Departament d'Estadística i I.O., Universitat de València.
Facultat de Matemàtiques, 46100–Burjassot, València, Spain.
Tel. 34.6.363.6048, Fax 34.6.363.6048 (direct), 34.6.386.4735 (office)
Internet: bernardo@uv.es, Web: http://www.uv.es/~bernardo/

Printed on December 9, 1996
Invited paper presented at the *Third Workshop on Bayesian Statistics in Science and Technology: Case Studies*, held at Carnegie Mellon University, Pittsburgh, U. S. A., October 5–7, 1995

Probing Public Opinion: the State of Valencia Experience

JOSÉ M. BERNARDO
Universitat de València, Spain

SUMMARY

This paper summarizes the procedures which have been set up during the last years at the Government of the State of Valencia, Spain, to systematically probe its public opinion as an important input into its decision processes.

After a brief description of the electoral setup, we (i) outline the use of a simulated annealing algorithm, designed to find a good design for sample surveys, which is based on the identification of representative electoral sections, (ii) describe the methods used to analyze the data obtained from sample surveys on politically relevant topics, (iii) outline the proceedings of election day —detailing the special problems posed by the analysis of exit poll, representative sections, and early returns data— and (iv) describe a solution to the problem of estimating the political transition matrices which identify the reallocation of the vote of each individual party between two political elections.

Throughout the paper, special attention is given to the illustration of the methods with real data. The arguments fall entirely within the Bayesian framework.

Keywords: BAYESIAN PREDICTION; HIERARCHICAL MODELLING; ELECTION FORECASTING; LOGARITHMIC DIVERGENCE; SAMPLE SURVEYS; SIMULATED ANNEALING.

1. INTRODUCTION

The elections held in the State of Valencia on May 28, 1995 gave the power to the Conservatives after sixteen years of Socialist government. During most of the socialist period, the author acted as a scientific advisor to the State President, introducing Bayesian inference and decision analysis to systematically probe the State's public opinion, with the stated aim of improving the democratic system, by closely taking into account the peoples' beliefs and preferences. This paper summarizes the methods used —always within the Bayesian framework— and illustrates their behaviour with real data.

Section 2 briefly describes the electoral setup, which allows a very detailed knowledge of the electoral results —at the level of polling stations,— and which uses Jefferson-d'Hondt

---

José M. Bernardo is Professor of Statistics at the University of Valencia. Research was partially funded with grant PB93-1204 of the DGICYT, Madrid, Spain.

algorithm for seat allocation. Section 3 focuses on data selection, describing the use of a simulated annealing algorithm in order to find a good design for sample surveys, which is based on the identification of a small subset of electoral sections that closely duplicates the State political behaviour.

Section 4 describes the methods which we have mostly used to analyze the data obtained from sample surveys, while Section 5 specializes on election day, describing the methods used to obtain election forecasts from exit poll, representative sections, and early returns data. Special attention is given to the actual performance of the methods described in the May 95 State election.

Section 6 describes a solution to the problem of estimating the political transition matrices which identify the reallocation of the vote of each individual party between two political elections. Finally, Section 7 contains some concluding remarks and suggests areas of additional research.

2. THE ELECTORAL SYSTEM

The State of Valencia is divided into three main electoral units, or provinces, Alicante, Castellón and Valencia, each of which elects a number or seats which is roughly proportional to its population. Thus, the State Parliament consists of a single House with 89 seats, 30 of which are elected by Alicante, 22 by Castellón and 37 by Valencia. The leader of the party or coalition that has a plurality of the seats is appointed by the King to be President of the State.

The seats in each province are divided among the parties that obtain at least 5% of the vote in the State according to a corrected proportional system, usually known as the d'Hondt rule —invented by Thomas Jefferson nearly a century before Victor d'Hondt rediscovered and popularized the system— and used, with variations, in most parliamentary democracies with proportional representation systems.

Table 1. *d'Hondt table for the results of province of Valencia in 1995 State elections*

	PP	PSOE	EU	UV
1	532524	429840	166676	137277
2	266262	214920	83338	68639
3	177508	143280	55559	45759
4	133131	107460	41669	**34319**
5	106505	85968	**33335**	27455
6	88754	71640	27779	—
7	76075	61406	—	—
8	66566	53730	—	—
9	59169	47760	—	—
10	53252	42984	—	—
11	48411	39076	—	—
12	44377	**35820**	—	—
13	40963	33065	—	—
14	38037	—	—	—
15	35502	—	—	—
16	**33283**	—	—	—
17	31325	—	—	—
18	—	—	—	—

According to d'Hondt rule, to distribute n_s seats among the, say, k parties that have overcome the 5% barrier, one (i) computes the $n_s \times k$ matrix of quotients with general element

$$z_{ij} = n_j/i, \quad i = 1, \ldots, n_s, \quad j = 1, \ldots, k,$$

where n_j is the number of valid votes obtained by the jth party, (ii) selects the largest n_s elements and (iii) allocates to party j a number of seats equal to the number of these n_s largest elements found in its corresponding column. Clearly, to apply d'Hondt rule, one may equivalently use the proportion of valid votes obtained by each party, rather than the absolute number of votes.

Thus if, for example, the 37 seats that corresponds to the province of Valencia are to be distributed among the four parties PP (conservatives), PSOE (socialists), EU (communists) and UV (conservative nationalists) who have respectively obtained (May 1995 results) 532524, 429840, 166676 and 137277 votes in the province of Valencia *and* over 5% of the State votes —the remaining 46094 counted votes being distributed among parties which did not make the overall 5% barrier— one forms the matrix in Table 1 and, according to the algorithm described, associates 16 seats to PP, 12 to PSOE, 5 to EU and 4 to UV.

It may be verified that the d'Hondt rule provides a corrected proportional system that enhances the representation of the big parties to the detriment of the smaller ones, but the correction becomes smaller as the number of seats increases, so that a pragmatically perfect proportional representation may be achieved with d'Hondt rule if the number of seats is sufficiently large. Indeed, if only one seat is allocated, d'Hondt rule obviously reduces to majority rule but, as the number of seats increases, d'Hondt rule rapidly converges to proportional rule: with the results described above, a proportional representation would yield 15.56, 12.56, 4.87 and 4.01, not far from the 16, 12, 5 and 4 integer partition provided by d'Hondt rule. Note that the last, 37th seat, was allocated to the conservative PP rather than to the socialist PSOE by only an small proportion, $(33283 - 33065) * 13 = 2836$ or 0.22/%, of the 1312411 votes counted

Since seats —and hence political power— are allocated by province results, and since there are some very noticeable differences in the political behaviour of the provinces —for instance the conservative nationalists UV are barely present outside the province of Valencia— most political analysis of the State are better done at province level, aggregating the provincial forecast in a final step.

Each province is divided into a variable number of electoral sections, each containing between 500 and 2000 electors living in a tiny, often socially homogeneous, geographical area. The State of Valencia is divided into 4484 electoral sections, 1483, 588 and 2410 of which respectively correspond to the provinces of Alicante, Castellón and Valencia. Votes are counted in public at each electoral section, just after the vote is closed at 8 pm. This means that at about 9 pm someone attending the counting may telephone to the analysis center the final results from that section; these data may be used to make early predictions of the results. Since the definition of the electoral sections has remained fairly stable since democracy was restored in Spain in 1976, this also means that a huge electoral data base, which contains the results of all elections (referendums, local, state, national and european elections) at electoral section level, is publicly available. In the next section we will describe how this is used at the design stage.

3. SURVEY DESIGN

In sample surveys, one typically has to obtain a representative sample from a human population, in order to determine the proportion $\boldsymbol{\psi} \equiv \{\psi_1, \ldots, \psi_k\}$, $(\psi_j > 0, \Sigma\psi_j = 1)$ of people who favor one of a set of, say k, possible alternative answers to a question. Naturally, most surveys contain more than one question, but we may safely ignore this fact in this discussion. Typically, the questionnaire also includes information on possible relevant *covariates*, such as sex, age, education, or political preferences.

Within the Bayesian framework, the analysis of the survey results essentially consists on the derivation of the posterior distribution of $\boldsymbol{\psi} = \{\psi_1, \ldots, \psi_k\}$. A particular case of frequent interest is that of *election forecasting*, where the ψ_j's, $j = 1, \ldots, k$ describe the proportion of the valid vote which each of the, say, k parties will eventually obtain.

The selection of the sample has traditionally been made by the use of the so-called "random" routes, which, regrettably, are often far from random. The problem lies in the fact that there is no way to guarantee that the attitudes of the population with respect to the question posed are homogeneous relative to the design of the "random" route. Indeed, this has produced a number of historical blunders.

An obvious alternative would be to use a real random sample, *i.e.*, to obtain a random sample from the population census —which is publicly available and contains name, address, age, sex and level of education of all citizens with the right to vote— and to interview the resulting people. The problem with this approach is that it produces highly scattered samples, what typically implies a very high cost. A practical alternative would be to determine a set of geographically small units who could jointly be considered to behave like the population as a whole, and to obtain the sample by simple random sampling within those units. Since the political spectrum of a democratic society is supposed to describe its diversity, and since the results of political elections are known for the small units defined by the electoral sections, a practical implementation of this idea would be to find a small set of electoral sections whose *joint* political behaviour has historically been as similar as possible to that of the whole population, and to use those as the basis for the selection of the samples. We now describe how did we formalize this idea.

To design a survey on a province with, say, n_p electoral sections —which on election day become n_p polling stations— may be seen as a *decision problem* where the action space is the class of the 2^{n_p} possible subsets of electoral sections, and where the loss function which describes the consequences of choosing the subset $\boldsymbol{s}$ should be a measure of the *discrepancy* $l(\hat{\boldsymbol{\psi}}_s, \boldsymbol{\psi})$ between the actual proportions $\boldsymbol{\psi} \equiv \{\psi_1, \ldots, \psi_k\}$ of people which favor each of the k alternatives considered, and the estimated proportions $\hat{\boldsymbol{\psi}}_s \equiv \{\hat{\psi}_{s1}, \ldots, \hat{\psi}_{sk}\}$ which would be obtained from a survey based of random sampling from the subset $\boldsymbol{s}$. The optimal choice would be that minimizing the expected loss

$$E[l(s) \,|\, D] = \int_\Psi l(\hat{\boldsymbol{\psi}}_s, \boldsymbol{\psi})\, p(\boldsymbol{\psi} \,|\, D)\, d\boldsymbol{\psi}, \tag{1}$$

where D is the database of relevant available information.

Since preferences within socially important questions may safely be assumed to be closely related with political preferences, the results of previous elections may be taken a as proxy for a random sample of questions, in order to approximate by Monte Carlo the integral above.

To be more specific, we have to introduce some notation. Let $\boldsymbol{\theta}_e = \{\theta_{e1}, \ldots, \theta_{ek(e)}\}$, for $e = 1, \ldots, n_e$, be the province results on n_e elections; thus, θ_{ej} is the proportion of the valid vote obtained by party j in election e, which was disputed among $k(e)$ parties. Similarly, let $\boldsymbol{w}_{el} = \{w_{el1}, \ldots, w_{elk(e)}\}$, $e = 1, \ldots, n_e$ and, $l = 1, \ldots, n_p$ be the results of the n_e elections in each of the n_p electoral sections in which the province is divided.

Each of the 2^{n_p} possible subsets may be represented by a sequence of 0's and 1's of length n_p, so that $\boldsymbol{s} \equiv \{s_1, \ldots, s_{n_p}\}$ is the subset of electoral sections for which $s_l = 1$. Taken as a whole, those electoral sections would produce an estimate of the provincial result for election e, which is simply given by the arithmetic average of the results obtained in them, *i.e.*,

$$\hat{\boldsymbol{\theta}}_{es} = \frac{1}{\sum_{l=1}^{n_p} s_l} \sum_{l=1}^{n_p} s_l \boldsymbol{w}_{el}. \tag{2}$$

Thus, if election preferences may be considered representative of the type of questions posed, a Monte Carlo approximation to the integral (1) is given by

$$E[l(s) \mid D] \simeq \frac{1}{n_e} \sum_{e=1}^{n_e} l(\hat{\boldsymbol{\theta}}_{es}, \boldsymbol{\theta}_e) \tag{3}$$

A large number of axiomatically based arguments (see *e.g.*, Good, 1952, and Bernardo, 1979) suggest that the most appropriate measure of discrepancy between probability distributions is the logarithmic divergence

$$\delta\{\hat{\boldsymbol{\theta}}_{es}, \boldsymbol{\theta}_e\} = \sum_{j=1}^{k(e)} \theta_{ej} \log \frac{\theta_{ej}}{\hat{\theta}_{sej}} \tag{4}$$

so that we have to minimize

$$\sum_{e=1}^{n_e} \sum_{j=1}^{k(e)} \theta_{ej} \log \frac{\theta_{ej}}{\hat{\theta}_{sej}} \,. \tag{5}$$

However, this is a really huge minimization problem. For instance, for the province of Alicante, the action space thus has 2^{1483} points, what absolutely forbids to compute them all. To obtain a solution, we decided to use a random optimization algorithm, known as *simulated annealing*.

Simulated annealing is an algorithm of random optimization which uses as a heuristic base the process of obtaining pure crystals (annealing), where the material is slowly cooled, giving time at each step for the atomic structure of the crystal to reach its lowest energy level at the current temperature. The method was described by Kirkpatrick, Gelatt and Vecchi (1983) and has seen some statistical applications, such as Lundy (1985) and Haines (1987).

Consider a function $f(\boldsymbol{x})$ to be *minimized* for $\boldsymbol{x} \in \boldsymbol{X}$. Starting from an origin $\boldsymbol{x}_0$ with value $f(\boldsymbol{x}_0)$ —maybe a possible first guess on where the minimum may lie—, the idea consists of computing the value $f(\boldsymbol{x}_{i+1})$ of the objective function f at a *random* point $\boldsymbol{x}_{i+1}$ at distance d of $\boldsymbol{x}_i$; one then moves to $\boldsymbol{x}_{i+1}$ with probability one if $f(\boldsymbol{x}_{i+1}) < f(\boldsymbol{x}_i)$, and with probability $\exp\{-\delta/t\}$ otherwise, where $\delta = f(\boldsymbol{x}_{i+1}) - f(\boldsymbol{x}_i)$, and where t is a parameter —initially set at a large value— which mimics the temperature in the physical process of crystallization.

Thus, at high temperature, *i.e.*, at the beginning of the process, it is not unlikely to move to points where the function actually increases, thus limiting the chances of getting trapped in local minima. This process is repeated until a temporary equilibrium situation is reached, where the objective value does not change for a while.

Once in temporary equilibrium, the value of t is reduced, and a new temporary equilibrium is obtained. The sequence is repeated until, for small t values, the algorithm reduces to a rapidly convergent non-random search. The method is applied to progressively smaller distances, until an acceptable precision in reached.

The optimization cycle is typically ended when the objective value does not change for a fixed number of consecutive tries. The iteration is finished when the final non-random search is concluded. The algorithm is terminated when the final search distance is smaller than the desired precision.

In order implement the simulated annealing algorithm is further necessary to define what it is understood by "distance" among sets of electoral sections. It is natural to define the class of sets which are at distance d of $\boldsymbol{s}_j$ as those which differ from $\boldsymbol{s}_j$ in precisely d electoral sections. Thus

$$d\{\boldsymbol{s}_i, \boldsymbol{s}_j\} = \sum_{l=1}^{n_p} ||s_{il} - s_{jl}|| \tag{6}$$

All which is left is to adjust the sequence of "temperatures" t —what we do interactively— and to choose a starting set $\boldsymbol{s}_0$ which may reasonably chosen to be that of the, say, n_0, polling stations which are closest in average to the global value, *i.e.*, those which minimize

$$\frac{1}{n_e} \sum_{i=1}^{n_e} \delta\{\boldsymbol{\omega}_{el}, \boldsymbol{\theta}_e\}. \tag{7}$$

To offer an idea of the practical power of the method, we conclude this section by describing the results actually obtained in the province of Alicante.

The province has 1483 electoral sections, so we have $2^{1483} \approx 10^{446}$ possible subsets. For these these 1483 sections we used the results obtained by the four major parties —PP, PSOE, EU and UV, grouping as "others" a large group of small, nearly testimonial parties— in four consecutive elections, local (1991), State (1991), national (1993) and european (1994). Thus, with the notation above we had $n_e = 4$, $n_p = 1483$ and $k(e) = 5$. For a mixture of economical and political considerations, we wanted to use at least 20 and no more than 40 electoral sections. Thus, starting with the set $\boldsymbol{s}_0$ of the 20 sections which, averaging over these four elections, where closest to the provincial result in a logarithmic divergence set, we run the annealing algorithm with imposed boundaries at sizes 20 and 40. The final solution —which took 7 hours on a Macintosh— was a set of 25 sections whose behaviour is described in Table 2.

For each of the four elections whose data were used, the table provides the actual results in the province of Alicante —in percentages of valid votes—, the estimators obtained as the arithmetic means of the results obtained in the 25 selected sections, and their absolute differences. It may be seen that those absolute differences are all between 0.01% 0.36%. The final block in Table 2 provides the corresponding data for the May 95 State elections, which were *not* used to find the design. The corresponding absolute errors —around 0.4, with corresponding *relative* errors of about 3%— are *much smaller* than the sampling errors which correspond to the sample sizes (about 400 in each province) which were typically used. Very similar results were obtained for the other provinces.

Our sample surveys have always been implemented with *home interviews* on citizens randomly selected from the representative sections using the electoral census. Thus, we could provide the interviewers with list of the people to be interviewed which included their name, address, and the covariates sex, age and level of education. The lists contained possible replacements with people of identical covariates, thus avoiding the danger of over representing the profiles which corresponded to people who are more often home.

Table 2. *Performance of the design algorithm for the province of Alicante in the 1995 State elections*

		PP	PSOE	EU	UV
Local 91	Results	31.50	43.17	7.23	1.22
	Estimators	31.30	43.32	7.24	1.32
	Abs. Dif.	0.20	0.15	0.01	0.09
State 91	Results	33.55	45.05	7.37	1.75
	Estimators	33.36	45.05	7.33	1.74
	Abs. Dif.	0.19	0.01	0.05	0.01
National 93	Results	43.87	39.94	10.32	0.57
	Estimators	43.64	39.75	10.62	0.49
	Abs. Dif.	0.22	0.19	0.30	0.07
European 94	Results	47.62	32.38	13.53	1.43
	Estimators	47.69	32.02	13.51	1.46
	Abs. Dif.	0.07	0.36	0.02	0.03
State 95	Results	47.24	36.30	11.06	2.11
	Estimators	48.26	36.33	10.50	1.79
	Abs. Dif.	1.02	0.03	0.56	0.32

4. SURVEY DATA ANALYSIS

The structure of the questionnaires we mostly used typically consisted of a sequence of *closed* questions –where a set of possible answers is given for each question, always leaving an “other options” possibility for those who do no agree with any of the stated alternatives, and a “non-response” option for those who refuse to answer a particular question. This was followed by a number of questions on the covariates which identify the social profile of the person interviewed; these typically include items such as age, sex, level of education, mother language or area of origin.

Let us consider one of the questions included in a survey and suppose that it is posed as a set of, say, k alternatives $\{\delta_1, \ldots, \delta_k\}$ (including the “other options” possibility) among which the person interviewed has to choose one and only one. The objective is to know the proportions of people which favor each of the alternatives, both globally, and in socially or politically interesting subsets of the population —that we shall call *classes*— as defined by either geographical or social characteristics. When the possible answers are *not* incompatible and the subject is allowed to mark more than one of them, we treated the multiple answer as a *uniform distribution* of the person’s preferences over the marked answers and randomly choose one of them, thus reducing the situation to one with incompatible answers.

Thus, if $\boldsymbol{x} = \{x_1, \ldots, x_v\}$ denotes the set of, say, v covariates used to define the population classes we may be interested in, the data D relevant to a particular question included in a survey may described as a matrix which contains in each row the value of the covariates and the answer to that question provided by each of the persons interviewed. Naturally, a certain proportion of the people interviewed —typically between 20% and 40%— refuse to answer some the questions; thus, if, say, n persons have actually answered and m have refused to answer a particular question its associated $(n+m)\times(v+1)$ matrix is defined to be

$$D = \begin{pmatrix} D_1 \\ \hline D_2 \end{pmatrix} = \left(\begin{array}{cccc} x_{1,1} & \dots & x_{1,v} & \delta_{(1)} \\ \vdots & \dots & \vdots & \vdots \\ x_{n,1} & \dots & x_{n,v} & \delta_{(n)} \\ \hline x_{n+1,1} & \dots & x_{n+1,v} & - \\ \vdots & \dots & \vdots & \vdots \\ x_{n+m,1} & \dots & x_{n+m,v} & - \end{array}\right) \tag{8}$$

where x_{ij} is the value of jth covariate for the ith subject, and $\delta_{(i)}$ denotes his or her preferences among the proposed alternatives.

Our main objective is the set of posterior probabilities

$$E[\boldsymbol{\psi} \mid D, c] = p(\boldsymbol{\delta} \mid D, c) = \{p(\delta_1 \mid D, c), \dots, p(\delta_k \mid D, c)\}, \quad c \in C, \tag{9}$$

which describe the probabilities that a person in class c prefers each of the k possible alternatives, for each of the classes $c \in C$ being considered. The particular class which contains all the citizens naturally provides the global results.

To compute these, we used the total probability theorem to 'extend the conversation' to include the covariates $\boldsymbol{x} = \{x_1, \dots, x_k\}$, so that

$$p(\boldsymbol{\delta} \mid D, c) = \int_{\boldsymbol{X}} p(\boldsymbol{\delta} \mid \boldsymbol{x}, D, c)\, p(\boldsymbol{x} \mid D, c)\, d\boldsymbol{x} \tag{10}$$

where $p(\boldsymbol{x} \mid D, c)$ is the predictive distribution of the covariates vector.

Usually, the joint predictive $p(\boldsymbol{x} \mid D, c)$ is too complicated to handle, so we introduce a *relevant function* $\boldsymbol{t} = \boldsymbol{t}(\boldsymbol{x})$ which could be thought to be *approximately sufficient* in the sense that, for all classes,

$$p(\boldsymbol{\delta} \mid \boldsymbol{x}, D, c) \approx p(\boldsymbol{\delta} \mid \boldsymbol{t}, D, c), \quad \boldsymbol{x} \in \boldsymbol{X} \tag{11}$$

and, thus, (10) may be rewritten as

$$p(\boldsymbol{\delta} \mid D, c) \approx \int_{\boldsymbol{T}} p(\boldsymbol{\delta} \mid \boldsymbol{t}, D, c)\, p(\boldsymbol{t} \mid D, c)\, d\boldsymbol{t}. \tag{12}$$

We pragmatically distinguished two different situations:

1. *Known marginal predictive.* In many situations, $\boldsymbol{t}$ has only a finite number of possible values with known distribution. For instance, we have often used as values for the relevant function $\boldsymbol{t}$ the cartesian product of sex, age group (less than 35, 35–65 and over 65) and level of education (no formal education, primary, high school and university); this produces a relevant function with 2*3*4=24 possible values, whose probability distribution within the more obvious classes, the politically relevant geographical areas, is precisely known from the electoral census. In this case,

$$p(\boldsymbol{\delta} \mid D, c) = \sum\nolimits_j p(\boldsymbol{\delta} \mid \boldsymbol{t}_j, D, c)\, w_{jc}, \quad \sum\nolimits_j w_{jc} = 1, \tag{13}$$

where w_{jc} denotes the weight within population class c of the subset of people with relevant function value $\boldsymbol{t}_j$.

2. *Unknown marginal predictive.* If the predictive distribution of $\boldsymbol{t}$ is unknown, or too difficult to handle, we used the $n+m$ random values of $\boldsymbol{t}$ included in the data matrix to approximate by Monte Carlo the integral (12), so that

$$p(\boldsymbol{\delta}\,|\,D, c) = \frac{1}{n+m}\sum_{j=1}^{n+m} p(\boldsymbol{\delta}\,|\,\boldsymbol{t}_j, D, c). \tag{14}$$

It is important to note that, in both cases, the 'extension of the conversation' to include the covariates automatically solved the otherwise complex problem of the *non-response*. Indeed, by expressing the required posterior distributions as weighted averages of posterior distributions *conditional* to the value of the relevant function, a *correct weight* was given to the different political sectors of the population —as described by their relevant $\boldsymbol{t}$ values— whether or not this distribution is the same within the non-response group and the rest of the population.

When the marginal predictive is known, those weights were directly input in (13), and only the data contained in D_1, *i.e.*, those which correspond to the people who answered the question, are relevant. When the marginal predictive is unknown, the weighting was done through (14) and the whole data matrix D become relevant.

The unknown predictive case is an interesting example of *probabilistic classification*. Indeed, it is *as if*, for each person with relevant function $\boldsymbol{t}$ who refuses to 'vote' for one of the alternatives $\{\delta_1, \ldots, \delta_k\}$, one would distribute his or her vote as

$$\{p(\delta_1\,|\,\boldsymbol{t}, D, c), \ldots, p(\delta_k\,|\,\boldsymbol{t}, D, c)\}, \quad \sum_{i=1}^{k} p(\delta_i\,|\,\boldsymbol{t}, D, c) = 1, \tag{15}$$

i.e., proportionally to the chance that a person, in the same class and with the same $\boldsymbol{t}$ value, would prefer each of the alternatives.

Equations (13) and (14) reformulate the original problem in terms of estimating the conditional posterior probabilities (15). But, by Bayes' theorem,

$$p(\delta_i\,|\,\boldsymbol{t}, D, c) \propto p(\boldsymbol{t}\,|\,\delta_i, D, c)\, p(\delta_i\,|\,D, c), \quad i = 1, \ldots, k. \tag{16}$$

Computable expressions for the two factors in (16) are now derived.

If, as one would expect, the $\boldsymbol{t}$'s may be considered exchangeable within each group of citizens who share the same class and the same preferences, the representation theorems (see *e.g.*, Bernardo and Smith, 1994, Chapter 4, and references therein) imply that, for each class c and preference δ_i, *there exists* a sampling model $p(\boldsymbol{t}\,|\,\boldsymbol{\theta}_{ic})$, indexed by a parameter $\boldsymbol{\theta}_{ic}$ which is some limiting form of the observable $\boldsymbol{t}$'s, *and* a prior distribution $p(\boldsymbol{\theta}_{ic})$ such that

$$p(\boldsymbol{t}\,|\,\delta_i, D, c) = \int_{\Theta_{ic}} p(\boldsymbol{t}\,|\,\boldsymbol{\theta}_{ic})\, p(\boldsymbol{\theta}_{ic}\,|\,D)\, d\boldsymbol{\theta}_{ic} \tag{17}$$

$$p(\boldsymbol{\theta}_{ic}\,|\,D) \propto \prod_{j=1}^{n_{ic}} p(\boldsymbol{t}_j\,|\,\boldsymbol{\theta}_{ic})\, p(\boldsymbol{\theta}_{ic}), \tag{18}$$

where, n_{ic} is the number of citizens in the survey which belong to class c and prefer option δ_i.

In practice, we have mostly worked with a *finite* number of $\boldsymbol{t}$ values. In this case, for each preference δ_i and class c, one typically has

$$p(\boldsymbol{t}_j \,|\, \boldsymbol{\theta}_{ic}) = \theta_{jic} \quad \sum\nolimits_j \theta_{jic} = 1, \quad i = 1, \ldots, k, \quad c \in C \tag{19}$$

where θ_{jic} is the chance that a person in class c who prefers the ith alternative would have relevant value $\boldsymbol{t}_j$, *i.e.*, a multinomial model for each pair $\{\delta_i, c\}$.

We were always requested to produce answers which would only depend on the survey results, without using any personal information that the politicians might have, or any prior knowledge which we could elicitate from previous work, so we systematically produced reference analyses. Using the multinomial reference prior, (Berger and Bernardo, 1992)

$$\pi(\boldsymbol{\theta}_{ic}) \propto \prod_j \left\{ \theta_{jic}^{-1/2} \left(1 - \sum\nolimits_{l=1}^{j} \theta_{lic} \right)^{-1/2} \right\}, \tag{20}$$

we find

$$\pi(\boldsymbol{\theta}_{ic} \,|\, D) \propto \prod_j \left\{ \theta_{jic}^{n_{jic}} \right\} \pi(\boldsymbol{\theta}_{ic}), \tag{21}$$

$$\begin{aligned} p(\boldsymbol{t}_j \,|\, \delta_i, D, c) &= \int_{\Theta_{ic}} \theta_{jic}\, \pi(\boldsymbol{\theta}_{ic} \,|\, D)\, d\boldsymbol{\theta}_{ic} \\ &= E[\theta_{jic} \,|\, D] = \frac{n_{jic} + 0.5}{n_{ic} + 1} \end{aligned} \tag{22}$$

where n_{jic} is the number of citizens in the survey which share the relevant value $\boldsymbol{t}_j$ among those which belong to class c and prefer option δ_i. Note that the reference analysis produces a result which is *independent of the actual number of different* **t** *values*, an important consequence of the use of the reference prior.

The second factor in (16) is the unconditional posterior probability that a person in class c would prefer option δ_i. With no other source of information, a similar reference multinomial analysis yields

$$p(\delta_i \,|\, D, c) = \frac{n_{ic} + 0.5}{n_c + 1}, \quad i = 1, \ldots, k, \tag{23}$$

where, again, n_{ic} is the number of citizens in the survey which belong to class c and prefer option δ_i, and n_c is the number of people in the survey that belong to class c and have answered the question. Note again that the reference prior produces a result which is *independent of the number of alternatives*, k.

Substituting (22) and (23) into (16) one finally has

$$p(\delta_i \,|\, \boldsymbol{t}_j, D, c) \propto \frac{n_{jic} + 0.5}{n_{ic} + 1} \, \frac{n_{ic} + 0.5}{n_c + 1}\,, \tag{24}$$

which is then used in either (13) or (14) to produce the final results.

Occasionally, we have used a more sophisticated hierarchical model, by assuming that for each preference δ_i, the $\{\theta_{1ic}, \theta_{2ic}, \ldots, \}$'s, $c \in C$, *i.e.*, the parameters which correspond to the classes actually used, are a random sample from some population of classes. In practice,

Prioridades de la Generalitat

De entre los diferentes servicios públicos que gestiona la ***Generalitat Valenciana*** ¿puede decirme los que en estos momentos deberían considerarse prioritarios?

1. Sanidad (ambulatorios, hospitales, control de alimentos, . . .).
2. Seguridad Ciudadana.
3. Vivienda (oferta y precios).
4. Educación (pública o subvencionada).
5. Medio Ambiente (humos, ruidos, basuras, . . .).
6. Tiempo Libre (instalaciones deportivas, espectáculos, exposiciones, . . .).
7. Infraestructuras viarias (autobuses, ferrocarriles, . . .).
8. Transporte público (autobuses, ferrocarriles,. . .)
9. Otras

		1	2	3	4	5	Otr	Totales
Comunidad Valenciana		34.9	19.1	13.6	14.2	11.4	6.8	1545
Provincia de Alicante		34.3	21.0	14.9	15.5	9.0	5.2	380
Provincia de Castellón		36.7	17.8	10.6	14.6	12.6	7.7	386
Provincia de Valencia		34.9	18.2	13.6	13.4	12.5	7.4	779
Ciudad de Valencia		34.1	17.6	15.6	14.3	10.5	8.0	389
Resto de Valencia		35.3	18.5	12.4	12.9	13.6	7.2	390
Intención voto	Abs	33.0	21.2	18.4	13.6	8.5	5.3	255
	PP	37.8	19.1	13.7	12.7	8.6	8.0	445
	PSOE	36.4	22.9	10.6	11.0	11.6	7.6	340
	EU	33.0	14.8	12.0	18.4	17.4	4.5	164
	UV	39.4	21.2	5.3	10.0	16.2	8.0	68

Figure 1. *Partial output of the analysis of one survey question*

however we have found few instances where a hierarchical structure of this type may safely be assured.

The methods described above were written in Pascal with the output formatted as a TEX file, with all the necessary code built in. This meant that we were able to produce reports of *presentation quality* only some *minutes* after the data were introduced, with the added important advantage of eliminating the possibility of clerical errors in the preparation of the reports.

Figure 1 is part of the actual output of such a file. It describes a fraction of the analysis of what the citizens of the State of Valencia thought the main *priorities of the State Government* should be at the time when the 1995 budget was being prepared. The first row of the table gives the mean of the posterior distribution of the proportions of the people over 18 in the State who favors each of the listed alternatives, and also includes the total number of responses over which the analysis is based. The other rows contain similar information relative to some conditional distributions (area of residence and favoured political party). The software combines together in ‘Others’ (Otr) all options which do not reach 5%. It may be seen from the table that it is estimateat that about 34.9% of the population believes the highest priority should be given to the health services, while 19.1% believes it should be given to law and order, and 14.2% believes it should be given to education; these estimates are based on the answers of the 1545 people who completed this question. The proportion of people who believe education should be the highest priority becomes 15.5% among the citizens of the province of Alicante, 13.6% among those who have no intention to vote, 11.0% among the socialist voters and 18.4% among the communist voters. The estimates provided were actually the means of the appropriate posterior distributions; the corresponding standard deviations were also computed, but not included in the reports in order to make those complex tables as readable as possble to politicians under stress.

Occasionally, we posed questions on a numerical scale, often the [0–10] scale used at Spanish schools. These included requests for an evaluation of the performance of a political leader, and questions on the level of agreement (0=total disagreement, 10=total agreement) with a sequence of statements designed to identify the people's values. The answers to these numerical questions were treated with the methods described above to produce probability distributions over the eleven $\{0, 1, \ldots, 10\}$ possible values. These distributions were graphically reported as histograms, together with their expected values. For instance, within the city of Valencia in late 1994, the statement "My children will have a better life than I" got an average level of agreement of 7.0, while "Sex is one of the more important things in life" got 5.0, "Spain should have never joined the European union" 3.2, and "Man should not enter the kitchen or look after the kids" only 2.0.

5. ELECTION NIGHT FORECASTING

On election days, we systematically produced several hours of evolving information. In this section we summarize the methods we used, and illustrate them with the results obtained at the May 28th, 1995 State election; the procedures used in other elections have been very similar.

Some weeks before any election we used the methodology described in Section 3 to obtain a set of representative electoral sections for each of the areas we wanted to produce specific results. In the May 95 election, a total of 100 sections were selected, in four groups of 25, respectively reproducing the political behaviour of the provinces of Alicante and Castellón, the city of Valencia, and the rest of the province of Valencia; these are the representative sections we will be referring to.

5.1. *The exit poll*

An exit poll was conducted from the moment the polling stations opened at 9 am. People were approached in their way out from the 100 representative polling stations. Interviewers handed simple forms to as many people as possible, where they were asked to mark by themselves their vote and a few covariates (sex, age, level of education, and vote in the previous election), and to introduce the completed forms in portable urns held by the interviewers.

Mobile supervisors collected the completed forms, each cycling through a few stations, and phoned their contents to the analysis center. Those answers (seven digits per person including the code to identify the polling station) where typed in, and a dedicated programme automatically updated the relevant sufficient statistics every few minutes.

The analysis was an extension of that described in Section 4. Each electoral section s was considered a class, and an estimation of the proportion of votes,

$$\{p(\delta_1 \mid D, s), \ldots, p(\delta_1 \mid D, s)\}, \quad s \in S, \tag{25}$$

that each of the parties $\delta_1, \ldots, \delta_k$ could expect in that section, given the relevant data D, was obtained by extending the conversation to include sex and age group, and using (13) rather than

(14), since the proportions of people within each sex and age group combination was known from the electoral census for all sections.

We had repeatdly observed that the logit transformations or the proportions are better behaved than the proportions themselves. A normal hierarchical model on the logit transformations of the section estimates was then used to integrate the results from all the sections in each province. Specifically, the logit transformations of the collection of k-variate vectors (25) where treated as a random sample from some k-variate normal distribution with an unknown mean vector $\boldsymbol{\mu} = \{\mu_1, \dots, \mu_k\}$ —which identify the logit transformation of the global results in the province— and were used to obtain the corresponding reference posterior distribution for $\boldsymbol{\mu}$, *i.e.*, the usual k-variate Student t (see *e.g.*, Bernardo and Smith, 1994, p. 441).

Monte Carlo integration was then used to obtain the corresponding probability distribution over the seat allocation in the province. This was done by simulating 2,000 observations from the posterior distribution of $\boldsymbol{\mu}$, using d'Hondt rule to obtain for each of those the corresponding seat allocation, and counting the results to obtain a probability distribution over the possible seat allocations and the corresponding marginal distributions on the number of seats which each party may expect to obtain in the province. The simulations from the three provinces were finally integrated to produce a forecast at State level.

The performance achieved by this type of forecast in practice is summarized in the first block of Table 3.

5.2. *The representative sections forecast*

By the time the polls closed (8 pm) the results of the exit poll could be made public. The interviewers located at the selected representative stations were then instructed to attend the scrutiny and to phone twice to the analysis center. They first transmitted the result of the first 200 counted votes, and then the final result.

The analysis of these data is much simpler than that of those from the exit poll. Indeed, we do not have here any covariates, nor any need for them, for these data do not have any non-response problems.

The results from each representative section were treated as a random sample from a multinomial model with a parameter vector describing the vote distribution within that section. Again, a hierarchical argument was invoked to treat the logit transformation of those parameters as a normal random sample centered in the logit transformation of a parameter vector describing the vote distribution in the province.

Numerical integration was then used to produce the reference posterior distribution of the province vote distribution and the implied reference posterior distribution on the seat allocation within that province. The simulations from the three provinces were then combined to produce a global forecast.

In the many elections we have tried, the technique just described produced very accurate forecasts of the final results about one hour after the stations closed. Figure 2 is a reproduction of the actual forecast made at 22h52 of May 28th, 1995, which was based on the 94 representative stations (from a total of 100) that had been received before we switched to the model which used the final returns.

Elecciones Autonómicas 1995
Comunidad Valenciana

Datos históricos relevantes

Autonómicas 1991	PP	PSOE	EU	UV	UPV	Otr
% votos	28.1	43.2	7.6	10.4	3.7	7.1
Escaños (89)	31	45	6	7	0	0

Datos procedentes del escrutinio de 94 mesas escogidas
Proyección a las 22 horas 52 min

	PP	PSOE	EU	UV	UPV	Otr
% votos válidos	**43.0**	**33.4**	**12.4**	**7.2**	2.8	1.1
Desviaciones	0.8	0.8	0.9	0.4	0.8	0.3
Escaños (89)	**42**	**32**	**10**	**5**	0	0
0.20	42	32	10	5	0	0
0.13	42	31	11	5	0	0
0.11	41	32	11	5	0	0
0.09	41	33	10	5	0	0
0.08	43	31	10	5	0	0
0.08	42	33	9	5	0	0
0.07	43	32	9	5	0	0
0.03	41	31	12	5	0	0
0.03	40	33	11	5	0	0
0.02	41	34	9	5	0	0

Distribución de diputados por partidos

PP	40	**41**	**42**	**43**	44
	0.05	0.28	0.46	0.20	0.02
PSOE	30	**31**	**32**	**33**	34
	0.03	0.26	0.42	0.24	0.04
EU	8	**9**	**10**	**11**	12
	0.03	0.18	0.42	0.30	0.06
UV	4	**5**	6		
	0.06	0.94	0.01		

Figure 2. *Actual forecast on election night, 1995*

5.3. *The early returns forecast*

By 11 pm, the return from the electoral sections which have been more efficient at the scrutiny started to come in through a modem line connected to the main computer where the official results were being accumulated. Unfortunately, one could not treat the available results as a random sample from all electoral sections; indeed, returns from small rural communities typically come in early, with a vote distribution which is far removed from the overall vote distribution.

Naturally, we expected a certain geographical consistency among elections in the sense that areas with, say, a proportionally high socialist vote in the last election will still have a proportionally high socialist vote in the present election. Since the results of the past election were available for each electoral section, each incoming result could be compared with the corresponding result in the past election in order to learn about the direction and magnitude of the swing for each party. Combining the results already known with a prediction of those yet to come, based on an estimation of the swings, we could hope to produce accurate forecasts of the final results.

Let r_{ij} be the proportion of the valid vote which was obtained in the last election by party i in electoral section j of a given province. Here, $i = 1, \ldots, k$, where k is the number of parties considered in the analysis, and $j = 1, \ldots, N$, where N is the number of electoral sections in the province. For convenience, let $\boldsymbol{r}$ generically denote the k-dimensional vector which contains the past results of a given electoral section. Similarly, let y_{ij} be the proportion of the valid vote which party i obtains in the present election in electoral section j of the province under study. As before, let $\boldsymbol{y}$ generically denote the k-dimensional vector which contains the incoming results of a given electoral section.

At any given moment, only some of the $\boldsymbol{y}$'s, say $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$, $0 \leq n \leq N$, will be known. An estimate of the final distribution of the vote $\boldsymbol{z} = \{z_1, \ldots, z_k\}$ will be given by

$$\hat{\boldsymbol{z}} = \sum_{j=1}^{n} \omega_j \boldsymbol{y}_j + \sum_{j=n+1}^{N} \omega_j \hat{\boldsymbol{y}}_j, \qquad \sum_{j=1}^{N} \omega_j = 1, \tag{26}$$

where ω_j is the relative number of voters in the electoral section j, known from the census, and the $\hat{\boldsymbol{y}}_j$'s are estimates of the $N - n$ unobserved $\boldsymbol{y}$'s, to be obtained from the n observed results.

The analysis of previous election results showed that the logit transformations of the proportion of the votes in consecutive elections were roughly linearly related. Moreover, within the province, one may expect a related political behaviour, so that it seems plausible to assume that the corresponding residuals should be exchangeable. Thus, we assumed

$$\begin{aligned} \log\left\{\frac{y_{ij}}{1-y_{ij}}\right\} &= \alpha_i \log\left\{\frac{r_{ij}}{1-r_{ij}}\right\} + \beta_i + \varepsilon_{ij}, \\ p(\varepsilon_{ij}) &= \mathrm{N}(\varepsilon_{ij} \,|\, 0, \sigma_i) \end{aligned} \qquad j = i, \ldots, k, \quad j = 1, \ldots, n, \tag{27}$$

and obtained the corresponding reference predictive distribution for the logit transformation of the $\boldsymbol{y}_{ij}$'s (Bernardo and Smith, 1994, p. 442) and hence, a reference predictive for $\boldsymbol{z}$.

Again, numerical integration was used to obtain the corresponding predictive distribution for the seat allocation in the province implied by the d'Hondt algorithm, and the simulations for the three provinces combined to obtain a forecast for the State Parliament.

The performance of this model in practice, summarized in the last two blocks of Table 3, is nearly as good as the considerably more complex model developed by Bernardo and Girón (1992), first tested in the 1991 State elections.

Table 3. *Vote distribution and seat allocation forecasts on election day 1995*

Parties	PP	PSOE	EU	UV	
Exit poll (14h29)	44.0±1.3 45	30.9±1.2 30	12.6±0.7 10	6.1±1.1 4	$p = 0.05$
Representative sections (22h52)	43.0±0.8 42	33.4±0.8 32	12.4±0.9 10	7.2±0.4 5	$p = 0.20$
First 77% scrutinized (23h58)	43.80±0.40 42	34.21±0.20 32	11.74±0.04 10	6.77±0.04 5	$p = 0.45$
First 91% scrutinized (00h53)	43.47±0.32 42	34.28±0.17 32	11.69±0.02 10	6.96±0.03 5	$p = 1.00$
Final	**43.3** **42**	**34.2** **32**	**11.6** **10**	**7.0** **5**	

5.4. *The communication of the results*

All the algorithms were programmed in Pascal with the output formatted as a TeX file which also included information on past relevant data to make easier its political analysis. A macro defined on a Macintosh chained the different programmes involved to capture the available data, perform the analysis, typeset the corresponding TeX file, print the output on a laser printer and fax a copy to the relevant authorities. The whole procedure needed about 12 minutes.

Table 3 summarizes the results obtained on May 95 election with the methods described. The timing was about one hour later than usual, because the counting for the local elections held on the same day was done before the counting for the State elections. Fore several forecasts, we reproduce the means and standard deviations of the posterior distribution of the percentages of valid vote at State level, and the mode and associated probability of the corresponding posterior distribution of the seat allocation. These include an exit poll forecast (at 14h29, with 5,683 answers introduced), a forecast based on the final results of the 94 representative sections received at 22h52 (when six of them were still missing), and two forecasts respectively based on the first 77% (reached at 23h58) and the first 91% (reached at 00h53) scrutinized stations. The final block of the table reproduces, for comparison, the official final results.

The analysis of Table 3 shows the progressive convergence of the forecasts to the final results. Pragmatically, the important qualitative outcome of the election, namely the conservative victory, was obvious from the very first forecast, in the early afternoon (when only about 60% of the people had actually voted!), but nothing precise could then be said about the actual seat distribution. The final seat allocation was already the mode of its posterior distribution with the forecast made with representative stations, but its probability was then only 0.20. That probability was 0.45 at midnight (with 77% of the results) and 1.00, to two decimal places, at 1 am (with 91%), about three hours before the scrutiny was actually finished (the scrutiny typically takes several hours to be actually completed because of bureaucratic problems always appearing at one place or another).

46250 Valencia

Partidos	1991		1995	
Censo	586 719	100.00	619 302	100.00
Abst.	218 112	37.17	169 588	27.38
psoe	134 943	36.60	116 034	25.80
pp	101 414	27.51	212 878	47.33
eu	30 761	8.34	64 469	14.33
uv	72 055	19.54	47 492	10.56
upv	8 568	2.32	4 610	1.02
cds	7 881	2.13	1 597	0.35
env	0	0.00	372	0.08
pra	0	0.00	457	0.10
pcpe	0	0.00	512	0.11
au	0	0.00	128	0.02
le-jons	0	0.00	293	0.06
pie	0	0.00	331	0.07
lae	0	0.00	194	0.04
r v	0	0.00	0	0.00
ph	0	0.00	347	0.07

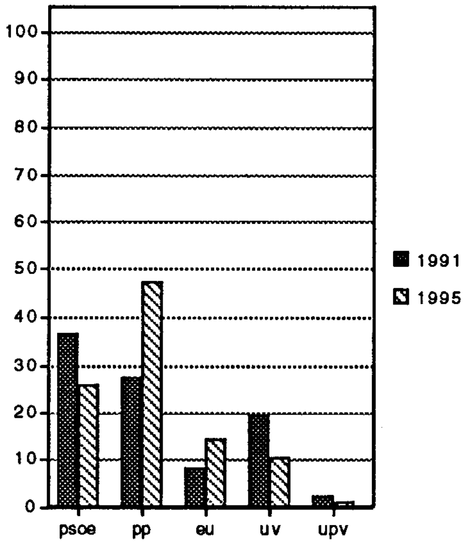

Figure 3. *Reproduction of the city of Valencia output from the 1995 election book*

By about 4 am, all the results were in, and have been automatically introduced into a relational data base (*4th Dimension* ™) which already contained the results from past elections. An script had been programmed to produce, format, and print, a graphical display of the elections results for each of the 442 counties in the State, including for comparison the results form the last, 1991, State election. Figure 3 reproduces the output which corresponds to the city of Valencia. Besides, the results where automatically aggregated to produce similar outputs for each of the 34 geographical regions of the State, for the 3 provinces, and for the State as a whole.

While this was being printed, a program in *Mathematica*™, using digital cartography of the State, produced colour maps where the geographical distribution of the vote was vividly described. Figure 4 is a black and white reproduction of a colour map of the province of Alicante, where each county is coded as a function the two parties coming first and second in the election. Meanwhile, the author prepared a short, introductory analysis to the election results.

Thus, at about 9 am, we had a camera-ready copy of a commercial quality, 397 pages book which, together with a diskette containing the detailed results, was printed, bounded and, *distributed* 24 hours later to the media and the relevant authorities, and immediately available to the public at bookshops.

6. THE DAY AFTER

After the elections have been held, both the media and the politicians' discussions often center on the *transition probabilities* $\Phi = \{\varphi_{ij}\}$ where

$$\varphi_{ij} = \Pr\{\text{a person has just voted for party } i \mid \text{ he (she) voted for party } j\}, \tag{28}$$

which describe the reallocation of the vote of each individual party between the present and the past election.

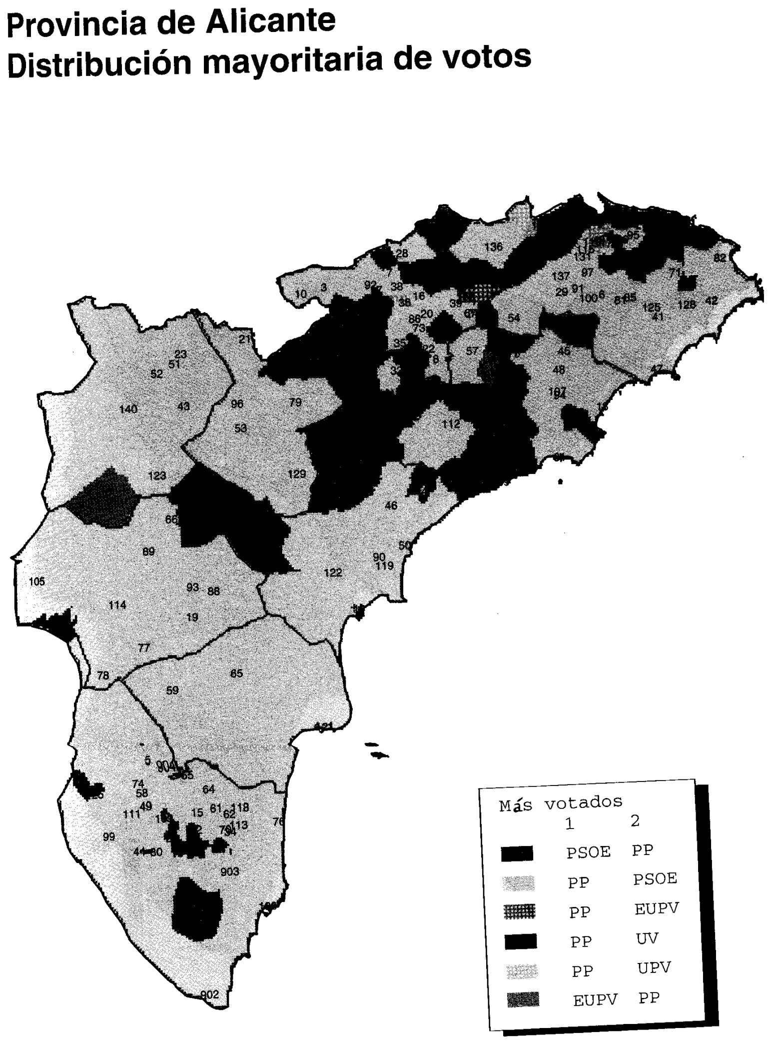

Figure 4. *Reproduction of a page on electoral geography from the 1995 election book*

Let N be the number of people included in either of the two electoral censuses involved. It is desired to analyze the aggregate behaviour of *all* those people, including those who never voted or only voted in one of the two elections. Let $\boldsymbol{p} = \{p_1, \dots, p_k\}$ describe the distribution of the behaviour of the people in the present election; thus, p_j is the proportion of those N people who have just voted for party j, and p_k is the proportion of those N people who did not vote, either because they decide to abstain or because they could not vote for whatever reason (business trip, illness, or whatever), including those who died between the two elections. Similarly, let $\boldsymbol{q} = \{q_1, \dots, q_m\}$ be the distribution of the people's behaviour in the previous election, including as specific categories not only what people voted for, if they did, but also whether they abstained in that election, or whether they were under 18 (and, hence, could not vote) at the time that election was held.

Obviously, by the total probability theorem, the transition matrix Φ has to satisfy

$$p_i = \sum_{j=1}^{m} \varphi_{ij}\, q_j, \quad i = 1, \dots, k. \tag{29}$$

A "global" estimation $\hat{\Phi}$ of the transition matrix Φ is most useful if it successfully "explains" the transference of vote in *each* politically interesting area, *i.e.*, if for each of these areas l,

$$p_{il} \simeq \sum_{j=1}^{m} \hat{\varphi_{ij}}\, q_{jl}, \quad j = 1, \ldots, m. \tag{30}$$

The exit poll had provided us with a politically representative sample of the entire population of, say, size n, for which

$$\begin{aligned} \boldsymbol{x} &= \{\text{NewVote, PastVote, Class}\} \\ \text{Class} &= \{\text{Sex, AgeGroup, Education}\} \end{aligned} \tag{31}$$

had been recorded, where Class is a discrete variable whose distribution in the population, say $p(c)$, is precisely known from the census.

For each pair $\{\text{PastVote} = j, \text{Class} = c\}$, the $\boldsymbol{x}$'s provide a multinomial random sample with parameters $\{\varphi_{1jc}, \ldots, \varphi_{kjc}, \}$ where φ_{ijc} is the probability that a person in class c had just voted party i, if he (she) voted j in the past election. The corresponding reference prior is

$$\pi(\boldsymbol{\varphi}_{jc}) \propto \prod_{i=1}^{k} \left\{ \varphi_{ijc}^{-1/2} \left(1 - \sum\nolimits_{l=1}^{i} \varphi_{ljc}\right)^{-1/2} \right\} . \tag{32}$$

Hence, for each pair (j, c) we obtain the modified Dirichlet reference posterior distribution

$$\pi(\boldsymbol{\varphi}_{jc} \,|\, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \propto \pi(\boldsymbol{\varphi}_{jc}) \prod_{i=1}^{k} \varphi_{ijc}^{n_{ijc}}, \tag{33}$$

where n_{ijc} is the number of citizens of type c in the exit poll survey who declared that have just voted i and that had voted j in the past election. The *global* posteriors for the transition probabilities $\{\pi(\varphi_{1j}, \ldots, \varphi_{kj}), j = 1, \ldots, m$ are then

$$\pi(\boldsymbol{\varphi}_j \,|\, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} = \sum_{c} \pi(\boldsymbol{\varphi}_{jc} \,|\, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) p(c), \tag{34}$$

where the $p(c)$'s are known from the census. The mean, standard deviation, and any other interesting functions of the transition probabilities φ_{ij}, may easily be obtained by simulation.

Equation (34) encapsulates the information about the transition probabilities provided by the exit poll data but, once the new results $p_1, \ldots, p_k$ are known, equation (29) has to be *exactly* satisfied. However, the (continuous) posterior distribution of the φ_{ij}'s cannot be updated using Bayes theorem, for this set of restrictions constitute an event of zero measure.

Deming and Stephan proposed in the forties an iterative adjustment of sampled frequency tables when expected marginal totals are known, which preserves the association structure and matches the marginal constraints; this is further analyzed in Bishop, Fienberg and Holland (1975). With a simulation technique, we may repeatedly use this algorithm to obtain a posterior sample of φ_{ij}'s which satisfy the conditions. Specifically, to obtain a simulated sample from each of the m conditional posterior distributions of the transition probabilities given the final results,

$$\pi(\boldsymbol{\varphi}_j \,|\, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{p}_j), \quad j = 1, \ldots, m, \tag{35}$$

	EU	PP	PSOE	...	Abs	Totales
EU	82327	11189	11796	...	41300	151242
	54.4	7.4	7.8	...	27.3	100.0
PP	2744	422648	8215	...	118082	558617
	0.5	75.7	1.5	...	21.1	100.0
PSOE	32735	85758	531739	...	192087	860429
	3.8	10.0	61.8	...	22.3	100.0
UV	7304	44056	6130	...	57728	208126
	3.5	21.2	2.9	...	27.7	100.0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Menores	10073	27046	15089	...	18314	75174
	13.4	36.0	20.1	...	24.4	100.0
Totales	**271606**	**1010702**	**798537**	...	**762419**	**3093574**
	8.8	32.7	25.8	...	**24.6**	100.0
	11.7	**43.4**	**34.3**	...	100.0	—

Figure 5. *Part of the transition matrix between the 1991 and the 1995 Valencia State Elections*

we (i) simulated from the unrestricted conditional posteriors a set of φ_{ij}'s, (ii) derived the corresponding joint probabilities $t_{ij} = \varphi_{ij}\, q_j$; (iii) applied the iterative algorithm to obtain a estimate $\hat{t}_{ij}$ which agrees with the marginals $\boldsymbol{p}$ and $\boldsymbol{q}$ and (iv) retransformed into the conditional probabilities $\hat{\varphi}_{ij} = \hat{t}_{ij}/q_j$.

The posterior mean, standard deviation, and any other interesting functions of the transition probabilities φ_{ij}, given the final electoral results $\boldsymbol{p}$, were then easily obtained from this simulated sample. Finally, we used the final estimates of the transition probabilities to derive estimates of the absolute vote transfers, obviously given by $v_{ij} = N\, \hat{\varphi}_{ij}\, q_j$, where N is the total population of the area analyzed.

Figure 5 reproduces some of the means of the posterior distribution of the transition probabilities between the 1991 and the 1995 elections in the State of Valencia, which were obtained with the methods just described. For instance, we estimated that the socialist PSOE retained 61.8% of its 1991 vote, and lost 10.0% (85,758 votes) to the conservative PP, and 22.3% (192,087) votes in people who did not vote.

7. FINAL REMARKS

Due to space limitations and to the nature of this meeting, we have concentrated on the methods we have mostly used in *practice*. Those have continually evolved since our first efforts at the national elections of 1982, described in Bernardo (1984). A number of interesting research issues have appeared however in connection with this work. A recent example (Bernardo, 1994) is the investigation of the optimal hierarchical strategy which could be used to predict election results based on early returns; this naturally leads to *Bayesian clustering* algorithms where, as one would expect from any Bayesian analysis, clearly specified preferences define the algorithm, thus avoiding the 'adhockeries' which plague standard cluster analysis.

ACKNOWLEDGEMENTS

In many senses, the work described in this paper has been joint with a team of people working under the author's supervision at *Presidència de la Generalitat*, the office of the State President. Explicit mention is at least required to Rafael Bellver and Rosa López, who respectively supervised the field work and the hardware setup and, —most specially— to Javier Muñoz, who did important parts of the necessary programming.

REFERENCES

Berger, J. O. and Bernardo, J. M. (1992). Ordered group reference priors with application to the multinomial problem. *Biometrika* **79**, 25–37.

Bernardo, J. M. (1979). Expected information as expected utility. *Ann. Statist.* **7**, 636–690.

Bernardo, J. M. (1984). Monitoring the 1982 Spanish Socialist victory: a Bayesian analysis. *J. Amer. Statist. Assoc.* **79**, 510-515.

Bernardo, J. M. (1994). Optimizing prediction with hierarchical models: Bayesian clustering. *Aspects of Uncertainty: a Tribute to D. V. Lindley* (P. R. Freeman and A. F. M. Smith, eds.). Chichester: Wiley, 67–76.

Bernardo, J. M. and Girón, F. J. (1992). Robust sequential prediction form non-random samples: the election night forecasting case. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 61–77, (with discussion).

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. Chichester: Wiley.

Bishop, Y. M., Fienberg, S. E. and Holland, P. W. (1975). *Discrete Multivariate Analysis*. Cambridge, Mass.: MIT Press.

Good, I. J. (1952). Rational decisions. *J. Roy. Statist. Soc. B* **14**, 107–114.

Haines, L. M. (1987). The application of the annealing algorithm to the construction of exact optimal designs for linear regression models. *Technometrics* **29**, 439–447.

Kirkpatrick, S., Gelatt, C. D. and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science* **220**, 671–680.

Lundy, M. (1985). Applications of the annealing algorithm to combinatorial problems in statistics. *Biometrika* **72**, 191–198.