Computers, Environment and Urban Systems 80 (2020) 101440

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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

Using household counts as ancillary information for areal interpolation of population: Comparing formal and informal, online data sources

Wen Zeng[a,b,*], Alexis Comber[b]

[a] Shandong University of Science and Technology, Qingdao, China
[b] University of Leeds, Leeds, United Kingdom

ARTICLE INFO

Keywords:
Areal interpolation
Household data
Household proportion method
Population estimation

ABSTRACT

Fine-scale population estimates are needed to support both public and private planning. Previous areal interpolation research has used various types and sources of data as ancillary information to guide and constrain the disaggregation from (usually) larger source zones to (usually) smaller target zones. Many new forms of open and free to access geo-located data are available, and as yet little research has evaluated the use of these data in areal interpolation. This study evaluates the effectiveness of household data as ancillary information from two sources: formal census household counts and informal data on residential (house) sales from commercial websites, applied to 2 case studies with different contexts - Leeds in UK and Qingdao in China. The proposed Household Proportion method uses household counts as ancillary information for areal interpolation of population. It is compared with other interpolation and the results show that HP method yields significantly better results than other interpolation approaches using ancillary data, with lower errors. This research also demonstrates that such data support the application of a suite of interpolation methods that make fewer assumptions about underlying spatial processes. The need to examine issues of representativeness and data coverage are identified and discussed, but the study demonstrates the opportunities for including freely available geo-located data to inform geographic analyses.

1. Introduction

Measures of populations over small areas are essential for a wide variety of public planning and commercial activities. Small areas include fine-scale geographical census units such as Output Areas (Martin, 1998) in the UK. The size and distribution of the population are key inputs for socioeconomic and planning studies, such as facility location-allocation analyses (Comber, Dickie, Jarvis, Phillips, & Tansey, 2015; Maliszewski, Kuby, & Horner, 2012), health care planning (Comber, Sasaki, Suzuki, & Brunsdon, 2011), disaster management (De Albuquerque, Herfort, Brenning, & Zipf, 2015) and analyses of environment inequality (Boyce, Zwickl, & Ash, 2016). Population information is also vital for the private sector to determine trade areas, evaluate retail trading performance, select potential business locations, predict retail sales, assess market shares, and so on (Church & Murray, 2009; Deng & Wu, 2013).

Areal interpolation is a commonly used method for estimating small area populations. It transforms data from *source* zones of known values to *target* zones with unknown ones (Goodchild & Lam, 1980). It is the process of re-distributing data reported over one set of geographic framework to another. Often this used to transform data from coarse, high-level geographic areas to finer-scale ones, or for areas with similar scales, but whose boundaries have changed. There are two reasons why areal interpolation methods are needed. First, many research, policy and commercial activities require fine resolution population information. However, to protect privacy, the population statistics provided by many national census agencies are not available at fine scales (Langford, 2013; Sridharan & Qiu, 2013), especially in countries with emerging economies such as China (Yang, Jiang, Luo, & Zheng, 2012). Second, area boundaries often change over time (e.g. Evans, 1996) resulting in mismatches between data from different censuses or captured at different times and causing problems for data integration and analysis. Geographic data boundary mismatches is a persistent problem in geography, planning, regional science, landscape ecology, and other fields (Zandbergen & Ignizio, 2010).

Researchers have continued to develop new procedures and strategies for improving areal interpolation estimates (Qiu & Cromley, 2013). These efforts have focused on improving analytical methods and on the ancillary information used to guide the allocation of data collected over source zones to target zones (Bakillah, Liang, Mobasheri, Jokar

* Corresponding author at: 579 Qianwangang Road, Huangdao District, Qingdao 266590, China.
E-mail address: Alvin_z@163.com (W. Zeng).

https://doi.org/10.1016/j.compenvurbsys.2019.101440
Received 25 June 2019; Received in revised form 3 November 2019; Accepted 6 November 2019
Available online 14 November 2019

Arsanjani, & Zipf, 2014; Haklay, 2010; Langford, 2013; Lin & Cromley, 2015; Yao et al., 2017). The number and variety of ancillary information used in areal interpolation have increased as new and different technologies have become available (Lin & Cromley, 2015). However, although household information is recognized as being related to population, and household-based methods have been used for estimating small populations (Deng & Wu, 2013; Smith & Cody, 2004; Smith, Nogle, & Cody, 2002), household data has seldom been used in areal interpolation, and few studies have evaluated its use in this context. This is despite an increased amount of household-related data available from websites that are open to the public, such as those for residential property sales and rentals. Research is needed to evaluate the effectiveness of such data sources in areal interpolation.

The purpose of this article is threefold. First, it aims to make a contribution to the literature on the use of household data as ancillary information in areal interpolation. In this study, *formal* household counts from the population census and *informal* property counts from websites listing the residential property for sale or rent were used as ancillary information to constrain areal interpolation, and the results were compared. As is shown below (in Section 3.1), household and property counts have a strong correlation and property counts can be used to represent household counts. Second, property sales data have seldom been used in previous studies, and the spatial inferential opportunities arising from these kinds of open big data have yet to be determined. This study used two case studies: a UK city and a Chinese city. In the UK case study, the presence of observed population counts at finer geographical scales (i.e. target zones) from the population census, allowed the interpolated populations to be validated. The aim was to identify suitable interpolation approaches and ancillary information for application to the China case study, where detailed census data are not publicly available with the aim of supporting location-allocation approaches.

The paper is organized as follows. Section 2 provides an overview of the literature on areal interpolation. In Section 3 the study area, data and methods are described. Section 4 presents the results, including the relationships between population and houses and properties under different areal interpolation methods. Section 5 discusses the results and highlights a number of issues related to the findings and the approach. Section 6 presents some conclusions.

2. Background: areal interpolation

Spatial interpolation is a widely applied method in geographical research. It is used to disaggregate spatial data of many different phenomena, processes and measurements such as population, hydrology, atmosphere, topography, agriculture, soil, land use, rainfall and temperature (Comber, Proctor, & Anthony, 2008; Goovaerts, 2000; Jia & Gaughan, 2016; Joseph, Sharif, Sunil, & Alamgir, 2013; Liao, Li, & Zhang, 2018; Mennis, 2003; Rigol, Jarvis, & Stuart, 2001; Shi & Tian, 2006). It uses values from known geographical locations to estimate (or predict) values at other unknown ones. There are many different approaches to spatial interpolation and methods can be broadly divided into the point and areal interpolation (Lam, 1983). Areal approaches can be further divided into methods that do not make use of ancillary data and those that do (Hawley & Moellering, 2005; Langford, 2006; Zhang & Qiu, 2011). A recent methodological advance is the use of the many new forms of data as ancillary information. Approaches to areal interpolation are reviewed with coded examples in Comber and Zeng (2019).

2.1. Methods without ancillary information

Areal weighting is an interpolation approach applied in situations where only source zone data are known. It interpolates source zone values proportionately from the areas of intersection between source and target zones (Goodchild & Lam, 1980; Lam, 1983). It can be implemented using polygon overlay operations in most GIS software packages (Xie, 1995) and is widely used if ancillary information is unavailable (Goplerud, 2016; Langford, 2006; Logan, Xu, & Stults, 2014). The disadvantage is that it assumes spatial homogeneity of the variable of interest within each source zone (Goodchild & Lam, 1980).

Pycnophylactic interpolation (Tobler, 1979) generates a smooth surface in the target zones, whilst preserving the overall mass or volume of the source zones counts. After an initial aggregation to target zones, it iteratively adjusts target values using the weighted average of nearest neighbours. The result is a smooth surface of target zone values (i.e. without discontinuities) under the assumption that no sharp boundaries exist in the distribution of the data (Hay, Noor, Nelson, & Tatem, 2005).

Point-based areal interpolation (Bracken & Martin, 1989; Martin, 1989) identifies a control point for each source zone (usually its centroid) which is assigned a density value. These values are interpolated to a regular grid using one of a number of methods (Martin, 1989) such as inverse distance weighting, kriging, etc., before being rescaled and combined to generate estimates over the target zones. Lam (1983) noted that the choice of the control point is critical as it has a significant impact on the resulting surface, particularly as the geometric centroid may be outside the polygon boundary (Tapp, 2010; Xie, 1995), or may only poorly represent the actual distribution of the feature as compared to, for example, a population-weighted centroid (Martin, 1989).

2.2. Methods using ancillary information

The processes of disaggregation can be spatially constrained to include or exclude certain within-target zone areas. For example, population distributions might be expected to be closely related to residential land use and housing density (Cromley, Hanink, & Bentley, 2012). Ancillary data of these features can be used to constrain population interpolation (Liu, Kyriakidis, & Goodchild, 2008) and such approaches are informed by an increasing variety of data (Langford, 2013). These methods have been extensively applied to population data (Cromley et al., 2012; Langford, 2007; Mennis, 2003; Reibel & Agrawal, 2007), socioeconomic variables (Eicher & Brewer, 2001; Goodchild, Anselin, & Deichmann, 1993; Mennis & Hultgren, 2006), agricultural census data (Comber et al., 2008) and time-variant population analysis with changing historical administrative boundaries (Gregory, 2002; Mennis, 2016).

Dasymetric interpolation is the most commonly applied method that uses ancillary information (Langford, 2013), although others exist such as street-weighting (Reibel & Bufalino, 2005; Xie, 1995). The approach is to create masks of areas to be included or excluded from the interpolation process and to guide the redistribution of values to target zones using the masks as spatial controls and constraints. The most commonly used auxiliary information is remotely sensed data as this can easily be used to create masks related to land use (Eicher & Brewer, 2001; Fisher & Langford, 1996; Langford, 2006; Mennis, 2003; Mennis & Hultgren, 2006). The allocation can be binary (e.g. Fisher & Langford, 1996) distinguishing only populated and unpopulated areas within the target zones, or categorical such that different population proportions are allocated to different land use classes (Eicher & Brewer, 2001; Langford, 2006; Mennis, 2003; Mennis & Hultgren, 2006). Dasymetric approaches provide a more spatially informed interpolation but land use data may not be available especially in developing countries (Yang et al., 2012) and an understanding of remote sensing techniques may be required that is outside of many GIS analyst skill sets (Langford, 2013). They also assume feature density (e.g. population) to be homogeneous within any given binary mask or land use class.

A second tranche of approaches uses point information to constrain areal interpolation (Zhang & Qiu, 2011). Tapp (2010) used US county address points as ancillary information to predict population and Harris and Chen (2005) similarly used postcode locations to estimate population density in the UK. The locations of schools, businesses, supermarkets, housing, and service sites are widely available and have been used to interpolate a variety of variables, as well as population. In this, the spatial distribution (proportion) of the point features in the target zones are used to proportionately allocate the source zone value.

A related approach uses target zone household counts to constrain population interpolation. This method is commonly applied and has been found to be one of the most accurate and cost-effective methods for small-area population estimation in demographic research (Smith & Cody, 2004). This approach first estimates household counts or proportions in target zones and then infers population from these based on the proportion of the source zone households in each target zone. Essentially, this assumes an average number of persons per household but effectively accounts for populations residing in groups or collective housing (Smith et al., 2002; Smith & Cody, 2004). Estimates of household counts and their relative fraction in each target zone, can be derived from a number of data sources such as building permits, certificates of occupancy, utilities customer databases (electricity, telephone etc), property tax records, and aerial photographs (Deng & Wu, 2013; Liu et al., 2008).

2.3. Methods using new forms of data

Traditionally, areal interpolation approaches with ancillary data have used data from formal sources, such as national mapping agencies and governments. Much of this high-quality data is now openly available and free to access in many countries (Langford, 2013). However, many new forms of ancillary data are also now available and are potentially able to provide information on the spatial patterns of socio-economic activity. These include mapped data on public-facing websites, different types of volunteered geographic information (VGI) and social media data. Some of these have been examined in the context of areal interpolation. Bakillah et al. (2014) used OpenStreetMap building and point of interest (PoI) data to interpolate population data to the building level and Kunze and Hecht (2015) used similar data to improve building-level population estimations by quantifying the extent of non-residential land. The use of spatial data generated by social networks as alternative ancillary data has also been explored (Kounadi, Ristea, Leitner, & Langford, 2018). Lin and Cromley (2015) evaluated the effectiveness of geo-located night-time tweets and found the approach to be less effective than traditional methods except for those age groups with a high percentage of Twitter users. Other approaches have used different forms of auxiliary information such as taxation data (Jia & Gaughan, 2016; Kar & Hodgson, 2012), social media data (Yu, Li, Zhu, & Plaza, 2018), mobile phone data (Liu, Peng, Wu, Jiao, & Yu, 2018) and point-of-interest data (Ye et al., 2019) as ancillary inputs to interpolation approaches. These suggest that non-traditional data may be useful and that there are opportunities to exploit new data sources. It is in this context that this research explored the utility of residential property sales data to inform areal interpolation of population.

3. Methods

This paper evaluated different areal interpolation approaches to determine the degree to which household data can be used as ancillary information in areal interpolation to estimate population. Here household count, a very common attribute in census data, was evaluated against residential property sales and rental data, collected from commercial websites, as ancillary information. The higher-level study objective was to determine the reliability of a simple but conceptually elegant areal interpolation approach using proxies for household counts, and thus population, from non-traditional sources.

3.1. Study areas and data

This study used Leeds in the UK and Qingdao in China as case studies. Leeds is the UK's third-largest Metropolitan District with a population of 751,500 people according to the 2011 Census. Its covers some 552 km^2 with a built-up area to the centre and south surrounded by a number of separate small towns and villages in a polycentric pattern (Meegan, 2015). Qingdao is a typical large city in the east coast of China, with a population of 3,779,000 in the central city area according to the 2010 Census. The area is about 1407 km^2 (Fig. 1).

Data from a number of sources were used. The UK population data was from the 2011 UK Census, reported over 3 scales: Middle Layer Super Output Area (MSOA, ~7000 people), Lower Layer Super Output Area (LSOA, ~1500 people) and Output Area (OA, ~300 people). It included the population and household counts for each census area. The population data for China was from the 2010 Census and included population and household count at District (~600,000 people) and Subdistrict levels (~60,000 people). Although residential property information is more closely associated with house (dwelling) count, such data is not available in the Chinese census. The correlation coefficients between dwelling count and household count (i.e. people living their daily lives together) across the Leeds MSOAs, LSOAs and OAs are all above 0.98 (p-value $< .000$) suggesting the appropriateness of using household count instead of dwelling count in this paper.

The property data for the UK was scraped from Zoopla, a residential property sales website (https://www.zoopla.co.uk/) in January 2019. This data describes the properties available for sale or rent in Leeds. Property data for Qingdao was scraped from Lianjia (https://qd.lianjia.com/) in June 2018, which is a large property agency in China. This lists properties available for sale or rent in Qingdao but does not include cheaper, affordable housing or properties in urban villages. The latitude and longitude of each property were included in their listing on both websites. There are large differences in data volumes, with >351,358 records for Leeds and 2849 records for Qingdao. This suggests the potential for over-estimation in Leeds and under-estimation (properties not listed on the website) in Qingdao. However, as the methods effectively use the proportions of source zone properties to allocate target zone population estimates, it is the relative rather than the absolute distribution of properties that is important.

A number of ancillary datasets were used to support different interpolation approaches. Land use data for dasymetric approaches was extracted from OS Open Map - Local (Ordnance Survey) for the UK and was generated from a Landsat 8 satellite image for China using an object-oriented classification with an overall accuracy of 0.927. In the UK data, *building* land types (i.e. structures with a roof) were used to indicate the places where people live, although this includes different kinds of building structures. The China land use data included a similar building class. The road network data for Leeds was from OS Open Roads and from Baidu (https://map.baidu.com/) for the Chinese case study. UK census area boundary data was from the UK Data Service and Chinese boundaries were obtained from the National Geomatics Centre of China (NGCC). The UK data were projected to British National Grid coordinates (EPSG: 27700) and Chinese data to WGS84 (EPSG: 4326). The data are listed in Appendix 1.

3.2. Analysis

Initial analyses examined the relationships between population, household counts as recorded in the census and property counts over different census reporting areas to confirm their correlation. Although a correlation between population and household/properties counts was expected, it was important to establish this relationship.

Then, four different areal interpolation methods were implemented and compared:

1. Areal Weighting allocates source zone counts to target zones based on the proportion of the source zone area intersected by each target zone. This is the simplest areal interpolation method without ancillary information (Fisher & Langford, 1996).
2. Dasymetric Interpolation is the most commonly used method. The binary dasymetric approach (Langford, 2007; Mennis, 2003;

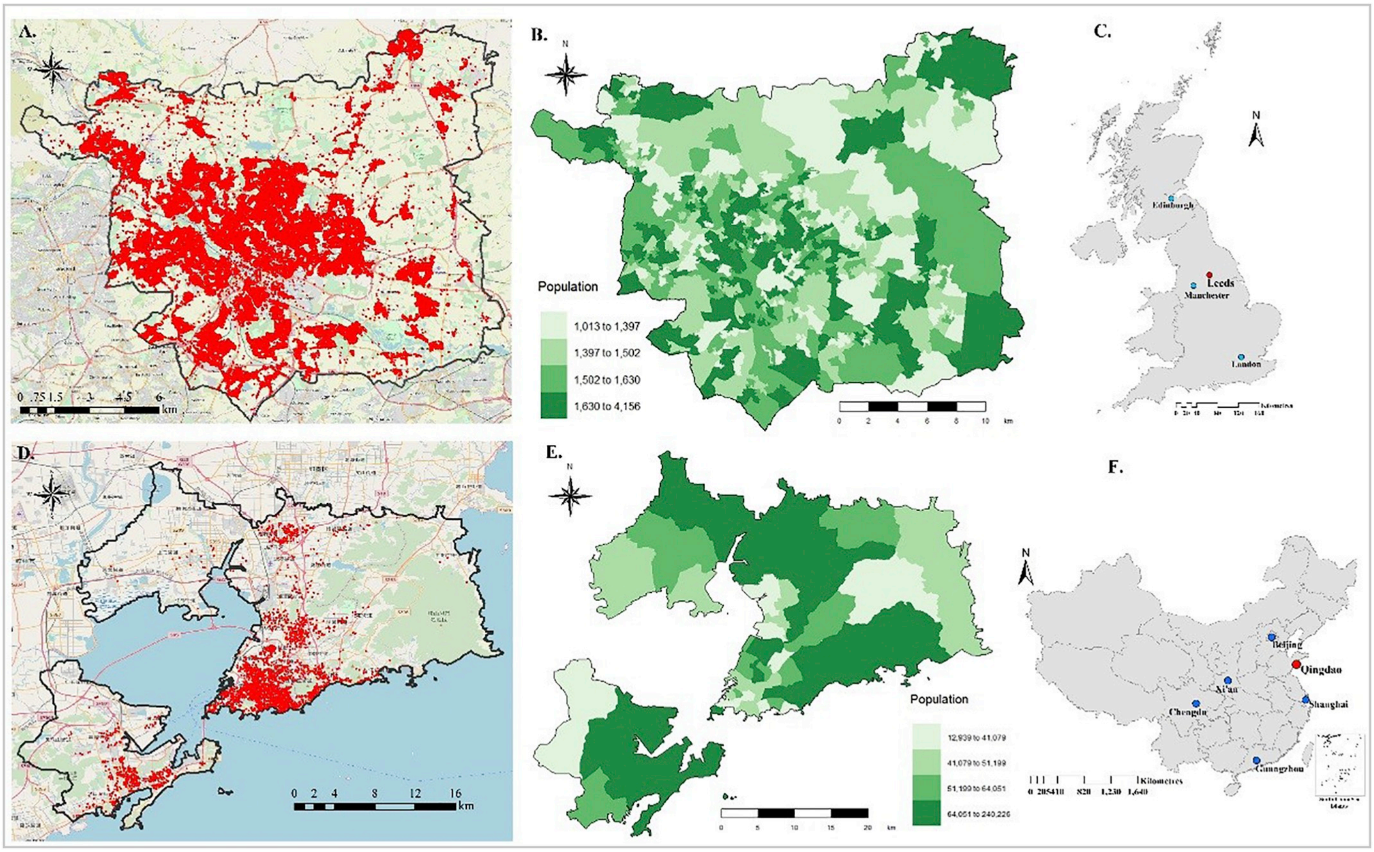

Fig. 1. Study areas and population distribution (A: Property sales point data in Leeds, UK; B: LSOA population distribution in Leeds; C: The location of Leeds in the UK; D: Property sales point data in Qingdao, China; E: Subdistrict population in Qingdao; F: The location of Qingdao in China).

Zandbergen & Ignizio, 2010) was used. In both case studies, the building area in the target zones was used to weight the allocation of population. The dasymetric masks are shown in Fig. 2. For the Leeds case study, the data contains fine-scaled building information, while the data for Qingdao is coarser, but the census areas parcels are larger (Fig. 2).

3. The Road Network method allocates population based on the proportion of the total length of source zone roads in each target zone. It assumes a homogeneous distribution of population along the road network (Reibel & Bufalino, 2005; Xie, 1995).
4. The Household Proportion method uses the proportions of source zone households or properties in each target zone, as determined through the spatial overlay. Counts of the household/property number in each source zone were generated, using a point in polygon operation. Then, the proportion of the total in each source zone was determined for each target zone. These proportions were used to weight the allocation of source zone populations. There are two variants of this method. One uses census household count as ancillary information (HP-census) and the other uses property counts from the house sales data (HP-sales).

The interpolation methods and their inputs parameters are summarised in Table 1. The basic operation of all four methods is some kind of spatial intersection of the source and target zones, with or without some kind of ancillary information. The logic and operation of these methods are similar, but with differences in the ancillary information used, if any. Thus, the performances of the different methods are mainly as a result of differences in the ancillary information. Based on this assumption, it is possible to examine the effectiveness of data from property websites as ancillary information relative to a similar but "authoritative" household data from populations censuses.

4. Results

4.1. Relationships between population and household

A correlation analysis of population and household counts was undertaken to establish their relationship over different scales for the 2 case study areas and the 2 sources of household information. The results are shown in Table 2 and broadly indicate strong, positive and significant correlations between total household and population counts. For the UK case study, the correlations are significant at each scale and similar for census household counts and the commercial properties data. In both cases the relationship weaken as the areal unit gets finer. This is likely to be because of local variations in household size, which varies more strongly at increasingly finer scales: the heterogeneity of household size is much greater for OAs than LSOAs or MSOAs, for example. As the geographical detail increases, the coefficient of variation of average household size and the local variation increases. The correlations at LSOA and OA scales in the UK suggest that house sales data may be a reasonable predictor of the population at smaller scales. For the China case study, the correlations are higher and more significant at Sub-District levels than District levels. This may be due to the large areas and the low number of Districts ($n = 6$). The correlation coefficients for households estimated from property data are lower than those derived from census households.

Table 2 suggests household data from the census or properties sales data can be used equally as ancillary information to support population estimation. The results also suggest similar relationships between household, population and scale in the UK and China, that in the UK this can be done at MSOA scales or above and that in China at Sub-District level with positive and significant correlations of (0.7–0.8).

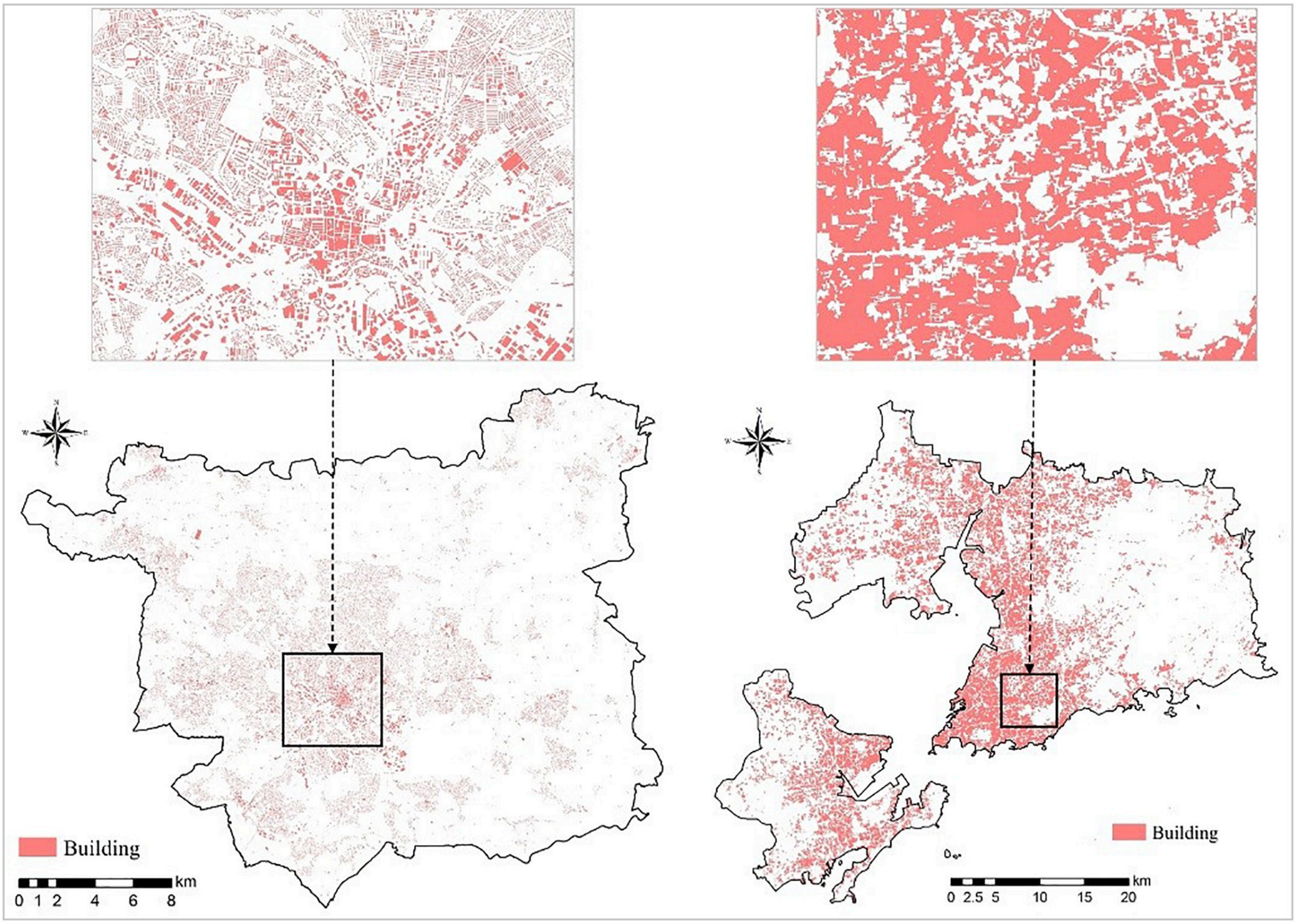

Fig. 2. Building areas in Leeds, UK (left-hand side) and in Qingdao, China (right-hand side).

Table 1
The different areal interpolation methods used in this study.

Method	Equation	Parameters
Areal weighting	$P_t = \sum_{r=1}^{q} \frac{A_{ts_r} P_{s_r}}{A_{s_r}}$	P_t estimated population of a target zone;
Dasymetric interpolation	$P_t = \sum_{r=1}^{q} \frac{B_{ts_r} P_{s_r}}{B_{s_r}}$	q number of source zones which overlap with the t^{th} target zone;
Road network	$P_t = \sum_{r=1}^{q} \frac{L_{ts_r} P_{s_r}}{L_{s_r}}$	P_s population of the r^{th} overlapping source zone;
Household proportion	$P_t = \sum_{r=1}^{q} \frac{H_{ts_r} P_{s_r}}{H_{s_r}}$	A_s area of the r^{th} source zone;
		A_{st} area of geometric overlap between the r^{th} source zone and the target zone.
		B_s building area of the r^{th} source zone;
		B_{st} building area of geometric overlap between the r^{th} source zone and the target zone.
		L_s total network length of the r^{th} source zone;
		L_{st} total network length of geometric overlap between the r^{th} source zone and the target zone.
		H_s household count of the r^{th} source zone;
		H_{st} household count of geometric overlap between r^{th} source zone and the target zone.

4.2. Comparisons of different areal interpolation methods

Having established the relationship between household data and population, it is possible to evaluate the results of the different interpolation methods described in Table 1. In order to compare the accuracy of different methods, the adjusted root mean square errors (RMSE) for each approach were compared (Table 4). These are generated from observed and estimated target zone populations. This is a standard approach and has been widely used in many previous studies (Comber et al., 2008; Eicher & Brewer, 2001; Fisher & Langford, 1996; Kounadi et al., 2018; Lin & Cromley, 2015; Mennis & Hultgren, 2006; Reibel & Bufalino, 2005; Tapp, 2010). The adjusted-RMSE is calculated as follows:

$$\text{Adjusted} - \text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{Z_i - \widetilde{Z}_i}{Z_i}\right)^2}$$

where Z_i is the actual population for zone i. $\widetilde{Z}_i$ is the estimated population for zone i.

Table 3 shows that the HP-census and HP-sales results have significantly lower errors than the other methods for the Leeds case study, at all levels, with the worst-performing (from Table 3, HP-sales at MSOA to OA) better than other methods. This indicates that the Household Proportion method with both census and sales data is a reliable interpolation method for this case study. For Qingdao, the Adjusted-RMSE for HP-census is much lower than other methods. HP-sales data is marginally higher than Network but still lower than Dasymetric. The high HP-sales error may be because the house sales data obtained

Table 2
Correlations of household counts with population counts over different census areas from different sources, with significance (p-values) and the associated coefficient of variation (CV).

Case study	Census area	Count source	Correlation	p-Values	Household CV
Leeds	MSOA	Census household	0.783	0.000	0.176
	LSOA	Census household	0.559	0.000	0.181
	OA	Census household	0.390	0.000	0.215
	MSOA	Property sales	0.724	0.000	0.200
	LSOA	Property sales	0.580	0.000	0.230
	OA	Property sales	0.444	0.000	0.321
Qingdao	District	Census household	0.976	0.001	0.439
	Subdistrict	Census household	0.978	0.000	0.589
	District	Property sales	0.662	0.152	0.553
	Subdistrict	Property sales	0.736	0.000	1.017

Table 3
The Adjusted-RMSE results from different areal interpolation methods.

Method	Leeds - UK			Qingdao - China
	MSOA to LSOA	MSOA to OA	LSOA to OA	District to subdistrict
Areal weighting	0.6561	1.7971	0.9051	1.0899
Network	0.3586	1.0438	0.6524	0.6304
Dasymetric	0.3704	0.9849	0.6196	0.7399
HP-census	0.1489	0.2042	0.2210	0.1195
HP-sales	0.1333	0.3070	0.2651	0.6572

from Qingdao reports only certain types of properties, that are more prevalent in certain areas of the city and does not include cheaper social housing. This bias may result in errors, however, the errors are still lower than those for the Dasymetric method and similar to the Network method.

The Adjusted-RMSE provides a measure of global fit, which indicates the overall effectiveness of interpolation approaches. In order to explore this in more detail, the estimation error and cumulative percentage of absolute error for different interpolation approaches are summarised in Figs. 3 and 4. Fig. 3 shows that the errors using HP-census or HP-sales in the Leeds case study, at all spatial levels, are distributed more closely around 0 than other methods. In contrast, the estimation errors from the other approaches have much wider distributions. This indicates that Household Proportion methods perform best in the Leeds case study. In the Qingdao case study, the estimation errors indicate that HP-census performs much better compared with other methods but that HP-sales tends to underestimate population with a much wider error distribution than other methods. This may be because there are fewer properties advertised in suburban areas (i.e. with more villages and farms) in Qingdao with the result that the population tends to be underestimated in these areas (Fig. 5).

Considering the absolute error values in Fig. 4, HP-census and HP-sales in Leeds consistently perform well at all levels (Fig. 4A–C). The cumulative percentage of absolute error for the Household Proportion approaches increases rapidly as the absolute error increases and then levels off as the cumulative percentage of absolute error approaches 80%. This indicates that the absolute errors generated from Household Proportion methods are lower than other methods. In the Qingdao case study (Fig. 4D), the cumulative percentage of absolute error from HP-census data increased to 100% very quickly, indicating that this method performs well. However, the cumulative percentage of absolute error

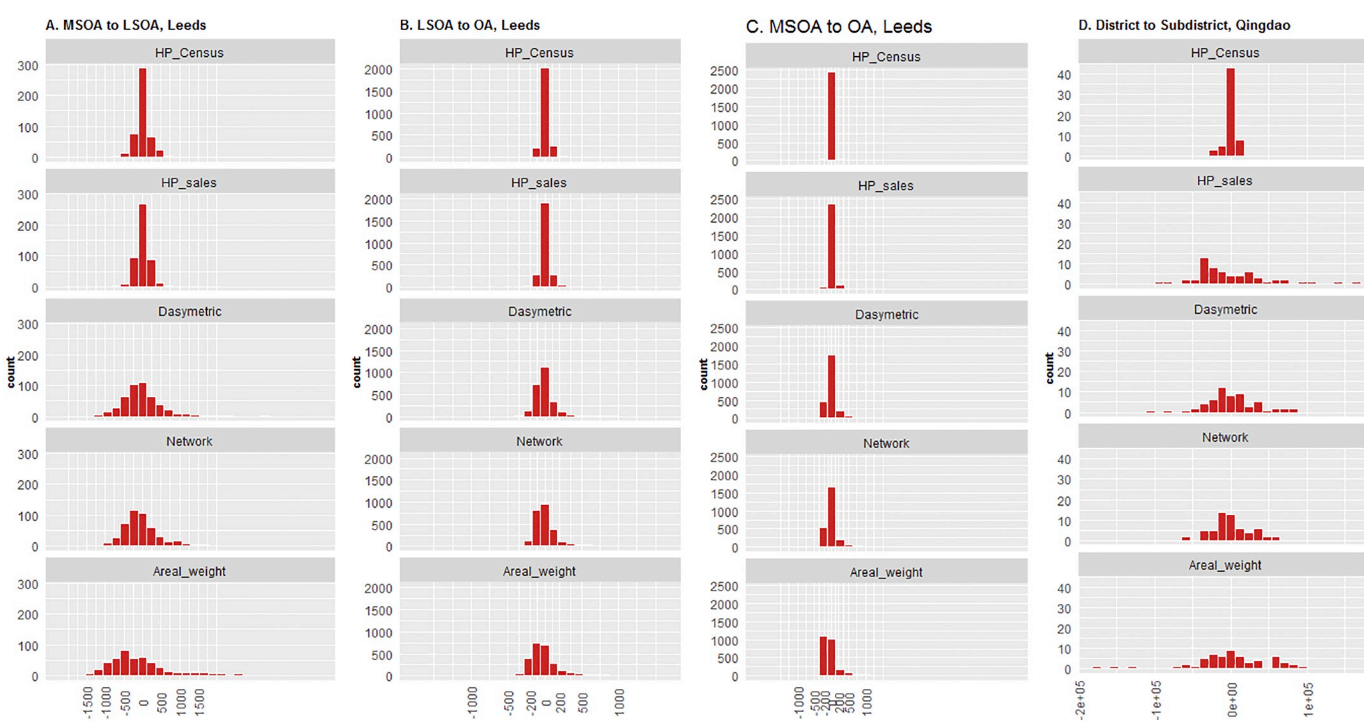

Fig. 3. The estimation error for different interpolation approaches in different cases. Note that all histograms have 30 bins and are vertically comparable.

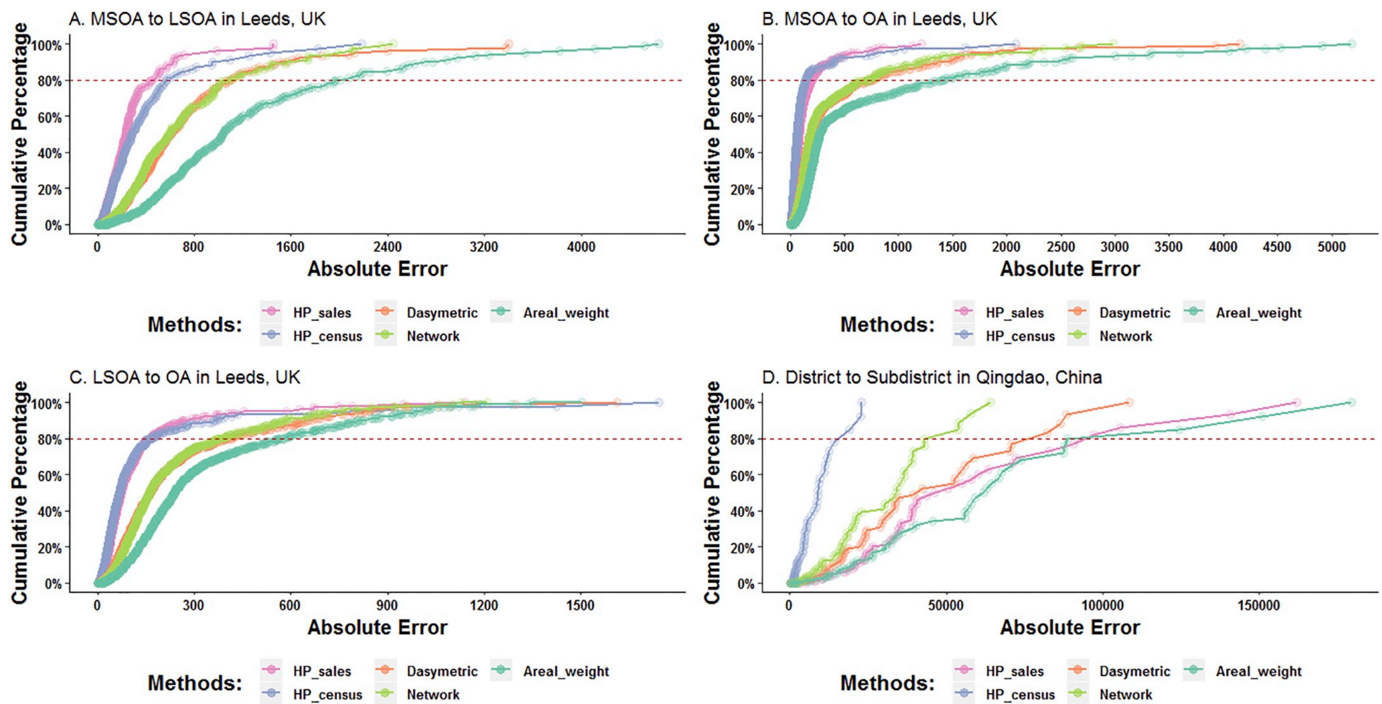

Fig. 4. The cumulative percentage of absolute error for the different interpolation approaches.

from HP-sales increased slowly compared with other methods and reached 80% at around same absolute error level as the areal interpolation method.

A Pearson correlation analysis was undertaken to compare estimated with observed populations (Table 4). The results indicate that the correlation coefficients for HP-census and HP-sales are highest, suggesting that these methods out-perform traditional areal interpolation methods.

In order to explore local variations, the error distributions arising from the different interpolation approaches are shown in Fig. 5. The error distributions for Household Proportion methods in Leeds are greater than in Qingdao. Most of the areas, at all levels in Leeds, have small errors using Household Proportion, and just a few areas have large over-estimation errors. This suggests that the surfaces generated by HP approaches more closely approximate to the observed population distribution in Leeds. In the Qingdao case study, HP-census is closer to

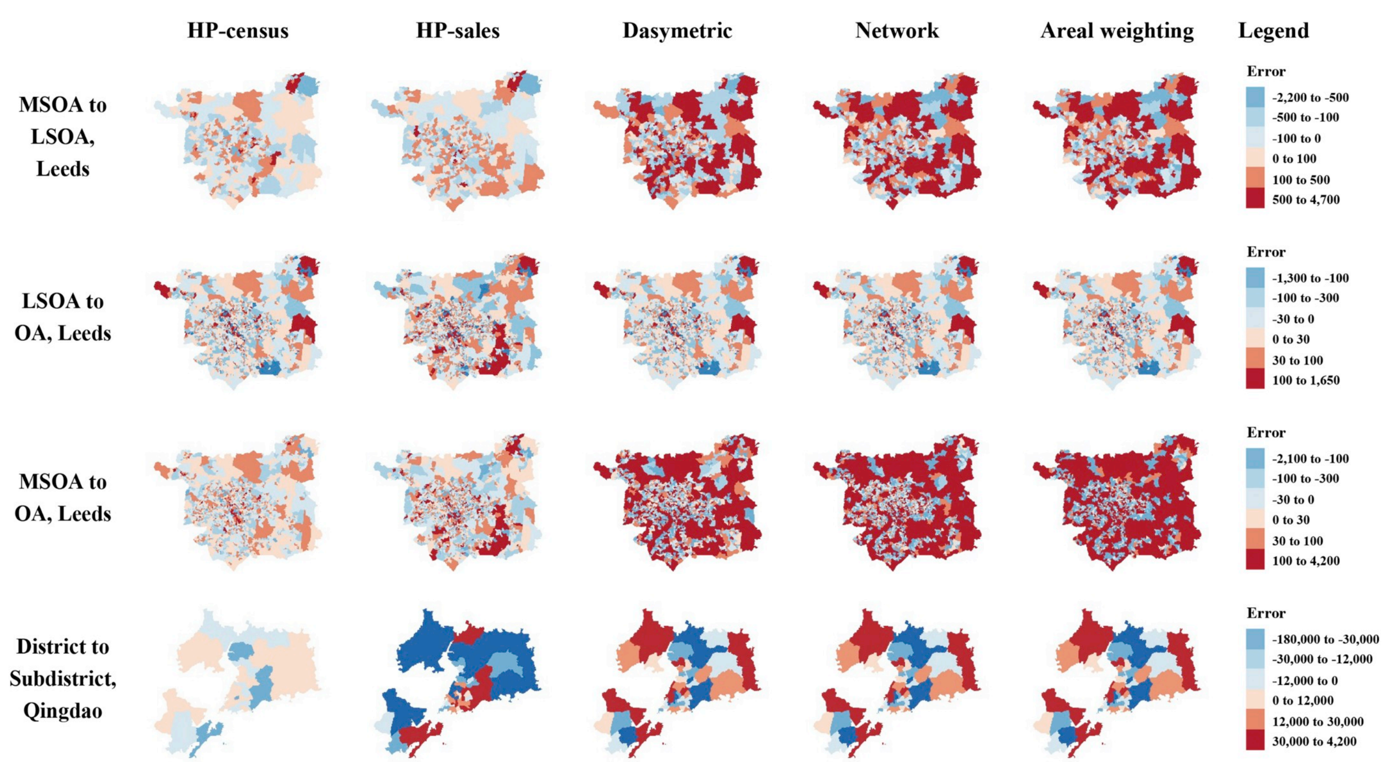

Fig. 5. The error distributions of results for different interpolation approaches.

Table 4
Pearson correlation coefficient between the estimated and actual population.

	Areal weighting	Dasymetric	HP-census	HP-sales	Network
MSOA to LSOA	0.2845	0.4446	0.6571	0.7738	0.4649
LSOA to OA	0.2738	0.3901	0.5563	0.6478	0.3847
MSOA to OA	0.1269	0.2247	0.4905	0.5319	0.2363
District to subdistrict	0.1753	0.6116	0.9879	0.8054	0.8226

Note: The p-values in all correlations are <0.001.

the observed population distribution than other approaches, and the HP-sales under-estimates population in some suburban target zones. This is because there are fewer properties for sale or rental in suburban areas. However, the errors tend to be small in the central city area, since the distribution of commodity housing is concentrated in these areas, with few social housing or urban villages. This suggests that HP-sales can be still applied in urban areas in the Qingdao case study and that overall, the error population surfaces indicate that the Household Proportion approach can be used to interpolate population.

5. Discussion

Previous areal interpolation research has used many different types and sources of data as ancillary information to guide and constrain interpolation from source zones to target zones. Few have focused on using household data to interpolate population. This research examined household data from two different sources, target zone household counts as collected in a formal population census and house sales data from residential property sales websites, as ancillary data for the areal interpolation of population. These were compared with other classic interpolation approaches (Dasymetric, Network, Areal Weighting) at different scales for 2 case studies. The best results were obtained using a simple constraint of the proportion of total household counts (Household Proportion) in each target zone. This method resulted in strong correlations between predicted population and observed population, with lower errors than other interpolation methods.

The HP-census method and HP-sales method in Leeds and HP-census method in Qingdao performed significantly better than other methods. Although HP-sales method in Qingdao performed somewhat poorly overall, the HP-sales error was lower than that of Dasymetric despite biases in the sample of properties listed on Chinese house sales websites. However, in the urban centre area in the Chinese case study, the errors tend to be smaller (Fig. 6), indicating that the method is sensitive to the representativeness of the data in the target zones, and that more

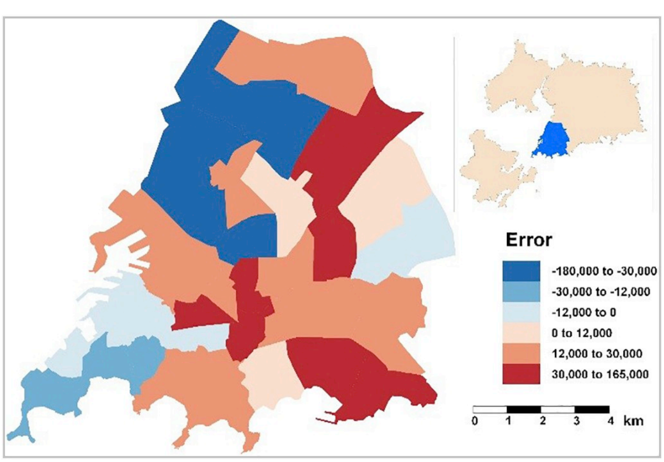

Fig. 6. The error distributions of results for Household Proportion method with sales data in the central urban area in Qingdao, China.

complete house information will generate better estimation results. This also indicates that, alongside an endorsement of the household counts method using data from diverse sources, there is a need for further evaluation in a range of environments and using different data sources.

Zandbergen and Ignizio (2010) noted that all interpolation methods have assumptions, flaws, and errors such that their performance may vary with location, with different data inputs etc., making it difficult to determine objectively a single “best method”. The inherent assumption of using household data as ancillary information is that household count has a significant and positive relationship with the population. The results of this study suggest that although this assumption may be violated, for example, the number of residents in each household may vary in different families and different places, the losses and gains balance each other out. The correlation results showed this relationship to be significant across target zone scales in Leeds and at subdistrict scales in Qingdao (Tables 2 and 3), which in turn indicated the suitability of this data as ancillary information for areal interpolation. This research also showed that this relationship gets weaker as the target zone scale gets finer. This is due to local variations in household size, which varies strongly at different target zone scales: the heterogeneity of household size is poorly represented by household counts at lower level geographical scales.

The heterogeneity of households size can also explain the variation in the performance of HP-sales method at different areas over the same geographical scales. Considering the error distributions at the same geographical scale, over-estimated populations tend to be found in the southeast and some central areas of Leeds (Fig. 5), due to the local variations of household size. For example, there are many farms and woodlands in the southeast of Leeds, and the household size there tends to be smaller than the average level (2.3 persons per household). The HP-sales method tends to over-estimate the population in these areas because of the assumption that the average household size is the same in all areas. To investigate this, the Pearson correlation between estimated population error and household size deviation was calculated and found to be -0.2822 (p-value $<.001$), which suggests that population estimated error and household size have a significant negative relationship. Thus, the population in areas with smaller household size tend to be over-estimated and vice versa (Fig. 7). This also indicates that the heterogeneity of household size is an important explanatory factor for the errors in the HP-sales method, which should be considered in future research. However, this approach was found to be better than other methods at all scales in Leeds, indicating its suitability as a method for areal interpolation of population.

A further issue is the different time stamps for the properties and census data, which may lead to population estimate biases and errors. For instance, the UK house sales data were collected in January 2019, while the census data was from 2011. New houses may be constructed and census areas may change in their composition over this period. Such differences may explain the population over-estimation in some of the central and suburban areas in Leeds (Fig. 5). Similarly, the well-documented and rapid urbanization in China over the last two decades may also explain poor population estimation in some areas.

The results of this study confirm the analytical potential of the many new forms of open and ubiquitously geo-referenced data that are available. Property sales data has the advantages of a fine geographical resolution, is easily obtained from open websites and requires only some kind of script to extract the data. It also obviates any concerns over the availability of ancillary information such as remote sensing images or land use data (Langford, 2013; Sadahiro, 2000), especially in developing countries (Yang et al., 2012). It has great potential to be applied in future studies in these and other areas.

The results of this research also indicate that local data context should be taken into account in areal interpolation. This research used two very different case studies: Leeds is a typical UK city while Qingdao is an eastern coastal city in China. Qingdao has a greater population and area but spatially coarser census areas. The finest census units in the UK

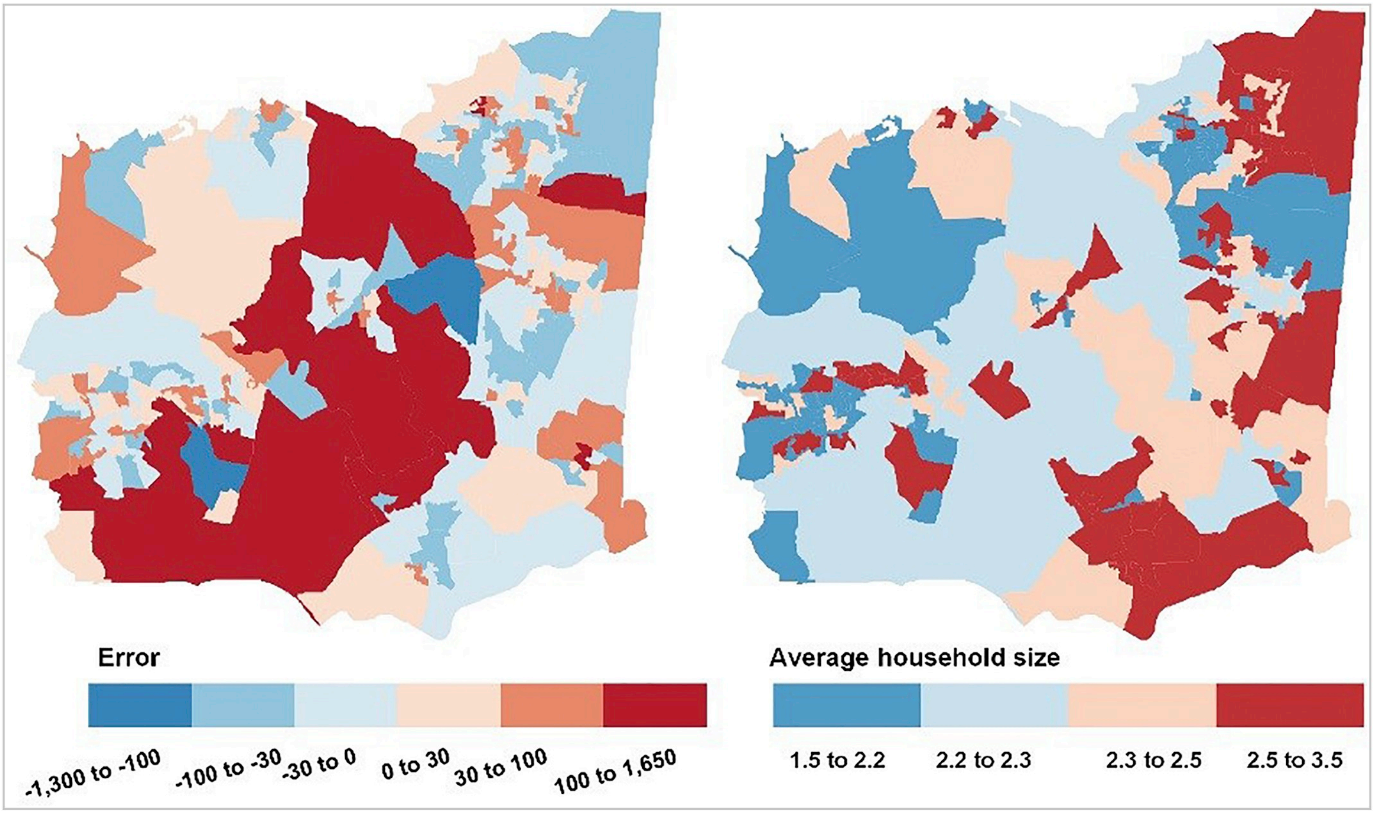

Fig. 7. The population estimation error distribution (left-hand) and average household size distribution in the southeast of Leeds.

are OAs (Output Areas) with around 300 persons, but the finest census units that are openly available in China are Sub-Districts with around 60,000 persons. The different contexts have an impact on areal interpolation results. The correlation coefficients between estimated population and observed population are stronger in Qingdao (Table 4), but the errors are smaller in Leeds (Table 3, Figs. 3–6), due to differences in the number of units and scales of generalisation.

Population data estimated over target zones using Household Proportion interpolation, using household locations as ancillary information, can support spatial planning when such data are not available. Detailed population distributions are needed to support service analysis and planning to optimise the spatial distribution of, for example, healthcare facilities. In countries such as China, where fine-scale population data is hard to obtain, data on household counts can be used directly to distribute populations over fine spatial grids to better support location-allocation.

6. Conclusions

This paper demonstrates how data from property sales websites can be used to guide and constrain areal interpolation. It evaluates the effectiveness of different types of household count data (from population census and from residential properties websites) as ancillary information. Different areal interpolation methods were compared with a Household Proportion approach using census household data and property sales data for two case studies in China and the UK. The results indicate that Household Proportion yielded the best population estimates, with significant improvements over other methods when error rates were compared, demonstrating the utility of both the approach as a new method for areal interpolation and of the data from new sources. Such data are geo-located, publicly accessible, low cost, with increasingly long runs and can be linked to a variety of socio-economic processes. They could be used to support a large number of geographical analyses, extending the developments that have been observed with volunteered geographic information, such as OpenStreetMap points-of-interest (Bakillah et al., 2014) or mobile phone data (Lin & Cromley, 2015). This paper demonstrates the value of using data of geo-located features, as are commonly found on public-facing websites and its effectiveness in supporting spatial disaggregation approaches. The need for such methods is particularly acute in environments and locations where detailed census area information is not available.

Declaration of Competing Interest

None.

Acknowledgements

We thank the editor and reviewers for careful review and insightful comments. This work was supported by the Natural Science Foundation of Shandong Province [grant numbers ZR201702170310], the State Scholarship Fund of China Scholarship Council [grant numbers 201808370092] and the Natural Environment Research Council (NERC) Newton Fund through the China- UK collaborative research on critical zone science [NE/S009124/1]. All of the data analyses and mappings were undertaken in R 3.5.1, the open source software. The R code and data used to generate the analyses are at https://github.com/Wen-Zeng/Areal-Interpolation.

Appendix A. Data sources

Data type	Data	Source	Period	Citation
Leeds:				
Census	2011 UK Census	UK data service	2011.3.27	http://infuse.ukdataservice.ac.uk/
House sales	Properties data	Zoopla	2019.15–2019.1.29	https://www.zoopla.co.uk/house-prices/browse/ls/
Land cover	OS Open Map – Local	Digimap	2018.7–2018.7	https://digimap.edina.ac.uk/webhelp/os/osdigimaphelp.htm#data_information/os_products/vectormap_local.htm
Road network	OS Open Roads	Digimap	2018.4–2018.4	https://digimap.edina.ac.uk/webhelp/os/osdigimaphelp.htm#data_information/os_products/os_open_roads.htm
Geographical administrative boundary	2011 Census Geography boundaries	UK data service	2011.3.27	https://www.statistics.digitalresources.jisc.ac.uk/search/field_topic/geography-64/type/dataset?sort_by=changed
Qingdao:				
Census	2010 China Census	National Bureau of Statistics	2010.11.1	www.stats.gov.cn/tjsj/pcsj/rkpc/6rp/indexch.htm
House sales	Properties data	Lianjia	2018–6.20-2018.6.25	https://qd.lianjia.com/xiaoqu/rs/
Land cover	Landsat8 satellite image	Landsat8 satellite image	2011.7–2011.7	https://earthexplorer.usgs.gov/
Road network	Baidu map	Baidu map	2015.12–2015.12	https://map.baidu.com/
Geographical administrative boundary	National Geomatic Centre of China	National Geomatic Centre of China	2010.11.1	http://www.gscloud.cn/

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. https://doi.org/10.1080/13658816.2014.909045.

Boyce, J. K., Zwickl, K., & Ash, M. (2016). Measuring environmental inequality. *Ecological Economics, 124*, 114–123. https://doi.org/10.1016/j.ecolecon.2016.01.014.

Bracken, I., & Martin, D. (1989). The generation of spatial population distributions from census centroid data. *Environment and Planning A, 21*(4), 537–543. https://doi.org/10.1068/a210537.

Church, R. L., & Murray, A. T. (2009). *Business site selection, location analysis and GIS.* New York: Wiley.

Comber, A., Dickie, J., Jarvis, C., Phillips, M., & Tansey, K. (2015). Locating bioenergy facilities using a modified GIS-based location–allocation-algorithm: Considering the spatial distribution of resource supply. *Applied Energy, 154*, 309–316. https://doi.org/10.1016/j.apenergy.2015.04.128.

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. https://doi.org/10.1111/j.1467-9671.2008.01130.x.

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*12465. https://doi.org/10.1111/gec3.12465.

Comber, A. J., Sasaki, S., Suzuki, H., & Brunsdon, C. (2011). A modified grouping genetic algorithm to select ambulance site locations. *International Journal of Geographical Information Science, 25*(5), 807–823. https://doi.org/10.1080/13658816.2010.501334.

Cromley, R. G., Hanink, D. M., & Bentley, G. C. (2012). A quantile regression approach to areal interpolation. *Annals of the Association of American Geographers, 102*(4), 763–777. https://doi.org/10.1080/00045608.2011.627054.

De Albuquerque, J. P., Herfort, B., Brenning, A., & Zipf, A. (2015). A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. *International Journal of Geographical Information Science, 29*(4), 667–689. https://doi.org/10.1080/13658816.2014.996567.

Deng, C., & Wu, C. (2013). Improving small-area population estimation: An integrated geographic and demographic approach. *Annals of the Association of American Geographers, 103*(5), 1123–1141. https://doi.org/10.1080/00045608.2013.770364.

Eicher, C. L., & Brewer, C. A. (2001). Dasymetric mapping and areal interpolation: Implementation and evaluation. *Cartography and Geographic Information Science, 28*(2), 125–138. https://doi.org/10.1559/152304001782173727.

Evans, N. (1996). Evaluating recent changes to the publication of UK agricultural census data. *Geography,* 225–234. https://www.jstor.org/stable/40572795.

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. https://doi.org/10.1111/j.0033-0124.1996.00299.x.

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. https://doi.org/10.1068/a250383.

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

Goovaerts, P. (2000). Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. *Journal of Hydrology, 228*(1–2), 113–129. https://doi.org/10.1016/S0022-1694(00)00144-X.

Goplerud, M. (2016). Crossing the boundaries: An implementation of two methods for projecting data across boundary changes. *Political Analysis, 24*(1), 121–129. https://doi.org/10.1093/pan/mpv029.

Gregory, I. N. (2002). The accuracy of areal interpolation techniques: Standardising 19th and 20th century census data to allow long-term comparisons. *Computers, Environment and Urban Systems, 26*(4), 293–314.

Haklay, M. (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and ordnance survey datasets. *Environment and Planning. B, Planning & Design, 37*(4), 682–703. https://doi.org/10.1068/b35097.

Harris, R., & Chen, Z. (2005). Giving dimension to point locations: Urban density profiling using population surface models. *Computers, Environment and Urban Systems, 29*(2), 115–132. https://doi.org/10.1016/j.compenvurbsys.2003.08.003.

Hawley, K., & Moellering, H. (2005). A comparative analysis of areal interpolation methods. *Cartography and Geographic Information Science, 32*(4), 411–423. https://doi.org/10.1559/152304005775194818.

Hay, S. I., Noor, A. M., Nelson, A., & Tatem, A. J. (2005). The accuracy of human population maps for public health application. *Tropical Medicine & International Health, 10*(10), 1073–1086. https://doi.org/10.1111/j.1365-3156.2005.01487.x.

Jia, P., & Gaughan, A. E. (2016). Dasymetric modeling: A hybrid approach using land cover and tax parcel data for mapping population in Alachua County, Florida. *Applied Geography, 66*, 100–108. https://doi.org/10.1016/j.apgeog.2015.11.006.

Joseph, J., Sharif, H. O., Sunil, T., & Alamgir, H. (2013). Application of validation data for assessing spatial interpolation methods for 8-h ozone or other sparsely monitored constituents. *Environmental Pollution, 178*, 411–418. https://doi.org/10.1016/j.envpol.2013.03.035.

Kar, B., & Hodgson, M. E. (2012). A process oriented areal interpolation technique: A coastal county example. *Cartography and Geographic Information Science, 39*(1), 3–16. https://doi.org/10.1559/152304063913.

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. https://doi.org/10.1080/15230406.2017.1304243.

Kunze, C., & Hecht, R. (2015). Semantic enrichment of building data with volunteered geographic information to improve mappings of dwelling units and population. *Computers, Environment and Urban Systems, 53*, 4–18. https://doi.org/10.1016/j.compenvurbsys.2015.04.002.

Lam, N. S. N. (1983). Spatial interpolation methods: A review. *The American Cartographer, 10*(2), 129–150. https://doi.org/10.1559/152304083783914958.

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. https://doi.org/10.1016/j.compenvurbsys.2004.07.001.

Langford, M. (2007). Rapid facilitation of dasymetric-based population interpolation by means of raster pixel maps. *Computers, Environment and Urban Systems, 31*(1), 19–32. https://doi.org/10.1016/j.compenvurbsys.2005.07.005.

Langford, M. (2013). An evaluation of small area population estimation techniques using open access ancillary data. *Geographical Analysis, 45*(3), 324–344. https://doi.org/10.1111/gean.12012.

Liao, Y., Li, D., & Zhang, N. (2018). Comparison of interpolation models for estimating heavy metals in soils under various spatial characteristics and sampling methods. *Transactions in GIS, 22*(2), 409–434. https://doi.org/10.1111/tgis.12319.

Lin, J., & Cromley, R. G. (2015). Evaluating geo-located twitter data as a control layer for

areal interpolation of population. *Applied Geography, 58*, 41–47. https://doi.org/10.1016/j.apgeog.2015.01.006.
Liu, L., Peng, Z., Wu, H., Jiao, H., & Yu, Y. (2018). Exploring urban spatial feature with dasymetric mapping based on mobile phone data and LUR-2SFCAe method. *Sustainability, 10*(7), 2432. https://doi.org/10.3390/su10072432.
Liu, X. H., Kyriakidis, P. C., & Goodchild, M. F. (2008). Population-density estimation using regression and area-to-point residual kriging. *International Journal of Geographical Information Science, 22*(4), 431–447. https://doi.org/10.1080/13658810701492225.
Logan, J. R., Xu, Z., & Stults, B. J. (2014). Interpolating US decennial census tract data from as early as 1970 to 2010: A longitudinal tract database. *The Professional Geographer, 66*(3), 412–420. https://doi.org/10.1080/00330124.2014.905156.
Maliszewski, P. J., Kuby, M. J., & Horner, M. W. (2012). A comparison of multi-objective spatial dispersion models for managing critical assets in urban areas. *Computers, Environment and Urban Systems, 36*(4), 331–341. https://doi.org/10.1016/j.compenvurbsys.2011.12.006.
Martin, D. (1989). Mapping population data from zone centroid locations. *Transactions of the Institute of British Geographers,* 90–97. https://doi.org/10.2307/622344.
Martin, D. (1998). Optimizing census geography: The separation of collection and output geographies. *International Journal of Geographical Information Science, 12*(7), 673–685. https://doi.org/10.1080/136588198241590.
Meegan, R. (2015). City profile–Leeds. *Cities, 42*, 42–53. https://doi.org/10.1016/j.cities.2014.08.003.
Mennis, J. (2003). Generating surface models of population using dasymetric mapping. *The Professional Geographer, 55*(1), 31–42. https://doi.org/10.1111/0033-0124.10042.
Mennis, J. (2016). Dasymetric spatiotemporal interpolation. *The Professional Geographer, 68*(1), 92–102. https://doi.org/10.1080/00330124.2015.1033669.
Mennis, J., & Hultgren, T. (2006). Intelligent dasymetric mapping and its application to areal interpolation. *Cartography and Geographic Information Science, 33*(3), 179–194. https://doi.org/10.1559/152304006779077309.
Qiu, F., & Cromley, R. (2013). Areal interpolation and dasymetric modeling. *Geographical Analysis, 45*(3), e213–e215. https://doi.org/10.1111/gean.12016.
Reibel, M., & Agrawal, A. (2007). Areal interpolation of population counts using pre-classified land cover data. *Population Research and Policy Review, 26*(5–6), 619–633. https://doi.org/10.1007/s11113-007-9050-9.
Reibel, M., & Bufalino, M. E. (2005). Street-weighted interpolation techniques for demographic count estimation in incompatible zone systems. *Environment and Planning A, 37*(1), 127–139. https://doi.org/10.1068/a36202.
Rigol, J. P., Jarvis, C. H., & Stuart, N. (2001). Artificial neural networks as a tool for spatial interpolation. *International Journal of Geographical Information Science, 15*(4), 323–343. https://doi.org/10.1080/13658810110038951.
Sadahiro, Y. (2000). Accuracy of count data estimated by the point-in-polygon method. *Geographical Analysis, 32*(1), 64–89. https://doi.org/10.1111/j.1538-4632.2000.tb00416.x.
Shi, W. Z., & Tian, Y. (2006). A hybrid interpolation method for the refinement of a regular grid digital elevation model. *International Journal of Geographical Information Science, 20*(1), 53–67. https://doi.org/10.1080/13658810500286943.
Smith, S. K., & Cody, S. (2004). An evaluation of population estimates in Florida: April 1, 2000. *Population Research and Policy Review, 23*(1), 1–24. https://doi.org/10.1023/B:POPU.0000019918.24143.c9.
Smith, S. K., Nogle, J., & Cody, S. (2002). A regression approach to estimating the average number of persons per household. *Demography, 39*(4), 697–712. https://doi.org/10.1353/dem.2002.0040.
Sridharan, H., & Qiu, F. (2013). A spatially disaggregated areal interpolation model using light detection and ranging-derived building volumes. *Geographical Analysis, 45*(3), 238–258. https://doi.org/10.1111/gean.12010.
Tapp, A. F. (2010). Areal interpolation and dasymetric mapping methods using local ancillary data sources. *Cartography and Geographic Information Science, 37*(3), 215–228. https://doi.org/10.1559/152304010792194976.
Tobler, W. R. (1979). Smooth pycnophylactic interpolation for geographical regions. *Journal of the American Statistical Association, 74*(367), 519–530. https://doi.org/10.1080/01621459.1979.10481647.
Xie, Y. (1995). The overlaid network algorithms for areal interpolation problem. *Computers, Environment and Urban Systems, 19*(4), 287–306. https://doi.org/10.1016/0198-9715(95)00028-3.
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. https://doi.org/10.1016/j.apgeog.2011.05.008.
Yao, Y., Liu, X., Li, X., Zhang, J., Liang, Z., Mai, K., & Zhang, Y. (2017). Mapping fine-scale population distributions at the building level by integrating multisource geospatial big data. *International Journal of Geographical Information Science, 31*(6), 1220–1244. https://doi.org/10.1080/13658816.2017.1290252.
Ye, T., Zhao, N., Yang, X., Ouyang, Z., Liu, X., Chen, Q., ... Jia, P. (2019). Improved population mapping for China using remotely sensed and points-of-interest data within a random forests model. *Science of the Total Environment, 658*, 936–946. https://doi.org/10.1016/j.scitotenv.2018.12.276.
Yu, Y., Li, J., Zhu, C., & Plaza, A. (2018). Urban impervious surface estimation from remote sensing and social data. *Photogrammetric Engineering & Remote Sensing, 84*(12), 771–780. https://doi.org/10.14358/PERS.84.12.771.
Zandbergen, P. A., & Ignizio, D. A. (2010). Comparison of dasymetric mapping techniques for small-area population estimates. *Cartography and Geographic Information Science, 37*(3), 199–214. https://doi.org/10.1559/152304010792194985.
Zhang, C., & Qiu, F. (2011). A point-based intelligent approach to areal interpolation. *The Professional Geographer, 63*(2), 262–276. https://doi.org/10.1080/00330124.2010.547792.