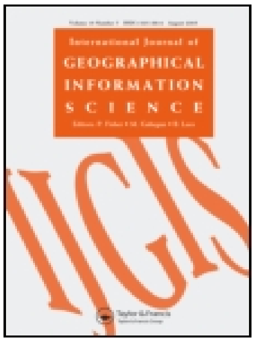

International Journal of Geographical Information Science

ISSN: 1365-8816 (Print) 1362-3087 (Online) Journal homepage: http://www.tandfonline.com/loi/tgis20

Error-sensitive historical GIS: Identifying areal interpolation errors in time-series data

Ian N. Gregory & Paul S. Ell

International Journal of Geographical Information Science
Vol. 20, No. 2, February 2006, 135–152

Research Article

Error-sensitive historical GIS: Identifying areal interpolation errors in time-series data

IAN N. GREGORY* and PAUL S. ELL

Centre for Data Digitization and Analysis, School of Geography, Archaeology and Palaeoecology, Queen's University Belfast, Belfast BT7 1NN, UK

(*Received 19 May 2005*)

Historical GIS has the potential to re-invigorate our use of statistics from historical censuses and related sources. In particular, areal interpolation can be used to create long-run time-series of spatially detailed data that will enable us to enhance significantly our understanding of geographical change over periods of a century or more. The difficulty with areal interpolation, however, is that the data that it generates are estimates which will inevitably contain some error. This paper describes a technique that allows the automated identification of possible errors at the level of the individual data values.

Keywords: Areal interpolation; Error; Historical GIS; Time series; Census

1. Introduction

Recent years have seen a flurry of activity in building national historical Geographical Information Systems, databases designed to hold information about a country's past and how it has changed over time. The Great Britain Historical GIS (Gregory *et al*. 2002, Gregory 2005) is a well-developed example, and other countries including the United States (McMaster and Noble 2005), China (Bol and Ge 2005), Belgium (De Moor and Wiedemann 2001), Ireland (Ell 2005), and Russia (Merzlyakova 2005) are also building or proposing similar systems. Knowles (2005a) provides brief descriptions of a range of such systems. At the core of most of these are spatial data representing the changing boundaries of the country's administrative units. These are linked to a large attribute database that contains census and other data published using these units (Gregory 2002a). Most systems attempt to cover the entire period over which the country has been regularly publishing census data, which often covers most of the last two centuries.

A key advantage of these systems is that they are integrated databases of census statistics over time that should allow us to explore long-term demographic change through attribute, space, and time. Although conventionally termed 'historical', the systems are relevant to researchers interested in the present as they allow questions such as 'how did we get to the situation we are in now?' or 'how did this phenomenon evolve?' to be answered. Thus, using these systems, census analysis need no longer be hamstrung by having the choice of either analysing a single snapshot of spatially detailed data, or looking at long-term change using massively aggregate spatial units such as British counties or US states that are wholly

*Corresponding author. Email: ian.gregory@qub.ac.uk

International Journal of Geographical Information Science
ISSN 1365-8816 print/ISSN 1362-3087 online © 2006 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/13658810500399589

inappropriate for spatial analysis purposes due to modifiable areal unit problem (Openshaw 1984).

Analysing spatially detailed long-term change over time requires us first to interpolate the relevant datasets onto a single set of administrative units to attempt to remove the impact of boundary changes. Having the changing boundaries of these units in the historical GIS database allows *areal interpolation* techniques to take data published using differing administrative units and estimate their values for a single set of standardized units termed the *target units* (Langford *et al.* 1991). In this way, we might, for example, take population data for England and Wales published from 1851 to 2001 and interpolate them all onto modern districts, 1851 registration districts, or any other suitable set of boundaries. This clearly has enormous potential, but data created by interpolation are estimates that will inevitably contain a certain degree of error. Careful choice of interpolation technique and target units can help to minimize this error, but it will still be present. Fully handling the error requires identification of which individual interpolated data values are suspected of containing error and, on a more positive note, which values can confidently be claimed to be free of significant errors. This paper describes a methodology that enables us to do this by identifying sudden changes in time-series of interpolated data where boundary changes have occurred and comparing these with changes where there have been no boundary changes. The paper focuses on census data, but the techniques would work equally well for any zone-based data published at regular intervals.

2. Literature review: Areal interpolation and error

Areal interpolation has long been recognized as a sphere where GIS and spatial analysis has much to offer (Goodchild and Lam 1980). It has been defined as 'the transfer of data from one set (source units) to a second set (target units) of overlapping, non-hierarchical, areal units' (Langford *et al.* 1991, p. 56). In its simplest form, known as *areal weighting*, the source units are overlaid onto the target units and the data values for the target units are estimated based on the assumption that the data are evenly distributed across the source units. This is implemented as follows:

$$\hat{y}_{\mathrm{t}} = \sum_{\mathrm{s}} \left(\frac{A_{\mathrm{st}}}{A_{\mathrm{s}}} \times y_{\mathrm{s}} \right) \tag{1}$$

where $\hat{y}_{\mathrm{t}}$ is the estimated population of the target zone, y_{s} is the population of the source zone, A_{s} is the area of the source zone, and A_{st} is the area of the zone of intersection between the source and target zones as calculated by the overlay operation (Goodchild and Lam 1980).

Obviously, the major problem with this approach is the assumption that y is evenly distributed across each source zone. With any variable associated with human geography, this is highly unlikely to be valid. To loosen this assumption, a variety of approaches have been developed that usually involve bringing in additional ancillary data that are thought to provide information on the distribution of y. Goodchild *et al.* (1993) demonstrate the use of a third set of 'control zones' that can be assumed to have an even population distribution. Langford *et al.* (1991) use satellite imagery classified into land-use types such as dense residential, residential, industrial, agricultural, or unpopulated. In both these cases, regression-based techniques are used to estimate the value of y for each target zone based on the ancillary information. Flowerdew and Green (1994) use a variety of data available for the target units to help in their interpolations. Rather than use regression, they use an iterative technique called the EM algorithm (Dempster *et al.* 1977) to determine what proportion of y should be allocated from each source zone to each target zone. Norman *et al.* (2003) use the density of British unit postcode centroids to provide ancillary information on the distribution of census data on the assumption that this provides a surrogate for the population distribution. This allows them to develop a dasymetric technique to standardize annual population estimates from 1990 to 1998 onto a 1998 ward geography. Reibel and Bufalino (2005) follow a similar approach using the density of streets as a surrogate to allow them to compare data from the 1990 and 2000 censuses for Los Angeles County. In all cases, the authors show that the additional information improves the accuracy of the interpolated estimates of y.

Gregory (2002b) explores techniques devised explicitly for the creation of long-run time-series of British historical data. He uses a variety of techniques including a dasymetric technique that makes use of the fact that, until 1971, censuses of England and Wales published the bulk of their data at district level but also published total population at the more spatially detailed parish level which, if y is not total population, may provide useful ancillary information. This is the approach used by the *Vision of Britain Through Time* project (Vision of Britain Through Time 2005). He also makes use of ancillary information from the target districts using the EM algorithm and additionally combines the two approaches. He shows that techniques making use of the EM algorithm are usually significantly more accurate than those that just use the dasymetric technique but, importantly, that the choice of the most accurate technique will vary according to the nature of the variable to be interpolated.

Only a limited number of studies have explicitly addressed the issue of error produced by areal interpolation. Sadahiro (2000) shows that error will be reduced where source zones are relatively small compared with the target zones, and that the relative shapes of the source and target zones will also affect accuracy. Sadahiro (1999) shows that the utility of ancillary information will vary according to its relevance to y and that adding inappropriate ancillary information may actually reduce the accuracy of interpolation compared with simple areal weighting. Similar conclusions have been drawn by Fisher and Langford (1995), Cockings *et al.* (1997), and Gregory (2002b). These studies demonstrate that the accuracy of areal interpolation will vary according to the nature of the variable being interpolated, the nature of the ancillary data, and the shape and size of both the source and target units. This leaves an important question: if the researcher is using real-world data, which approach will produce the most accurate results, and how accurate will these be?

Simpson (2002) goes some way to answering this question by developing two measures to quantify the relationship between the source and target units. He refers to these as the *degree of hierarchy* and the *degree of fit*. The degree of hierarchy attempts to quantify the extent to which the target zones nest into the source zones. It is expressed as the total number of source zones whose entire population is allocated to a single target zone as a proportion of the total number of source zones. The degree of fit exploits the idea that the greater the similarity between the source and target zones, the lower the error. It is measured as:

$$\frac{100}{S}\sum_{s=1}^{S}(\max(w_{\mathrm{st}})) \tag{2}$$

where max(w_{st}) is the largest weighting factor associated with each source zone S for S=1,..., S. The weighting factor w_{st} is the proportion of a source zone's population allocated to each target zone, so in simple areal weighting, the weighting factor would be the area of the zone of the intersection divided by the area of the source zone, or A_{st}/A_s (see equation (1)). An alternative approach is given by Gregory and Ell (2005), who examine the relative spatial distributions of the source and the ancillary data using regression techniques to determine which ancillary data are most appropriate to enhance the accuracy of the interpolation.

While both of these approaches give an indication of the best choice of target zone and of ancillary data at the global level (i.e. for the entire study area), they give little idea of which individual data values can be regarded as reliable and which must be treated with suspicion. In other words, we do not want to simply identify globally that error may be occurring somewhere on the study area; we want to identify locally where (and perhaps when) error may be occurring.

3. Identifying possible interpolation errors

The key to identifying possible error caused by interpolation is first to interpolate a time series of data onto a single set of target units and then to examine the inter-censal population changes for each target unit to identify sudden changes in population change. To do this, inter-censal population changes are calculated as normalized rates using the equation:

$$\Delta y = 100 \times \left(\frac{y_{\text{end}} - y_{\text{start}}}{y_{\text{end}} + y_{\text{start}}} \right) \qquad (3)$$

where Δy is the inter-censal population change between the population at the end of a decade, y_{end}, and the population at its start, y_{start}. The reason for using a measure with y_{end} + y_{start} as the denominator is that this gives us a symmetrical measure from −100.0%, complete depopulation of a previously populated area, to 100.0%, new population moving into a previously unpopulated area, with 0.0 indicating no change (Bracken and Martin 1995). Conventional rates, which just use y_{start} as the denominator, do not give symmetrical measures because while the maximum loss is −100.0%, there is no theoretical maximum gain. Thus, using the normalized rates, a gain of 20% is the exact opposite of a loss of 20%, which is not the case with a conventional rate.

Figure 1 shows the interpolated total population for a hypothetical target unit. The target units are defined as 2001 districts, and data for 1951–1991 have been interpolated onto these. The total population time-series graph and normalized population changes are shown in table 1. The target unit's population is rising at a slowly increasing rate from 2.4% in the 1950s to 5.9% in the 1990s. The trend is bucked by a sudden increase in the 1970s. If we know that a boundary change occurred in the 1970s, then clearly this gives the suspicion that this change is not being handled well by the interpolation and that the values from 1951 to 1971 are all likely to contain errors.

The population change values, shown in figure 2, highlight how unusual the change in the 1970s is. Although there could be an alternative cause, this peak on the population change graph is a distinctive feature of interpolation error. We refer to this as a *spike* which can be defined as an inter-censal population change value that is either greater than the population change values before and after it or less than the population change values before it and after it. Thus, from these data, there is a

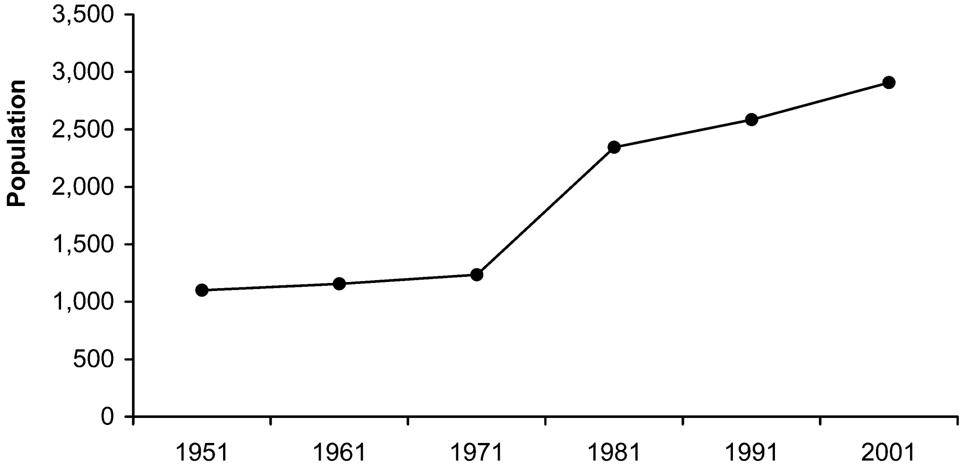

Figure 1. Time-series graph for the total population of a hypothetical interpolated variable. 2001 units are used as the target units, and a boundary change is known to have occurred in the 1970s. This raises the suspicion that the 1971 value and those preceding it contain error.

spike in the 1970s but in no other decade. The *spike size* is defined as the shorter distance between the population change for the spike decade and the population change to whichever is the nearer in value of the preceding and following values. More formally, spike size, *ss*, is:

$$\begin{gathered} ss = \Delta y_{\mathrm{t}} - \Delta y_{\mathrm{t}+1} \text{ or } ss = \Delta y_{\mathrm{t}} - \Delta y_{\mathrm{t}-1} \text{ whichever } ss \text{ is the nearest to zero} \\ \text{where } ((\Delta y_{\mathrm{t}} > \Delta y_{\mathrm{t}+1} \text{ and } \Delta y_{\mathrm{t}} > \Delta y_{\mathrm{t}-1}) \text{ or } (\Delta y_{\mathrm{t}} < \Delta y_{\mathrm{t}+1} \text{ and } \Delta y_{\mathrm{t}} < \Delta y_{\mathrm{t}-1})) \end{gathered} \tag{4}$$

where *ss* is the spike size, Δy_t is the normalized population change in the spike period, Δy_{t+1} is the population change in the inter-censal period following the spike, and $\Delta y_{t\text{-}1}$ is the population change in the inter-censal period immediately preceding it. Thus, using the data in table 1, there is a spike in the 1970s, and its spike size is 31.0 – 4.8=26.2. The concept of a spike is important, as, if there is no spike, it is unlikely that there is boundary change induced error in the interpolated values. Although spikes will occur naturally in a time series, large spikes become increasingly suspicious. The issue of what constitutes a large spike will be returned to later in the paper.

One problem with the concept of the spike is that it cannot be used on either the first or last inter-censal period. In this case, all values must be considered as worthy of further investigation, and a *change in change* measure is used whereby the first value is subtracted from the second, and the one preceding the last is subtracted

Table 1. Total population and inter-censal population change (as a normalized percentage) for the hypothetical variable shown in figure 1.

	1951		1961		1971		1981		1991		2001
Total population	1100		1155		1236		2345		2583		2906
Population change		2.4		3.4		31.0		4.8		5.9	

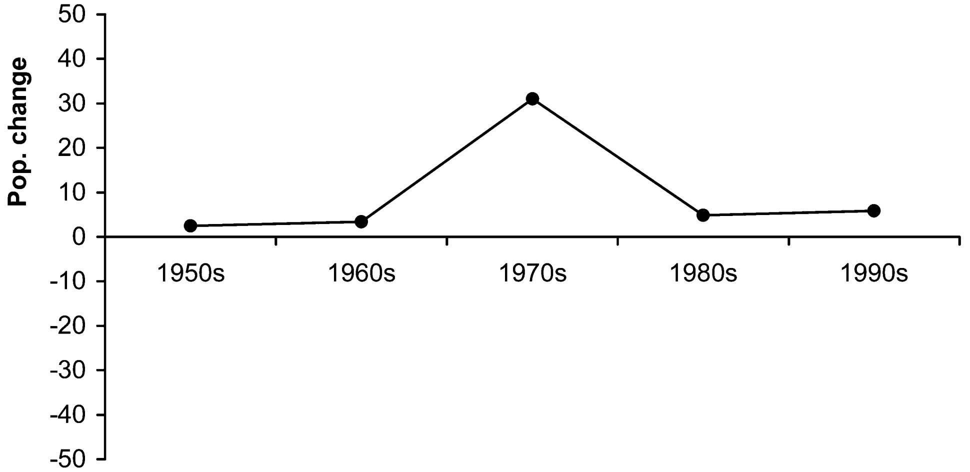

Figure 2. Time series of normalized population change for the data shown in figure 1.

from the last. If there are n inter-censal periods, t_1, t_2...t_n, then the two change in changes, *cc*, will be:

$$cc = \Delta y_{t2} - \Delta y_{t1} \text{ and } cc = \Delta y_{tn} - \Delta y_{t(n-1)} \tag{5}$$

Thus, for the data in table 1, the two change in changes are: 3.4 – 2.4=1.0 for the 1950s and 5.9 – 4.8=1.1 for the 1990s. Although the change in change measure is mathematically similar to the spike size, they are conceptually very different. The spike size only consists of a subset of the data values, whereas all possible values are included in the change in change data. As a result, the two measures must be considered separately in the subsequent analyses.

A second challenge is to identify where boundary changes have occurred. This could be done by referring to boundary change reports, but a simpler approach is to incorporate it into the interpolation process. Almost all techniques rely on calculating the area of the zones of intersection between each source zone and each target zone, and comparing this with the area of the appropriate source zone. In simple areal weighting, as shown in equation (1), these values are represented by A_{st} and A_s, respectively. If the area of the zone of intersection is the same as the area of the source zone, i.e. $A_{st}=A_s$, then the boundary of the source zone has not been affected by a change, so no error has been made in the calculation of the population for this zone of intersection. If this is true for all of the zones of intersection that make up a target zone, then the target zone has not been affected by interpolation error. In this way, interpolated data that may contain error can be isolated from those that have had no boundary change, or those where there has only been aggregation.

The final issue is to identify the degree of suspicion that each spike or change in change should be held in, taking into account that it is possible for a spike to occur within a time series at the same time as a boundary change without this necessarily meaning that the spike is caused by interpolation error. The challenge is thus to identify spikes that seem suspiciously large. To do this, we make use of the fact that we can compare spikes that coincide with a boundary change, termed *suspect spikes,* with those that occurred without a boundary change, termed *natural spikes*. The size of each suspect spike can be compared with the size of the sample of natural spikes to give an indication of whether the spike appears to be a result of interpolation error or may simply be natural change along the time series. The change in change values from the beginning and end of the time series form a separate set of suspect and natural values.

Once we have a natural sample for comparison, quantifying the uncertainty of each spike size or change in change becomes relatively straightforward. First, the mean and standard deviations of the natural sample are calculated, and these are used to calculate a z-score for each suspect value using:

$$z = \left(\frac{x - \bar{x}_{\mathrm{n}}}{s_{\mathrm{n}}} \right) \tag{6}$$

where x is the suspect value, and $\bar{x}_{\mathrm{n}}$ and s_{n} are the mean and the standard deviation respectively of the spike size's natural sample. The same principal can be used for the change in changes. A z-score tells us how close to the mean a value lies with 0.0 being exactly on the mean, 1.0 being one standard deviation above the mean and -1.0 being one standard deviation below it. As a rule of thumb, 95% of values will have a z-score of between -1.96 and 1.96, and 99% will be between -2.58 and 2.58. It is over-simplistic to simply state that changes with z-scores over 1.96 or 2.58 are definitely errors. Instead, strategies for handling high z-scores (positive or negative) will be discussed later in the paper. It is also worth noting that it is likely that there may be values in the natural sample that will have high z-scores. While these cannot be caused by interpolation error, their values are worth checking to see if they are caused by other errors such as transcription errors, or perhaps boundary changes that occurred in real life but that have been missed from the GIS. This is, therefore, a useful technique for checking for errors in addition to those thought to have occurred because of interpolation.

To return to the example from table 1, if we found that the mean natural spike size for the entire study area is 1.0, and the standard deviation is 1.5, the z-score associated with the 1970s spike would be $(26.2-1.0)/1.5=16.8$. The probability of this having occurred randomly is a tiny fraction of 1%. If a boundary change had also occurred in the 1950s, the change in change for this decade (as no spike is available) would be compared with the natural sample of change in changes. If this sample has a mean of 1.25 and a standard deviation of 1.75, this gives us $(1.0-1.25)/1.75=-0.43$, which is likely to have occurred at random and suggests no significant error from the interpolation process.

4. Real-world example: Population change in Warwickshire, 1851–1951

In order to investigate the effectiveness of this technique, a real-world example was used. This involved parish-level total population data for the county of Warwickshire in central England from 1851 to 1951. The reason for this choice was that parishes are a very difficult type of administrative unit to use for handling statistical data. Their shapes, areas, and total populations vary enormously, and their boundaries were subject to frequent changes. Warwickshire gives us an area that is very diverse. Prior to the reforms of the early 1970s, the county included the city of Birmingham, one of England's largest, a variety of other urban areas including Coventry, Rugby, and Stratford-upon-Avon, and several sparsely populated rural areas. Therefore, we have a type of administrative unit and a study area that presents significant challenges to any interpolation methodology.

Data from each census from 1851 to 1931 were used as source data to be interpolated onto 245 parishes as configured in 1951, the target units. Note that there was no census in 1941, but this makes little difference to this analysis, and for the remainder of this paper, the period 1931–1951 will simply be considered as a single inter-censal period. The raw data, in the form of population density for 1851, 1901, and 1951, are shown in figure 3. The diagram has been simplified from the census reports which used three noticeably different definitions of Warwickshire: the Ancient County in 1851, the Registration County in 1901, and the Administrative County in 1951. In the diagram, all parishes intersecting the 1951 Administrative County boundary have been included regardless of which county they were in at the appropriate census year, and all parishes lying outside this boundary have been excluded. In 1851, 360 000 people were enumerated in this area sub-divided among 346 administrative units with a maximum density of 28 100 persons km^2. By 1901, the population had risen to 1 200 000, but changes to the parish structure had reduced the number of parishes to 292 with a maximum density of 20 000 persons km^2. This trend of increasing population in a smaller number of parishes with a subsequently lower maximum population density continued. In 1951, there were 1 860 000 people in 245 parishes with a maximum density of only 7300 persons km^2. Clearly, even the most basic summary statistics about changing parish population densities over this time period will be highly misleading due to changes in the administrative geography.

The EM algorithm with total population in 1951 as ancillary information was used to interpolate the source data. This means that we are assuming that the 1951 population distribution provides relevant information about the population distribution in earlier periods.

Interpolating the data onto the 245 target parishes gives the population densities shown in figure 4. As can be seen, this immediately gives a pattern that is easier to understand and where the higher values are generally found on the 1951 map, as would be expected given the rising population of the county.

The interpolated dataset gives us a total of 2205 inter-censal population changes. Of these, only 301 changes, 13%, were affected by boundary changes and are thus considered suspect. The mean of the natural sample of spike sizes is 0.66 percentage points, and the standard deviation is 8.30. The change in changes have a mean of 6.02 percentage points and a standard deviation of 16.9.

The utility of the technique at the level of the individual target unit can be demonstrated by examining three separate parishes: Water Orton, Sherbourne, and Birmingham. Water Orton is a small parish on the eastern fringes of Birmingham. It had an unusual history in that it had a small area of just over 2 km^2 in 1851; it was then amalgamated into Aston in the 1850s, a much larger parish that had an area of 57 km^2 and a population that grew rapidly to around 250 000 in the late 19th century. Water Orton was then re-formed as a separate parish in the 1890s with close to its original boundaries and a population that was still under 2000 in 1951. Sherbourne parish is a rural parish a few kilometres south-west of Warwick. It was a very small parish with a population of only 124 in 1951. It was affected by boundary changes in the 1910s and the 1850s. The parish of Birmingham provides a contrasting example. In 1851, this was a tiny administrative area covering less than 1 km^2 and with a population of only 919. Over the century that followed, the parish

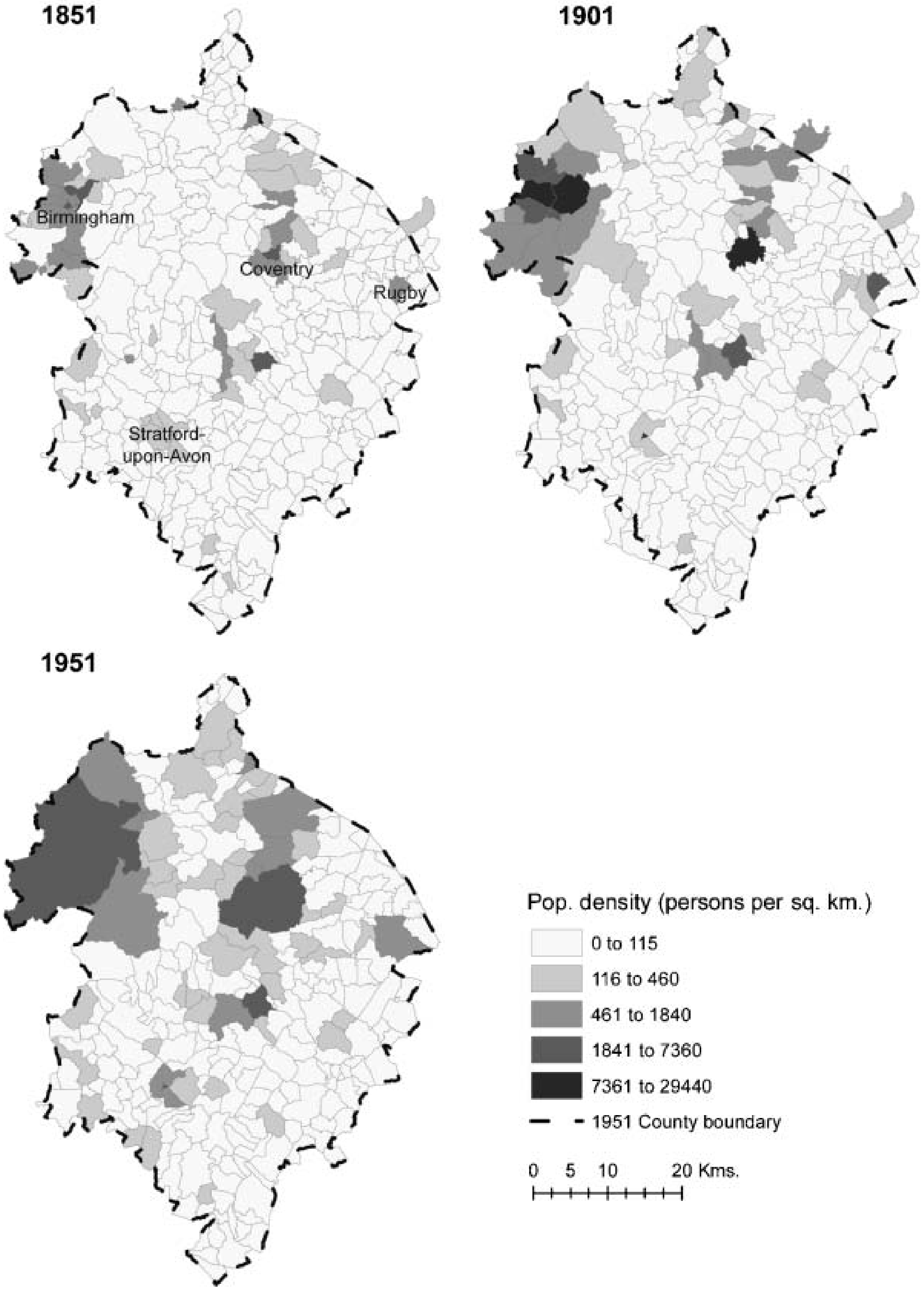

Figure 3. Population density in Warwickshire: 1851, 1901, and 1951 (source: printed census reports). Note that parishes intersecting the 1951 Administrative County of Warwickshire have been included; parishes that were part of Warwickshire but lie outside the 1951 boundary have been excluded. The legend uses a geometric progression from 115 persons km^2 increasing by a factor of 3. The maximum population density was 28 113 persons km^2.

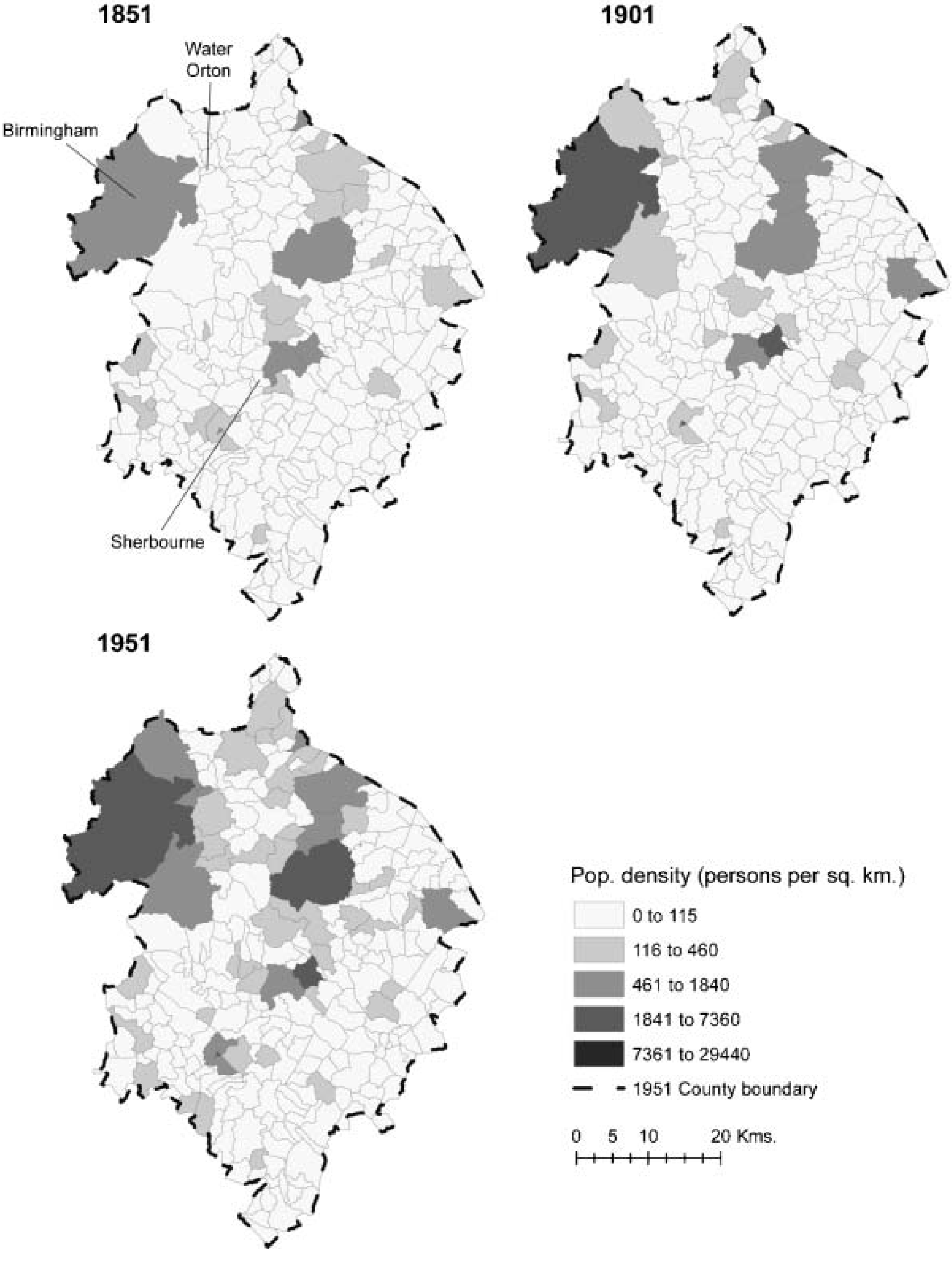

Figure 4. Population densities in Warwickshire: 1851, 1901, and 1951, interpolated onto 1951 parish boundaries. Class intervals are the same as those used in figure 3.

grew in response to the rapid urban growth of the city of Birmingham, such that by 1951, the parish covered an area of over 200 km^2 and had a population of 1.1 million. This growth was achieved by population growth in the city combined with a constant series of boundary changes and amalgamations that expanded the parish boundary. The total interpolated population and inter-censal population changes are shown in figures 5–7.

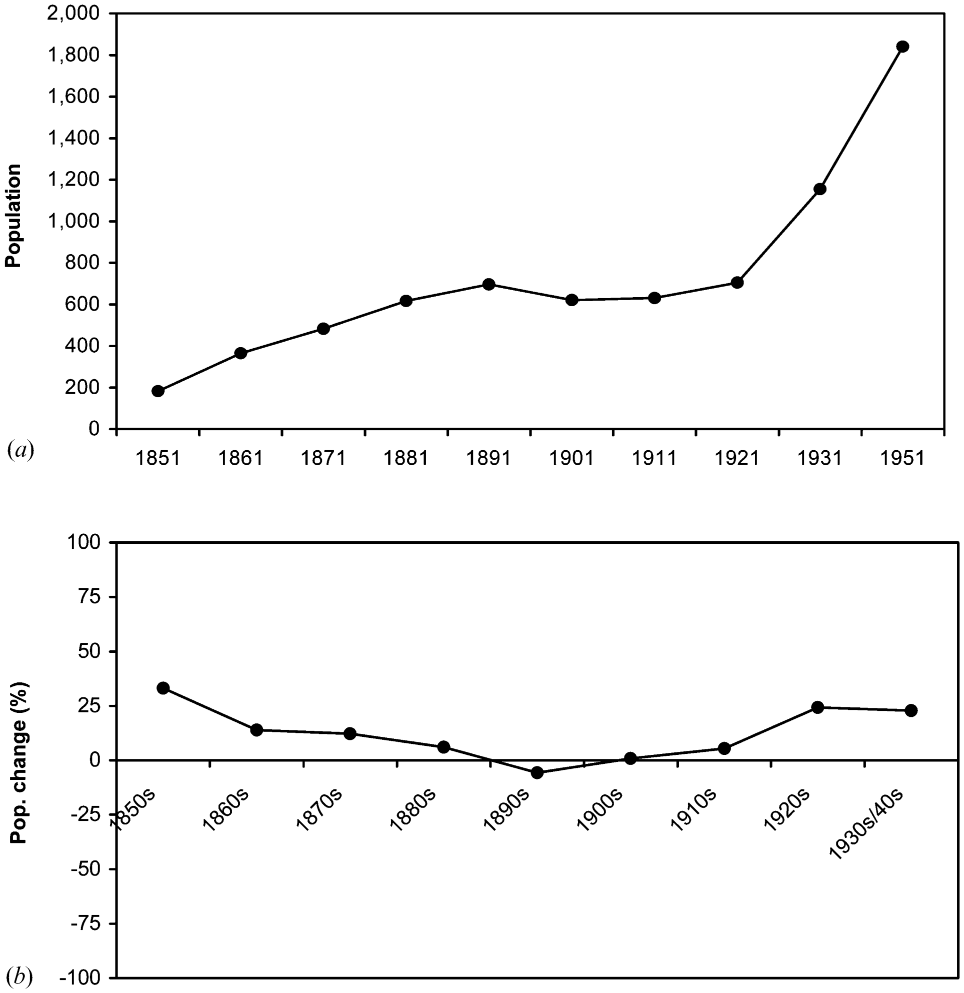

Figure 5. Total population and population change in Water Orton target unit, 1851–1951. (*a*) Total population. (*b*) Population change (normalized percentage).

Table 2 shows the population changes, spike sizes and change in changes, and *z*-scores for the three parishes. In Water Orton two spikes occur, one in the 1890s at the same time as a major boundary change and one in the 1920s when there is no boundary change. The spike sizes are, however, quite small, at −6.7 and 1.4 percentage points, respectively (see equation (4)). The change in changes are −19.3 and −1.4 percentage points for the 1850s and 1930s/1940s, respectively (see equation (5)). The spike sizes give *z*-scores of only −0.89 and 0.09, respectively (equation (6)), suggesting that these are well within the acceptable range, while the change in changes also give low *z*-scores of −1.50 and −0.44, respectively. From this, we can conclude that boundary changes are not contributing any noticeable error to the interpolated data and that other values where there are no boundary changes also seem robust.

The time series of population changes for Sherbourne, shown in figure 6(*b*), contains three spikes: in the 1860s there is a spike of 1.9 percentage points, in the 1900s there is a spike of −3.5 percentage points, and in the 1910s there is a spike of

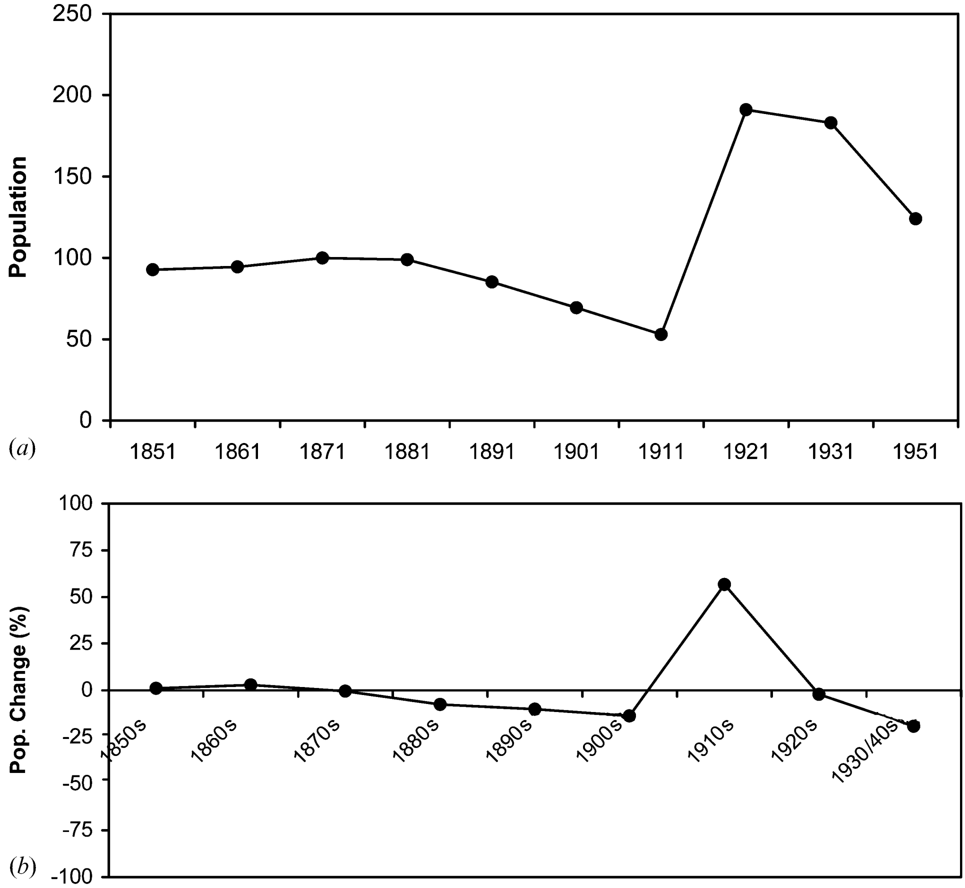

Figure 6. Total population and population change in Sherbourne target unit, 1851–1951. (*a*) Total population. (*b*) Population change (normalized percentage).

58.8 percentage points. In addition, the 1850s change in change is 1.9 percentage points, and the 1930s/1940s change is −17.1 percentage points. Table 2(*b*) shows how these are used to calculate *z*-scores. For the 1860s and 1900s, these are 0.15 and −0.50, respectively. There is no boundary change in either of these years and no reason to be suspicious of these values, as they have very low *z*-scores. The 1910s, however, where there is a boundary change, show a *z*-score of 7.00, a highly unusual value that suggests that large amounts of error are likely to have been introduced as a result of the interpolation process. Neither of the *z*-scores for the change in changes arouses any suspicion. From these results, therefore, we can take it that the boundary change in the 1850s has not introduced significant error into the estimates of total population for 1851. The 1910s boundary change, however, almost certainly has introduced error, and thus all of the estimates of total population from 1911 to 1851 seem to be seriously under-estimated.

Figure 7 shows the interpolated total population and population growth of the Birmingham target unit. Although there are a large number of spikes on this distribution, they are all small with one key exception: the 1850s change in change which, as table 2(*c*) shows, has a *z*-score of −2.80, and is thus unlikely to have occurred randomly. It must be noted, however, that given the rapid growth of

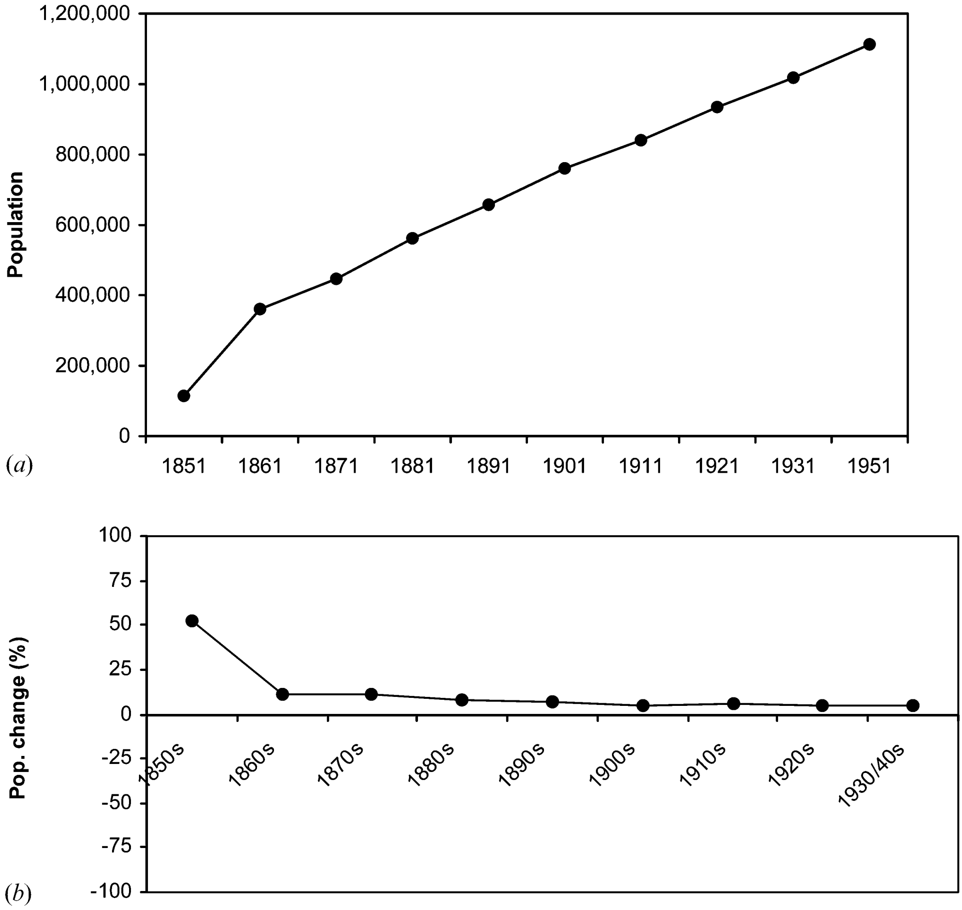

Figure 7. Total population and population change in Birmingham target unit, 1851–1951. (*a*) Total population. (*b*) Population change (normalized percentage).

Birmingham around this time, it is possible that this value represents a genuine population increase, but the technique draws attention to the fact that it is at the least suspicious. Although it seems likely that there has been an error in calculating the 1851 population, we can have a high degree of confidence in the other values.

If required, this approach could be used to provide a global summary of how well an interpolation has worked. One approach would be to simply perform a *t*-test to compare the natural sample of spikes and change in changes with the suspicious sample to give a global measure of the similarity between the two samples. A more powerful approach is to take an arbitrary cutoff, such as *z*-scores above 1.96 or below -1.96, where boundary changes are known to have occurred, and explore a summary of these. Table 3 shows this for the parishes in Warwickshire for every 20 years giving the number of parishes affected by boundary changes that have *z*-scores in this range. The total number of target units with a suspicious spike size or change in change between each year and 1951 is shown, along with this as a percentage of the total number of target units and their population as a percentage of the total population of Warwickshire. What emerges from this is that it is clear that for most years, the interpolation has led to errors in target units with small populations, and

Table 2. Evaluating potential error for the parishes of: (*a*) Water Orton, (*b*) Sherbourne, and (*c*) Birmingham[a].

	(*a*) Water Orton			(*b*) Sherbourne			(*c*) Birmingham		
Decade	Population change	Spike/ C. in C.	*z*-score	Population change	Spike/ C. in C.	*z*-score	Population change	Spike/ C. in C.	*z*-score
1850s	33.2	−19.3	−1.50	0.9	1.9	−0.24	51.9	−41.3	**−2.80**
1860s	13.9	–	–	2.8	1.9	0.15	10.6	−0.8	−0.18
1870s	12.2	–	–	−0.5	–	–	11.4	0.8	0.02
1880s	6.0	–	–	−7.6	–	–	7.8	–	–
1890s	−5.8	−6.7	−0.89	−10.1	–	–	7.3	–	–
1900s	0.9	–	–	−13.6	−3.5	−0.50	5.0	−0.4	−0.13
1910s	5.5	–	–	56.7	58.8	**7.00**	5.4	0.4	−0.03
1920s	24.3	1.4	0.09	−2.1	–	–	4.2	0.3	−0.04
1930/ 1940s	22.9	−1.4	−0.44	−19.2	−17.1	−1.37	4.5	0.3	−0.34

[a] The columns contain population change (in normalized percent), the spike (1860s to 1920s) or change in change (1850s and 1930/1940s) size (in percentage points) and *z*-scores for each parish. Dashes indicate that the population change was not a spike. Suspicious *z*-scores are shown in bold; both of these coincide with boundary changes.

as the total population affected is small, it may be felt that these errors are acceptable. The difficulty comes with the 1851 data. These have noticeably more suspicious values than later dates and the total population affected by suspicious results has increased to 63%. This is largely because of Birmingham's suspicious result which, as Birmingham has such a large population, has seriously affected the percentages.

5. Implications for research

The key implication of this work is that it allows us to use national historical GISs to produce long-term time-series of interpolated data and then identify which individual data values we believe may have had significant amounts of error introduced to them as a result of the interpolation process. A by-product of the methodology is that we may also identify errors in data values that are not due to interpolation but are nevertheless important. The key output of the methodology is that it provides each individual data value with an indication of the degree of

Table 3. Suspected error in interpolated parish populations in Warwickshire[a].

Year	Suspicious values	Percentage of total targets	Percentage of total population
1931	11	4.5	3.5
1911	12	4.9	1.4
1891	13	5.3	2.2
1871	13	5.3	2.1
1851	31	12.7	62.7

[a] The total population from 1851 to 1931 has been interpolated onto 245 target zones. Suspicious values are those where there is a spike or change in change with a *z*-score of more than ±1.96. These are expressed as a percentage of the total number of target parishes and their population as a percentage of the total population of Warwickshire.

confidence that we can have in it. This means that we can now identify where and when error may be occurring. How we then handle the error is a further issue and may depend on the purpose to which the data are put.

Traditionally, where interpolated data are used, there have been two main strategies for handling error. The first has been simply to ignore it, and the second has been to produce caveats on either the dataset or on the results of the research that use it. In many ways, this second approach is as bad as the first. If interpolated data are disseminated, the caveats give the user little or no idea of what impact the interpolation error may have on their research. While accompanying documentation may state that the data must be used with caution, this is an abdication of responsibility on behalf of the data provider, as the user will have little idea as to what this means and is thus likely to have little choice but to ignore any error except perhaps where it has an obvious impact on the results. Only by providing an indication of uncertainty at the level of the individual data values can we work round this problem.

Many issues still remain. The key one is how to deal with error once we have identified where we believe it may be occurring. The answer to this will depend on the purpose to which the interpolated data are to be put and will be a combination of two approaches: the first is to minimize the amount of error in the interpolated dataset, and the second is to explicitly handle the error in any subsequent analyses or visualizations.

To minimize the error, global summaries such as those presented in table 3 can be used to evaluate issues such as choice of target zones and choice of techniques and ancillary data to be used. At a local level, it allows the researcher to explore particular target areas that are found to be seriously error-prone and perhaps simply aggregate these areas into neighbouring areas and redo the interpolation. While this is undesirable in some ways as it leaves the dataset open to the allegation of gerrymandering, it is often the case that the highest rates of error are found in very small target units that may not be important to an analysis.

Highlighting uncertain data values also opens the option of performing a secondary interpolation. In this, spikes are smoothed out of time-series graphs by allocating population to adjoining target units which share a common source unit with the suspect target unit. This needs to be done in an iterative manner to ensure that the problem is not simply moved from one unit to another, but it is certainly possible to devise techniques that do this. Using the example of Sherbourne shown in figure 6, this would ensure that more data are allocated to Sherbourne target parish in 1911 than before at the expense of the adjacent parish or parishes with which there have been boundary changes.

However well the interpolation is performed, there will still be errors in the results, so any analysis that is performed with these data needs to address this explicitly. Although there have been many calls for error-sensitive GISs (for example, Unwin 1995), little work in this area to date has focused on subjects such as error-sensitive spatial analysis or error-sensitive visualizations of polygon data. Much work remains to be done in this area, but some obvious basic ways to explicitly handle error in analysis include exploring the relationship between values suspected of containing error and residuals and outliers found in an analysis. Where local analysis techniques are used (Fotheringham 1997), spatial variation in error can be compared with spatial variations in the results.

Visualization through maps is in many ways more complicated involving complex issues of perception and users' understanding, but it is clear that this is an issue that must be addressed. Lessons learned from this would have the potential to be applied to any areas where there is uncertainty in the data being mapped, not simply where interpolated data are used.

6. Conclusions

This paper has explored one approach to identifying error in areally interpolated data. It is most applicable in historical GIS which is a rapidly growing field and one in which there have been large amounts of investment in many countries (Ell & Gregory 2001, Knowles 2005b). If this investment is to reap its just rewards, techniques such as this are essential when using areally interpolated data. A particular danger is that error-filled data will be used inappropriately to the discredit of the field as a whole. The methodology described here allows researchers to avoid this mistake.

Variations on the methodology should also be more widely applicable. Significant effort is being put into comparing 'modern' censuses, usually those from 1971 on in Great Britain (e.g. Martin *et al.* 2002, Martin 2003, Norman *et al.* 2003), and there are clearly implications from this methodology that can be used in this type of work. In addition, although the paper has concentrated on decennial census data, the methodology could equally well be used on annual data or data published at any other interval.

This methodology works well with types of data that tend to follow broad trends. Many census data are of this nature, but for other forms of data, such as those concerned with epidemic diseases, they may be less effective. There is also the problem that when multiple variables are to be interpolated, each variable is likely to need to be tested separately. For variables that are likely to have a very similar distribution, such as males aged 35–44 and males aged 45–54, this may be excessively cautious, but for others, such as differing employment types including both manufacturing and agriculture, it will be essential.

The method is not able to identify whether a specific data value is definitely an error; it merely highlights those that appear suspicious. There will be cases where 'genuine' data are highlighted as suspect and where data with errors will escape unnoticed. Nevertheless, the case study of Warwickshire parishes, an example specifically chosen to be challenging, illustrates that it is effective at highlighting suspect data values. This is, therefore, a major step forward in identifying interpolation error that represents a significant improvement from global measures such as those of Simpson (2002) and Gregory and Ell (2005). We are now in a position to make appropriate use of areal interpolation within national historical GISs and similar systems. This in turn means that we will be increasingly able to use GIS to unlock a new understanding of long-term change in British and other societies.

Acknowledgements

This work benefited from funding from the Economic and Social Research Council's Research Methods Programme (Ref. 333250016) and from a Leverhulme Early Career Fellowship awarded to Ian Gregory (Ref. ECF/40115). The 1851 spatial data for Warwickshire are from N. Burton, J. Westwood, and P. Carter *Ancient Parishes of England and Wales, 1500–1850* distributed by AHDS History, UK Data Archive, Study Number 4828.

References

Bol, P. and Ge, J., 2005, China Historical GIS. *Historical Geography*, **33**, pp. 150–152.

Bracken, I. and Martin, D., 1995, Linkage of the 1981 and 1991 UK census using surface modeling concepts. *Environment and Planning A*, **27**, pp. 379–390.

Cockings, S., Fisher, P.F. and Langford, M., 1997, Parameterization and visualisation of the errors in areal interpolation. *Geographical Analysis*, **29**, pp. 314–328.

De Moor, M. and Wiedemann, T., 2001, Reconstructing Belgian territorial units and hierarchies: An example from Belgium. *History and Computing*, **13**, pp. 71–97.

Dempster, A., Laird, N. and Rubin, D., 1977, Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **39**, pp. 1–38.

Ell, P.S., 2005, Towards a comprehensive GIS for Ireland. *Historical Geography*, **33**, pp. 138–140.

Ell, P.S. and Gregory, I.N., 2001, Adding a new dimension to historical research with GIS. *History and Computing*, **13**, pp. 1–6.

Fisher, P.F. and Langford, M., 1995, Modelling the errors in areal interpolation between zonal systems by Monte Carlo simulation. *Environment and Planning A*, **27**, pp. 211–224.

Flowerdew, R. and Green, M., 1994, Areal interpolation and types of data. In *Spatial Analysis and GIS*, A.S. Fotheringham and P.A. Rogerson (Eds), pp. 121–145 (London: Taylor & Francis).

Fotheringham, A.S., 1997, Trends in quantitative methods I: Stressing the local. *Progress in Human Geography*, **21**, pp. 88–96.

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

Goodchild, M.F., Anselin, L. and Deichmann, U., 1993, A framework for the areal interpolation of socio-economic data. *Environment & Planning A*, **25**, pp. 383–397.

Gregory, I.N., 2002a, Time variant databases of changing historical administrative boundaries: A European comparison. *Transactions in GIS*, **6**, pp. 161–178.

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

Gregory, I.N., 2005, The Great Britain Historical GIS. *Historical Geography*, **33**, pp. 136–138.

Gregory, I.N., Bennett, C., Gilham, V.L. and Southall, H.R., 2002, The Great Britain Historical GIS Project: From maps to changing human geography. *Cartographic Journal*, **39**, pp. 37–49.

Gregory, I.N. and Ell, P.S., 2005, Breaking the boundaries: Integrating 200 years of the Census using GIS. *Journal of the Royal Statistical Society, Series A*, **168**, pp. 419–437.

Knowles, A.K. (Ed.), 2005a, Reports on National Historical GIS Projects. *Historical Geography*, **33**, pp. 134–158.

Knowles, A.K., 2005b, Emerging trends in historical GIS. *Historical Geography*, **33**, pp. 7–13.

Langford, M., Maguire, D. and Unwin, D.J., 1991, The areal interpolation problem: Estimating population using remote sensing in a GIS framework. In *Handling Geographical Information: Methodology and Potential Applications*, I. Masser and M. Blakemore (Eds), pp. 55–77 (Harlow, UK: Longman).

Martin, D., 2003, Extending the automated zoning procedure to reconcile incompatible zoning systems. *International Journal of Geographical Information Science*, **17**, pp. 181–196.

Martin, D., Dorling, D. and Mitchell, R., 2002, Linking censuses through time: problems and solutions. *Area*, **34**, pp. 82–91.

McMaster, R.B. and Noble, P., 2005, *The U.S. National Historical Geographical Information System, Historical Geography*, **33**, pp. 134–136.

MERZLYAKOVA, I.A., 2005, Historical GIS initiative in Russia. *Historical Geography*, **33**, pp. 147–149.

NORMAN, P., REES, P. and BOYLE, P., 2003, Achieving data compatibility over space and time: Creating consistent geographical zones. *International Journal of Population Geography*, **9**, pp. 365–86.

OPENSHAW, S., 1984, *The Modifiable Areal Unit Problem. Concepts and Techniques in Modern Geography 38* (Norwich, UK: Geobooks).

REIBEL, M. and BUFALINO, M.E., 2005, Street-weighted interpolation techniques for demographic count estimation in incompatible zone systems. *Environment and Planning A*, **37**, pp. 127–39.

SADAHIRO, Y., 1999, Accuracy of areal interpolation: A comparison of alternative methods. *Journal of Geographical Systems*, **1**, pp. 323–346.

SADAHIRO, Y., 2000, Accuracy of count data transferred through the areal weighting interpolation method. *International Journal of Geographical Information Science*, **14**, pp. 25–50.

SIMPSON, L., 2002, Geography Conversion Tables: A framework for conversion of data between geographical units. *International Journal of Population Geography*, **8**, pp. 69–82.

UNWIN, D., 1995, Geographic Information Systems and the problem of 'error and uncertainty'. *Progress in Human Geography*, **19**, pp. 549–558.

Vision of Britain Through Time 2005, Available online at: http://www.visionofbritain.org.uk (accessed 25 April 2005).