Lin. (2010). "Multi-Layer Multi-class Dasymetric Mapping to Estimate Population Distribution." *Science of the Total Environment* 408, 4807–16.
Tapp, A. (2010). "Areal Interpolation and Dasymetric Mapping Methods Using Local Ancillary Data Sources." *Cartography and Geographic Information Science* 37, 215–28.
Tobler, W. (1979). "Smooth Pycnophylactic Interpolation for Geographical Regions." *Journal of the American Statistical Association* 74, 518–36.
U.S. Census Bureau. (1993). *TIGER/Line 1992 Files*. Washington, DC: Department of Commerce, U.S. Census Bureau, Geography Division.
Wright, J. (1936). "A Method of Mapping Densities of Population: With Cape Cod as An Example." *Geographical Review* 26, 103–10.
Xie, Y. (1995). "The Overlaid Network Algorithms for the Areal Interpolation Problem." *Computers, Environment and Urban Systems* 19, 287–306.
Yuan, Y., R. Smith, and W. Limp. (1997). "Remodelling Census Population with Spatial Information from Landsat TM Imagery." *Computers Environment and Urban Systems* 21, 245–58.
Zandbergen, P., and D. Ignizio. (2010). "Comparison of Dasymetric Mapping Techniques for Small-area Population Estimates." *Cartography and Geographic Information Science* 37, 199–214.
Zhang, C., and F. Qiu. (2011). "A Point-based Intelligent Approach to Areal Interpolation." *The Professional Geographer* 63, 262–76.

View publication stats