Voting behaviour, individual data and the ecological fallacy

A. Forcina, Dipartimento di Economia, University of Perugia, Italy
M. Gnaldi, Dipartimento di Scienze Politiche, University of Perugia, Italy
V. Tomaselli, Dipartimento di Scienze Politiche e Sociali, University of Catania

November 18, 2014

Abstract

We analyse a rather unique set of individual data on voting behaviour for an investigation into the ecological bias. We review and clarify the conditions under which ecological inference will produce biased estimates and explain why this is going to affect different estimation methods. Our investigation highlights that only the ecological inference methods which allow to model the effect of covariates may produce unbiased estimates. However, as the application to our data shows, even the use of covariates and more sophisticated methods of ecological inference may fail to produce satisfactory estimates. A logistic multilevel analysis is applied to assess the true effect of covariates and to estimate the importance of additional sources of random variations.

Keywords: Ecological inference, voting behaviour, logistic models, multilevel analysis.

1 Introduction

In studies of voting behaviour, an important issue is the shape and strength of association between the choices of voters in a given election and that in a previous election; one may also be interested to study how much voters’ choices depend on their sex, age, education and social class. However, due to the special nature of electoral data, the true joint distributions between such pairs of variables cannot usually be observed and the only source of information can be extracted either from official electoral data aggregated at the level of polling stations or from sample surveys.

Inference based on aggregated data has the advantage, relative to those based on sample surveys, to be unexpensive and to refer to the whole population under investigation; sample surveys, on the other hand, are based on direct answers given by voters. An attempt to compare the estimates provided by the two approaches in the study of the association between the choices made in two related elections is presented in [Russo (2014)]; however, because in her data the proportions of voters for different parties in the elections seem to be substantially different from the corresponding proportions in the official data, it seems difficult to draw valid conclusions from her study where the data about the true joint distribution are not available. [Liu (2007)] also compared the performance of sample surveys to that of several ecological inference methods for estimating the association between race and propensitum to register in a New Orleans majoral election. He showed that, relative to the proportions of registered voters provided by true data, most ecological inference methods did better than survey data in that specific context.
Since Robinson (1950)'s seminal paper, it is well known that the association between two variables estimated from data aggregated within geographical units, like polling stations, may be substantially biased relative to the association that would emerge if data recorded at the individual level were available. The phenomenon, which came to be known as the ecological fallacy, was used by Robinson as an argument for banning ecological inference from sociological investigations. Nowadays this recommendation has not, perhaps, many followers due to a better understanding of the conditions which may produce an ecological fallacy and the emergence of more sophisticated methods of ecological inference. In addition, as Subramanian et al (2009) pointed out, in many contexts, the degree of association at the individual level depends itself on modelling assumptions and is not such an objective quantity as Robinson seemed to believe.

In this paper we use a data set from the city of Palermo, Italy, which gives the proportion of eligible voters who turned out to vote at the 2012 Democratic Party primary election, distinctly for sex and six different age groups. Though the data are aggregated at the level of polling stations, small local units with an average of about 950 voters, all the ecological inference methods that we applied provide estimates which are substantially different from those provided by the true data. This has an explanation along the lines of Wakefield (2004): the proportions of people who turn out to vote within each group, by sex and age, are highly correlated with the relative size of these groups. Estimates continue to be biases even if we fit ecological inference models where the relative size of such groups are used as covariates, a feature which, as we explain, goes beyond ecological fallacy.

In Section 2 we recall the main features of ecological data, review some methods of ecological inference and explain why, even when individual data are available, the assessment of the nature of the association may depend on modelling assumptions. In Section 3 we analyse a data set on voting behaviour for the city of Palermo, clarify why ecological bias is to be expected in this context and why, even the use of appropriate covariates, cannot solve the problem.

2 Methodological aspects

Let \(Y\) denote the set of options \(C\) available to voters in a give election and \(X\) a discrete variable with \(R\) categories which we expect to be associated with \(Y\), like, for instance, age groups, sex, social class or the choice made in a previous election. Suppose we are interested in the association between \(X\) and \(Y\) within a given area, like a borough, which is divided into a collection of \(N\) polling stations. Let \(n_{uij}\) denote the number of voter in polling station \(u\) with \(X = i\) and \(Y = j\); in most cases these data, which we may call individual, are not known and we have to relay on the aggregated data: \(x_{ui}\), the number of voters with \(X = i\) and \(y_{uj}\) those who voted option \(j\) in the new election. Let \(p_{uij}\) denote the proportion of voters who choose option \(j\) among those with \(X = i\) in polling station \(u\); if \(X\) and \(Y\) were binary variables, like for instance when \(X\) denotes sex and \(Y\) is whether one goes to vote or not, the association may be measured by the correlation coefficient or the odds ratio. With \(R \times C\) tables, the degree of association is 0 when the probabilities \(p_{uij}\) do not change with \(i\). Though there are many summary measures of association, in the following we will simply look at how much \(p_{uij}\) changes with \(i\).
2.1 Ecological inference methods

The so-called Goodman (1953) model is essentially based on two assumptions: (i) voters with $X = i$ choose option $Y = j$ with probability $\pi_{ij}$ which is constant across polling stations and (ii) voters choose independently from each other. These assumptions imply that the $n_u$ voters in polling station $u$ split among the $C$ options according to a sum of $R$ multinomial distributions. Under these assumptions no ecological fallacy can arise and an unbiased estimate of the $\pi_{ij}$ parameters can be computed by ordinary least squares applied to the following set of equations where we write $v_{uj} = y_{uj}/n_u$ and $t_{ui} = x_{ui}/n_u$ to denote the marginal proportions.

Because the marginal proportions $t_{ui}$ sum to 1, the proportion for $I = R$, that is $t_{iR}$, may be obtained from $1 - \sum_{i=1}^{R-1} t_{ui}$

$$v_{uj} = \sum_{i=1}^{R} t_{ui} \pi_{ij} + \epsilon_{uj} = \pi_{Rj} + \sum_{i=1}^{R-1} (t_{ui} - t_{uR}) \pi_{ij} + \epsilon_{uj}; (1)$$

Note also that, because the probabilities $\pi_{ij}$ also sum to 1 within each row, the equation for the last column where $j = C$ is redundant.

The model proposed by Brown and Payne (1986) may be seen as a refinement of the Goodman model both in the assumptions on voting behaviour and in the method used to estimate the unknown parameters. In this model the probability that a voter in polling station $u$ and with $X = i$ chooses $Y = j$ is no longer assumed to be the same in all polling stations, but it is allowed to vary at random across polling stations as in a Dirichlet distribution whose average, which we may call again $\pi_{ij}$, is the same for all polling stations when covariates are not available. Because parameters are estimated by maximum likelihood on a logistic scale, the estimates are more efficient and lie always between 0 and 1. Recently, Forcina, Gnaldi and Bracalente (2012) have proposed a modified version of the above model by assuming that voters tend to cluster at random within each polling station in smaller groups whose size is itself random. Voters within the same cluster are assumed to share the same probabilities of selecting each $Y$ option. These assumptions imply a different variance structure for the observations.

A collection of methods proposed by Gary King and his coworkers, see for instance King (1997), King et al. (1999) and Rosen et al. (2001), have become very popular within certain scientific communities where they are considered to be the most advanced methods of ecological inference, though their merits are debated, see Liu (2007, pp. 6-7) for a review. In spite of their sophisticated Bayesian framework, the underlying assumptions, when translated into models of voting behaviour, are rather simple. For instance, in (Rosen et al., 2001, p. 137) it is assumed that the probability that a voter in polling station $u$ chooses option $j$ equals $\sum_i n_{ui} p_{uij}$; this implies that all voters in polling station $u$ are assumed to behave as an homogeneous group, irrespective of their value of $X$, an assumption which seems rather unrealistic, especially in the context of voting behaviour when $X$ is the party voted in the previous election; a similar criticism was raised by Greiner and Quinn (2009).

2.2 The ecological fallacy revisited

In spite of the impact that Robinson’s paper had on the scientific community, the true nature of ecological bias is not always well understood. For instance, according to Russo (2014) the ecological fallacy is due to the “incorrect assumption that individual members of a group have
the same characteristics as those of the group taken as a whole”, a statement which could be only vaguely correct if one replaced "individuals” with local units. Though the conditions required to prevent ecological bias in King’s models are clearly stated (see for instance [King et al. 1999], p. 67), the title of one of his first papers [King 1997]: “A solution to the ecological inference problem” may have lead some people to believe that these methods have some kind of intrinsic protection against the ecological fallacy; for instance [Seligson, 2002, p. 273] seems to be convinced that Gary King has advanced "towards solving the ecological inference problem” as long as we have ”relatively homogeneous ecological units”, a rather vague statement, though, somehow, closer to the truth.

Perhaps, one of the simplest explanations of the mechanism underlaying the ecological fallacy is provided by ([Wakefield 2004] Sec. 3.3) in the case where \(X\) and \(Y\) are binary variables and there are only two local units (for example polling stations); a numerical example along those lines is given in Table 1: while the proportion of females increases from 0.2 to 0.6, the overall proportion of those who turn out to vote decreases from 0.7 to 0.4; thus, any method of ecological inference would lead to conclude that the proportion of those who turn out to vote is smaller among females relative to males. Instead it can be seen that, within each table, the proportion of those who go to vote is higher for females relative to males. The apparent paradox has been produced intentionally as follows: when we go from the first to the second table the proportions of people who turn out to vote decrease for both males and females and, at the same time, the marginal proportion of females increases.

An explanation for the general case where we have a collection of \(N R \times C\) tables is easily provided for the Goodman linear regression model where, because of the accounting identity which imply that \(v_{ij} = \sum_i p_{uij} t_{ui}\), the error term may be written as

\[
\epsilon_{uj} = \sum_{i=1}^R t_{ui} (p_{uij} - \pi_{ij})
\]

The condition for the least square estimates to be unbiased is that \(\epsilon_{uj}\) is uncorrelated with the \(t_{ui}\) proportions which are the explanatory variables in the regression model; this condition, in turn, implies that

\[
E [ e_{uj} | t_{u} ] = 0
\]

where \(e_{uj}\) denotes the vector with elements \((p_{uij} - \pi_{ij})\) and \(t_{u}\) is the vector with the row proportions, (see, for example [Wooldridge 2004] Sec. 2.2). The same condition must hold for the Brown and Payne model and the frequentist version of King’s model as described in ([Rosen et al. 2001] Sec. 4) in order for the estimates to be consistent (see [Wooldridge 2004].
Sec. 12.2): though these models are based on a logistic regression, when covariates are not included, the error term has the same structure as in the Goodman model. As we will see in Section 3, where we analyse the Palermo data set, the estimates produced by these two more sophisticated models are not much closer to the truth than those of the Goodman model.

Whether or not the above condition is satisfied, when we fit a regression model like Goodman’s, the residuals will always be uncorrelated with the marginal row proportions $t_{ui}$, thus, when only ecological data are available, it is not possible to check the condition for no ecological bias. This is possible, instead, when the joint distribution of $X$ and $Y$ is observed in each polling station. In particular, when, like in our case, $Y$ is binary, a logistic regression model can be fit to test whether the observed proportions $p_{uij}$ depend on the marginal proportions $t_{ui}$.

### 2.3 Association in the individual data

We now try to explain why, even when the individual data are available, there may not be an obvious estimate of the overall association between $X$ and $Y$. Suppose first that no covariates are available and that, though the proportions $t_{ui}$ of voters with $X = i$ vary at random among polling stations, the voting behaviour conditionally on $X$ is the same. In this simple context we may sum the joint frequencies $n_{uij}$ with respect to polling stations and estimate the conditional probabilities as

\[
\hat{\pi}_{ij} = \frac{\sum_u n_{uij}}{\sum_j \sum_u n_{uij}},
\]

these estimates may then be used to analyse the association between $X$ and $Y$. This is essentially what Robinson did with his census data on race and illiteracy: he computed an estimate of the conditional probabilities as in equation (2), computed the correlation coefficient and compared it with the one based on estimates provided by an ecological regression with the states as local units.

When the conditional probabilities of choosing $Y = j$ given $X = i$ are not constant but vary as a logistic function of covariates measured at the level of polling stations, the degree of association will also vary across polling stations. In such cases one could try to fit a logistic model and then estimate the degree of association conditionally on a specific value of the covariates which is of interest, like their average, or simply average across polling stations the probabilities estimated with the logistic model.

Subramanian et al (2009) has studied in detail the case where, in addition to covariates, there is also substantial random variation across local units. They fit several different logistic multilevel models to individual data similar to those used by Robinson, except that race has three categories are used for race and an historical covariate is also used. Their results indicate that the models which take into account the additional variability across states (their local units) fit much better than the model based on the naive estimates given in (2); the more sophisticated models give a substantially different picture of association between race and illiteracy.

### 3 The Palermo data

In 2012, before the city borough election, the Democratic Party held his own primary election to choose his candidates. The city of Palermo is divided into about 600 polling stations; after
removing those located into temporary communities like the local hospitals and the prison, we
are left with a collection of 593 local units. The headquarters of the Party made available for
research a table where, for each polling station, voters are classified by sex, age and whether
they voted or not. A raw summary of the data is given in Table 2. Though there are more

Table 2: Proportion of eligible voters within males and females and proportion of those who
turned out to vote within each sex by age group

| Age Group | 18-25 | 25-30 | 30-45 | 45-65 | 65-75 | over 75 | Totals |
|-----------|-------|-------|-------|-------|-------|--------|-------|
| Males     | 0.122 | 0.083 | 0.264 | 0.338 | 0.112 | 0.081  | 265,632|
| Females   | 0.102 | 0.071 | 0.246 | 0.333 | 0.121 | 0.127  | 298,773|
| Males     | 0.0442| 0.0448| 0.0478| 0.0688| 0.0654| 0.0345 | 14,640 |
| Females   | 0.0435| 0.0467| 0.0445| 0.0647| 0.0427| 0.0156 | 14,156 |

females among eligible voters, overall there are more male voters. The propensity to vote
within the different age groups is not much different for males relative to females, however, it
is 1.53 times higher for males among voters aged 65-75 and 2.23 times higher among voters
over 75. Also, the tendency to go to vote for those aged 45-65 relative to those aged 18-25 is
about 1.56 times higher among males and 1.48 among females.

3.1 Logistic multilevel models

We also fitted a set of logistic regression models similar to those in Subramanian et al. (2009) to
estimate the size and structure of random variations among polling stations in the propensity
to vote and to see how this affects the estimates of the fixed effects. For each polling station
voters are divided according to sex and to 6 age groups as in Table 2; for the voters in each
sex by age group we know the size of the group and the number of those who voted. Polling
stations were grouped by the Democratic Party into 31 gazebos, these are collections of polling
stations located in the same surroundings. We also used four covariates measured for each
polling station which turned out to be significant:

- $pd$, the proportion of those who voted the Democratic Party at the municipal election
  held a month later;
- $idv$, the proportion of those who voted the Italia dei valori Party at the municipal
  election held a month later;
- $mol$, the proportion of males aged between 45 and 75;
- $fol$, the proportion of females aged between 45 and 75.

For the multilevel analysis, our data consist of 12 observations within each polling stations
and 593 polling stations grouped into 31 gazebos. In this model, there are three sources of
variation: (i) binomial within polling stations, (ii) among polling stations within gazebos with a standard deviation of 0.2311 and (iii) among gazebos with a standard deviation of 0.2547.

The multilevel logistic model above provides estimates of the probability to go to vote within each sex by age group which depends on the covariates and thus are different in each polling station. An overall estimate may be computed in two ways:

1. first compute the estimates within each of the 593 polling stations and then average;
2. compute the predicted values of the probabilities to go to vote for an hypothetical polling station whose covariates are equal to the average for the whole town.

The three different estimates (as in (2), average across polling stations and for an average polling station) are displayed in Figure 1.

![Figure 1: Estimates of voting probabilities, age groups on the x axis, F in the left panel, M on the right; row proportions on the blue line, averaged logistic multilevel estimates on the green line and multilevel logistic estimates at the average value of covariates on the red line](image)

It can be seen that the three different estimates of voting probabilities are very similar within each sex by age group. Thus, in this context, there is not much ambiguity in determining the pattern of association between sex-age and propensity to vote on the basis of the data at the individual level. This may be due to the fact that, while we are working with a medium sized town Robinson was dealing with a big country, in addition, our local units (polling stations) are of very small size while in Robinson’s case they were states. As an additional check, we also fitted a model similar to Model 4 in Subramanian et al (2009): in our context this is equivalent to assume that the random variation among polling stations and gazebos is different for each age group. The estimates of the standard deviation within each age group are displayed in Table 3: they do not seem to be too much different from each other though for some age groups there is more variation within polling stations rather than gazebos, though this does not seems to have a clear trend.
Table 3: Variability among polling stations and gazebos separately for each age group

|       | 18-25 | 25-30 | 30-45 | 45-65 | 65-75 | over 75 | all together |
|-------|-------|-------|-------|-------|-------|---------|--------------|
| Polling st. | 0.2482 | 0.2588 | 0.2628 | 0.2450 | 0.2957 | 0.3339 | 0.2311       |
| Gazebos  | 0.2589 | 0.1906 | 0.2321 | 0.2730 | 0.2782 | 0.2591 | 0.2547       |

3.2 Ecological inference

The estimates of voting probabilities provided by three different ecological inference methods without covariates are displayed in Table 4. It is easily seen that they are substantially different from those based on individual data in Table 2: for certain age groups estimated probabilities are close to 0 while those for other age groups are much too higher relative to those provided by all methods of estimation based on individual data displayed in Fig.1. Note also that the estimates provided by King OLS and the revised Brown-Payne methods are rather similar to one another.

Table 4: Ecological inference estimates of the probability of voting by sex and age groups

| Method  | sex | age groups | 18-25 | 25-30 | 30-45 | 45-65 | 65-75 | over 75 |
|---------|-----|------------|-------|-------|-------|-------|-------|--------|
| Goodman | F   | 0.000      | 0.000 | 0.000 | 0.124 | 0.000 | 0.128 |
|         | M   | 0.000      | 0.000 | 0.000 | 0.123 | 0.147 | 0.000 |
| BP rev. | F   | 0.000      | 0.000 | 0.000 | 0.148 | 0.000 | 0.152 |
|         | M   | 0.000      | 0.000 | 0.000 | 0.278 | 0.000 | 0.000 |
| King OLS| F   | 0.000      | 0.000 | 0.000 | 0.148 | 0.000 | 0.164 |
|         | M   | 0.000      | 0.000 | 0.000 | 0.267 | 0.000 | 0.000 |

As we saw in Sec. 2.2, we may use the individual data to check whether there are the conditions for obtaining estimates from ecological data which are, at least, consistent. To test for this, we used the following procedure:

- we fitted a logistic multilevel model separately for each age group, where the number of voters and non voters, for males and females within each polling station are the observations at the first level, nested within polling stations which, as before, are nested within gazebos;

- as potential covariates we considered the proportion of eligible voters belonging to each age group separately for males and female, in addition to the pd and idv covariates described above;

- an informal model selection procedure was used to select which covariates should be included.

Though the proportion of voters aged 45-65 and 65-75 were significant most of the times, when pd and idv were also used, some of the previous covariates appeared to have no longer a
significant effect. This could be due to the fact that \(pd\) and \(idv\) are closely related to the age distribution within each polling station. More precisely, when either \(pd\) or \(idv\) increases, the proportion of eligible voters in the age range from 18 to 45 decreases while the proportion in the age range from 45 to over 75 increases. The parameter estimates are displayed in Table 5. The fact that, most of the times, two or more of the covariates are highly significant indicates that the probabilities to vote for the corresponding age group are strongly correlated with the marginal proportions and thus consistent estimates cannot be obtained by ecological inference in this specific context. The same conclusion is implied by the significance of the \(pd\) or \(idv\) covariates which are correlated with the marginal distribution of eligible voters.

One may wonder whether ecological estimates that are closer to the truth may be obtained by using appropriate covariates; [Liu (2007)] found that the estimates from King’s model improved substantially by including certain covariates. We have fitted a revised Brown and Payne model where we allow the probability of voting for each sex by age group to depend on the same set of covariates which appeared to have a significant effect in the logistic multilevel models described above. Unfortunately, the resulting estimates are not much better than those given in Table 4. To gain a better understanding of what is happening, we tried to compare the predictions from ecological inference modelling with covariates with those obtained from the logistic multilevel models based on individual data. The actual procedure is describe below:

- For each polling station we compute the overall number of voters predicted by both the Brown and Payne model with covariates and by the set of multilevel models (one for each age group). These predictions were compared with the observed values and the standard deviation of the error for the two methods was computed. This is equal to 15.45 for the Brown-Payne model and 15.94 for the multilevel models, thus, ecological inference with suitable covariates predicts the overall number of voters more accurately than the multilevel models which use individual data.

- Both estimation methods provide also an estimate of the number of voters classified by age and sex within each polling station and we may compare again these predictions with the observed frequencies, compute errors and the standard deviation within each sex by age group for the two methods. Here the picture is quite different: the multilevel

Table 5: Estimated parameters for the multilevel logistic models for the propensity to vote; \(F\) is the intercept within females, \(M - F\) is the difference in intercept between males and females; \(\circ =\) non significant, \(\star = 5\%\) significant, \(\ast = 1\%\) significant, \(\bullet = p\text{-value smaller than 0.001}

| Parameters | Age groups | 18-25 | 25-30 | 30-45 | 45-65 | 65-75 | over 75 |
|------------|------------|-------|-------|-------|-------|-------|---------|
| \(F\)     |            | -4.9207\(\ast\) | -4.5096| -4.4335\(\ast\) | -3.6369\(\ast\) | -4.7301\(\ast\) | -5.0218\(\ast\) |
| \(ps\)    |            | 13.1413\(\ast\) | 14.8040\(\ast\) | 8.7711\(\ast\) | 12.4372\(\ast\) | 7.9891\(\ast\) | 7.1119\(\ast\) |
| \(idv\)   |            | 4.5733\(\ast\) | 0.0000° | 4.2559\(\ast\) | 5.2272\(\ast\) | 4.0019\(\ast\) | 10.3799\(\ast\) |
| \(P(45 - 65)\) |            | 2.3815\(\ast\) | 2.6580\(\ast\) | 1.6830\(\ast\) | 0.0000° | 2.0258\(\ast\) | 0.0000° |
| \(P(65 - 75)\) |            | 2.3888\(\ast\) | 0.0000° | 1.9564\(\ast\) | 1.3247\(\ast\) | 2.9426\(\ast\) | 0.0000° |
| \(M - F\) |            | 0.0256° | -0.0570° | 0.0853\(\bullet\) | 0.0898\(\bullet\) | 0.4785\(\bullet\) | 0.8171\(\bullet\) |
models provide a better prediction within each group of the number of voters classified by sex and age, as displayed in Figure 2: though for certain groups the two errors are similar in others ecological inference estimates have a much larger error.

![Figure 2: Standard deviation in the prediction of the number of voters in each sex by age groups; age groups are on the x axis, females on the eft panel](image)

The results above suggest that, even if we manage to adjust for the sources of ecological fallacy by fitting ecological inference models with covariates, this may not be sufficient to obtain reliable estimates. It may happen, like in our data, that while the probability to vote for certain categories of \( X \) increases with a covariate, it may decrease for other categories of \( X \) in such a way that these two effect partly compensate at the aggregated level. It follows that our ecological inference model may predict very well the variations of the aggregate proportion and, at the same time provide rather poor estimates of the probability of voting within each age by sex group.

### References

Brown, P., Payne, C., 1986. Aggregate data, ecological regression and voting transitions. *Journal of the American Statistical Association*, 81, 453–460.

Forcina, A., Gnaldi, M. and Bracalente, B., 2012. A revised Brown and Payne model of voting behaviour applied to the 2009 elections in Italy. *Statistical Methods and Applications*, 21, 109–119.

Goodman, L. A., 1953. Ecological regression and the behaviour of individuals. *American Sociological Review*, 18, 351–367.

Greiner, D.J. and Quinn, K.M., 2009. R x C ecological inference: bounds, correlations, flexibility and transparency of assumptions. *J. Royal Statist. Soc. A*, 172, 67–81.

King, G., 1997. A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data. Princeton University Press, Princeton, NJ.

King, G., Rosen, O. and Tanner, M. A., 1999. Binomial-Beta Hierarchical Models for Ecological Inference. *Sociological Methods & Research*, 28, 61–30.
Liu, B., 2007. EI extended model and the fear of ecological fallacy. Sociological Methods and Research, 20, 1–23.

Robinson, W.S., 1950. Ecological correlations and the behaviour of individuals. Amer. Sociological Rev., 15, 351–357.

Rosen, O., Jiang, W., King, G. and Tanner, M.A., 2001. Bayesian and frequentist inference for ecological inference: the RxR case. Statistica Neerlandica, 55, 134–156.

Russo, L., 2014. Estimating floating voters: a comparison between the ecological inference and the survey methods. Quality and Quantity, 48, 1667-1683.

Seligson, M.A., 2002. The renaissance of political culture or the renaissance of the ecological fallacy? Comparative Politics, 34, 273–292.

Subramanian, SV., Jones, K., Kaddour, A. and Kreiger, N., 2009. Rvisiting Robinson: the perils of individualistic and ecologic fallacy. Intern. Journ. of Epidemiology, 38, 342–360.

Wakefield, J., 2004. Ecological inference for $2 \times 2$ tables. Journal of the Royal Statist. Soc. A, 167, 1–42.

Wooldridge, J.M., 2004. Econometric analysis of cross section and panel data. The MIT Press, Cambridge, Massachusetts.