Journal of the
Royal Statistical Society
SERIES A
Statistics
in Society

J. R. Statist. Soc. A (2013)
176, *Part* 3, *pp.* 655–678

Spatial vote redistribution in redrawn polling units

Jose M. Pavía and Antonio López-Quílez

University of Valencia, Spain

[Received June 2011. Final revision April 2012]

Summary. A large proportion of electoral analyses using geography are performed on a small area basis. In each new election there are always modifications to the previously existing polling units. The use of past voting results in small area aggregate data electoral forecasting models and political analyses therefore requires establishing a correspondence between old and new polling units. Traditionally, the task of tracking changes to assign an electoral history to the new units properly has been carried out by hand, comparing unit codes and census figures. This is an extremely cumbersome task that cannot always be performed, as when a massive (geographically intense) reorganization of polling unit boundaries takes place. Nowadays, however, assisted by the increasing availability of geographical data, this chore could be easily automated and even improved with the help of spatial statistical software. The paper suggests several methods for allocating votes by using geographical information systems tools and shows the effectiveness of spatial strategies. These approaches will permit electoral pollsters and forecasters to solve the issue efficiently and to apply the most successful electoral forecasting techniques that are currently in use and will help electoral geographers with the problem of comparing spatial aggregate electoral data from different elections. The relevance of the analysis, nevertheless, goes beyond electoral data, as the reallocation of data from one set of administrative units onto another arises in many applications. The geometric approach is proposed as a natural substitute for the classical approach and three additional approaches (centroid, surface and compositional) are also suggested, exploiting the spatial patterns that electoral outcomes display. The relative performance of the various methods is assessed in three real data instances. The results suggest that the surface approach, which obtains past voting outcomes in each polling unit by averaging their vote proportion interpolations, is the most suitable procedure.

Keywords: Boundary changes; Electoral forecasts; Geographic information systems; Lattice; Modifiable areal unit problem; Small size electoral analysis; Spatial patterns

1. Introduction

Previous electoral results play a key role in electoral analyses. They are used by political journalists and political parties' teams to perform quick evaluations of outcomes, by political researchers and electoral geographers to complete more detailed analyses and by electoral pollsters and forecasters to predict electoral results. In fact, a great variety of models proposed in the literature to explain and predict electoral outcomes use previous results as an explanatory variable (e.g. Clark (2009) and Curtice and Firth (2008)). Between elections, however, there are always changes in the composition and magnitude of the electorate; therefore, theoretically, previous and current electoral outcomes cannot be directly compared. Even so, as ballot secrecy conceals individual votes and great changes do not usually occur between consecutive elections, empirical analyses in political science and electoral forecasting ordinarily assume that the same or equivalent populations are voting in both elections when carrying out aggregate data comparisons.

Address for correspondence: Jose M. Pavía, Departamento de Economía Aplicada, Facultad de Economía, Universidad de Valencia, Campus Els Tarongers, 46022 Valencia, Spain.
E-mail: pavia@uv.es

 0964–1998/13/176655

Sometimes, however, major shifts take place between elections (such as during the 10-yearly redrawing of state legislative and congressional district boundaries following the US census) and the hypothesis of stationary electorates cannot be assumed. In these cases, analysts must track any changes to assign an electoral history to its units of analysis properly. If electoral analyses are performed on a large scale (e.g. Burden and Kimball (1998) and Kim *et al.* (2003)), using consolidated administrative units (such as cities, counties and provinces) or large districts and constituencies where major shifts are uncommon, it is usually not too complex a task to track votes. In the worst case, to obtain a practical approximation of the previous electoral results for each new redrawn district, it is generally sufficient to consider the most disaggregated level of electoral data available (polling units) and to add or subtract the votes of respectively the new or old polling units included or removed to reach an estimate.

The geography of elections varies from country to country and between elections in the same country. Therefore, given the great variety of names that electoral units receive, it is useful to fix some terms. Whatever the country, electoral authorities distribute voters by using a geographic administrative hierarchical structure. The electorate is split into constituencies (the smallest geographical unit for which representative(s) are elected) and those again into usually several levels of smaller units (precincts, wards, polling districts or sections, voting locations, polling stations and even ballot boxes). Depending on the country, however, election results are counted and declared at different levels. So, in this paper, the term polling unit will denote the smallest geographical unit for which voter data are available (e.g. wards in the UK or precincts in the USA).

A large proportion of electoral analyses using geography are performed on a small area basis such as polling units. Electoral geographers and forecasters are regular users of this kind of data. On the one hand, since Cox's seminal work (Cox, 1969) recycled Key's ideas (Key, 1949) and conjectured that, in addition to personal characteristics, people's political behaviour was influenced by their social contacts and place was also recognized as an important dimension to understand voting decisions (Agnew, 1987), many studies have attempted to provide evidence of this by exploiting the geography of electoral outcomes (e.g. Pattie and Johnston (2000), Macallister *et al.* (2001) and Khofeld and Sprague (2002)). On the other hand, some of the most successful techniques for predicting electoral outcomes rely on small area data to make projections (e.g. Bernardo (1984), Mitofsky and Edelman (2002), Pavía (2010) and Curtice *et al.* (2011)).

Unfortunately, the complexity of the problem of matching old and new polling units escalates as we go down the scale of elector aggregation. Shifts are more frequent in small polling units but the smaller the polling units the more difficult it is to track changes. These problems have bothered analysts for decades. The importance and difficulty of establishing matches has been acknowledged since the very beginning of electoral polling and still prevents researchers from using the preferred approaches as frequently as desired. For example, in the report that analysed the failure of the 1948 US Presidential pre-election polls, it was admitted that taking representative or pinpoint sampling 'in areas selected on the basis of past voting history' would have been a really accurate option for sampling; but they were 'employed on a limited scale . . . [b]ecause of shifting election boundaries, and the difficulty of defining them in many communities', which made pinpoint sampling 'almost impossible to apply in many states' (Mosteller *et al.* (1949), page 341). Nevertheless, despite the great difficulties that are associated with the issue, many electoral researchers and political agents need to have these matches to perform their analyses properly. Hence, regardless of the complexities involved, the issue has been frequently addressed—especially within the election forecasting framework (e.g. Bernardo (1997), Pavía-Miralles (2005) and Kyle *et al.* (2007))—employing plenty of patience, common sense, expert judgement and, exceptionally, heroic assumptions. The mapping of relationships between prior

polling units and current units has been traditionally constructed by hand, comparing prior and current geographical administrative codes and census figures based on features such as relative location and number of voters. This strategy cannot always be performed, as when a massive (geographically intense) reorganization of polling unit boundaries takes place.

Fortunately, assisted by the increasing availability of geographical data that electoral authorities are offering to the public, which provides the boundary, code, name information and maps of voting units, the criteria that are followed to undertake this cumbersome task can be automated with the help of geographic information system and statistical spatial software. What is more, taking into account the spatial patterns that electoral results clearly show (e.g. O'Loughlin (2002), Sui and Hugill (2002) and Pavía *et al.* (2008)), spatial strategies could improve the process of redistributing votes and also provide solutions when the manual approach is not affordable.

The rest of the paper is organized as follows. Section 2 sets out the criteria that are traditionally followed to assign previous election results in polling units. Section 3 suggests a strategy to automate the ideas outlined in the previous section and offers several options exploiting the spatial auto-correlation of electoral outcomes. In Section 4, the procedures proposed are applied to Goteborg (Sweden) and Barcelona (Spain), where a complete reorganization of polling units has taken place recently, and also to Västra Götalands läns (Goteborg county), and their relative performance is compared. Finally, Section 5 summarizes and discusses findings.

2. Classical approach: matching polling units manually

As a result of population shifts and administrative decisions, at each new election there are nearly always modifications in the previously existing small area polling units. New polling units arise due to creations, fusions, divisions and the reorganization of existing units and, at the same time, changes also occur in the polling units that are apparently stable due to incoming and outgoing voters. In these circumstances, a series of reasonably grounded empirical rules has been proposed to establish matches between old and new polling units, making it easier to incorporate past historical outcomes into political and electoral small area applications.

Following Pavía-Miralles (2005), pages 1117–1118, the basic rules can be summarized as follows:

(a) a direct match is established between polling units that have apparently not changed under the assumption that the relatively small number of entrances and exits in their voter lists are random;
(b) when either two or more units are combined to create one (or more) new unit(s), the aggregate outcomes of the original units are considered as historical data for the new unit(s);
(c) for those new units which stem from the division of a previously existing unit, the vote proportions of the original unit are assigned as their past vote proportions;
(d) either neighbourhood, city or even constituency average vote proportions are assigned as historical data for new (or practically new) polling units, because they are usually in the expansion areas of the cities.

The small list of rules described above generally makes it possible to assign a past to each unit by using only polling unit identifiers (codes, names, etc.) and elector figures. Almost all the polling units in almost all elections can be manually matched by using exclusively these data and with a high level of confidence. From time to time the cases that analysts face are not as simple and clear and more than one compatible match is plausible, raising doubts regarding how to establish the correspondences (for example, there are cases in which significant redistributions

of voters occur between polling units, situations in which a disappearing unit splits its electors between two or more units, several redistributions taking place at the same time in the same area making voter tracking more difficult, some units that disappear from one municipality turn up in another, or even some cities crossing a constituency border). Fortunately, this situation usually only occurs for a negligible percentage of units and, in these cases, the analyst's judgement, along with the implementation of some simplifying assumptions (and maybe the use of some extra information such as the postal addresses of the buildings in which voting occurs) is enough to produce accurate distributions. Applied with flexibility, these rules have been found to be empirically sound and useful in a large number of elections.

The problem arises when a complete redrawing of polling unit boundaries takes place in a geographically concentrated number of units—such as occurred before the 2006 Riksdag elections with the districts in Goteborg (Sweden) and in 2009 with the sections in Barcelona (Spain). (Districts and sections are the names that are used by Swedish and Spanish electoral authorities respectively for their smallest polling units and they are the equivalent of American precincts.) In these cases, owing to the massive reorganization of polling units, it is impossible to allocate properly the votes in a considerable number of the new units exclusively of the foregoing data. Therefore, unless we agree on using some broad and perhaps not very realistic assumptions (such as extending rule (b) to the whole area), extra information is absolutely necessary to track changes properly.

So far, the spatial component of polling units has not been exploited explicitly. However, each polling unit is unequivocally related to an area in space (called a polygon in geographical information system terminology). Therefore, an alternative way to establish the correspondence between old and new units could be based on comparing the depiction on a map of the polygons in both elections (Fig. 1). This approach could provide a solution for this issue in the case of voting area shapes being drastically redrawn.

To illustrate how this approach could be put into practice, Fig. 1 (which depicts the polygons of the polling units for both the 2002 and the 2006 Riksdag elections for the same area of Goteborg) will be used. As can be observed in Fig. 1, a profound restructuring of the spatial polling map of Goteborg took place between the 2002 and the 2006 elections and, although the numbers of polling units in both elections were quite similar (286 in 2002 and 279 in 2006), no satisfactory solution can be provided to track the votes of the new units using exclusively the codes of the units and their number of electors. This chore could be accomplished under

Fig. 1. Extract of the Goteborg (Sweden) division in polling units for (a) the 2002 and (b) the 2006 Riksdag elections: the same area is depicted in both figures and the same 2006 polygon is shaded in both figures

the assumption of votes uniformly distributed in each polling area with the help of Fig. 1. In particular, focusing on the 2006 polling unit that is shaded in both maps, it is observed that this unit has its roots in four polling units in the 2002 elections (labelled A, B, C and D in Fig. 1) and covers approximately half of the areas of the former A and B units, a third of the C unit and around four-fifths of the D unit. These fractions, by virtue of the above assumption, would be automatically translated into polling unit figures and would make it possible, once combined, to reach an approximation of the past electoral results for this new polling unit.

3. Spatial-based methods

The previous four rules along with the last extension based on a map representation of polling units allows us to generate approximations of past voting behaviour for every polling unit provided that the required information is available. This task, however, is extremely cumbersome if performed manually. Fortunately, if the files providing the polygon attribute tables are available, the area that is shared by each new and old polling unit can be determined accurately with the help of spatial software and the operations that are required to assign previous results to every voting unit can be computerized (e.g. by using packages such as `maptools`, `sp` or `spdep` of the well-known free statistical software R; see the Web page of the spatial task view, `http://cran.r-project.org/web/views/Spatial.html`). This approach would only require the use of the spatial commands of intersection and union to yield results and is the simplest form of areal interpolation proposed in the literature (Gregory and Ell, 2005).

Human societies, however, are not arranged in a statistically independent manner (O'Loughlin, 2002). Housing and labour market operations tend to produce social differentiation between areas. Moreover, the socio-economic effects of policies and political actions vary across space given the impression that perceptions and opinions change depending on location (Johnston and Pattie, 2006). Therefore, as both individual backgrounds and local contexts interact to determine the political behaviour of voters (Pavía *et al.*, 2008), some geographical structure and spatial patterns are obtained from election results. Despite this, the previous approach (which is hereafter referred to as the geometric approach) ignores these facts. It assumes the same distribution of votes in every subarea of the polling unit and does not take into account the types of voters who are being shuffled in and out of the area. Furthermore, it implicitly assumes a uniform distribution of voters in the polling unit, when, in the same way as other social variables, population density and voter turnout record geographical trends. To adjust for these features, three other methods, also founded on geographical considerations, are suggested in this paper. All these strategies are point-based approaches and use spatial interpolation as a basis but differ in the number of points and variables that they interpolate. The first alternative (from now on referred to as the centroid approach) identifies each polling area by its geometric point centre, the centroid, and uses the values that are assigned to the centroids of old polling units to interpolate the values corresponding to the centroids of new polling units. The second alternative (the surface approach) extends interpolation to every point of the surface and obtains an approximation by averaging the values interpolated in all the points of the polling unit. Finally, the third option (called the compositional approach) uses a weighted voter density approach. In this case, voter density is also interpolated at each point and used as a weighting variable. The interpretation and details of these four approaches (geometric, centroid, surface and compositional) are provided in the following subsections and some critical comments about potential refinements are discussed in the final subsection.

3.1. Geometric approach

Any division of the electoral space into polling units is (see, for example, Fig. 1, and Fig. 2 in Section 4) a partition of a two-dimensional set into non-overlapping and non-empty elements that cover the whole set. Given two different partitions, any element B of the second partition could be expressed unequivocally by the union of its intersections with all the elements of the first partition. The geometric approach assumes a uniform spatial distribution of votes within each polling unit, so the area that is shared between each new and old polling unit is used to determine the proportion of votes that are shuffled from each old unit to each new unit. More specifically, let $D_1, D_2, \ldots, D_m$ be the polygons of the m units of the previous elections and let $B_1, B_2, \ldots, B_n$ be the polygons of the n units of the current elections. Let $v_{1j}, v_{2j}, \ldots, v_{pj}$ be the votes counted for each of the p political options in D_j at the previous election. Then, given that

$$B_i = \bigcup_{j=1}^{m} (B_i \cap D_j),$$

the geometric estimation of the past vote proportion for the kth political option in the ith new unit, $\hat{p}_{ki}^{\mathrm{G}}$, is obtained from

$$\hat{p}_{ki}^{\mathrm{G}} = \frac{\sum_{j=1}^{m} v_{kj} |B_i \cap D_j| |D_j|^{-1}}{\sum_{h=1}^{p} \sum_{j=1}^{m} v_{hj} |B_i \cap D_j| |D_j|^{-1}} = \frac{\sum_{j=1}^{m} v_{kj} \omega_{ij}}{\sum_{h=1}^{p} \sum_{j=1}^{m} v_{hj} \omega_{ij}}, \tag{1}$$

where $|D|$ represents the area of polygon D and ω_{ij} is the fraction of polygon D_j that intersects with polygon B_i.

Although this approach was initially proposed to overcome the difficulties that are posed by a massive modification of district shapes, it could be employed to generate historical values for every polling unit automatically. This does not imply abruptly breaking away from the classical approach. This method is quite loyal to classical rules. Indeed, except for rule (d) in Section 2, both the classical and the geometric approach will produce basically the same results.

3.2. Centroid approach

The geometric approach offers an intuitive solution to solve the problem in the spirit of the classical approach but misses the point with its discrete conceptualization of space. The spatial division of the electoral territory into polling units is only one of the possible partitions that could be implemented (see Figs 1 and 2). Dealing with continuity usually implies a point-based approach, and we are handling aggregate data with an areal reference. Hence, to bridge this gap, the centroid approach *transforms* area-based data into point-based data by identifying each small area unit with its geometric centre and constructs its estimates by regarding the proportions of votes as continuous processes in space. In particular, if C_j ($j = 1, \ldots, m$) represents the centroid of the polygon D_j and p_{kj} denotes the proportion of votes counted in the previous elections for political option k in unit j, the centroid approach considers the values p_{kj} as a sample (observed in the points $C_1, C_2, \ldots, C_m$) of the underlying processes and, by point interpolation, estimates past vote proportions, $\hat{p}_{ki}^{\mathrm{C}}$, through equation (2), in the centroids O_i (for $i = 1, \ldots, n$) of the new polling areas, which are used to represent the whole new units, i.e. the past vote proportions in each new ith polling unit are obtained as a weighted average, with weights λ_{ij}, of the vote proportions registered in the old polling units p_{kj}. The values of the assigned proportions depend on the weights λ_{ij} that are used:

$$\hat{p}_{ki}^{C} = \sum_{j=1}^{m} \lambda_{ij} p_{kj} \Big/ \sum_{j=1}^{m} \lambda_{ij}. \tag{2}$$

Several interpolation techniques have been proposed in the literature (Cressie, 1993): deterministic and geostatistical, global and local, exact and inexact. In this work, it was considered that a local, point-based respectful interpolator (a technique that predicts identical values at the sampled locations to those measured) should be used. Among these, the two major alternatives are the inverse-distance-weighted interpolation method and the kriging procedure, both of which generate weights from surrounding measured values to predict values at unmeasured locations, the closest measures having the most influence. Kriging weights come from a semivariogram that must be estimated by looking at the spatial structure of the data. The solution therefore depends on the analyst's skills and preferences. Hence, in this work, it was decided that the most automatic option would be used, provided by the inverse-distance-weighted interpolation method, to make the technique more accessible for the average analyst. The so-called inverse distance squared weighted interpolation (the inverse-distance-weighted interpolation with Euclidean distance) has been used (as default in ArcGIS® 9) as the interpolation technique (Johnston *et al.*, 2003), where the weight λ_{ij} measures the inverse of the Euclidean distance between the pair of centroids C_j and O_i.

The centroid $C = (x_c, y_c)$ of a polling unit polygon with vertices $(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)$—where $(x_0, y_0) = (x_n, y_n)$—and area

$$|A| = \frac{1}{2} \sum_{i=0}^{n-1} (x_i y_{i+1} - x_{i+1} y_i)$$

is obtained through

$$x_c = \frac{1}{6|A|} \sum_{i=0}^{n-1} (x_i + x_{i+1})(x_i y_{i+1} - x_{i+1} y_i)$$

and

$$y_c = \frac{1}{6|A|} \sum_{i=0}^{n-1} (y_i + y_{i+1})(x_i y_{i+1} - x_{i+1} y_i).$$

Extending this procedure to all units will entail significant changes with respect to the classical method. On this occasion, except for the case discussed in rule (a)—where both methods will generate basically the same estimates—the solutions of the classical and centroid approaches will be different. The results for rules (c) and (d) will more than likely improve with this method, whereas they will probably be worse for the cases that are discussed in rule (b).

3.3. Surface approach

The centroid approach makes approximations by interpolating at a unique point per polling unit, irrespective of what the polygon is like. However, there is no isomorphic match between polygons and centroids, and the same centroid could be obtained from different forms. Recognizing that propensity to support a particular party or candidate is not uniformly distributed in each voting unit, the surface approach takes into account the particular shape of the polling area as a way of ascertaining where voters are being shuffled in from. This approach is based on a raster representation (see, for example, chapter 3 in Longley *et al.* (2001)) of space and attains vote proportion estimates of past results, $\hat{p}_{ki}^{S}$, in polling unit i by averaging the point estimates on the whole surface of the polygon B_i, where the approximations of the vote proportions $\hat{p}_k(s)$ are reached by employing equation (2) at every point s of the electoral space, i.e. after obtaining *centroid* approximations for each point of the electoral space, the surface allocations are

obtained as the mean of these approximations in each polling unit:

$$\hat{p}_{ki}^{\mathrm{S}} = \int_{B_i} \hat{p}_k(s)\,\mathrm{d}s \Big/ |B_i|. \tag{3}$$

Although the number of interpolated pixels is usually very large, the raster image is always discrete. Therefore, the integral is approximated by a sum $\hat{p}_{ki}^{\mathrm{S}}$ obtained by

$$\hat{p}_{ki}^{\mathrm{S}} = f_i^{-1} \sum_{h=1}^{f_i} \hat{p}_k(s_h),$$

where f_i is the number of pixels in polygon B_i.

Using this strategy to assign past proportions to all voting areas would lead to results that are different from those obtained by way of the classical approach in all cases. This does not pose special concerns, except in case (a) where there is apparently no logical reason to abandon the rule. Nevertheless, in the rest of the cases, it seems reasonable to expect better allocations to be achieved with this approach. Thus, taking into account this likely trade-off and that usually the most common case in practice is that discussed in rule (a), the opportunity of extending this procedure to computerize the issue would be judged from empirical evaluations.

3.4. Compositional approach

The surface approach gives equal worth to all the point interpolations of the polling unit. However, not all unit subareas are equally populated. Hence, in line with Martin (1989) and Braken and Martin (1995), the compositional approach seeks to take advantage of this fact to improve assignments theoretically, albeit following a different strategy. In particular, using a procedure similar to that employed to attain the function $\hat{p}_k(s)$, a voter density function is derived for all the points of the electoral region being considered and used to weight vote proportion interpolations, i.e., by defining d_j as the voter density in the centroid C_j of the polling unit D_j, as given by equation (4), it follows that the compositional vote proportion estimates $\hat{p}_{ki}^{\mathrm{D}}$ are obtained through equation (5), i.e. the vote proportion allocations in each polling unit are obtained as a weighted average of the *centroid* interpolations at the points of the polling area by using the population densities at the points as weights:

$$d_j = |D_j|^{-1} \sum_{k=1}^{p} v_{kj}, \tag{4}$$

$$\hat{p}_{ki}^{\mathrm{D}} = \int_{B_i} \hat{p}_k(s)\,\hat{d}(s)\,\mathrm{d}s \Big/ \int_{B_i} \hat{d}(s)\,\mathrm{d}s, \tag{5}$$

where the function $\hat{d}(s)$ values are obtained by

$$\hat{d}(s) = \sum_{j=1}^{m} \lambda_j(s) d_j \Big/ \sum_{j=1}^{m} \lambda_j(s), \tag{6}$$

and $\lambda_j(s)$ measures the inverse of the Euclidean distance between the points C_j and s. Again, because of the raster discrete representation of the surface, the compositional estimates are, in practice, obtained by a sum:

$$\hat{p}_{ki}^{\mathrm{D}} = \sum_{h=1}^{f_i} \hat{p}_k(s_h)\hat{d}(s_h) \Big/ \sum_{h=1}^{f_i} \hat{d}(s_h).$$

Although surface and compositional approaches are different, equation (5) would collapse into equation (3) in the case of independence between turnout and party votes. As occurs with

the surface procedure, extending this method to obtain estimates in all polling units would produce approximations with similar properties but also differing from the allocations that would be obtained by way of the classical approach. Therefore, as in the previous approach, it should mainly be assessed empirically.

3.5. Potential refinements

The spatial-based procedures suggested are only intended to be a sample of the possibilities within a spatial framework. The spatial redistribution of votes is a particular case of the more general problem of reallocating data from a set of geographical administrative units onto another. It will be worth testing in this context other solutions that have been implemented in other frameworks (see Gregory and Ell (2005) and the references therein). A possible avenue of research to be explored in the future would be to study how the use of dasymetric mapping and related techniques would enhance the quality of approximations. Using ancillary sources of information, such as information about land uses (Flowerdew and Green, 1994), the spatial distribution of built structures (Longley and Mesev, 1997) or satellite imagery (Robinson *et al.*, 2002), would probably enhance the accuracy of mainly the geometric and centroid assignments.

Simple and logical improvements could be introduced in the above approaches for countries, like the UK or the USA, where census figures are available for geographical administrative areas that are below polling unit level, such as output areas or enumeration districts in the UK or block census in the USA.

For example, in the case of the UK, using the intersection function, the common area between each electoral ward and each enumeration district could be calculated and employed to estimate the spatial distribution of the electorate within each electoral ward by combining enumeration district population figures and ward electoral data and this information could subsequently be used to construct weighted versions of the above estimators. Likewise, given that US voting precinct borders 'always follow a census block boundary' (US Census Bureau (2000), chapter 8), block census figures could be employed to obtain a finer compositional approximation of each precinct by weighting (using census figures) within each precinct its block vote proportion and turnout interpolations attained after applying the surface approach. Obviously, these two examples do not exhaust the potential that population figures could offer in this context. Some of the proposals of Simpson (2002) and Martin (2003) could also be adapted as alternatives.

In our analyses, none of these population-based refinements were implemented because in both countries, Spain and Sweden, polling units are also the smallest geographical administrative areas for which population variables are published.

4. Redistributing votes: three illustrative examples

In this section the four procedures that were described in Section 3 are used in Västra Götalands läns (Goteborg's county) to assess the methods in a situation where the usual changes in polling units are made (see Section 4.3), whereas the cases of the cities of Goteborg in Sweden (Section 4.1) and Barcelona in Spain (see Section 4.2) are analysed as examples of profound reorganizations of polling unit areas.

In both Sweden and Spain, a hierarchical codification is employed to identify each polling unit at the different levels. With the help of a sequence of up to 11 elements, each polling unit in the country is uniquely coded. The units are reached after dividing every municipality into small areas that vary in size but comprise as a rule around 1500 people who are entitled to vote.

In Goteborg between the 2002 and 2006 Riksdag elections (the Swedish General elections), polling units were completely restructured (see Fig. 1), with 286 units in 2002 becoming 279 new

Fig. 2. Extract of the Barcelona (Spain) division in voting sections for (a) the 2008 Spanish general election and (b) the 2009 European Parliament election: the same area (Gracia neighbourhood) is depicted in both figures

units in 2006. Meanwhile, in Barcelona, before the 2009 European Parliamentary elections, the boundaries of its 1482 polling units were redrawn (see Fig. 2), yielding a partition of Barcelona's land into 1061 new units. Both instances are therefore perfect candidates to test the methods proposed when manual matching is less advisable. In the case of Goteborg, nevertheless, the geometrical extension of the manual approach is also gauged. In Barcelona, the manual option was ruled out because of the large number of polling units involved. Likewise, to complete the study and taking advantage of the authors' access to the geographical data of all the polling units in Goteborg county (Västra Götalands läns), the relative merits of the five alternatives have also been evaluated (in a context of election night forecasting) in a case where every polling unit is matched by using the spatial approaches. In Västra Götalands county, except for Goteborg districts, no other manually untraced changes existed.

Whatever the technique that is used, however, a problem arises when assessing the approximations. The proportions that are assigned to each new polling unit refer to their past electoral results whereas the observed data for these units are those from the current election. Therefore, they cannot be directly compared because there are electoral swings from one party and candidate to another and between elections. So, to gauge the differences between actual results and approximations, the latter should be time *transferred* to the moment of the posterior elections. To do so, we used the model-based prediction approach that is detailed in Pavía-Miralles (2005), appendix A. Pavía-Miralles proposed that, at polling area level and for each party or candidate, the current and past election proportions of votes are linearly related and suggested normally and zero-mean-distributed random disturbances with correlations constant between parties and independence between polling units, i.e. if π_{ki} and p_{ki} denote the proportion of votes counted for party k in polling unit i in the current and past election respectively then

$$\pi_{ki} = \alpha_k + \beta_k p_{ki} + \varepsilon_{ki}, \qquad k = 1, \ldots, p, \quad i = 1, \ldots, N, \tag{7}$$

ε_{ki} being zero-mean disturbances and α_k and β_k unknown parameters. Note that the multiequation system given by expression (7) is a seemingly unrelated regression model, but with linearly dependent disturbances (the sum of the proportion of votes obtained for the p political options in each polling unit is 1), which can be properly estimated with the help of Moore–Penrose generalized inverse matrices following the iterative algorithm that was proposed in Pavía-Miralles (2005), page 1121.

Other specifications have been employed in the literature to predict the share of votes—for example, Bernardo (1997) used a linear model on the logit transformations of the proportions and Curtice and Firth (2008) proposed a linear hierarchy 'nested' change shares model. We have opted for a linear specification, with the response variables in their 'raw' form, after observing the strong linear relationships that in our examples link current and previous proportions.

As a general strategy to assess the relative performance of the various options of allocating votes, first, the model parameters have been estimated by using the rest of the polling units, second, these estimates have been employed to predict the current proportions from the past allocated proportions and, finally, the predictions have been compared with actual results both at polling unit and aggregate level. To perform the aggregate comparisons, polling unit predictions have been combined (with weights based on unit electors and turnouts) to achieve an adequate estimator of the aggregate outcomes of all the polling units, i.e., if $\hat{\pi}_{ki}^X$ represents the proportion of votes estimated by using approach X for option k in each new polygon B_i after equation (7), then a natural predictor of the estimated proportion of votes for option k in the population as a whole, $\hat{\pi}_k^X$ (with $X \equiv$ M (manual), G (geometric), C (centroid), S (surface), D (compositional)), is

$$\hat{\pi}_k^X = \sum_{i=1}^{n} e_i \hat{t}_i^X \hat{\pi}_{ki}^X \Big/ \sum_{i=1}^{n} e_i \hat{t}_i^X, \qquad k = 1, \ldots, p, \tag{8}$$

where e_i denotes the electors who are entitled to vote in the ith unit and $\hat{t}_i^X$ is an approximation of the polling unit turnout that is attained after regressing past and current unit turnouts (Pavía-Miralles and Larraz-Iribas, 2008) and applying the parameter estimates to the corresponding allocated past turnout. Once aggregate estimates have been obtained, the statistic that is given by equation (9), based on the concept of entropy, is used to evaluate the degree of adjustment between aggregate estimates and real aggregate values:

$$H^X = -100 \sum_k \pi_k \log(1 - |\pi_k - \hat{\pi}_k^X|), \tag{9}$$

where π_k is the actual proportion of votes counted on the whole population for option k. The H-statistic is symmetrical, grows with the number of proportions to be estimated, punishes more errors in the highest proportions and is 0 with a perfect fit. Another measure used was the difference in the absolute values between the predicted and actual percentages $E^X = \Sigma_k |\pi_k - \hat{\pi}_k^X|$.

4.1. Goteborg

Goteborg, which is on the west coast of Sweden, is the second-largest city in the country and is the third-largest constituency (*Göteborgs kommun* with 17 permanent seats) out of the 29 that Sweden is divided into for Riksdag elections. For the 2006 Riksdag election, Goteborg redistributed its 373 836 electors into 279 units following a marked restructuring of its 2002 polling unit division map (see Fig. 1), making it impossible to assign historical electoral results by the classical approach. The four automated approaches that were described in the previous section were therefore used to allocate past vote proportions to these new polling units. Helped by the relatively small number of units in Goteborg, the manual version of the geometric approach was also used to carry out this task. The rest of the units in Sweden were matched by hand using the rules that were stated in Section 2. Two analyses using equation (7) were obtained to assess the approximations. First, the model was fitted by using only the rest of Goteborg's county polling units and, second, a model with all the rest of Swedish polling units was also estimated. Once the regression parameters had been estimated and used to transfer the past allocations to the instant when actual results were observed, the predictions and real data were compared both jointly (Table 1) and separately (Fig. 3 and Table 2).

Table 1. Estimates of 2006 Riksdag elections for Goteborg after regressing on spatial allocations

Approach	*Turnout† (%)*	*% for the following parties‡:*								*Error§*	*Entropy§§*
		Social Democrats	*Moderate Party*	*Liberal Party*	*Left Party*	*Green Party*	*Christian Democrats*	*Centre Party*	*Other parties*		
2002 results*	75.21	33.31	17.32	17.99	11.87	6.41	8.62	1.87	2.59	—	—
Manual**	77.20	30.08	28.95	10.48	7.72	6.66	5.88	4.28	5.95	7.64	1.25
Geometric**	76.74	30.33	28.41	10.37	7.94	6.80	5.82	4.34	5.99	6.93	1.14
Centroid**	76.37	31.46	27.13	10.36	8.55	6.60	5.61	4.18	6.12	6.86	1.12
Surface**	76.23	31.13	27.71	10.12	8.19	6.88	5.58	4.32	6.05	6.96	1.17
Compositional**	76.08	31.29	27.54	10.08	8.26	6.91	5.52	4.31	6.07	6.99	1.18
Manual††	77.00	28.33	28.82	10.37	8.34	7.74	6.19	4.75	5.46	5.34	1.01
Geometric††	76.54	28.91	28.23	10.22	8.62	7.79	6.04	4.72	5.48	3.79	0.65
Centroid††	76.18	29.41	27.64	9.98	8.85	7.99	5.91	4.72	5.50	3.36	0.50
Surface††	76.05	29.69	27.54	9.95	8.92	7.91	5.79	4.69	5.50	3.80	0.58
Compositional††	75.90	29.85	27.35	9.89	9.00	7.94	5.73	4.68	5.51	3.92	0.59
2006 results*	75.96	29.18	26.59	10.22	8.72	8.38	6.80	4.46	5.65	—	—

†The result rows show actual turnouts measured as the ratio of party votes and total electors (blank and null votes excluded). The approach rows portray the turnout estimates obtained after applying the corresponding strategy.
‡The result rows show the proportions of valid votes won by each political party. The approach rows portray the forecasts obtained after transferring the corresponding past approximations to 2006.
§The error column displays the sum of the differences in absolute values between the percentages of predictions and actual outcomes.
§§The entropy column provides the values obtained after applying equation (9).
*Advance votes that did not reach the polling stations on election day are excluded. Only the outcomes of the polling units with a geographical reference have been considered.
**Forecasts were obtained after fitting equation (7) using the rest of polling units of Goteborg county.
††After use of the rest of Swedish units.

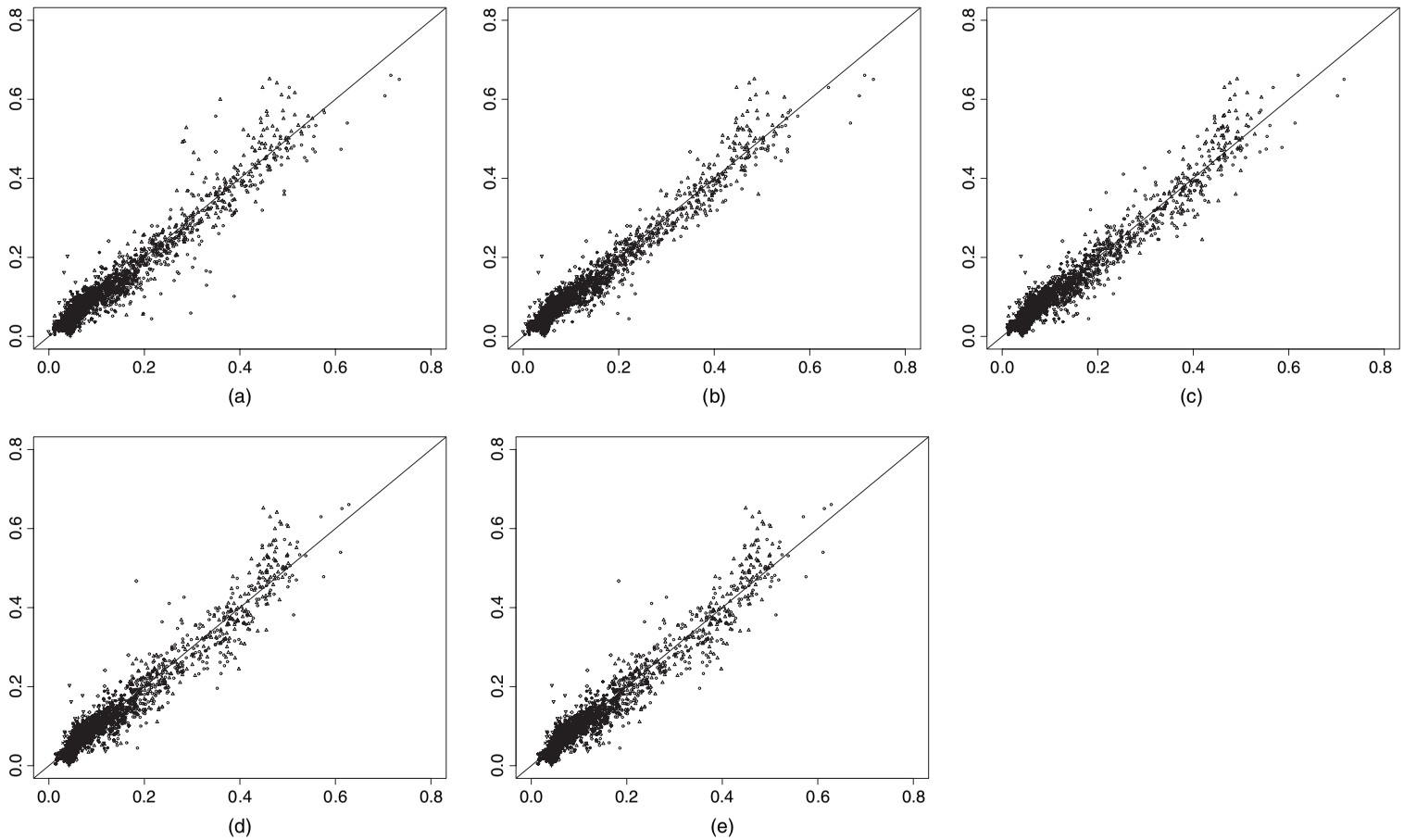

Fig. 3. Comparing real proportions (vertical axes) and approximations (horizontal axes) at polling level unit in Goteborg (Sweden) regressed by using the rest of the polling units of Sweden (the distance from the 45° line indicates how far apart estimates and outcomes are; the number of data points in each scatter plot is 2332 (279 districts by eight political options: Δ, Social Democrats; ○, Moderate Party; ×, Liberal Party; ⊕, Left Party; ◇, Christian Democrats; ∇, Green Party; +, Centre Party; ⊗, rest of the parties)): (a) manual; (b) geometric; (c) centroid; (d) surface; (e) compositional

Table 2. Polling unit median absolute errors by party for Goteborg

Approach†	*Error for the following parties:*								*Overall*‡
	Social Democrats	*Moderate Party*	*Liberal Party*	*Left Party*	*Green Party*	*Christian Democrats*	*Centre Party*	*Other parties*	
Manual	2.86	3.81	1.25	1.30	1.27	0.94	0.72	1.05	1.65
Geometric	2.59	3.23	1.21	1.09	1.17	0.86	0.76	1.02	1.49
Centroid	2.61	3.13	1.12	1.13	1.18	0.84	0.68	0.98	1.46
Surface	2.60	3.40	1.16	1.22	1.19	0.87	0.73	1.06	1.53
Compositional	2.86	3.47	1.22	1.23	1.20	0.89	0.72	1.05	1.58

†After use of the rest of Swedish units.
‡Polling unit median absolute errors for all parties.

In the 2006 Riksdag elections more than a dozen parties presented candidates, but only seven won representation. Thus, for presentation purposes, the analysis has focused only on those seven parties: Social Democrats, the Moderate Party, the Liberal Party, the Left Party, Christian Democrats, the Green Party and the Centre Party. An eighth option, which represents the outcomes of the rest of the parties, is also considered.

Table 1 presents the aggregate results that were obtained after assigning a past vote history to each district of Goteborg and transferring them through regression to 2006. As can be observed, despite the large swings that were registered between the 2002 and 2006 elections, all the aggregate party approximations obtained are quite accurate. A large proportion of the deviations between actual results and approximations cannot be attributed to the approach that was followed to allocate votes, but rather to the model employed to translate the proportions temporally. In fact, when all the polling units of Sweden (excluding Goteborg) are used to adjust the regression model the predictions are far better than the forecasts that were attained when only the rest of the polling units of Goteborg county are used to fit the regression. Table 1 also shows how the automatic options produce better results than the manual approach in terms of absolute error and entropy. Although all the automatic procedures seem to generate similar results in terms of accuracy, the centroid method records the smallest entropy and error coefficients in both cases.

At polling unit level the results are also satisfactory. The graphical comparison in Fig. 3 shows overall a good level of agreement between actual proportions and estimates using the rest of the Swedish units. The smallest median individual deviations are observed for the centroid approach, followed by the geometric approximation, whereas the manual approach displays the largest differences (see Table 2). The surface and compositional approaches are ranked in an intermediate position, displaying close outcomes. By parties, as a rule the larger the party the higher is the error. This result is not surprising given that deviations are measured in absolute terms. The deviations for the Moderate Party, however, stand out. The relatively high error for this party does not seem to be a consequence of allocating procedures but rather should be attributed to the regression model: the swing for the Moderate Party almost reaches 11 percentage points in the whole of Sweden when in Goteborg the figure was only around 9 percentage points. The correlation coefficients between actual shares and predictions fluctuate between a minimum of 0.9624 (registered by the manual estimates) to a maximum of 0.9753 (achieved by the geometric predictions).

4.2. Barcelona

Barcelona is the second-largest city (with a population of more than 1 600 000 inhabitants) in

Spain. At the beginning of 2009, Barcelona's authorities restructured the way that the city was divided for elections (see Fig. 2), with the number of polling units dropping from 1482 to 1061. The first elections to be held in this new situation were the 2009 European Parliament elections. Despite the relatively atypical behaviour that voters normally show in these kinds of election (for example, turnout rates are usually abnormally low), the four automated approaches that were described in Section 3 were applied to these elections. Manual allocation was discarded in this case because of the large number of sections involved. In this case, in contrast with Pavía (2010), who recommended using data from the same kind of elections to use equation (7), the 2008 Spanish general election outcomes were used to assign past historical results.

In the 2009 European Parliament elections, 35 political parties presented candidates in Spain, of which only five parties received relevant support in Barcelona—the Socialist Party, a right-wing coalition of regional parties, the Conservative Party, a national coalition of left-wing parties and a left-wing coalition of regional parties—therefore, the analysis has focused on these five parties, plus a sixth option which aggregates the results of the remaining political alternatives. The data from the rest of the polling units (2514) of Barcelona province, matched by using the classical approach, were used to fit equation (7) to transfer the 2008 allocated results to 2009. The comparisons of the transferred estimates and the real data are displayed in Table 3 (together) and in Fig. 4 and Table 4 (separately).

Table 3 presents the aggregate results that were obtained after assigning a past vote history to each polling unit of the city of Barcelona and transferring them to 2009. As can be observed, despite the marked swings that were registered between the 2008 and 2009 elections, the forecasts obtained are quite accurate. Almost certainly, a large proportion of the deviations between actual results and approximations should be attributed to the regression model that is used. Indeed, the relatively higher level of mobilization registered among Conservative Party supporters, which in the province of Barcelona are relatively more numerous in the capital, probably explains the deviations in Socialist Party, Conservative Party and turnout forecasts. As in Goteborg, all the procedures seem to produce very similar results in terms of accuracy, the surface approach yielding the best in this case.

At disaggregated level the results are also satisfactory. The polling unit level graphical comparison in Fig. 4 and the summary statistics of Table 4 reveal that the actual proportions and transferred forecasts are quite similar. Overall, the smallest individual deviations are observed for the surface approach with the compositional approach displaying the largest differences and the geometric and the centroid approaches in an intermediate position. Notwithstanding this, all the approaches generate very similar outcomes, with correlation coefficients above 0.95. By parties, the smallest deviations are registered again for the surface approach, except for the Socialist Party. Furthermore, as expected, the larger parties tend to show the higher deviations. The correlation analysis also shows that allocated shares could be employed as good approximations of past proportions in small area models.

4.3. Västra Götalands läns

In the two previous subsections, the performance of the approaches proposed has been analysed in cases of a massive (geographically intense) redrawing of polling unit boundaries, employing the rest of the polling units, matched by hand, to estimate the regression functions. However, the problem of assigning past vote history to polling units is not exclusive to situations of massive redraws. Also, it seems a little strange to employ a function that is estimated by using only manually matched units to assess the performance of *competing* allocating approaches. Thus, to complete this study, the analysis has been extended to a more common case, where every polling unit in the constituency is matched by using every approach and the regression model

Table 3. Estimates for the 2009 European election for Barcelona after regressing on spatial forecasts

Approach	*Turnout*† *(%)*	*% for the following parties*‡						*Error*§	*Entropy*§§
		Socialist Party	*Right-wing regional coalition*	*Conser-vative Party*	*Left-wing regional coalition*	*Left-wing national coalition*	*Other parties*		
2008 results*	71.85	42.84	20.66	18.33	6.96	6.37	4.84	—	—
Geometric	37.96	33.73	21.85	19.73	8.96	7.29	8.45	3.47	0.29
Centroid	37.97	33.57	21.76	19.66	8.94	7.23	8.85	3.54	0.28
Surface	37.93	33.54	21.72	19.64	8.92	7.33	8.84	3.36	0.27
Compositional	37.99	33.86	21.55	19.57	8.89	7.30	8.83	3.88	0.33
2009 results	39.72	32.71	21.71	20.68	8.37	7.97	8.54	—	—

†The result rows show actual turnouts measured as the ratio of valid votes and total electors (null votes excluded). The spatial approach lines portray the turnout estimates obtained after applying the corresponding strategy.
‡The result rows show the proportions of valid votes won by each political party. The approach rows portray the forecasts obtained after transferring the corresponding past approximations to 2009.
§The error column displays the sum of the differences in absolute values between percentages of predictions and actual outcomes: $E = \Sigma_k |\pi_k - \hat{\pi}_k^X|$.
§§The entropy column offers the values obtained after applying equation (9).
*Only the outcomes of the polling units with a geographical reference have been considered.

is estimated by using the values that are allocated with the corresponding approach. More specifically, the 2006 Riskdag elections in *Västra Götalands* (Sweden) were studied and the relative merits of each approach were compared in the context of election night forecasting. This line of attack allows the procedures to be tested on a problem where these techniques could be useful and avoids inducing additional noise by transferring the same units that have been used to fit the regression models.

On election night, at a given time t of the count, only data from $n(t)$ polling units (with $0 \leqslant n(t) \leqslant n$) are observed. Especially in the early stages, available data are not a random or representative sample of election results. Therefore, one possible strategy to make a prediction for the whole population consists of

(a) using the proportions of the observed units to estimate the parameters of expression (7),
(b) obtaining forecasts for the $n - n(t)$ unobserved polling units conditional on the parameter estimates and
(c) aggregating all available proportions (observed and forecast) to obtain a prediction of final outcomes (Pavía-Miralles, 2005).

Assuming that the observed polling units correspond to the first $n(t)$ districts, a predictor of the final outcome proportions is obtained (Pavía *et al.*, 2008) by

$$\hat{\pi}_k^X = \frac{\sum_{i=1}^{n(t)} e_i t_i \pi_{k,i} + \sum_{i=n(t)+1}^{n} e_i \hat{t}_i^X \hat{\pi}_{ki}^X}{\sum_{i=1}^{n(t)} e_i t_i + \sum_{i=n(t)+1}^{n} e_i \hat{t}_i^X}. \tag{10}$$

Five different time points were selected to assess the forecasting power of the various approaches. Table 5 presents, for approximately 1%, 2.5%, 5%, 10% and 25% of the electorate polled,

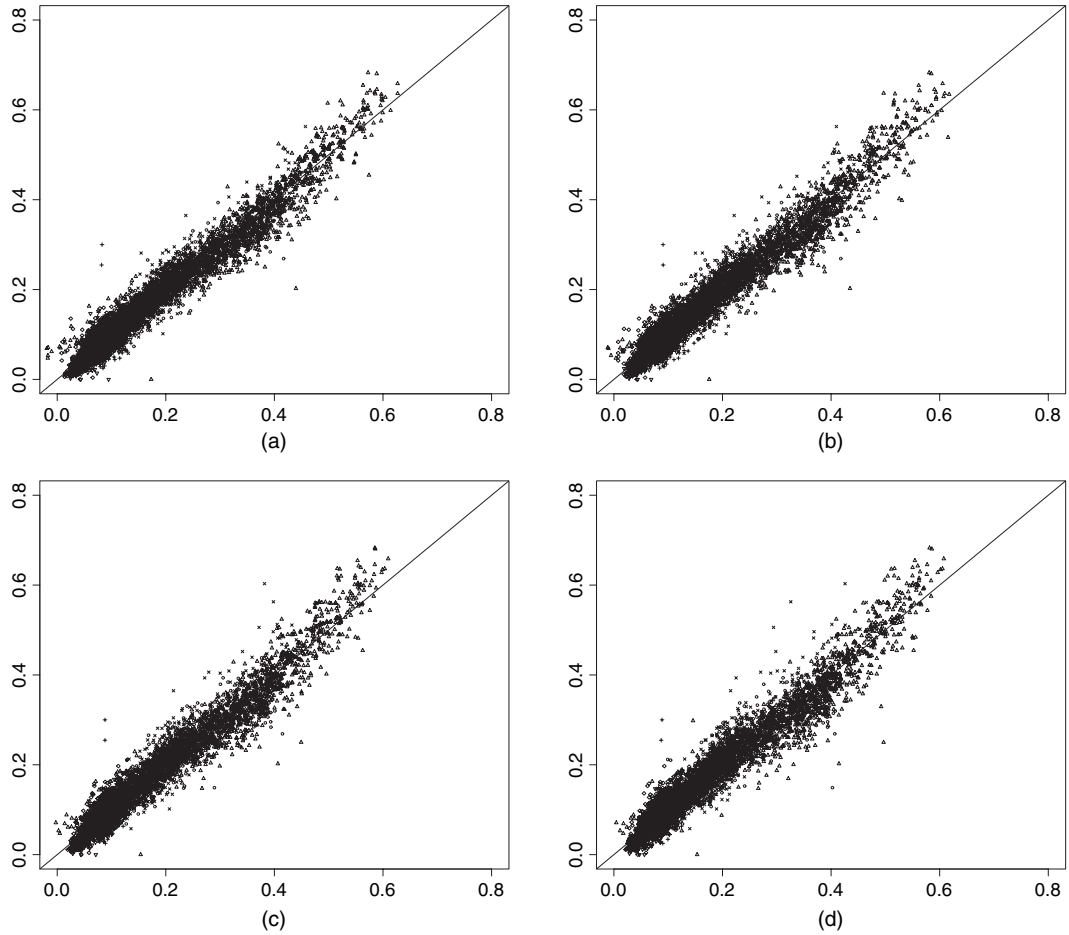

Fig. 4. Comparing real proportions (vertical axes) and forecasts (horizontal axes) at polling unit level in Barcelona (Spain) regressed by using the rest of the polling units of Barcelona province (the distance from the 45° line indicates how far apart estimates and outcomes are; the number of data points in each scatter plot is 6366 (1061 sections per six political options: Δ, Socialist Party; ○, right-wing coalition of regional parties; ×, Conservative Party; ◇, national coalition of left-wing parties; ∇, left-wing coalition of regional parties; +, rest of the parties)): (a) geometric; (b) centroid; (c) surface; (d) compositional

Table 4. Polling unit median absolute errors by party for Barcelona

Approach	*Errors for the following parties:*						*Overall†*
	Socialist Party	*Right-wing regional coalition*	*Conservative Party*	*Left-wing regional coalition*	*Left-wing national coalition*	*Other parties*	
Geometric	2.17	1.78	2.02	1.22	1.28	1.37	1.58
Centroid	2.02	1.67	1.87	1.25	1.23	1.37	1.53
Surface	2.06	1.58	1.74	1.16	1.22	1.29	1.43
Compositional	2.23	1.83	2.04	1.23	1.30	1.36	1.59

†Polling unit median absolute errors for all parties.

Table 5. Forecasts of the 2006 Riksdag election final results for the county of Västra Götalands

Approach	*Polled*†	*Units*‡	*Turnout*§ *(%)*	*% for the following parties*§§*:*								*Entropy**
				Social Democrats	*Moderate Party*	*Liberal Party*	*Christian Democrats*	*Centre Party*	*Green Party*	*Left Party*	*Other parties*	
2002 results**	100.00	1009	77.70	37.82	14.84	14.27	10.93	5.82	4.77	9.10	2.45	—
Provisional	0.98	29	80.20	29.38	22.86	6.00	10.85	16.41	3.82	5.00	5.68	3.39
Manual			79.14	36.15	24.61	8.43	6.54	7.24	4.07	6.41	6.56	1.16
Geometric			79.17	36.26	24.51	8.42	6.49	7.21	4.08	6.47	6.57	1.23
Centroid			79.21	36.20	24.37	8.37	6.56	7.21	4.15	6.57	6.57	1.24
Surface			78.62	34.62	25.38	8.49	6.23	7.00	4.38	7.31	6.58	0.86
Compositional			78.76	33.32	26.08	8.25	6.65	7.43	4.52	7.30	6.45	0.93
Provisional	2.47	55	78.38	33.72	21.10	6.04	10.17	13.57	3.99	5.41	5.99	2.02
Manual			79.09	35.62	24.16	8.49	7.20	7.37	4.36	6.54	6.26	1.00
Geometric			79.20	35.76	24.05	8.47	7.14	7.34	4.37	6.60	6.26	1.08
Centroid			79.18	35.87	24.09	8.34	7.16	7.40	4.48	6.69	6.18	1.09
Surface			78.82	34.95	24.36	8.56	6.81	7.09	4.79	7.09	6.36	0.79
Compositional			78.84	33.99	24.50	8.35	7.02	7.71	4.98	7.08	6.37	0.36
Provisional	5.01	94	78.96	34.89	21.74	6.08	9.58	12.43	4.04	5.43	5.81	1.93
Manual			79.03	35.48	24.39	8.26	7.42	7.15	4.54	6.64	6.12	0.87
Geometric			79.15	35.65	24.24	8.25	7.36	7.12	4.54	6.70	6.15	0.98
Centroid			79.06	35.66	24.18	8.32	7.35	7.19	4.52	6.67	6.11	0.99
Surface			78.91	34.43	24.76	8.76	7.19	7.07	4.82	6.77	6.21	0.50
Compositional			78.94	33.71	24.93	8.71	7.37	7.44	4.87	6.70	6.28	0.42
Provisional	9.99	159	79.02	35.35	22.37	6.56	8.88	11.13	4.32	5.46	5.93	1.72
Manual			78.99	35.11	24.70	8.23	7.48	7.15	4.72	6.33	6.28	0.65
Geometric			79.09	35.28	24.55	8.21	7.43	7.12	4.73	6.39	6.29	0.75
Centroid			78.97	35.24	24.57	8.21	7.42	7.16	4.74	6.40	6.26	0.73
Surface			79.04	33.83	25.07	8.48	7.44	7.24	5.13	6.47	6.33	0.37
Compositional			79.19	33.07	25.21	8.50	7.61	7.55	5.23	6.44	6.39	0.63

(*continued*)

Table 5 (*continued*)

Approach	*Polled*†	*Units*‡	*Turnout*§ *(%)*	*% for the following parties*§§*:*								*Entropy**
				Social Democrats	*Moderate Party*	*Liberal Party*	*Christian Democrats*	*Centre Party*	*Green Party*	*Left Party*	*Other parties*	
Provisional	25.00	333	78.16	35.70	22.60	7.06	8.26	9.49	4.82	6.13	5.95	1.56
Manual			78.73	34.81	24.44	8.25	7.65	7.13	5.11	6.48	6.12	0.57
Geometric			78.86	34.88	24.37	8.25	7.62	7.13	5.11	6.52	6.14	0.61
Centroid			78.84	34.83	24.38	8.26	7.62	7.14	5.11	6.54	6.13	0.59
Surface			78.84	33.72	24.77	8.35	7.68	7.32	5.39	6.61	6.16	0.28
Compositional			78.94	33.30	24.82	8.35	7.80	7.53	5.44	6.59	6.17	0.41
2006 results**	100.00	985	78.84	33.96	24.70	8.32	7.86	7.52	6.41	5.57	5.66	—

†The percentage of the census polled at each moment of the scrutiny appears in the provisional rows of this column.
‡The number of polling units for each election appears in the result rows of this column. The number of units polled at each moment during the count appears in the provisional rows of this column.
§The result rows show actual turnouts measured as a ratio of party votes and total electors (blank and null votes excluded). The provisional rows show the turnout at that moment of the count. The approach rows portray the turnout forecasts obtained after applying the corresponding approach.
§§The provisional rows show the share of valid votes that each political party was receiving at that moment of the count. The approach rows portray the final vote forecasts obtained after transferring the corresponding past approximations to 2006.
*The entropy column displays the values obtained after applying equation (9).
**Advanced votes that did not reach the polling stations on election day have been excluded. Only the outcomes of the polling units with a geographical reference have been considered.

the forecasts that would be obtained if the approximations obtained after allocating past voting results by using each of the approaches had been used on the 2006 Riksdag election night in Västra Götalands county (1 164 890 electors in 985 polling units) to predict final results. The number of polling units that are available at each point in time, the proportion of the electorate polled and the predicted turnout are also included in Table 5. Moreover, the entropy statistic is also shown to facilitate the evaluation of the forecasts.

In spite of the environment of drastic political change and lag in vote convergence, all spatial techniques used to match polling units seem to yield results that are comparable with those obtained after manual matching and significantly improve provisional results, even when only a very small proportion of votes had been polled. Computerizing the process of assigning past voting history seems to improve the quality of allocations in this case also. Indeed, in addition to the cumbersome extra work that is saved, in view of these findings, using either the surface or compositional approach can even lead to a systematic improvement in county forecasts. These methods show for the whole Västra Götalands läns smaller entropy values in every stage of the count of votes.

Aggregate and polling unit analyses are interesting, but what really determines the results of an election are the outcomes that are registered in each constituency. For Riskdag elections, Västra Götalands splits its electorate into five constituencies—Göteborgs kommun (373 836 electors), Västra Götalands läns västra (253 306 electors), Västra Götalands läns norra (200 285 electors), Västra Götalands läns södra (140 416 electors) and Västra Götalands läns östra (197 047 electors). Therefore, to assess the quality of the previous forecasts, the analysis has been extended to this level.

Sweden uses a complex proportional electoral system with permanent and adjustment seats. Permanent seats are distributed within each constituency almost exclusively considering the outcomes of the corresponding constituency, whereas adjustment seats are distributed taking into account both national and constituency outcomes. Constituency proportions of votes and turnouts are therefore relevant to assign seats between parties. Hence, the goodness of fit of both types of estimate has been analysed. To study the degree of global adherence of party forecasts to real outcomes, the weighted average of the corresponding constituency H-statistics has been calculated for each stage of the count of votes by using the number of electors in the constituency as a weighting variable. The discrepancies between observed and actual turnouts have also been summarized via weighted averages of constituency turnout absolute errors (Table 6).

Table 6 shows that provisional outcomes are substantially improved by forecasts. All manual and spatial strategies generate very accurate results. The surface approach clearly yields the best results, particularly in the early stages of counting votes when forecasts are more valuable. Furthermore, although all forecasting strategies converge as the amount of population polled increases, the surface approach still maintains its advantage (combining proportion and turnout predictions) even after a quarter of the votes have been scrutinized.

5. Discussion and concluding remarks

Traditionally, the issue of providing small size polling units with a past voting history has been a very tedious and cumbersome task performed (where possible) manually by comparing previous and current geographical administrative codes and census figures. This unpleasant task, which is not always possible to execute, has dampened analysts' interest. Nowadays, however, provided that spatial data for polling units are available, this chore could be made easier and even improved with the help of spatial software. In particular, using a situation in which a massive (geographically intense) reorganization of polling unit boundaries occurs as a basis

Table 6. Summary of the goodness of fit of the 2006 Riksdag election night forecasts at constituency level

Approach	*Constituency-weighted entropies for the following % of electorate polled†:*					*Constituency-weighted turnout errors for the following % of electorate polled‡:*				
	0.98	*2.47*	*5.01*	*9.99*	*25.00*	*0.98*	*2.47*	*5.01*	*9.99*	*25.00*
Provisional	5.41	5.41	5.31	5.16	5.16	0.72	0.68	0.60	0.55	0.46
Manual	0.32	0.29	0.25	0.31	0.31	0.57	0.62	0.55	0.48	0.41
Geometric	0.35	0.32	0.29	0.22	0.25	0.62	0.62	0.49	0.45	0.41
Centroid	0.38	0.35	0.30	0.23	0.25	0.62	0.62	0.49	0.45	0.41
Surface	0.20	0.23	0.24	0.20	0.25	0.47	0.31	0.25	0.41	0.20
Compositional	0.41	0.31	0.33	0.45	0.38	0.52	0.34	0.28	0.63	0.26

†Weighted average of constituency entropies: $\Sigma_c \tau_c H_c^X$, where τ_c is the relative number of electors in constituency c and H_c^X the adjustment measure obtained after applying equation (9) to the forecasts achieved in constituency c with X allocations.
‡Weighted average of constituency turnout absolute errors: $\Sigma_c \tau_c \ |\hat{t}_c^X - t_c|$, being $\hat{t}_c^X$ and t_c the predicted and actual turnout in c respectively.

(where the manual approach would be almost impossible to apply), this paper proposes exploiting the spatial patterns that electoral outcomes display and suggests several methods to reassign votes by using spatial strategies. The geometric approach emerges as a natural substitute for the classical approach in a geographical information system environment and proposes comparing the depiction of old and new polling units on a map as a strategy to establish matches between units. This approach, however, assumes uniform spatial distribution of votes within each polling unit as if units were geographical units *per se* and overlooks the fact that, in the same way as other social variables, electoral results have spatial patterns within polling units. Hence, three additional approaches are also proposed exploiting the fact that party supporters are not randomly distributed in space: the centroid, surface and compositional approaches.

Three real data examples are used to evaluate the performance of the various methods. The cases of Goteborg in Sweden and Barcelona in Spain are studied as examples of places that have recently undergone a complete restructuring of polling units, whereas the example of Västra Götalands (Sweden) is analysed to see how these procedures would work in a wider situation. The outcomes clearly show that all spatial-based approaches yield accurate approximations, which are at least as good as those recorded by using the classical procedure. Therefore, in addition to the cumbersome extra work that is saved, the automation of the process frequently produces improvements in quality. Although the four methods produce comparable results, the surface approach registers (mainly in the Västra Götalands example) the best outcomes. Thus, taking into account that the compositional approach entails more computation and that the other three methods imply quite similar computational burdens, the surface procedure seems to be the best to promote.

In addition to the potential benefits that these techniques have for local political party teams, small size political analysts and electoral forecasters, we believe that these approaches could also help in the issue of combining small scale electoral aggregate figures and data taken from surveys with geographical markers (such as exit polls or cluster surveys) in the models that are used by electoral geographers, political scholars and pollsters (Jacobs and Spierings, 2010; Greiner and Quinn, 2010; Pavía and Larraz, 2012). Although electoral geography and political science share the aim of understanding why voters cast their ballots the way that they do, these

two fields of electoral study tend to rely on completely different methodologies and data sets. Political scientists use survey data from individuals, whereas geographers use aggregate data, often censuses of small areas. Both data sets have a significant weakness. Political analyses fail to capture contextual influences and electoral geographers cannot attribute socio-economic characteristics to voters. Although some reasons can be proffered for this neglect on behalf of political researchers, according to O'Loughlin (2002) many have tended to eschew aggregate data that are collected for geographical units partly because of the difficulties of inferring across levels. Assisted by the increasingly widespread use of geographical information systems and the increasing availability of data (not only electoral) with geographical references, this paper provides a solution to the complexities that are involved in dealing with constantly shifting polling areas and the difficulties that this poses for the introduction of past voting results (and census figures) in, for instance, multilevel and longitudinal multilevel political models (e.g. Steele (2008)) or in spatial panel forecasting models (Baltagi *et al.*, 2012). Thus, for example, as US census tract areas remain fairly constant from census to census (US Census Bureau, 2000), these procedures could be combined, once estimates have been obtained at this level, to implement longitudinal analysis and to study the relationships between party or candidate swings and compositional factors (such as religion, class, occupation, gender and demographic structure).

Furthermore, these approaches could also be employed to audit and supervise the processes of redrawing constituencies, state legislative and congressional boundaries. For instance, at present, US state legislatures draw their congressional boundaries on the basis of a complex mixture of partisan considerations, incumbency protection and race. The resulting boundaries are often the result of political battles and sometimes represent nothing more than a compromise partition of the state that reflects the balance of power between the two major parties. By way of example, the 2002 redistricting plan that was devised by the Republican-dominated Pennsylvania state legislature produced Republican victories in 12 of the 19 congressional districts, even though a Democrat won the governorship and Democrats had a slight majority in party registration advantage (Turner, 2005). These techniques could therefore be used to alert about extreme partisan gerrymandering and population instability (Yoshinaka and Murphy, 2009) that could bias the outcomes of US congressional elections. We think that starting with precinct level data the geometric approach should be used, as a rule, to evaluate the predictable consequences of possible redistricting.

Finally, it should be highlighted that, although we have focused on the issue of spatial redistribution of electoral results, the problem of restructuring small area boundaries also affects other socio-economic variables observed as aggregates, which also display spatial patterns (Myint, 2008). Therefore, these techniques could also be applied to non-electoral variables exploiting the advantages that current software offers by integrating statistical and spatial methodologies (Bivand *et al.*, 2008).

Acknowledgements

The authors thank Tor Lundberg and Henrik Hannebo from Valmyndigheten (the Swedish electoral authority) for their first-rate assistance in providing all the Swedish electoral and geographical data handled in this paper, the staff of Informació de Base i Cartografia de l'Institut Municipal d'Informàtica from the City Council of Barcelona (l'Ajuntament de Barcelona) for providing Barcelona's division in sections and the members of Area de Processos Electorals from the Departament de Governació i Administracions Públiques of Generalitat de Catalunya and the Subdirección General de Política Interior y Procesos Electorales of the Spanish Government for supplying Spanish electoral data. We also thank a reviewer and mainly the Associate

Editor for their helpful suggestions and comments, and Tony Little for correcting the English of the manuscript. This research has been supported by the Spanish Ministerio de Ciencia e Innovación through project CSO2009-11246.

References

Agnew, J. A. (1987) *Place and Politics*. London: Allen and Unwin.
Balgati, B. H., Bresson, G. and Pirotte, A. (2012) Forecasting with spatial panel data. *Computnl Statist. Data Anal.*, **56**, 3381–3397.
Bernardo, J. M. (1984) Monitoring the 1982 Spanish socialist victory: a Bayesian analysis. *J. Am. Statist. Ass.*, **79**, 510–515.
Bernardo, J. M. (1997) Probing public opinion: the State of Valencia experience. In *Bayesian Case Studies* (eds C. Gatsonis, J. S. Hodges, R. E. Kass, R. McCulloch, P. Rossi and N. D. Singpurwalla), pp. 3–21. New York: Springer.
Bivand, R. S., Pebesma, E. J. and Gómez-Rubio, V. (2008) *Applied Spatial Data Analysis with R*. New York: Springer.
Bracken, I. and Martin, D. (1995) Linkage of the 1981 and 1991 UK census using surface modeling concepts. *Environ. Planng* A, **27**, 379–390.
Burden, B. C. and Kimball, D. C. (1998) A new approach to the study of ticket splitting. *Am. Polit. Sci. Rev.*, **92**, 533–544.
Clark, M. (2009) Valence and electoral outcomes in Western Europe, 1976–1998. *Elect. Stud.*, **28**, 111–122.
Cox, K. R. (1969) The voting decision in a spatial context. In *Progress in Geography*, vol. 1 (eds C. Board, R. J. Chorley, P. Haggett and D. R. Stoodart), pp. 81–117. London: Arnold.
Cressie, N. A. C. (1993) *Statistics for Spatial Data*. New York: Wiley.
Curtice, J. and Firth, D. (2008) Exit polling in a cold climate: the BBC–ITV experience in Britain in 2005 (with discussion). *J. R. Statist. Soc.* A, **171**, 509–539.
Curtice, J., Fisher, S. D. and Kuha, J. (2011) Confounding the commentators: how the 2010 exit poll got it (more or less) right. *J. Electns Publ. Opin. Parties*, **21**, 211–235.
Flowerdew, R. and Green, M. (1994) Areal interpolation and types of data. In *Spatial Analysis and GIS* (eds A. S. Fotheringham and P. A. Rogerson), pp. 121–145. London: Taylor and Francis.
Gregory, I. N. and Ell, P. S. (2005) Breaking the boundaries: geographical approaches to integrating 200 years of the census. *J. R. Statist. Soc.* A, **168**, 419–437.
Greiner, D. J. and Quinn, K. M. (2010) Exit polling and racial bloc voting: combining individual-level and RxC ecological data. *Ann. Appl. Statist.*, **4**, 1774–1796.
Jacobs, K. and Spierings, N. (2010) District magnitude and voter turnout: a multi-level analysis of self-reported voting in the 32 Dominican Republic districts. *Elect. Stud.*, **29**, 704–718.
Johnston, K., ver Hoef, J. M., Krivoruchko, K. and Lucas, N. (2003) *ArcGIS® 9: Using ArcGIS® Geostatistical Analyst*. Redlands: Esri.
Johnston, R. and Pattie, C. (2006) *Putting Voters in Their Place: Geography and Elections in Great Britain*. Oxford: Oxford University Press.
Key, V. O. (1949) *Southern Politics in State and Nation*. New York: Knopf.
Khofeld, C. W. and Sprague, J. (2002) Race, space, and turnout. *Polit. Geog.*, **21**, 173–195.
Kim, J., Elliot, E. and Wang, D.-W. (2003) A spatial analysis of county-level outcomes in US Presidential elections: 1988–2000. *Elect. Stud.*, **22**, 741–746.
Kyle, S., Samuelson, D. A., Scheuren, F., Vicinanza, N. and Dingman, S. (2007) Explaining discrepancies between official votes and exit polls in the 2004 Presidential election. *Chance*, **20**, 36–45.
Longley, P. A., Goodchild, M. F., Maguire, D. J. and Rhind, D. W. (2001) *Geographic Information Systems and Science*. Chichester: Wiley.
Longley, P. A. and Mesev, T. V. (1997) The use of diverse RS-GIS sources to measure and model urban morphology. *Geog. Syst.*, **4**, 5–18.
Macallister, I., Johnston, R. J., Pattie, C. J., Tunstall, H., Dorling, D. F. L. and Rossiter, D. J. (2001) Class dealignment and the neighbourhood effect: Miller revisited. *Br. J. Polit. Sci.*, **31**, 41–59.
Martin, D. (1989) Mapping population data from zone centroid locations. *Trans. Inst. Br. Geog.*, **14**, 90–97.
Martin, D. (2003) Extending the automated zoning procedure to reconcile incompatible zoning systems. *Int. J. Geog. Inform. Sci.*, **17**, 181–196.
Mitofsky, W. J. and Edelman, M. (2002) Election night estimation. *J. Off. Statist.*, **18**, 165–179.
Mosteller, F., Hyman, H., McCarthy, P. J., Marks, E. S. and Truman, D. B. (1949) *The Pre-election Polls of 1948*. New York: Social Science Research Council.
Myint, S. W. (2008) An exploration of spatial dispersion, pattern, and association of socio-economic functional units in an urban system. *Appl. Geog.*, **28**, 168–188.
O'Loughlin, J. (2002) The electoral geography of Weimar Germany: exploratory spatial data analysis ESDA of Protestant support for the Nazi party. *Polit. Anal.*, **10**, 217–243.

Pattie, C. and Johnston, R. (2000) People who talk together vote together: an exploration of contextual effects in Great Britain. *Ann. Ass. Am. Geog.*, **90**, 41–66.
Pavía, J. M. (2010) Improving predictive accuracy of exit polls. *Int. J. Forecast.*, **26**, 68–81.
Pavía, J. M. and Larraz, B. (2012) Nonresponse bias and superpopulation models in electoral polls. *Rev. Espan. Invest. Sociol.*, **137**, 237–264.
Pavía, J. M., Larraz, B. and Montero, J. M. (2008) Election forecasts using spatiotemporal models. *J. Am. Statist. Ass.*, **103**, 1050–1059.
Pavía-Miralles, J. M. (2005) Forecasts from non-random samples: the election night case. *J. Am. Statist. Ass.*, **100**, 1113–1122.
Pavía-Miralles, J. M. and Larraz-Iribas, B. (2008) Quick counts from non-selected polling stations. *J. Appl. Statist.*, **35**, 383–405.
Robinson, S., Langford, M. and Tate, N. (2002) Modelling population distribution with OS LandLine.Plus data and Landsat imagery. In *Proc. GIS Research UK 11th A. Conf*. (eds S. Wise, P. Brindley, Y.-H. Kim and C. Openshaw), pp. 320–325. Sheffield: University of Sheffield.
Simpson, L. (2002) Geography Conversion Tables: a framework for conversion of data between geographical units. *Int. J. Popln Geog.*, **8**, 69–82.
Steele, F. (2008) Multilevel models for longitudinal data. *J. R. Statist. Soc.* A, **171**, 5–19.
Sui, D. Z. and Hugill, P. J. (2002) A GIS-based spatial analysis on neighborhood effects and voter turn-out: a case study in College Station, Texas. *Polit. Geog.*, **21**, 159–173.
Turner, R. C. (2005) The contemporary presidency: do Nebraska and Maine have the right idea?: the political and partisan implications of the district system. *Pres. Stud. Q.*, **35**, 116–137.
US Census Bureau (2000) *Strength in Numbers; Your Guide to Census 2000 Redistricting Data From the U.S. Census Bureau*. Washington DC: US Census Bureau.
Yoshinaka, A. and Murphy, C. (2009) Partisan Gerrymandering and population instability: completing the redistricting puzzle. *Polit. Geog.*, **28**, 451–462.