J. R. Statist. Soc. A (2005)
168, *Part* 2, *pp.* 387–399

Estimating Parliamentary composition through electoral polls

Frederic Udina

Universitat Pompeu Fabra, Barcelona, Spain

and Pedro Delicado

Universitat Politècnica de Catalunya, Barcelona, Spain

[Received February 2003. Revised August 2004]

Summary. All electoral systems have an *electoral formula* that converts proportions of votes into Parliamentary seats. Pre-electoral polls usually focus on estimating proportions of votes and then apply the electoral formula to give a forecast of Parliamentary composition. We describe the problems that arise from this approach: there will typically be a bias in the forecast. We study the origin of the bias and some methods for evaluating and reducing it. We propose a bootstrap algorithm for computing confidence intervals for the allocation of seats. We show, by Monte Carlo simulation, the performance of the proposed methods using data from Spanish elections in previous years. We also propose graphical methods for visualizing how electoral formulae and Parliamentary forecasts work (or fail).

Keywords: Electoral forecast; Electoral formula; d'Hondt rule; Monte Carlo simulation; Seat allocation

1. Introduction

Designing and conducting electoral polls involves several well-known steps. In this paper we study problems relating to the estimation of Parliamentary composition from a statistical point of view.

Let K parties be contending for a total of M seats in a Parliament. Let C be the number of electoral districts or constituencies, and let M_j be the number of seats that are decided by district j, with $\Sigma_j M_j = M$. After the elections, the proportion of valid votes f_{ij} that is obtained by party i in district j is made known. The electoral formula in use (see Section 4 for an analysis of some of the more usual ones) is applied to these proportions to distribute the seats between the parties. Let m_{ij} be the number of seats that are obtained by party i in district j.

The effect of different electoral formulae from the political point of view has been studied at length (see Cox (1997), Taagepera and Shugart (1989) or Benoït (2000)). However, to the best of our knowledge, the intrinsic statistical problems relating to electoral forecasting have not been fully investigated. Brown and Payne (1984) described the methods that were used by the British Broadcasting Corporation for election night forecasting of the 1983 British general election. The methods are very specific because of the special electoral rules in Britain: each of a large number of constituencies (650 in 1983) decides a seat by the majority rule. Katz and King (1999) proposed a statistical model for multiparty district level electoral data based on covariates

Address for correspondence: Frederic Udina, Departament d'Economia i Empresa, Universitat Pompeu Fabra, Ramon Trias Fargas 25, 08005 Barcelona, Spain.
E-mail: udina@upf.es

 0964–1998/05/168387

rather than pre-electoral polls. Although the model seems to work well in contexts where a few parties contend in a few hundred districts, we found it less useful for forecasting based on pre-electoral polls, where covariates are not commonly used, and in cases such as ours, with more parties, fewer districts and greater variability between districts. Bernardo (1984) described a Bayesian hierarchical model that was used in the Spanish general elections of October 1982. It focuses on forecasting the proportion of votes and does not study in detail the problems that we study here, namely those related to forecasting the number of seats that are allocated to each party.

A typical pre-electoral poll attempts to estimate the proportions f_{ij} by fixing a total sample size N and distributing it between the districts. The distribution rule is usually somewhere between one consisting of the simple division $n_j = N/C$ and the proportional distribution according to the number of potential voters in each district. After conducting the poll, the sample proportions $\hat{f}_{ij}$ will be the estimators for the unknown proportions. From these estimated proportions, the estimated number of seats $\hat{m}_{ij}$ can be computed through the electoral formula in use.

The main point of this paper is to show that a significant bias appears when the Parliamentary composition is estimated by adding up the $\hat{m}_{ij}$. We show it both graphically and numerically. We design and describe a graphical method based on principal components to visualize the forecasting of several electoral polls once the final results are known.

The bias in the prediction of the Parliamentary composition depends on the actual proportions of votes (which are unknown when the poll is being conducted) and the sample size that is used in each district. In most cases, the bias vanishes when the sample size increases, but there are some values of the proportions that make forecasting the results impossible: the simplest case occurs when two parties each have 50% of voters and they are contending for a single seat. We also study methods for computing confidence intervals for seat allocation based on the parametric bootstrap. We study, by Monte Carlo simulation, the performance of the proposed methods using data from electoral polls and elections in Spain from the year 2000, and in Catalonia for the regional Parliament in 1999.

In this paper, we do not address the non-sampling errors that seriously affect polls: misresponses, detection of abstention and no response or missing data (see Voss *et al.* (1995) for details on these issues). We shall see that even under ideal sampling conditions complex and interesting problems remain in the estimation of the Parliamentary composition.

The paper is organized as follows. Section 2 describes the Spanish electoral systems. In Section 3 we describe the main problem of bias in the Parliamentary composition forecasts, by using a graphical method. Such a method can also be used to show the discrepancy between the forecast of any real electoral poll and those resulting from Monte Carlo simulation. In Section 4 we briefly describe and analyse the most commonly used electoral formulae and we study the origin and consequences of the estimation bias. After the conclusions, the mathematical details are included in Appendix A.

2. The Spanish electoral system

Results from the 1999 Catalan regional Parliament election and 2000 Spanish general election are used throughout the paper to illustrate the performance of the methods proposed. In this section we describe the electoral system and circumstances concerning both electoral processes.

In the 2000 Spanish general election (see Colomer (2001) for a description of the political context) 350 seats were to be shared out. The number of constituencies (referred to hereafter as districts) was 52. The number of seats in a district is approximately proportional to the electoral roll. The distribution of the number of seats per district is represented in Table 1, as a stem-and-leaf

Table 1. Stem-and-leaf display for the number of seats in the constituencies, Spain 2000†

19	0	1133333333344444444
(27)	0	555555555556666777777899999
6	1	013
3	1	6
2	2	
2	2	
2	3	14

†Leaf unit 1, 0.

display (the first column shows cumulative frequencies from the top and from the bottom to the stem that contains the median, whose absolute frequency is shown in parentheses). There is a large majority of medium or small constituencies; only Madrid and Barcelona have more than 30 seats.

12 parties obtained at least one seat in the 2000 Spanish general election. Three of them are well established nationwide: the Popular Party (denoted PP; 183 seats in 2000), the Socialist Party (denoted PSOE; 125 seats) and the ex-Communist Party (denoted IU; eight seats). The Popular Party and the Socialist Party obtained seats in (almost) every district, but the ex-Communist Party could take a seat only in large districts. The other nine parties are regional parties, mainly from Catalonia, the Basque Country or Galicia (denoted BNG).

The electoral rule that is used in Spain is the d'Hondt rule applied to each district, with a required threshold of 3% of votes. Given the relatively small size of many districts, it is difficult for the small parties (the ex-Communist Party, for instance) to win a seat. As an example of this, the ex-Communist Party had about 30% more votes than the main Catalan nationalist party (denoted CIU), but the Catalan nationalist party obtained 15 seats and the ex-Communist Party only eight, because the votes of the Catalan nationalist party were concentrated in only four districts.

The Catalan Parliament has 135 seats that correspond to four districts: Barcelona (85 seats), Tarragona (18), Girona (17) and Lleida (15). Five parties could obtain at least one seat: the centre–right nationalist party (denoted Cill; 56 seats in 1999), the Catalan Socialist Party (denoted PSC-CpC; 52 seats), the Popular Party (12 seats), the left nationalist party (denoted ERC; 12 seats) and the left ecologist party (denoted IC; three seats). The electoral rule that is used in the Catalan elections is the d'Hondt rule applied in each district, with a required minimum of 3% of votes to obtain a seat.

Let us now consider electoral polls in Spain. A week before election day it is usual for all the national newspapers to include the results of an electoral poll. The main newspapers commission large polls: sample sizes of around 15 000 in the 2000 general elections, and around 3000 for the 1999 Catalan elections. The sample size is broken down to provide estimates in individual districts. Moreover, there is an official Centre for Sociological Research that conducts its own electoral poll (it usually works with larger sample sizes than private organizations do).

3. Visualizing results and polls

To represent the results of some Parliamentary elections graphically and to compare them with the forecast of pre-electoral polls we use a principal component biplot. We illustrate the technique by using data from the aforementioned election results. Working on the basis of the known final results, we use a computer program to generate B samples (typically $B = 2000$)

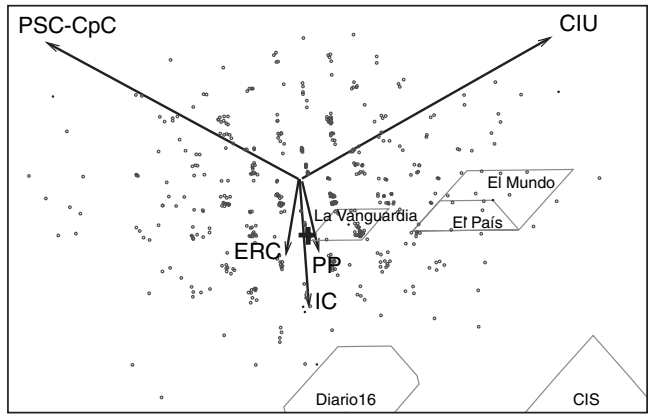

Fig. 1. Biplot for the Catalan Parliamentary elections in 1999

drawn from C multinomial distributions, one for each district. We use the same sample sizes as certain published pre-electoral polls which present sufficient technical details. Applying the electoral formula to each simulated vector of proportions, we obtain a set of forecasted Parliaments, vectors of K integers (K is the number of parties). Using the first and second principal components of this data set we can best represent the results in a plane, where the simulated Parliaments are seen as a cloud of points (Figs 1 and 2). In the same graphics we represent the vectors corresponding to the parties by projecting the variable vectors onto the representation plane. We draw them with their origins at the average point of the cloud of virtual Parliaments.

Fig. 1 is a biplot based on principal components that shows polls and results for the Catalan Parliamentary elections in 1999, in terms of seats allocated to the parties. The points in the scatterplot represent (a sample of) 2000 Parliaments obtained from simulated polls based on the final results. The arrows, starting from the average point of forecasted Parliaments, represent the directions favouring the respective parties. The cross, located above the label 'PP', marks the position of the real Parliament and shows visually that there is a significant bias in the estimation of Parliamentary composition. The polygons represent the forecast that was given by polls published in several newspapers a week before the elections; see below. There is a discernible pattern in the cloud of points: feasible Parliaments have integer co-ordinates that add up to M so they are arranged in hyperplanes that are still visible when projected onto our principal components plane. In fact, for better visualization we draw only some of the points (for instance 500), selected at random. The percentage of variance that is captured by the two dimensions represented in the plot is $55.2\% + 27.1\% = 82.3\%$.

Fig. 2 shows polls and results for general elections in Spain for the year 2000. It is similar to Fig. 1. The real Parliament is marked with a small cross to the left of the centre of the arrows' origin. The bias is clearly apparent. Some published pre-electoral polls are also displayed; see

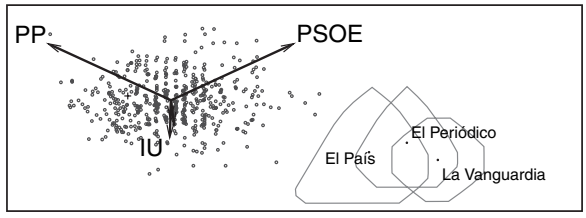

Fig. 2. Biplot for the Spanish Parliamentary elections in 2000

below. In the cases that are depicted in Figs 1 and 2 we can evaluate the bias roughly as being between a half and a third of the sample variability. Numerical evaluation confirms this impression (see Section 4.1). In this case, the captured variance is $65.8\% + 15.1\% = 80.9\%$.

To incorporate the Parliaments that were forecasted by pre-electoral polls we use the confidence intervals that they give for the number of seats forecasted for each party. We compute all the feasible Parliaments that fit the given confidence intervals and draw the convex hull of the projected points. If any of the polls use a different global sample size from the sample size that we use to generate the virtual Parliaments, we adjust the distance to the origin of the corresponding polygon accordingly.

4. Electoral formulae: an estimation bias problem

We concentrate here on a single district, where K parties obtained proportions of votes $(f_1, f_2, \ldots, f_K)$ and there are M seats to be allocated between them. An electoral formula is a rule for transferring these proportions to a seat allocation $(m_1, m_2, \ldots, m_K)$ such that $\Sigma m_i = M$. Most electoral formulae (see Taagepera and Shugart (1989)) are *proportional rules* that attempt to make the averages f_i/m_i similar between the parties. The most frequently used proportional rules work as follows.

(a) Exclude from the seat distribution parties that have proportions of the votes that are lower than a fixed threshold $\delta \geqslant 0$.
(b) Choose a non-decreasing sequence of denominators $d = (d_1, d_2, \ldots d_M)$.
(c) Form all the $K \times M$ quotients f_i/d_j, $i = 1, \ldots, K$, $j = 1, \ldots, M$.
(d) Select the M largest quotients and give the corresponding parties a seat for each of their largest quotients.

The choice of the denominator sequence d controls the proportionality of the rule (see Benoït (2000) for a study of proportionality). In an extreme case, if $d = (1, 1, 1, \ldots)$ the rule gives all the seats to the most voted for party and thus has no proportionality at all. The so-called *d'Hondt rule* takes $d = (1, 2, 3, \ldots)$. This is the rule that is used in Spain (with $\delta = 0.03$) and other European countries. It was also used in the USA to distribute seats in the House of Representatives among states according to the population size. It gives medium-sized parties more chances of obtaining seats, and small parties fewer. The *Sainte-Lagüe rule* takes $d = (1, 3, 5, 7, \ldots)$ and makes it easier for small parties to obtain their first seat. Among the commonly used formulae, the *modified Sainte-Lagüe rule* has maximum proportionality (see Benoït (2000)). It takes the sequence $d = (1.4, 3, 5, 7 \ldots)$. To ensure that every party having more votes than δ has at least one seat, we could use the sequence $d_j = 1 + M(j - 1)$. In the remaining part of our paper we use the d'Hondt rule (except in Fig. 4 for comparison) but a similar analysis could be performed with any other rule.

In Appendix A we include a mathematical formulation of these proportional rules and in Figs 3 and 4 we give a graphical idea of how they work. In Fig. 3 we show the case of two parties contending for five seats, such as in Cáceres in the Spanish 2000 elections. The horizontal axis is the proportion f_1 of votes for one of them (the Popular Party); the Socialist Party has the rest. The vertical axis represents the number of seats that are allocated to the Popular Party. For a population proportion of 52%, the real result there, the small bell-shaped curve, shows the approximate sampling distribution of the sample proportion for a sample size of $n_i = 199$, as used for Cáceres in one of the main electoral polls. Fig. 3 shows that there are some values of f_1 that make the number of seats jump, and that there are intervals in which the allocation of seats is constant. The rule used there is the d'Hondt rule.

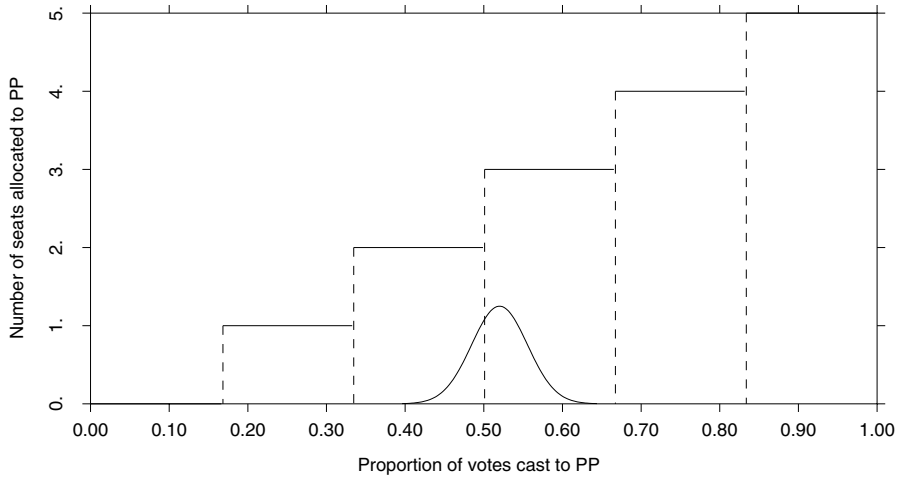

Fig. 3. Allocation of five seats to two contending parties and the sampling distribution of the sample proportion

Fig. 4 shows the allocation of seats when there are three parties in the game. The triangles depicted are the R^3-simplex $\{(f_1, f_2, f_3) \in R^3 : f_i \geqslant 0, i = 1, 2, 3; f_1 + f_2 + f_3 = 1\}$ in a so-called *ternary diagram* or *barycentric co-ordinate space*; see Aitchison (1986). It has been used in the voting context by Saari (1995) and Katz and King (1999). Triangular co-ordinates are used: a given point has proportions that are measured by the distances to the sides of the triangle. Thus, each point in the triangle corresponds to a feasible combination of proportions of votes. The regions of constant seat allocation (we call them *constant seat allocation cells*) are convex polygons with 4–6 sides. See Appendix A for the mathematical details. The only difference between the triangles in Fig. 4 is the proportional rule that is used to allocate the seats. The marked point is for the results in Ourense in the 2000 Spanish elections. Using the d'Hondt rule (Fig. 4(a)), three seats go to the Popular Party, one to the Socialist Party and none to the regional party. Other rules (the Sainte-Lagüe rule in Fig. 4(b) and the modified Sainte-Lagüe rule in Fig. 4(c)) give slightly different seat apportions, but in all cases the sampling variability will be quite important.

4.1. The bias in estimating the number of seats

The origin of the bias that is shown in Figs 1 and 2 can be clearly seen in simple cases. When there are only two parties, as in Fig. 3, and the real proportion in favour of the first is close to one of the jumps, a significant part of the samples that are drawn from the population would predict the wrong number of seats. In the situation that is depicted in Fig. 3, as many as 28.6% of the samples would predict two seats, so the expected number of seats predicted by sampling would be $0.714 \times 3 + 0.286 \times 2 = 2.714$ with a bias of -0.286. Here we use the normal approximation to the binomial distribution and a similar approximation to the multinomial distribution is used in the following discussion.

Fig. 5 shows a case in which there are three parties with proportions of votes 49.7%, 39.3% and 11.0%. It is the case of Asturias in the 2000 Spanish elections. The marked point is very close to two edges of its constant seat allocation cell (the one that is below it, corresponding to five seats for the Popular Party, three for the Socialist Party and one for the ex-Communist Party). Sampling from these proportions is very unstable from the point of view of seat alloca-

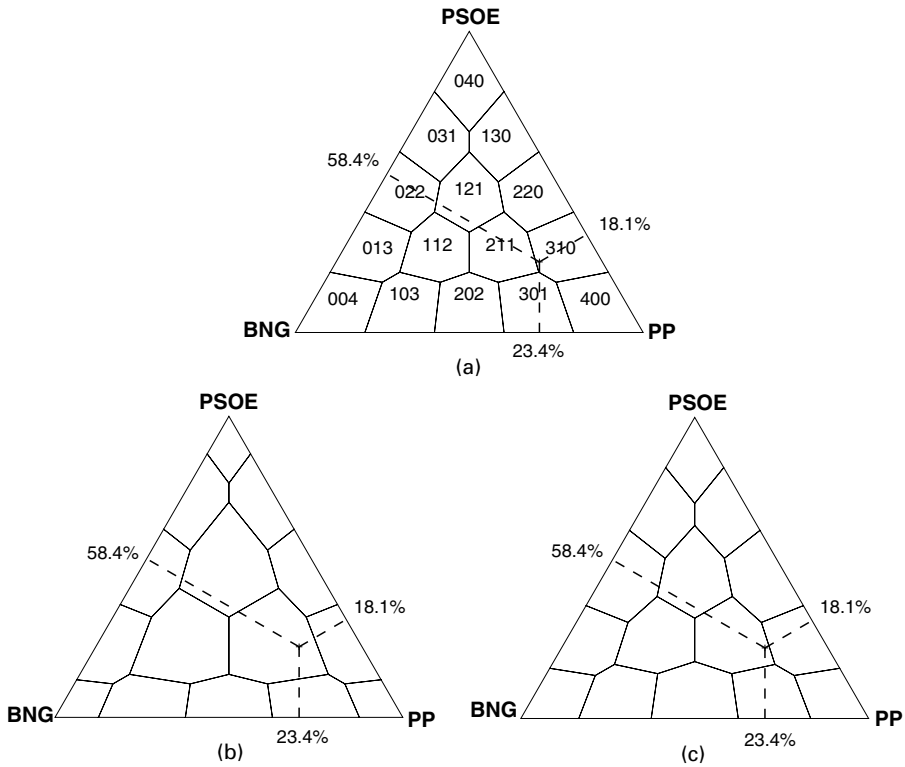

Fig. 4. Three parties contending four seats (in triangular co-ordinates): the polygons are regions with constant seat allocation (the allocation rules are described in the text)

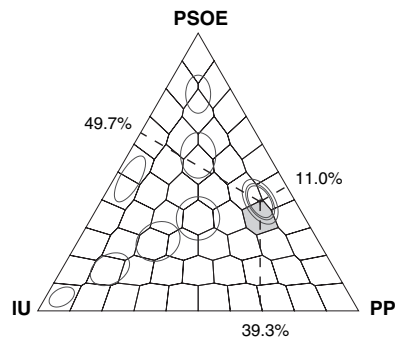

Fig. 5. Three parties contending nine seats: (◯), level curves of the sampling distribution of the sample proportions (see the text)

tion: more than 50% of the samples would give a wrong forecast. The ellipses that are centred on the point with these co-ordinates are level curves of the joint density of the sample proportions when the sample size $n_i = 337$ (which was used in one of the main electoral polls published before the elections). The levels have been chosen so that the probabilities inside the ellipses are 0.90, 0.95 and 0.99. Most of the samples give proportions that fall outside the correct constant seat allocation cell. The average predicted seat allocation will give more seats to the Socialist Party and ex-Communist Party, and fewer to the Popular Party than the right

predictions. The other ellipses contain 0.99 of the probability for other values of the population proportions.

The magnitude of the bias vector, defined as the difference between the average seat allocation by the samples and the real seat allocation, depends partly on the sample size but mainly on the real proportion of votes: if the point with real proportions falls close to the centre of its constant seat allocation cell, a small sample size can be enough to obtain a good forecast, but if it falls close to any cell edge a larger sample size is needed. In the particular cases where these proportions fall on a cell edge, the bias will not disappear even if the sample size increases to ∞.

When there are many districts, we might expect the biases to cancel one another out, and that the estimation for the whole Parliament would have no bias. This is so in some cases, but data from real elections in Spain show that it is not usually the case. One reason why cancelling may not occur is that some locally important parties contend only in a few districts (between one and four for some parties in the Spanish 2000 elections). Another reason is that several districts may have a similar bias, e.g. districts that have the same contending parties, the same number of seats and a similar proportion of votes. The proportion for votes for all these districts will be close to the same cell edge and will have a similar bias.

In Table 2 we describe the distribution of the bias in the estimated number of seats for each of the main parties contending in the 2000 Spanish election. We generate 10000 simulated polls covering all the districts and observe the bias between the number of seats that are allocated by each poll and the real Parliament. In the upper part of Table 2 we list the proportion of polls that miss the right number of seats for each party. In the lower part, the bias average, standard deviation and mean-squared error are computed over this set of 10000 simulated polls. Note that the magnitude of the bias has the same order as the standard deviation in the estimation of the number of seats for individual parties. Considering the overall estimation of the Parliament, we obtain that the bias in the estimation is still of the same order of magnitude as the variability: the squared bias is 12.02 and the variance is 16.03. Repeating the study by using data from the 2004 Spanish election, we obtain 4.48 and 20.09 respectively, so the bias still accounts for 18% of the mean-squared error.

Table 2. Distribution and summary of the bias in seat estimation (main parties) from 10000 simulated electoral polls based on the Spanish 2000 election results†

Number of missapportioned seats	*Distribution for the following parties:*							
	PP	*PSOE*	*CIU*	*IU*	*PNV*	*CC*	*BNG*	*ERC*
$\leqslant -5$	0.00	0.15	0.00	0.00	0.00	0.00	0.00	0.00
From -4 to -2	0.05	0.43	0.09	0.01	0.00	0.00	0.23	0.00
From -1 to 1	0.30	0.34	0.91	0.74	0.98	1.00	0.77	1.00
2–4	0.45	0.08	0.00	0.25	0.02	0.00	0.00	0.00
$\geqslant 5$	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Average	-2.45	2.03	0.45	-0.84	-0.27	0.38	0.84	0.20
Standard deviation	2.50	2.45	0.82	1.03	0.53	0.53	0.91	0.40
Mean-squared error	12.23	10.14	0.88	1.76	0.35	0.43	1.55	0.20
Number of seats	183	125	15	8	7	4	3	1

†PP, Popular Party; PSOE, Socialist Party; CIU, Catalan nationalist party; IU, ex-Communist Party; PNV, Basque nationalist party; CC, Canazian Coalition; BNG and ERC, Galician and Catalan left nationalist parties.

Estimation bias is a serious problem for Parliamentary forecasting. Some opinion poll firms conclude that Parliamentary forecasts ought not to be published. Others publish them in conjunction with confidence intervals for the seats that are assigned to each party; in the examples that we have studied, however, the confidence level is not clearly stated. Good confidence intervals could be computed by simulation as described in Section 4.3.

4.2. Estimating the bias by parametric bootstrap

One possible way to correct the bias or to compute confidence intervals when the real proportions are unknown is by means of a parametric bootstrap. We apply this technique, described in Efron and Tibshirani (1993), in the following way.

(a) Let S_0 be an electoral poll conducted in every district with sample sizes n_i for a total sample size N. Let $\hat{f}_{ij}$, $i=1,\ldots,K$, $j=1,\ldots,C$, be the resulting sample proportions for each party in each district. Let $\hat{m}_{ij}$ be the number of seats that are allocated to party i in district j according to these results.

(b) Repeat B_1 times:
   (i) for each district, draw a sample S_1, using the same sample sizes as in S_0, from a multinomial distribution with parameters $\hat{f}_{ij}$;
   (ii) use the resulting sample proportions to allocate seats $\hat{m}^*_{ij}$.

(c) Using the bootstrapped seats $\hat{m}^*_{ij}$ that are obtained in step (b) compute their mean $\bar{m}^*_{ij}$ through all the B_1 bootstrapped samples. Then, $\bar{m}^*_{ij}-\hat{m}_{ij}$ is an estimator of the unknown bias $E(\hat{m}_{ij})-m_{ij}$. The new bootstrap estimate of m_{ij} is then

$$\hat{m}^{\mathrm{B}}_{ij}=\hat{m}_{ij}-(\bar{m}^*_{ij}-\hat{m}_{ij}).$$

We conducted Monte Carlo simulations to evaluate the performance of this bias estimation method by using data from the 2000 Spanish general elections. There were $K=12$ parties and $C=52$ districts (most of the parties have no real presence in many districts). We used a total sample size of $N=15\,000$ distributed between districts by giving a fixed quota of 100 to each and distributing the rest in proportion to the size of the electoral census (these were the sizes that were used by one of the main electoral polls). Using the results of the elections, we produced $B_0=1000$ polls like S_0 described in step (a) above. We applied the procedure described above to each of these 1000 polls using $B_1=1000$. Table 3 lists the average bias that is given by these polls in the first row. This row numerically reflects the same bias that can be seen graphically in Fig. 2. In the second row we list the average (over B_0 samples) of the bootstrap estimates of the bias. Note that the magnitude of the bias is severely underestimated, but the direction of the estimation is mostly correct: the figures in Table 3 have a correlation of

Table 3. Bias and estimated bias for the total of seats for each of 12 parties†

	Results for the following parties:											
	PP	*PSOE*	*CIU*	*IU*	*PNV*	*CC*	*BNG*	*PA*	*ERC*	*ICV*	*EA*	*CHA*
Bias	−2.51	2.04	0.45	−0.85	−0.26	0.39	0.86	−0.08	0.19	−0.02	−0.11	−0.09
Estimated bias	−0.45	0.48	0.01	0.03	−0.12	0.00	0.08	0.01	0.12	−0.04	−0.03	−0.09

†PP, Popular Party; PSOE, Socialist Party; CIU, Catalan nationalist party; IU and ICV, ex-Communist Parties; PNV, Basque nationalist party; CC and PA, regional parties; BNG, ERC, EA and CHA, left nationalist parties.

0.91. Also note that the bias figures that are shown in Table 2 are slightly different from those listed here because they are computed in different simulation runs and Monte Carlo sample sizes.

4.3. Bootstrap confidence intervals for seat allocations

Now we deal with the problem of giving confidence intervals for the seat allocation m_i of each party i, $i = 1, \ldots, K$, from the results of an electoral poll. We use a parametric bootstrap procedure.

Let $\hat{\mathcal{P}} = (\hat{m}_1, \hat{m}_2, \ldots, \hat{m}_K)$ be the composition of the Parliament obtained from an electoral poll S conducted in all C districts. Let $\hat{f}_{ij}$ be the resulting proportion of votes that is obtained by S for party i in district j, and let $\hat{m}_{ij}$ be the corresponding allocated seats. So, $\hat{m}_i = \Sigma_j \hat{f}_{ij}$.

The confidence interval procedure for m_i based on a parametric bootstrap operates as follows, for a given confidence level $1 - \alpha$.

(a) Let $b = 0$ and repeat B times:
   (i) let $b = b + 1$;
   (ii) for each district, draw a sample from a multinomial distribution with parameters $\hat{f}_{ij}$;
   (iii) use the resulting sample proportions to allocate seats $\hat{m}^*_{ij}$, and compute the total seats that are allocated to each party, $\hat{m}^*_i = \Sigma_{j=1}^{C} \hat{m}^*_{ij}$; let $\hat{\mathcal{P}}^*_b = (\hat{m}^*_1, \ldots, \hat{m}^*_K)$, the bth simulated Parliament;
   (iv) compute the distance d_b from $\hat{\mathcal{P}}$ to $\hat{\mathcal{P}}^*_b$.

(b) Compute the minimum distance from $\hat{\mathcal{P}}$ that covers at least $1 - \alpha$ of the simulated Parliaments, i.e.

$$d_\alpha = \min_b \{d_b : \#\{d_k | d_k \leqslant d_b\} \geqslant (1 - \alpha)B\}.$$

(c) Take all the Parliaments $\hat{\mathcal{P}}^*_b$ that have $d_b \leqslant d_\alpha$ and for each party compute the interval that covers all the seats assigned to that party in these Parliaments. This process finally gives K intervals, which we call the confidence interval for the real Parliament $\mathcal{P}_0$.

We report Monte Carlo results for nominal confidence levels of 90% and 60% (which give intervals of width comparable with those of the main published pre-electoral polls: typical interval widths were 6, 7, 2, 1, 1, 2 for the six largest parties in the 2000 Spanish elections; note that none of these published polls included any statement about the confidence level of the intervals given).

The Monte Carlo experiment is as follows. We take the actual results (the real proportion f_{ij} of voters for party i in district j) and we simulate 1000 electoral polls according to them. For each simulated poll we computed a confidence interval for the Parliament $\mathcal{P}_0$ by using the parametric bootstrap. A proportion of at least $1 - \alpha$ of these confidence intervals for the Parliament $\mathcal{P}_0$ is expected effectively to contain the real Parliament. Table 4 shows that this is actually so for data from the 2000 Spanish elections. We used the Euclidean distance (similar results were obtained for other distances). The mean width of the intervals for the largest parties is also reported.

5. Concluding remarks

We have presented graphical tools to evaluate the results of pre-electoral polls in terms of estimated seats in Parliament. The problem of bias in the allocation of seats estimation has been pointed out as fundamental. The study of the bias problem indicates that the difficulty in estimating the allocation of seats in a district depends on several parameters beyond the sample size.

Table 4. Coverage and average bootstrapped confidence intervals for Parliaments over 1000 simulated polls†

Nominal coverage $1-\alpha$ *(%)*	*Real coverage (%)*	*Average interval for the following main parties:*					
		PP	*PSOE*	*CIU*	*IU*	*PNV*	*CC*
90	92.7	±5.4	±5.4	±2.5	±3.4	±1.6	±1.4
60	62.3	±3.3	±3.3	±2.3	±2.8	±1.8	±1.3

†PP, Popular Party; PSOE, Socialist Party; CIU, Catalan nationalist party; IU, ex-Communist Party; PNV, Basque nationalist party; CC, regional party.

Two districts with similar numbers of seats can have different numbers of contending parties and, even if the number of parties is the same, they can differ in the proportions of voters for each party. From our point of view, when a pre-electoral poll is designed, regional sample sizes should be assigned according to the difficulty of estimation in each electoral region. Electoral polls, however, use sample sizes that are based on the size of the electorate in each district. Alternative rules for sample size assignment deserve further consideration.

Finally, we discuss the important issue of generalization: what electoral systems may present the problems described above? First, it should be noted that we are dealing with a particular electoral poll structure such that a nationwide sample is broken down into constituencies to provide district level estimates to be added up to produce the national level estimates. This is not the method that is used in countries with a large number (several hundreds) of small constituencies, typically with one seat per constituency, known as a *single-member constituency*. Examples are general elections in France (about 550 single-member constituencies) and the UK (about 650 single-member constituencies). Second, the problem of electoral poll bias appears at first at the level of single districts, and then national level bias may or may not occur. Individual constituency bias, as has been shown in Section 4, depends on the sample size and the share of the parties, and in general, for a given sample size, is favoured by a small number of seats per district. Third, the national level bias does not appear under certain conditions. A large number of heterogeneous constituencies should present heterogeneous district level bias that cancels out when individual predictions are summed across districts. However, when a party is present in only a few constituencies (the case of regionalist parties) the corresponding district level bias cannot cancel out at the aggregate level. Therefore, homogeneity in constituencies, a small number of districts and the existence of regionalist parties cause the formation of national level bias (homogeneity here means the same parties, the same number of seats and similar shares for the parties).

As an illustration of the possibility of generalization of our results, we consider the lower house electoral systems in the 15 countries constituting the European Union before May 2004. The details of these electoral systems can be found in European Centre for Parliamentary Research and Documentation (2000) and Colomer (2004a, b). Four countries out of 15 have a large number of single-member constituencies: France, Germany, Italy and the UK. Moreover, the electoral system in Germany and Italy is mixed: it complements the single-member majority system with a proportional system to apportion part of the seats (half of the total number of seats in Germany, and a quarter in Italy, are chosen by the proportional rule). The electoral system in the Netherlands uses a proportional rule with closed party lists. There is a single district with 150 seats. Therefore it follows that electoral polls in these five countries will not share with the Spanish case the structure or the problems that were pointed out in this paper.

The other 10 countries (Austria, Belgium, Denmark, Finland, Greece, Ireland, Luxembourg, Portugal, Spain and Sweden) have proportional electoral systems with an approximately similar distribution of sizes of constituency, even if the number of constituencies varies with the size of the country. It is likely that electoral polls in these countries present the drawbacks that were dealt with in this paper.

Acknowledgements

We acknowledge comments from participants at a seminar at the Universitat Pompeu Fabra and at the Royal Statistical Society's conference in 2002. We also acknowledge useful comments on the first version of the paper from the Joint Editor and the referees.

The work of Frederic Udina was supported by grant BFM-2003-03324, and the work of Pedro Delicado was supported by grant BFM-2001-2327.

Appendix A: Mathematical details of proportional rules

Let K parties be competing for M seats (in a single district). Let δ be the minimum proportion to obtain any seat. Let $(f_1, f_2, \ldots, f_K)$ be the proportion of votes that is obtained by the parties. We have

$$0 \leqslant f_i \leqslant 1 \qquad (i = 1, \ldots, K),$$
$$\sum_{i=1}^{K} f_i = 1.$$

Let d_j, $j = 1, \ldots, M$, be a non-decreasing sequence. Define the quotients $q_{i,j}$, for $i = 1, \ldots, K$ and $j = 1, \ldots, M$, by the rules

(a) if $f_i < \delta$, $q_{i,j} = 0$ and
(b) if $f_i \geqslant \delta$, $q_{i,j} = f_i/d_j$.

A quotient $q_{i,j}$ deserves a seat if and only if it is one of the M greatest among all the KM quotients or, equivalently, if there are more than $M(K-M)$ quotients that are smaller than itself. Ties between quotients are a set of very low probability. In such improbable cases, electoral laws usually give the seat to the party with the greatest absolute number of votes (and by draw if these are equal). Thus the rule is an application S from the simplex

$$\left\{ \mathbf{f} \in R^K \,\middle|\, \sum_{i=1}^{K} f_i \leqslant 1 \right\}$$

onto the discrete set of possible Parliamentary configurations

$$\left\{ \mathbf{m} \in Z^K \,\middle|\, \sum_{j=1}^{K} m_j = M \right\}$$

defined by

$$\begin{aligned} &S_d(f_1, \ldots, f_K) = (m_1, \ldots, m_K) \Leftrightarrow \\ &\quad \forall i = 1, \ldots, K,\ m_i = \max\{j = 1, \ldots, M \mid Q(i, j) > KN - N\} \\ &\quad \text{where } Q(i, j) = \#\{q_{k,l} < q_{i,j} : k = 1, \ldots, K, l = 1, \ldots, M\}. \end{aligned} \tag{1}$$

Note that $Q(i, j)$ is the number of quotients that are less than $q_{i,j}$. The definition can be written in the following non-closed form that may be more useful:

$$\begin{aligned} &S_d(f_1, \ldots, f_K) = (m_1, \ldots, m_K) \Leftrightarrow \\ &\quad \forall i, j \in \{1, \ldots, K\},\ i \neq j,\ m_i = 0 \text{ or } f_i/d_{m_i} > f_j/d_{m_j+1}, \end{aligned} \tag{2}$$

which states that the last quotient of party i that gained a seat must be larger than the first quotient of party j that did not.

With this definition it is easy to understand that the regions with constant seat allocation (which we call constant seat allocation cells) are limited by hyperplanes and thus are convex polyhedrons. Each is limited by the inequalities appearing in expression (2). Up to $K(K-1)$ of them can be effective. In the case $m_i = 0$ the effective inequality is simply the boundary of the simplex. In Figs 3–5 we have seen examples of such regions in the plane, when $K = 3$.

References

Aitchison, J. (1986) *The Statistical Analysis of Compositional Data*. London: Chapman and Hall.

Benoït, K. (2000) Which electoral formula is the most proportional?: a new look with new evidence. *Polit. Anal.*, **8**, 381–388.

Bernardo, J. M. (1984) Monitoring the 1982 Spanish socialist victory: a bayesian analysis. *J. Am. Statist. Ass.*, **79**, 510–515.

Brown, P. and Payne, C. (1984) Forecasting the 1983 British general election. *Statistician*, **33**, 217–228.

Colomer, J. M. (2001) The 2000 general election in Spain. *Elect. Stud.*, **20**, 463–501.

Colomer, J. M. (2004a) *Cómo Votamos. Los Sistemas Electorales del Mundo: Pasado Presente y Futuro*. Barcelona: Gedisa.

Colomer, J. M. (ed.) (2004b) *Handbook of Electoral System Choice*. Basingstoke: Palgrave–Macmillan.

Cox, G. W. (1997) *Making Votes Count: Strategic Coordination in the World's Electoral Systems*. Cambridge: Cambridge University Press.

Efron, B. and Tibshirani, R. J. (1993) *An Introduction to the Bootstrap*. New York: Chapman and Hall.

European Centre for Parliamentary Research and Documentation (2000) Electoral systems in Europe: an overview. *Technical Report*. European Centre for Parliamentary Research and Documentation, Brussels. (Available from `http://www.ecprd.org/Doc/publica/OTH/elect_system.html`.)

Katz, J. N. and King, G. (1999) A statistical model for multiparty electoral data. *Am. Polit. Sci. Rev.*, **93**, 15–32.

Saari, D. G. (1995) *Geometry of Voting*. Berlin: Springer.

Taagepera, R. and Shugart, M. S. (1989) *Seats and Votes: the Effects and Determinants of Electoral Systems*. New Haven: Yale University Press.

Voss, D. S., Gelmann, A. and King, G. (1995) The polls—a review. Preelection survey methodology: details from eight polling organizations, 1988 and 1992. *Publ. Opin. Q.*, **59**, 98–132.