Areal interpolation and the UK's referendum on EU membership

Abstract

I show how results from the United Kingdom's referendum on membership of the European Union can be remapped from local authority level to parliamentary constituency level through the use of a scaled Poisson regression model which incorporates demographic information from lower level geographies. I use these estimates to show how the geographic distribution of signatures to a petition for a second referendum was strongly associated with how constituencies voted in the actual referendum.

1 Introduction

On the 23rd June 2016 the United Kingdom held a referendum on its membership of the European Union.[1] Results were announced in 382 different local counting areas. In England, Wales and Scotland these local counting areas coincided with local authority areas.[2] This followed the pattern set by previous referendums on Scottish independence (2014) and the alternative vote (2011), which also used local authority areas as counting areas.

Unfortunately, local authority areas are not the most important geographical unit in the study of British politics. In part, this is because these areas vary wildly in size. There were 1799 registered voters in the UK's smallest local authority area (Isles of Scilly): the figure for the largest local authority (Birmingham) was almost four hundred times greater, at

[1] The referendum was also held in Gibraltar, which is not part of the United Kingdom, but is rather a British Overseas Teritory. Results from Gibraltar nevertheless counted towards the UK total.

[2] Northern Ireland acted as a single counting area, but results were made available at the level of parliamentary constituencies.

707,293 voters. For most purposes – and in particular for the purpose of examining dyadic representation (Hanretty, Lauderdale, and Vivyan 2016) – results at the level of Westminster parliamentary constituencies would be far more useful than results from local authority areas. Unfortunately, these two geographics do not coincide. Only twenty-six Westminster constituencies are perfectly homologous with local authority areas. Most local authorities combine two or more parliamentary constituencies.

It is therefore difficult to remap results at a local authority level on to the level of Westminster constituencies. If results were available at a lower level (say, at the level of council wards) it would be possible to aggregate results to the level of Westminster constituencies – but although some local authorities have published results at a ward or constituency level, most have not. Given this, it becomes necessary to find a principled method of *areal interpolation* in order to project from our source geography (local authority areas) to our target geography (Westminster constituencies).

In this paper, I set out such a method. This method takes advantage of the fact that although results at not available at a lower level, demographic variables known strongly to predict the results are available. I estimate two separate Poisson regression models of the count of voters for Leave and Remain in each local authority area, taking the population of that area as an offset, and modelling the count as a function of certain demographic variables derived from the 2011 Census. I then identify areas which result from the intersection of source and target geographies (areas which are composites of Census Output Areas), and for which the same demographic information also exists. For each intersection, I generate predicted counts of Leave and Remain voters. I then scale these counts so that the predicted counts sum to the observed counts at the level of the local authority. Finally, I re-aggregate these scaled counts to the level of

the parliamentary constituency. I subsequently use these parliamentary constituency level estimates to explore the relationship between support for Remain in the referendum, and the number of signatures on a parliamentary petition calling for a second referendum.

2 UK electoral geography and demographic information

In order to map between source and target geographies, it is useful to have information on lower-level geographies which can be aggregated to *either* the source or the target geography. The lowest geography for which demographic information is available is the Census Output Area. Census output areas are tracts with populations of between 100 and 625 individuals. Key statistics on these output areas are available from the Office of National Statistics (in the case of output areas in England and Wales) or from Scotland's Census (in the case of output areas in Scotland). Most key statistics are comparable between England and Wales, and Scotland, though some variables available in the English census are not available in the Scottish census.[3]

The Office of National Statistics and Scotland's Census have published lookup tables which enable us to map between (i) output areas and local authority areas and (ii) output areas and Westminster constituencies. By merging these two lookup tables, we can identify groups of output areas which result from the intersection of local authority areas and Westminster parliamentary constituencies. We can also calculate values of our de-

[3] The English and Welsh census contains information on the percentage of the population without a passport; the Scottish census does not. The percentage without a passport has been found to be a powerful predictor of the Leave vote. See Burn-Murdoch, John, "Brexit: voter turnout by age", *Financial Times*, 24th June 2016, available online at `http://blogs.ft.com/ftdata/2016/06/24/brexit-demographic-divide-eu-referendum-results/`.

mographic variables for these intersections by creating sums of counts or population-weighted means.

There are 851 such intersections. For each of the 380 local authority areas and each of the 851 intersections I have calculatd the percentage of the population who:

- have Level 1 educational qualifications (1-4 GCSEs or equivalent);
- have Level 2 educational qualifications (5+ GCSEs or equivalent);
- have Level 3 educational qualifications (2+ A levels or equivalent);
- have Level 4 educational qualifications or greater (university degree or equivalent);
- own their home (with or without a mortage)
- who have White British as their ethnicity
- work in higher managerial, administrative, or professional occupations;
- work in lower managerial, administrative, or professional occupations;
- work in lower supervisory or technical occupations;
- work in semi-routine occupations;
- work in routine occupations;
- have never worked or are long-term unemployed;

as well as

- the median age of residents
- the population of the area
- the region in which the area lies.

All of these variables will eventually feature as independent variables in a regression equation. The variables omitted for reasons of collinearity are (for education) the proportion of the population with no educational qualifications, and (for occupation) the proportion of the popula-

tion working in intermediate occupations, as small employers, or in jobs not otherwise classifiable.

These variables were chosen on the basis that previous research has identified them as important predictors of Euroskepticism, or support for Euroskeptic parties (Ford and Goodwin 2014, ch. 4). There are some important predictors which is it not possible to include. The vote share received by UKIP in the 2014 European Parliament elections is an important predictor of the Leave vote in the 2016 referendum – but since it is not a demographic variable included in the census, it cannot be used for this kind of areal interpolation.

3 Areal interpolation

Areal interpolation is a process by which the values of variables originally measured using one set of geographical units or zones can be estimated for a different and incompatible (or misaligned) set of geographical units. Many different methods have been used for areal interpolation. These methods can be classified in different ways (Thomas-Agnan, Vanhems, and others 2015, Table 1). Three methods are particularly important: areal weighting, dasymmetric interpolation, and regression with auxiliary information. The method that I use here is a particular type of regression-based method. Because this method is more complicated than either areal weighting or dasymmetric interpolation, I must first describe why these two methods are likely to yield poorer estimates in the present case.

3.1 Areal weighting

Areal weighting is a method for areal interpolation which requires three types of information:

- information on the *source* geographies (the geographical units over which our variable is measured; in this case, local authority areas);
- information on the *target* geographies (the geographical units for which we wish to produce estimates); and
- the values of the variable of interest measured on the source geography.

The estimate of the values of the variable of interest for the target geography is a weighted mean of the values of the variable in the source geography. These weights depend on the degree of overlap between different source and target geographies. If 20% of the area of a particular target unit t comes from some source unit s_1, and 80% from source unit s_2, then our estimate for t is simply twenty percent of the value for s_1 plus eighty percent of the value for s_2.

3.2 Dasymmetric interpolation

Dasymmetric interpolation is like areal weighting, but requires one more type of information, namely information on a control variable. In the common form of dasymmetric interpolation (Thomas-Agnan, Vanhems, and others 2015), this control variable is population, and it is measured using a smaller system of geographical units formed by the intersection of source and target geographies. Using dasymmetric interpolation the estimate is a weighted mean of the values of the variable in different source geographies, where the weights depend on the share of the target population contributed by each source unit. If 20% of the population of a target unit live in source s_1, and 80% in source s_2, then our estimate is simply twenty percent of the value for s_1, plus eighty percent of the value of s_2.

Both areal weighting and dasymmetric interpolation could be used to produce constituency level estimates of EU referendum vote shares. Goplerud (2015) has argued that both methods work well when interpolat-

ing election results under "old" and "new" boundaries for six different countries, in the sense that the mean absolute error typically ranges between two and three percent. The question therefore is whether a different method would provide better results.

3.3 Regression-based methods

Regression-based methods require additional information to produce estimates. In particular, they require information on additional covariates which are known, or presumed to be, related to the variable of interest. These covariates must at least be measured at the level of the source and target geographies, but may also be measured at the level of intersections between these geographies.

In the case of referendum voting, the covariates described in the previous section can give us additional information about likely outcomes, information which is not used by either areal interpolation or dasymmetric interpolation, and which allows us to generate more plausible results. Take as examples the two London constituencies of Hornsey and Wood Green and Tottenham. Both constituencies are contained within the London Borough of Haringey. 75% of people who voted in Haringey voted for the United Kingdom to remain part of the European Union. We might therefore ascribe this figure to both constituencies. However, Hornsey and Wood Green is very different in many important respects from Tottenham.

In Hornsey and Wood Green, around half of the population have a university degree of equivalent (level 4) qualification; in Tottenham the figure is closer to a quarter. In Hornsey and Wood Green, twenty percent of people are engaged in higher managerial, administrative, or professional occupations; the figure in Tottenham is half that. Given these differences, and given the fact that education was a very important predictor

of how people would vote in the referendum, it is very likely that opinion in Tottenham is different from opinion in Hornsey and Wood Green.

The challenge lies in incorporating this additional information whilst at the same time respecting the particular constraints implied by misaligned data. Any estimates produced must, in order to be credible, satisfy the *pycnophylactic property*. That is:

- predictions for target areas must *equal* the known value from source areas where source and target overlap perfectly;
- predictions for target areas must *add up to* the known value from source areas where the source is entirely composed of two or more target areas;
- predictions for groups of target areas (regions) must *add up to* the known value from groups of source areas

It is this pycnophylactic property which motivates the use of a scaled Poisson regression model. The model is a Poisson model because a Poisson model is an appropriate model for count data, and modelling counts of voters who voted Leave or Remain makes it very easy to check whether the pycnophylactic property is satisfied; and the model is a *scaled* model because predictions from the model are scaled in order to ensure that the pycnophylactic property is satisfied in this way. This model is essentially the same as that used by Flowerdew and Green (1989).

3.4 The scaled Poisson regression model

To describe the model, it will be useful to establish notation in order to refer to these geographies in the abstract, and to give a more detailed justification for using count data. I will use s to refer to units in the source geography – in this case, local authority areas. I will use t to refer to units in the target geography (Westminster parliamentary constituencies). I use

st to refer to the intersection of source area s and target area t. I use y_s to refer to the value of the variable of interest in the source area. y_s is always known. I use y_t to refer to the value of the variable of interest in the target area. y_t is not known and must be estimated. Finally, I use y_{st} to refer to the value of y in the area formed by the intersection of areas s and t. With a slight abuse of notation, I will talk about intersections being in areas s or t by talking about areas st such that $st \in s$ or $st \in t$.

Values in areal interpolation may be of two kinds: intensive, or extensive (Lam and Goodchild 1980). Intensive variables are variables such that the value of the variable in the source (target) geography is equal to the sum of the values in all constituent intersections:

$$y_s = \sum_{st \in s} y_{st}$$

Count variables are intensive variables. The number of people who voted Leave in a local authority area is simply the sum of the number of people who voted leave in the output areas which make up that local authority.

Extensive variables are variables such that the value of the variable in the source (target) geography is a weighted mean of values in constituent intersections. Thus, for some weighting scheme with weights w,

$$y_s = \sum_{st \in s} w_{st} y_{st}$$

Rates and proportions are common forms of extensive variables.

This distinction is not rigid. Extensive variables can be modelled as intensive variables. Conversely, where intensive variables have a theoretical maximum (the number of people who voted Leave cannot be greater than the number of people), they can be expressed as extensive variables. I introduce this distinction because it is related to my choice of outcome

variable. I model *counts* of people who voted Leave and Remain (an intensive variable). This is perhaps different to how the problem would normally be approached as a question of modelling the percentage of voters who voted Leave or Remain in each area.

There are four steps to the procedure. *First*, I model the count of people who voted Leave Y_s^L (or Remain: Y_s^R) in each local authority area using information on the demographics of each area (X_s) and the population in that area (p_s).

$$Y_s^L = f(X_s, p_s) \tag{1}$$

Second, I use this model to generate predicted counts for each intersection.

$$\hat{Y}_{st}^L = f(X_{st}, p_{st}) \tag{2}$$

Because no model is perfect, these predicted counts will not equal known results when aggregated to the level of the local authority area. As it stands, they cannot be aggregated to the level of the constituency, because they would then give demonstrably wrong answers for those constituencies which are perfectly homologous with local authority areas. This is an undesirable characteristic of a method of areal interpolation. It is necessary therefore to scale these results in order that the method produces correct results for these areas (and better results elsewhere).

The *third* step therefore involves scaling predicted counts in each local authority area by multiplying them by a scale factor which is equal to the known result divided by the sum of the predicted counts of all intersections in that area. Call these scaled counts Y^*.

$$Y^{*L}_{st} = \frac{Y_s}{\sum\limits_{st \in s} \hat{Y}_{st}^L} \hat{Y}_{st}^L \tag{3}$$

The *fourth and final* step involves aggregating these scaled counts to the level of the parliamentary constituency. This step just requires addition.

$$Y^{*L}_{t} = \sum_{st \in s} \hat{Y}^{L}_{st} \tag{4}$$

This final step gives the counts of Leave and Remain voters in each constituency. With these counts it is simple to calculate the proportion of all voters who voted Leave or Remain.

Although I have described a model for counts, many of the steps described above could be altered to produce a model for percentages. A model for percentages might seem more attractive. Very often we do not care about the number of votes cast for each option in an area, but only about the proportion of votes cast won by each option, and in particular whether or not a particular option in a two-way contest has secured more than 50% of the vote. However, it is not possible to alter the procedure above to model Leave (or Remain) votes as a percentage *of votes cast*, or even as a percentage of *all eligible voters*, but only Leave (or Remain) votes as a percentage *of the total population*. This is because the number of eligible voters or votes cast is not known at the level of the intersection between source and target geographies. In order to combine percentages from different units, we need to know something about the denominator in those percentages. Yet we do not know how many votes were cast in each intersection, or how many eligible voters reside there. We do know how many people live there, and so we could alter the procedure above to model votes cast as a percentage of the population. But these percentages are not of direct interest in the same way that percentages of votes cast are.

This reasoning does not suggest that a model for percentages is worse – simply that it does not have the intuitive appeal that it might appear to have on the face of things. A model for counts might be preferred on

other grounds. After all, aggregating is simpler for counts, requiring only addition. Aggregating for percentages is more complicated, because it requires division (each intersection's population must be divided by the total higher level area's population to create a set of weights), multiplication (each percentage must be multiplied by its weight), and addition (each product must be added together to produce an aggregate total). Additionally, a model for counts which includes an offset might (in a particular case) provide more accurate estimates than a model for percentages.[4]

3.5 Requirements and assumptions

There are certain assumptions implicit in this method which it is important to note, and one requirement.

First, the model requires detailed information about small geographical units which can be aggregated up to the level of intersections between source and target units. Where national censuses are conducted, this information may be measured at the level of the census tract, or enumeration district, or (as in this case) output area. Where no national census is conducted, or where it is released only at levels comparable in size to source of target units, this model will not be viable.

Second, the method assumes that the relationship between the additional covariates and the outcome is a strong relationship. Under certain circumstances, adding additional information can lead to lower accuracy (Sadahiro 1999). In this application, the relationship, as measured by different model fit statistics, is very strong. In other applications, it may be difficult for researchers to judge whether the relationship is "strong enough."[5]

[4] An online appendix demonstrates that this is the case for a small number of constituencies for which the result is known exactly.

[5] In an online appendix I provide simulation evidence to suggest that models with fit statistics comparable to the fit statistics reported in the following section almost always yield better estimates than estimates from dasymmetric interpolation.

Third, the method assumes that the same relationship found at the level of the source geography also holds at the level of the intersections. This assumption can be fallacious, and when it is fallacious it is closely related to the ecological fallacy. Just as a relationship measured at the level of a district may not hold at the individual level, a relationship measured at the level of the source geography may not hold at the level of the intersection geography. The more the source and intersection geographies differ in scale, the more likely this is to be true, though the effect is not restricted to such aggregation effects. Although the intersection geographies are smaller than local authority areas, they are not very much smaller: the median population in a local authority area is roughly 125,000; the median population in an intersection area is 84,000.[6]

Finally, the method assumes a particular functional form. I assume that the counts of Leave and Remain voters are Poisson distributed. Other functional forms are possible. Two alternatives are a negative binomial model and a linear model of the logged number of counts. In practice, and because of the scaling step, these alternative models deliver very similar estimates.[7]

4 The models

Table 1 shows the results of these two models. The models include all of the variables mentioned above, plus an interaction between age and education. The table does not report the coefficients for the regional dummies. Note that the coefficients represent the change in rates of voting Leave (Remain) as a percentage of the population. Variables that would ordinarily show a positive association with voting Leave (Remain) may

[6] To some extent the assumption can be assessed by simulation, and in the online appendix I provide such simulations.

[7] The average correlation between different functional forms is > 0.97. See the online appendix.

have the opposite sign if those variables depress turnout, and variables whose effects on turnout are greater than their effects on voting Leave or Remain may have the same sign across both models.

Table: Poisson regression models of Leave and Remain vote

	Leave	Remain
(Intercept)	-1.095***	2.543***
	(0.062)	(0.060)
Pct L1 qual.s	-4.911***	-6.298***
	(0.238)	(0.212)
Pct L2 qual.s	7.662***	-11.707***
	(0.191)	(0.181)
Pct L3 qual.s	-2.197***	-7.980***
	(0.234)	(0.221)
Pct L4 qual.s	-4.402***	-3.627***
	(0.086)	(0.083)
Median age	0.014***	-0.125***
	(0.002)	(0.002)
Pct owning house	0.474***	0.934***
	(0.022)	(0.021)
Pct White British	0.152***	0.322***
	(0.007)	(0.006)
Pct. higher managerial	1.785***	-2.424***
	(0.049)	(0.044)
Pct. lower managerial	-2.346***	2.549***
	(0.044)	(0.039)
Pct. lower supervisory	4.405***	-4.987***
	(0.078)	(0.085)
Semi routine	-2.743***	1.099***
	(0.054)	(0.059)
Routine	-0.900***	-0.911***
	(0.039)	(0.044)
Never worked or long-term unemployed	-4.475***	-1.209***
	(0.052)	(0.053)
Level.1.qualifications x MedianAge	0.139***	0.195***
	(0.006)	(0.005)
Level.2.qualifications x MedianAge	-0.158***	0.249***
	(0.005)	(0.004)
Level.3.qualifications x MedianAge	0.003	0.226***
	(0.006)	(0.006)
Level.4.qualifications.and.above x MedianAge	0.049***	0.158***
	(0.002)	(0.002)
McFadden R-sq.	0.95	0.85
N	380	380

The coefficients in the model are not particularly interesting, both because they reflect a mixture of effects on turnout and effects on vote choice, and because they are not intended to capture causal effects. The purpose of this model is simply to explain a high proportion of the variance in

rates at which people turn out to vote for either option, so that the model can then be used to make projections on to a different geography.

We can assess the fit of the model by using a variety of pseudo-R-squared measures. Not all of these measures cope equally well with the presence of an offset in a model. The measure I find most useful is McFadden's R-squared, which is equal to one minus the log likelihood of the fitted model divided by the log-likelihood of the null (offset-only) model. As the table shows, on this measure both models perform extremely well. As such, both of these models can be used to make predictions at the level of the target geography.[8]

5 Projections

The models shown in the previous table can be used to generate predictions of the votes cast for each option in each of the 851 areas formed by the intersection of local authority areas and Westminster constituencies. These predictions can then be scaled in order to ensure that they add up to the known results at local authority levels. These scaled predictions can finally be aggregated to the level of the 632 Westminster parliamentary constituencies in Great Britain, providing us with an estimate of the likely outcome in each seat.

Table 2: Estimated outcome by party holding seat

	Leave	Remain
Conservative	245	85
Green	0	1
Labour	149	83
Liberal Democrat	2	6
Other	0	1
Plaid Cymru	1	2
Scottish National Party	2	54
UKIP	1	0

[8] If pseudo R-squareds are judged unhelpful, then an alternative way of evaluating the fit of both models is to calculate the mean absolute error on the Leave share of the vote, which works out at 2.5%.

Leave	Remain
400	232

Table 2 provides a count of the estimated number of seats which voted for Leave or for Remain, according to the party which won the seat in the 2015 election. Overall, 400 (63%) of seats in Great Britain were "won" by the Leave campaign; this figure increases to 407 (63%) if we include the (known) results from Northern Irish constituencies. Leave was the most popular outcome in both Labour and Conservative-held seats. This poses a problem for the Labour party, which campaigned in favour of Remain. Although the Conservatives were the more divided party, their muddled, divided position more closely reflected the position of the country as a whole.

One natural question concerning these estimates is: are they are any good? We can compare these estimates to the known figures for 27 constituencies. These figures are known because local councils in these areas provided detailed breakdowns of the vote by constituency or by ward.[9]

These constituencies are not representative of the UK as a whole. All are urban. More than half are Scottish. None overlap entirely with local authorities. This means that the mean error reported for these constituencies will be greater than the mean error across all constituencies, since the mean error for all constituencies will include 35 perfect estimates where constituency boundaries perfectly match local authority boundaries. For this same reason, errors calculated on the basis of these constituencies likely over-state the degree to which a scaled Poisson regression model out-performs dasymmetric interpolation.

[9] These breakdowns are not perfect guides to the result in each ward. The result in each local authority area is a combination of votes cast on the day and postal votes. To the best of my knowledge, most councils did not allocate postal votes to specific wards or specific "mini-counts". Accordingly, the counting of postal votes was distributed between the different "mini-counts". The result for particular wards therefore represents a combination of the votes cast in that ward on the day, and a non-random allocation of postal votes from across the local authority area. In the general election of the previous year, one-fifth of votes were cast by post (Rallings and Thrasher 2015).

With these qualifications in mind, the mean absolute error across these 27 constituencies was 2.17 percentage points for the scaled Poisson regression model, and 6.29 percentage points for dasymmetric interpolation. The median absolute error was smaller. For the scaled Poisson regression model, half of constituencies had errors equal to or less than 1.62%, compared to an equivalent figure of 5.22% for dasymmetric interpolation.

6 Link to referendum signatures

A month before the referendum, a petition was created on the parliamentary petitions website which called for a second referendum in the event that the vote for either Leave or Remain was less than sixty percent, or if turnout was lower than seventy-five percent. Despite being created by a Leave supporter who anticipated defeat, the petition was repurposed by many Remain voters. After Remain's defeat in the referendum, the number of signatories increased rapidly. Within one week of the referendum, it had accumulated four million signatures.

The parliamentary petitions website provides data not just on the total number of signatories, but on the number of signatories per parliamentary constituency. These figures can be expressed as a fraction of the electorate in each constituency. The highest rates were found in Cities of London and Westminster, Hornsey and Wood Green, Kensington, the lowest rates in West Tyrone, Upper Bann, Strangford.

The correlation between the rate at which the petition was signed and the estimated share of the Remain vote in each area is high, at $r = 0.73$. This correlation – and the much lower rates observed in Scottish constituencies – can be seen in Figure 1.

This correlation may over-state the actual impact of Remain votes on signing behaviour if, for example, the areas which tend to vote Remain are areas which generally have high signing rates for most petitions (which

might in turn result from higher rates of internet use). I therefore model the rate of signing as a function of (a) the estimated Remain share in each constituency; (b) the total number of signatures on any petition, per constituency, as of December 2015, divided by the electorate in each constituency, expressed as a percentage (mean = 12.2; SD = 2.9); and (c) a dummy for constituencies in Scotland.

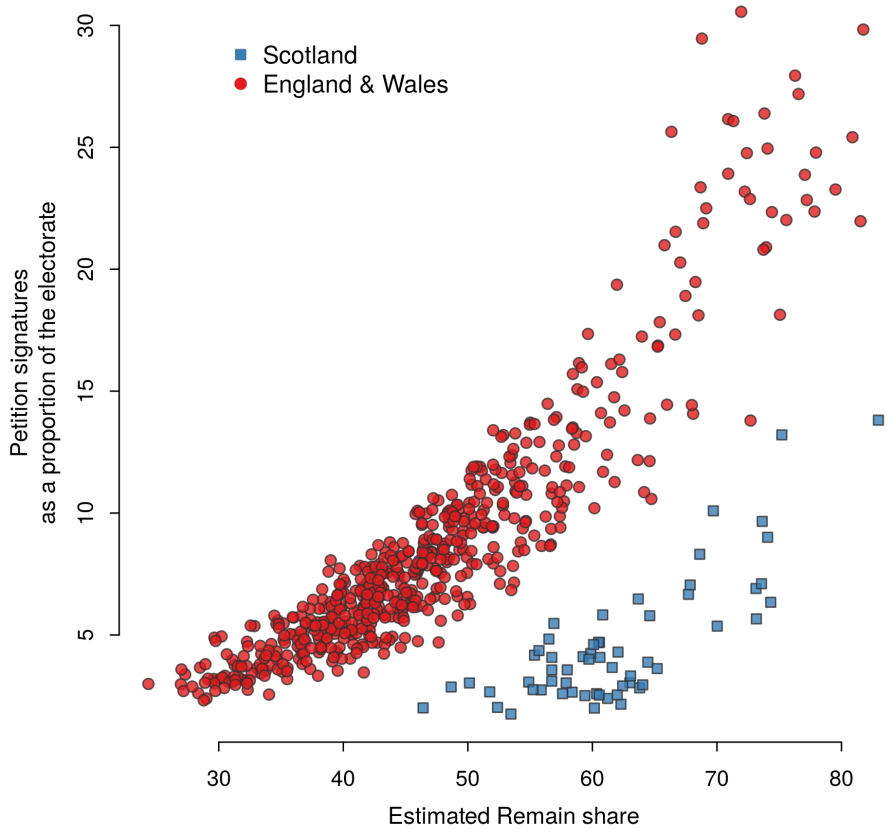

Figure 1: Remain vote share against rate of petition signatures

Table 3: Regression model of signatures to referendum petition

	Petition signatures
(Intercept)	-11.927***
	(0.379)
Rate of signing petitions in general	0.249***
	(0.036)
Remain share of the vote	0.374***
	(0.010)
Scottish seat	-9.314***
	(0.323)
R-squared	0.85
N	632

The model suggests that the number of signatories to the petition went up in line with the share of the Remain vote in the constituency, and in particular the number of signatures grew by an amount equal to four-tenths of a percentage point of the electorate in that constituency. I cannot conclude that these signatories were Remain voters. EU citizens resident in Britain can petition their MP but were not eligible to vote in the election. However, the results of the regression model do show that most of the signatures to this petition came from areas which supported Remain, and that this pattern cannot be explained away by pointing to the generally higher rates at which these constituencies sign online petitions.

7 Conclusion

In this short note I have demonstrated a method to recover estimates of the Leave and Remain share of the vote in Westminster constituencies. The method of areal interpolation I have used will be useful for other jurisdictions which, like the UK, use multiple, overlapping electoral geographies, and which either do not release detailed (ward-level) data, or release it on an irregular basis. The method does, however, require that both the source and the target geography can be represented in terms of smaller administrative units, and that Census data (or other demographic data) be available for these smaller administrative units.

The estimates I have produced – which form a supplementary appendix to the note – will be useful for researchers interested in investigating the future consequences of the Brexit. I have demonstrated one such application, where information on the relevant outcome (petition signatures) was only available at the level of the parliamentary constituency, rather than the local authority area. This is likely also to be the case for many other future outcomes of interest.

8 Appendix

8.1 Alternative functional forms

In the main body of the paper I noted that the method I use assumes a particular functional form, and that other functional forms would be possible. The purpose of this section is to show that other functional forms generate very similar estimates.

Table 4 shows the correlation between estimates of the Leave share of the vote from the following different model forms:

- the model used in the paper, which uses a scaled Poisson regression model
- a model which uses a scaled negative binomial model;
- an ordinary least squares model which uses as dependent variable the log of the number of voters for Leave (Remain), which includes the log population as a model term rather than as an offset; but which is scaled in the same way as the first two models
- an ordinary least squares model which uses as dependent variable the percentage of the population who voted for Leave (Remain), which includes the log population as a model term rather than as an offset, and which is scaled in the same way as the first two models
- (as a comparison) dasymmetric interpolation

Table 4: Correlation between different model forms

	Poisson	Negative binomial	Log-linear	Percent	Dasymmetric
Poisson	1	0.999	0.982	0.996	0.928
Negative binomial	0.999	1	0.983	0.998	0.929
Log-linear	0.982	0.983	1	0.981	0.9
Percent	0.996	0.998	0.981	1	0.929
Dasymmetric	0.928	0.929	0.9	0.929	1

As the table demonstrates, the correlation between all of the different functional forms is extremely high, and above 0.9 in all cases. However, correlations between the model-based methods are always higher than the correlations between any model-based method and dasymmetric interpolation.

High correlations between different model forms do not imply that the estimates have low error. It is possible to compare the estimates from all of these models with the know results from the 24 constituencies mentioned in the body of the article. Table 5 shows the mean absolute error and the 90% range for these different methods.

Table 5: Absolute error by method

Method	MAE	90% of errors within…
Negative binomial	2.135	(0.38, 4.69)
Poisson	2.172	(0.31, 4.49)
Percentage	3.22	(0.61, 8.57)
Dasymmetric	6.287	(1.57, 12.26)
Log-linear	14.17	(3.23, 25.27)

The dasymmetric model is not the worst, being beaten by an ordinary least squares model where the dependent variable is the log number of voters for each option. Of the different model forms, the negative binomial model performs best on the basis of mean absolute error, but given (a) the small difference in MAE, equal to one twentieth of a percentage point; (b) the non-representative nature of the constituencies selected; and (c) the greater parsimony of the Poisson model, I continue to use the Poisson model.

8.2 Ecological fallacy

In the main body of the article I noted that the method I use assumes that the same relationships found at the level of local authorities also obtain at the level of intersections between local authorities and Westminster con-

stituencies, and that strictly speaking this assumption is fallacious. This fallacy is not just theoretical: many papers over the years have demonstrated that different bivariate correlations may be obtained depending on the way units are aggregated (Openshaw and Rao 1995; Openshaw and Taylor 1979).

In order to test whether this assumption was met, I carried out simulations. I drew one set of coefficients from the Leave model shown in Table 1, and one set of coefficients from the Remain model, and used these coefficients to simulate outcomes at the level of the intersection. I then aggregated these outcomes to the level of the local authority, and estimated the same model as that shown in Table 1, saving the coefficients. I was then able to compare the "known" coefficients with the estimated coefficients. Across 1000 simulations, the average correlation was 0.993. This suggests that the aggregation of intersections to local authority areas does not markedly change the recovered relationship.

8.3 "Good enough" models

In the main body of the article I noted that adding auxiliary information need not always improve the accuracy of estimates relative to simpler methods. In Appendix Table 5 I showed that the scaled Poisson regression produced better estimates than dasymmetric interpolation for a small number of constituencies for which information was available. It is therefore not clear whether the method I have set out in the article produces estimates that are better than dasymmetric interpolation.

Once again, I turn to simulation to assess whether the model is good enough to provide better estimates than can be provided by dasymmetric interpolation.

I proceed as follows:

- I sample from the multivariate distribution of coefficients as reported in Table 1;
- I generate counts of Leave and Remain voters at the level of the intersection;
- I then aggregate these up to (a) local authority level and (b) parliamentary constituency level;
- I draw from a uniform distribution between one and eighteen. Call this number v;
- I then randomly select v terms from the list of model terms found in Table 1
- With these v terms, I carry out a scaled Poisson regression to estimate Leave and Remain vote shares
- I also carry out dasymmetric mapping to estimate Leave and Remain shares
- For each simulation, I calculate the mean absolute error for both methods

The results of these simulations are shown in Figure 2. The MAE for dasymmetric interpolation is almost constant, as this merely reflects the variation in the coefficients used to generate the known results at intersection level. The variation in the MAE for the Poisson model reflects the success of the model, which in turn depends on the number and identity of the variables randomly selected to be part of the model. Although there are simulations where the Poisson regression model delivers worse results than dasymmetric interpolation, these are few in number, and typically occur where the performance of the model is poor compared to the models used in the main body of the article.

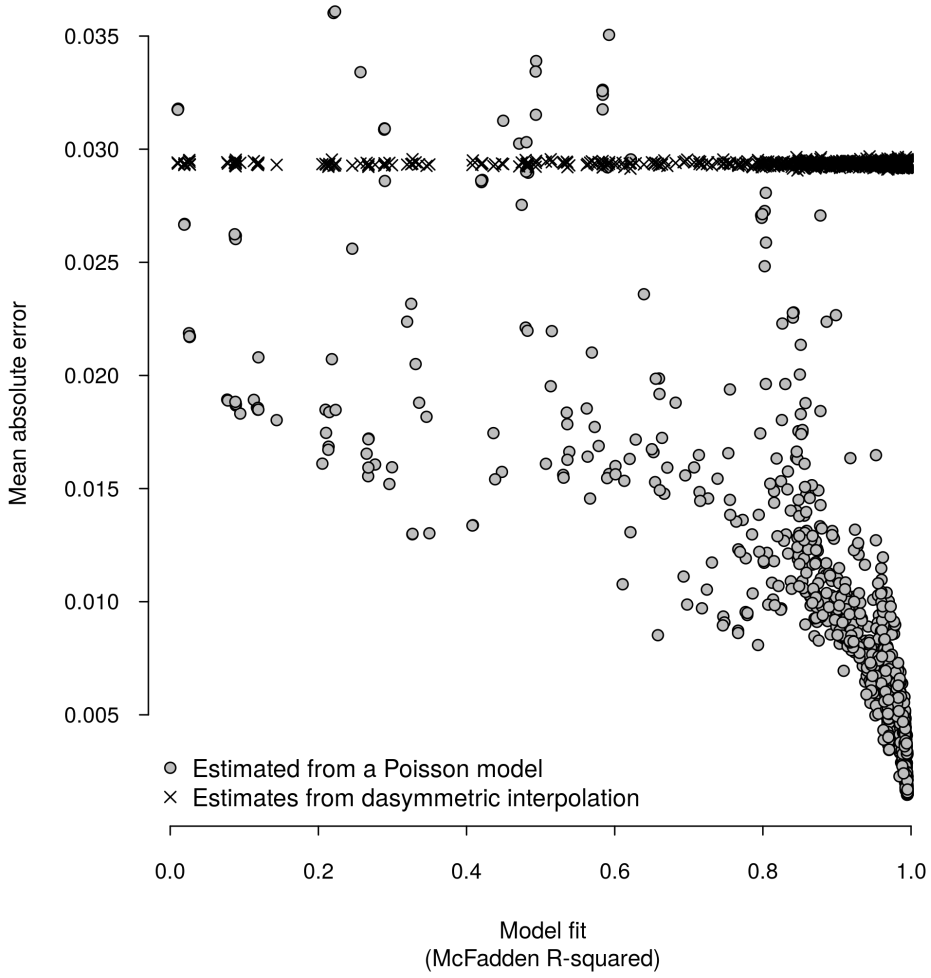

Figure 2: Mean error by model performance, simulated results

References

Flowerdew, Robin, and Mick Green. 1989. "Statistical methods for inference between incompatible zonal systems." *Accuracy of spatial databases*: 239–247.

Ford, Robert, and Matthew Goodwin. 2014. *Revolt on the right: Explaining support for the radical right in britain*. Routledge.

Goplerud, Max. 2015. "Crossing the boundaries: An implementation of two methods for projecting data across boundary changes." *Political Analysis*: mpv029.

Hanretty, Chris, Benjamin E Lauderdale, and Nick Vivyan. 2016. "Dyadic representation in a westminster system." *Legislative Studies Quarterly* Online early-access. `http://dx.doi.org/10.1111/lsq.12148`.

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

Openshaw, Stan, and Liang Rao. 1995. "Algorithms for reengineering 1991 census geography." *Environment and planning A* 27(3): 425–446.

Openshaw, Stan, and Peter J Taylor. 1979. "A million or so correlation coefficients: Three experiments on the modifiable areal unit problem." *Statistical applications in the spatial sciences* 21: 127–144.

Rallings, Colin, and Michael Thrasher. 2015. *The 2015 general election: Aspects of participation and administration*. Elections Centre, Plymouth University.

Sadahiro, Yukio. 1999. "Accuracy of areal interpolation: A comparison of alternative methods." *Journal of Geographical Systems* 1(4): 323–346.

Thomas-Agnan, Christine, Anne Vanhems, and others. 2015. "Spatial reallocation of areal data–another look at basic methods." *Revue d'Économie Régionale & Urbaine* (1): 27–58.