Universitat Politècnica de Catalunya
Departament d'Estadística i Investigació Operativa

Tesi Doctoral

Anàlisi de Dades Discretes:
Freqüència de Freqüències i Dades Multinomials

Presentada per: Xavier Puig i Oriol
Director: Josep Ginebra i Molins

octubre de 2009

Índex

Resum

La Tesi la integren dues parts molt diferenciades que tenen en comú tractar de l'anàlisi de dades discretes i l'utilitzar conjunts de dades com a punt de partida.

La primera part està escrita en anglès i s'adapta al format d'una tesi escrita per articles. Aquesta part gira al voltant del modelat i l'anàlisi de freqüències de freqüències fent servir models de barreja de Poisson truncats a zero. Primer es mostra com al truncar l'espai mostral del model Inversa Gaussiana-Poisson, es pot ampliar l'espai de paràmetres del model i es comprova els avantatges de fer-ho. A continuació es comprova que una generalització del model Inversa Gaussiana-Poisson ajusta molt bé aquest tipus de dades, i explora què passa si intercanvies l'ordre entre barrejar i truncar la distribució de Poisson. L'últim capítol d'aquesta primera part defensa que la gràcia de fer servir el truncament de la barreja de Poissons per ajustar aquest tipus de dades és que permet estimar la densitat de la freqüència de paraules del vocabulari de l'autor. També proposa estimar mesures de diversitat a través de la variabilitat d'aquestes estimacions de la freqüència de paraula del vocabulari. Aquests models permeten estimar la distribució de vocabulari d'un autor i donen peu a comparar la riquesa i diversitat de vocabulari entre autors.

La segona part de la tesi, escrita en català, segueix el format de tesi tradicional i està motivada al voltant de l'anàlisi dels resultats a les últimes cinc eleccions al Parlament de Catalunya. Mitjançant models Bayesians per a l'anàlisi cluster per a dades categòriques identificarem l'existència de patrons de vot, veurem quines àrees geogràfiques pertanyen a cada patró de vot i estudiarem com aquests patrons han anat variant al llarg de les diferents eleccions. L'objectiu d'aquesta segona part és doble. Per un cantó ajudem a desenvolupar metodologia per comparar i validar models Bayesians en el context de l'anàlisi cluster de resultats electorals fent servir eines de representació gràfica. Per un altre cantó analitzem l'evolució dels resultats electorals observats. Queda pendent estendre els models Bayesians seleccionats de forma que permetin estimar les matrius de transició de vot entre eleccions consecutives.

Summary

This phD thesis is composed of two very different parts that have in common the fact that they deal with the analysis of discrete data and the use data as the starting point.

The first part is written in English and it is formatted as a thesis written by articles. This part focuses on the modeling and the analysis of frequencies of frequencies using zero truncated Poisson mixture models. First, it shows that by truncating the sample space of the inverse Gaussian-Poisson model one is allowed to extend its parameter space and in that way improve its fit. A three parameter generalization of this model is the zero truncated generalized inverse Gaussian-Poisson mixture model. In this thesis we also check that this three parameter model provides excellent fits for these type of data, and also we compare the fit of the truncated generalized inverse Gaussian-Poisson mixture model with the fit of the model that results from switching the order of the mixing and truncation stages. The last chapter of this first part argues that using zero truncated Poisson mixture models to fit this type of data allows one to estimate the density of the frequency of words in the vocabulary of the author. It also proposes to estimate measures of diversity through the variability of these estimates of the word frequencies of vocabulary. These models allow one to estimate the distribution of the vocabulary of an author and in that way allow one to compare the richness and diversity of vocabulary among authors.

The second part of the thesis, written in Catalan, follows the traditional phD thesis format, and it is motivated by the analysis of the results on the last five elections to the Parliament of Catalonia. Through the use of Bayesian models for the cluster analysis of categorical data we identify the existence of voting patterns, we allocate the areas to each patterns of vote and we study how these patterns have varied along the different elections. The aim of this second part is double. On one hand we help develop methodology to compare and validate Bayesian models in the context of the cluster analysis of electoral data using graphical tools. On the other hand, we analyze the evolution of the observed electoral results. In the near future we plan to extend these Bayesian models in order to estimate the vote transition matrices from one election to the next.

Part I

Analysis of Frequency of Frequencies and Stylometry

Capítol 1

Introduction

Some of the most useful tools in authorship attribution studies and in ecology rely on the analysis of word or species frequency count data. In the first case for example, texts are treated as samples from the vocabulary of their author and the word frequency counts in them are used to learn about his style and, in particular, about the size, evenness and diversity of his vocabulary, which might help distinguish his style from the style of other authors.

There has been a long lasting debate on which statistical models are most useful for word or species frequency count data. Given that most words (species) appear very few times and very few words (species) are repeated many times, word and species frequency count data typically have reverse J-shaped distributions with long upper tails.

The inverse gaussian-Poisson mixture model is very useful when modelling highly skewed non-negative integer data like word or species frequency count data in linguistics and in ecology. When using this statistical model on the frequency of word or species frequency data, one typically truncates its sample space at 0 to accommodate for the ignorance about the number of words or species that are not observed.

In Chapter 2 it is shown that by truncating the sample space of the inverse Gaussian-Poisson model one is allowed to extend its parameter space and in that way improve its fit when the frequency of one is larger and the right tail is heavier than is allowed by the unextended model. By fitting the extended model to word frequency count data we find many instances where the maximum likelihood estimates fall in the extension of the parameter space. A three parameter generalization of this model is the zero truncated generalized inverse Gaussian-Poisson mixture model.

In Chapter 3 it is found that this three parameter model provides excellent fits for the word frequency counts of very long texts, where the truncated inverse Gaussian-Poisson special case fails because it does not allow for the large degree of over-dispersion in the data. The role played by the three parameters of this truncated GIG-Poisson model is also explored. Our second goal in that Chapter is to compare the fit of the truncated GIG-Poisson mixture model with the fit of the model that results from switching the order of the mixing and truncation stages. An heuristic interpretation of the mixing distribution estimates obtained under this alternative GIG-Truncated Poisson mixture model is also provided.

In Chapter 4 it is first argued that modelling word or species frequency count data through zero truncated Poisson mixture models allows one to interpret the model mixing distribution as the distribution of the word or species frequencies of the vocabulary or population. As a consequence, estimates of their mixing density serve as estimates of the density of the word frequencies of the vocabulary of the author, and can be used as fingerprints of the style of the author in his texts. It is also proposed that the measures of the evenness and of the diversity of a vocabulary or population be approximated through the expectation of these measures under the word or species frequency distribution. That leads to the assessment of the lack of diversity through measures of the variability of the mixing frequency distribution estimates described above.

Chapter 2 has appeared published in *Statistical Modelling* and Chapter 3 will appear published in the *Journal of Applied Statistics*. Chapter 4 is being considered for publication, and it includes Section 4.3 and Subsection 4.4.1 which is work made only by the adviser of this thesis, Josep Ginebra, on the definition and the interpretation of measure of the evenness and of measure of the diversity within a vocabulary or population. Hence that should not be considered as part of the thesis. We included that material to make sure that the chapter was self-contained and made sense.

Capítol 2

Extended truncated inverse Gaussian-Poisson model

2.1 Introduction

The inverse Gaussian-Poisson mixture model was introduced by Holla (1966) to model highly skewed non-negative integer data in the study of repeated accidents and of recurrent disease symptoms. It has also been very successful in the analysis of the number of larvae on corn bean plants in Sankaran (1968), in the analysis of sentence length in Sichel (1974), in developing a theory of repeat-buying to understand the purchasing behavior of consumers in Sichel (1982b), in the analysis of library book circulation, of the dispersion of papers among journals, of the number of authors listed in indexed abstracts and of the number of references in scientific papers in Sichel (1985, 1991, 1992) and Burrell and Fenton (1993), in the analysis of the number of head lice in prisoners in Stein et al. (1987), when modelling insurance data ever since Willmot (1987) and Tremblay (1992), in the analysis of statistical disclosure control data in Carlson (2002), in linear models with inverse Gaussian-Poisson errors in Stein and Juritz (1988) and in Shoukri et al.(2004) and to carry out market analysis of web sites in Ajiferuke et al. (2006).

The inverse Gaussian-Poisson model, denoted from now on as the IG-Poisson model, has also been used with great success in the analysis of the frequency of word or species frequency data in linguistics and ecology ever since Sichel (1971, 1975) and Ord and Whitmore (1986). Given that the size of an author's theoretical vocabulary and the total number of species in an ecosystem are unknown, one can not count the words or

species that are not observed and it is necessary to use the zero truncated version of the IG-Poisson model. There are also many instances in other fields of application where one observes the frequency of 0's but one needs to model the frequency of zeros apart through zero-inflated or zero-deflated models and thus also require the use of the positive version of the IG-Poisson model.

It turns that by truncating the IG-Poisson model at zero one is entitled to extend its parameter space. This chapter first describes this phenomenon and then illustrates it in the specific context of the analysis of the frequency of word frequency data in stylometry, even though it applies more generally in the analysis of many other kinds of positive integer valued data with a reverse J-shaped distribution with long upper tail.

Section 2.2 describes the word frequency count data and it motivates the usefulness of the zero truncated IG-Poisson model in the analysis of this type of data. The IG-Poisson model considered here is a special case of the three parameter Sichel model advocated for in Sichel (1971, 1975). Besides providing very good fits, these models allow for a simple mechanistic explanation of the data generation process that lets one interpret the mixing distribution as the word frequency distribution of the vocabulary of the author of the text. That can be very useful when characterizing the literary style of an author.

Section 2.3 describes how one can extend the parameter space of the zero truncated IG-Poisson model. Section 2.4 illustrates the usefulness of this extended model by presenting many instances where its maximum likelihood estimates fall in the extension of the parameter space. The extended model turns out to be helpful when the frequency of ones is larger and the upper tail is heavier than allowed by the unextended model, and thus when word frequency count data sets are more over-dispersed than allowed by the unextended model.

The kind of extension that we describe for the truncated IG-Poisson model is analogous to the one reported in Griffiths (1973) for the truncated beta binomial model and in Engen (1974) for the truncated negative binomial model. It is not unlikely that similar phenomena occur for other statistical models.

Remark 2.1: The term 'extended model' here is not being used in the sense of generalizing a model by adding parameters to it, but in the sense of making it more flexible and useful by expanding its parameter space in a natural way.

2.2 Word frequency count data and the IG-Poisson model

2.2.1 Word frequency count data

To characterize the style of an author through its vocabulary the basic assumption made is that the author has available a list of words and that each word in that list, i, is characterized through the proportion of times that i would be found in a text of infinite length by that author, which is denoted by π_i. The set of probabilities π_j when j ranges over all the v words known by an author, $(\pi_1, \ldots, \pi_v)$, constitute the probability function of the theoretical vocabulary of that author. For convenience one treats π_j as a continuous variable with a density function $\psi(\pi)$ that will sometimes be denoted as the density of vocabulary.

As an approximation, texts written by an author are then treated as if they were random samples drawn from his theoretical vocabulary. If one denotes the total number of words (tokens) in a given text by n, the number of occurrences of word i by n_i, and the proportion of occurrences of that word in that text by $\hat{\pi}_i = n_i/n$, the expected value of $\hat{\pi}_i$ is π_i.

Let v_n denote the number of different words (types) in a text of size n, and let $v_{r:n}$ denote the number of different words appearing exactly r times in it, which in Good (1953) is recognized as 'the frequency of the frequency r'. The proportion of different words appearing exactly r times in a text of size n will be denoted by $\hat{p}_{r:n} = v_{r:n}/v_n$ and its expectation, which depends on n, will be denoted by $p_{r:n}$. By counting the number of words used once, $v_{1:n}$, the number of words used twice, $v_{2:n}$, and so on, one obtains the vector $(v_{1:n}, v_{2:n}, \ldots, v_{n:n})$ of word frequency count data set, and the vector of proportions $(\hat{p}_{1:n}, \hat{p}_{2:n}, \ldots, \hat{p}_{n:n})$.

The total number of words in a given text, n, the number of different words in it, v_n, and the number of different words appearing exactly r times in that text, $v_{r:n}$, are related through $v_n = \sum_{r=1}^{n} v_{r:n}$ and $n = \sum_{r=1}^{n} r v_{r:n}$.

Table 2.1 presents part of the word frequency count data sets for some chapters of *Tirant lo Blanc*, which is a chivalry book written in catalan and analyzed in Riba (2002), Giron et al. (2005) and Riba and Ginebra (2006). For example, the first row in that table indicates that Chapter 1 has a total of 255 words, out of which 142 are different; in it

	$v_{1:n}$	$v_{2:n}$	$v_{3:n}$	$v_{4:n}$	$v_{5:n}$	$v_{6:n}$	$v_{7:n}$	$v_{8:n}$	$v_{9:n}$	$v_{10:n}$	$v_{11:n}$	$v_{12:n}$	...	n	v_n
Chapter 1	107	16	6	2	2	2	2	1	1	1	0	1	...	255	142
Chapter 2	172	26	19	7	2	2	2	2	1	1	1	1	...	476	239
Chapter 3	299	70	32	16	10	5	4	2	5	1	2	0	...	1174	459
Chapter 4	205	52	20	7	10	3	2	2	1	0	1	1	...	670	310
Chapter 5	302	54	27	18	7	4	4	1	1	1	1	2	...	1089	435
Chapter 6	238	37	18	6	2	2	1	1	0	1	2	0	...	615	315
Chapter 7	123	20	7	3	3	1	0	0	0	1	0	0	...	283	161
Chapter 8	97	11	9	4	1	3	0	1	2	0	0	0	...	237	130
Chapter 9	111	16	2	6	0	1	2	0	2	0	0	1	...	223	141
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Chapter 485	213	49	14	6	5	4	4	2	0	3	1	1	...	741	309
Chapter 486	108	25	10	13	2	2	0	1	0	2	0	0	...	402	169
Chapter 487	129	29	10	6	1	1	0	2	2	2	0	0	...	348	184

Taula 2.1: Part of the word frequency counts in the 425 chapters of *Tirant lo Blanc* with more than 200 words, obtained from Riba (2002).

107 words appear once, 16 words appear twice, 6 words appear three times and so on, with the most frequent word appearing 15 times. The longest chapter in that book has a total of $n = 6521$ words, it has $v_n = 1365$ different words and its most frequent word appears 354 times.

Table 2.2 presents the word frequency count for the nouns in the Macaulay's essay on Bacon, considered in Sichel (1975), of *Alice in Wonderland* and of *Through the Looking Glass* by Lewis Carroll, of *The War of the Worlds* by H.G. Wells, of *Max Havelaar* which is in Dutch and by Douwes Dekker and of a turkish archeology text, all considered in Baayen (2001).

	$v_{1:n}$	$v_{2:n}$	$v_{3:n}$	$v_{4:n}$	$v_{5:n}$	$v_{6:n}$	$v_{7:n}$	$v_{8:n}$	$v_{9:n}$	$v_{10:n}$	...	n	v_n
Essays on Bacon	990	367	173	112	72	47	41	31	34	17	...	8045	2048
Alice in Wonderland	1176	402	233	154	99	57	65	52	32	36	...	26505	2651
Through the Looking	1491	460	259	148	113	78	61	47	28	26	...	28767	3085
War of the Worlds	3613	1138	567	340	250	177	135	93	72	67	...	59938	7112
Max Havelaar	6004	1731	819	491	368	258	168	137	123	108	...	99767	11161
Turkish Archeology	2326	477	178	107	53	33	22	26	7	7	...	6939	3302

Taula 2.2: Part of the word frequency counts of the nouns in Macaulay's essay on Bacon, and of all the words in *Alice in Wonderland*, of *Through the Looking Glass*, of the *War of the Worlds*, of *Max Havelaar* and of a Turkish archeology text.

Characterizing the vocabulary of an author by modelling the word frequency counts in its texts, $(v_{1:n}, v_{2:n}, \ldots, v_{n:n})$, is more convenient than by modelling the word counts in them, $(n_1, n_2, \ldots, n_v)$, because the support of this later distribution encompasses the words of both the observed and the unobserved vocabulary and it is thus only partially known. On the down side, note that the distribution of $(v_{1:n}, v_{2:n}, \ldots, v_{n:n})$ depends on text size n in ways that are more complicated than the distribution of $(n_1, n_2, \ldots, n_v)$. In particular, note that the expectation of $(\hat{p}_{1:n}, \hat{p}_{2:n}, \ldots, \hat{p}_{n:n})$ depends on n, which is not the case for $(\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_v)$.

Given that most of the words appear only very few times and only very few words are repeated many times, the distribution of $(v_{1:n}, v_{2:n}, \ldots, v_{n:n})$ is reverse J-shaped with an extraordinarily long upper tail. Zipf (1932) models word frequency count data through the Riemann distribution under which $p_{r:n}$ is made proportional to r^{-s} with $s > 1$, but its fits to word frequency count data are unsatisfactory. Yule (1944) and Good (1953) conjecture that the strong skewness of these distributions should be modelled through Poisson mixtures. Herdan (1961, 1964) considers the Waring distribution, but it is not flexible enough because its mode is always at $r = 1$ and its parameters are not related to the text size n. This leads Herdan to conclude that the structure of word frequency distributions is indeed the one of a Poisson mixture.

Following Sichel (1975), the probabilities that a word is repeated exactly r times in a text of size n, $(p_{1:n}, p_{2:n}, \ldots, p_{n:n})$, are related next to the density of vocabulary $\psi(\pi_i)$ and to n through the zero truncated IG-Poisson model.

2.2.2 The zero truncated IG-Poisson model

If a specific word, i, has a probability π_i of being used each time that an author writes a word, the number of times that this word appears in a text by that author with a total of n words would be distributed as a binomial(n, π_i). Hence, if the density of the theoretical vocabulary of the author was $\psi(\pi)$ the probability that a word from that vocabulary appears exactly r times in a text of size n, $p_{r:n}$, can be modelled through a $\psi(\pi)$-binomial mixture model.

Usually n will be large and all the π_i will be very small, and therefore one can approximate the model for $p_{r:n}$ through a $\psi(\pi)$-Poisson mixture model,

$$p_{r:n}^{pm} = \int_{R^+} \frac{(n\pi)^r e^{-n\pi}}{r!} \psi(\pi) d\pi, \text{ for } r = 0, 1, \ldots, n. \tag{2.1}$$

This argument provides a simple mechanistic description of the word frequency count

generation process as a Poisson mixture, which is lacking in the ad-hoc models often considered for this kind of data (see, e.g., Baayen 2001), and which entitles one to interpret the model mixing density $\psi(\pi)$ as the density of the word frequency of the vocabulary of the author. Hence, that mixing distribution can be very helpful when characterizing the literary style of an author starting from the word frequency count data sets of his texts.

Note that the same argument applies when one models the probability of observing r individuals of an specie in a given area in ecology, or the probability of an insurance customer making r claims or of a library book being borrowed r times during a given time period. In those instances where one samples from populations with a very small total number of individuals, v, it might be better to use discrete Poisson mixture models like the ones recently considered in Böhning and Kuhnert (2006).

Following a recommendation in Good (1953), Sichel (1971, 1975) models the mixing vocabulary distribution through an inverse-gaussian distribution, defined on R^+ with density function

$$\psi(\pi|b,c) = \frac{b}{2}\sqrt{\frac{c}{pi}}e^{b}\pi^{-3/2}e^{-\frac{\pi}{c}-\frac{b^2c}{4\pi}}, \tag{2.2}$$

where b is in $(0,\infty)$, c is in $(0,\infty)$, and pi is the known irrational number. Even though the support of (2.2) is $(0,\infty)$, under the values of (b,c) that one considers in practice (2.2) decreases very fast with increasing π and it approximates the distribution of the theoretical vocabulary very well. For details on this distribution see Chikara and Folks (1989), Seshadri (1993, 1999) and chapter 15 in Johnson and Kotz (1994).

By replacing (2.2) in (2.1) and solving the integral one obtains the probability mass function,

$$p^{igp}_{r:n}(b,c) = \sqrt{\frac{2}{pi}}\sqrt{b(1+cn)^{1/2}}\frac{e^b}{r!}\Big(\frac{bcn}{2\{1+cn\}^{1/2}}\Big)^r K_{r-\frac{1}{2}}(b\{1+cn\}^{1/2}),\ \text{for } r=0,1,\ldots, \tag{2.3}$$

where $K_\gamma(\cdot)$ is the modified Bessel function of the third kind of order γ. The support of (2.3) is unbounded but in practice $p^{igp}_{r:n}(b,c)$ dies out very fast with increasing r. Sichel (1975, 1986a) reparametrizes this model through

$$\alpha = b\{1+cn\}^{1/2} \ \text{ and } \ \theta = \frac{cn}{1+cn}, \tag{2.4}$$

where (α,θ) is in $(0,\infty)\times(0,1)$ and it captures the dependence of $p_{r:n}$ both on text size, n, as well as on the mixing IG(b,c) distribution. Note that unless the text size n is very small, θ will be near one. Under the new parametrization, (2.3) becomes

$$p^{igp}_{r:n}(\alpha,\theta) = e^{\alpha\sqrt{1-\theta}}\sqrt{\frac{2\alpha}{pi}}\frac{(\frac{1}{2}\alpha\theta)^r}{r!}K_{r-\frac{1}{2}}(\alpha), \qquad \text{for} \quad r=0,1,2,\ldots. \tag{2.5}$$

For any given n there is a one-to-one relationship between (α, θ) in $(0, \infty) \times (0, 1)$ and (b, c) in $(0, \infty) \times (0, \infty)$, with $b = \alpha\sqrt{1-\theta}$ and $c = \theta/n(1-\theta)$.

This two parameter model first considered in Holla (1966) and denoted as the IG-Poisson model, is a special case of a three parameter generalized IG-Poisson mixture model proposed in Sichel (1975) for word frequency count data of texts with large n. It is also a special case of the compound generalized Poisson Pascal model considered in Willmot (1988b), which allows one to interpret this mixture model as a Poisson stopped sum of extended truncated negative binomial random variables (see Willmot, 1986, and Klugman et al., 1998, p.260). The IG-Poisson is also a member of the Sundt-Jewell family of distributions (see Willmot, 1988c).

Given that the size v of an author's theoretical vocabulary is unknown, one can not count the words that are known by the author but are not observed in the text, and that makes it necessary to consider the positive version of (2.5),

$$p_{r:n}^{tigp}(\alpha, \theta) = \frac{e^{\alpha}}{e^{\alpha[1-(1-\theta)^{1/2}]} - 1} \sqrt{\frac{2\alpha}{pi}} \frac{(\frac{1}{2}\alpha\theta)^r}{r!} K_{r-\frac{1}{2}}(\alpha), \qquad \text{for} \quad r = 1, 2, \ldots, \tag{2.6}$$

which is recognized as the *zero truncated IG-Poisson*(α, θ) model. To avoid the need to evaluate the Bessel function, Sichel (1975) computes these probabilities recursively through

$$p_{1:n}^{tigp}(\alpha, \theta) = \frac{1}{2} \frac{\alpha\theta}{e^{\alpha(1-\sqrt{1-\theta})} - 1}, \tag{2.7}$$

$$p_{2:n}^{tigp}(\alpha, \theta) = \frac{1}{4}\theta(1+\alpha) p_{1:n}^{tigp}(\alpha, \theta), \tag{2.8}$$

and

$$p_{r:n}^{tigp}(\alpha, \theta) = \theta(1 - \frac{3}{2r}) p_{r-1:n}^{tigp}(\alpha, \theta) + \frac{(\alpha\theta)^2}{4r(r-1)} p_{r-2:n}^{tigp}(\alpha, \theta), \text{ for } r > 2. \tag{2.9}$$

From Sichel (1992) follows that the limiting model in (2.6) when α tends to 0 is:

$$p_{r:n}^{tigp}(\alpha = 0, \theta) = \frac{1}{2} \frac{1}{1-\sqrt{1-\theta}} \frac{\Gamma(r-\frac{1}{2})}{\Gamma(\frac{1}{2})} \frac{\theta^r}{r!}, \text{ for } r = 1, 2, \ldots, \tag{2.10}$$

which is a special case of the extended truncated negative binomial model considered in Engen (1974), Willmot (1988c) and Hoshino (2005). This limiting model allows one to extend the parameter space of the truncated IG-Poisson(α, θ) model from $(0, \infty) \times (0, 1)$ to $[0, \infty) \times (0, 1)$.

Sichel (1971, 1974, 1975, 1982a, 1986a, b), Burrell and Fenton (1993), Heller (1997) and Karlis (2001) propose various estimation methods for (2.5) or (2.6) including maximum likelihood, the method of moments, minimizing the χ^2 statistic and equating the observed and the theoretical probabilities for $r = 1, 2$. The maximum likelihood estimating

equations for the untruncated model, in (2.5), can be found in Sichel (1971,1986a) and in Atkinson and Yeh (1982), and the ones for alternative parametrizations of the model in Stein et al.(1987) and in Willmot (1988a).

In this chapter all the estimation will be done through the maximization of the likelihood function by direct search of the likelihood surface, which nowadays has become a straightforward computational exercise that avoids the need to deal with the complicated estimating equations of the zero truncated version of the model, in (2.6).

Sichel (1975, 1986a), Pollatschek and Radday (1981), Holmes (1992), Holmes and Forsyth (1995), Baayen (2001) and Riba and Ginebra (2006) fit the zero truncated IG-Poisson(α, θ) model to word frequency count data and find that unless n is large it fits the data reasonably well. Ord and Whitmore (1986) fit this model on species frequency count data, while Sichel (1992a, b) uses it on bibliometric data in which one observes the frequency of zeros but needs to model $p_{0:n}$ apart from $p_{r:n}$ for $r > 0$.

2.3 Extended truncated IG-Poisson model

In the literature, the parameter space for the zero truncated IG-Poisson model in (2.6) is always assumed to be $(0, \infty) \times (0, 1)$, or $[0, \infty) \times (0, 1)$ if one includes the limiting model in (2.10) as we do. It is not unusual to find instances where the estimate of α is 0 or close to 0.

In fact, it turns that for any given value of (α, θ) in $(-1, 0) \times (0, 1)$ the set of probabilities $p^{tigp}_{r:n}(\alpha, \theta)$ defined recursively through (2.7), (2.8) and (2.9) are all positive and such that $\sum_r p^{tigp}_{r:n}(\alpha, \theta) = 1$, and hence when α is in $(-1, 0)$ these recursive equations also define a probability distribution on the positive integers.

As a consequence, when one truncates the sample space of the IG-Poisson model at zero one can extend the parameter space for (α, θ) from $[0, \infty) \times (0, 1)$, which is the parameter space of the un-truncated model, into $(-1, \infty) \times (0, 1)$, and in this way gain flexibility when modelling word or species frequency count data. We denote the probability mass function defined through (2.7), (2.8) and (2.9) on this extended parameter space by $p^{etigp}_{r:n}(\alpha, \theta)$.

It is important to note that this extension of the parameter space is only feasible after one truncates at zero the sample space of the IG-Poisson model and it is therefore directly

related to that truncation. In fact, the values of α in $(-1, 0)$ extending the parameter space of the truncated model are not allowed for the untruncated IG-Poisson model because $p_{1:n}^{igp}(\alpha, \theta)$ is negative whenever α is negative. This phenomenon is analogous to the one observed in Griffiths (1973) when one truncates the beta-binomial model and to the one observed in Engen (1974) when one truncates the negative binomial model.

Figure 2.1 presents contour plots of $p_{r:n}^{etigp}(\alpha, \theta)$ as a function of (α, θ) for $r = 1, \ldots, 10, 15, 20$. When $\alpha > 0$ these contour plots coincide with the ones of (2.6), when $\alpha = 0$ they correspond to (2.10) but for α in $(-1, 0)$ they are only defined through (2.7), (2.8) and (2.9). Figure 2.2 presents the probability mass function of this model for sixteen pairs of values of (α, θ).

Figure 2.1 indicates that $p_{1:n}^{etigp}(\alpha, \theta)$ is a decreasing function of both α and θ. That this is always the case can be proved by differentiating (2.7) with respect to α and θ and using the Taylor series expansion of the exponential function. For that reason, for the unextended model the maximum value for the probability at one corresponds to $\alpha = 0$, and from (2.10) it is equal to:

$$p_{1:n}^{tigp}(\alpha = 0, \theta) = \frac{1}{2} \frac{\theta}{1 - \sqrt{1 - \theta}}, \tag{2.11}$$

which for θ near 1 is close to one half. Unless n is very small, θ is bound to adopt values close to 1 because of (2.4), and therefore the only way to increase the probability at one is by considering negative values of α.

Observe also that $p_{2:n}^{etigp}(\alpha, \theta)$ is an increasing function of α, and that away from $\theta = 1$, when $r > 2$ the probability $p_{r:n}^{etigp}(\alpha, \theta)$ is an increasing function of both α and θ.

In fact observe that α mostly determines the value of $p_{r:n}^{etigp}(\alpha, \theta)$ for small r while θ determines the value of that probability function for large r, as anticipated in Sichel (1982) for the unextended part of the model. As a consequence if one decreased α and increased θ simultaneously the frequency of ones and the tail probabilities would simultaneously increase and one would obtain a more dispersed word frequency count set distribution.

By extending the parameter space to include values of α in $(-1, 0)$ the statistical model gains flexibility because it accommodates for word frequency count data sets with larger values of $p_{1:n}$ which will improve the fit for texts from a theoretical vocabulary with a large total number of words v and thus with small word frequencies, π_i.

Whenever the estimate of α falls in $(-1, 0)$ this extended model improves the fit of the unextended model that restricts α to be in $[0, \infty)$. When that happens, it indicates

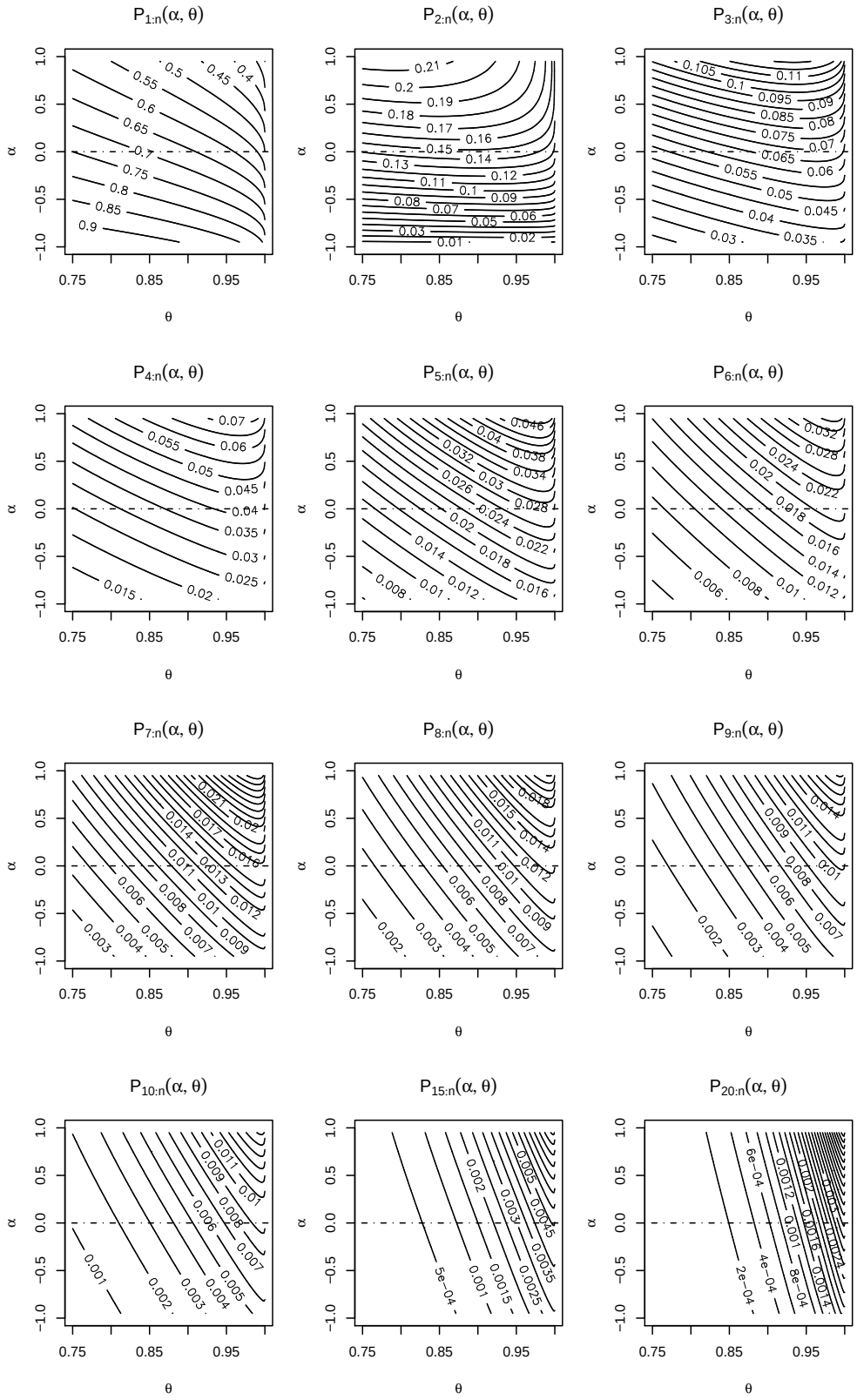

Figura 2.1: Contour plots for the probability function of the extended truncated IG-Poisson(α, θ) model, $p^{etigp}_{r:n}(\alpha, \theta)$, computed recursively through (2.7), (2.8) and (2.9) for $r = 1, \ldots, 10, 15, 20$.

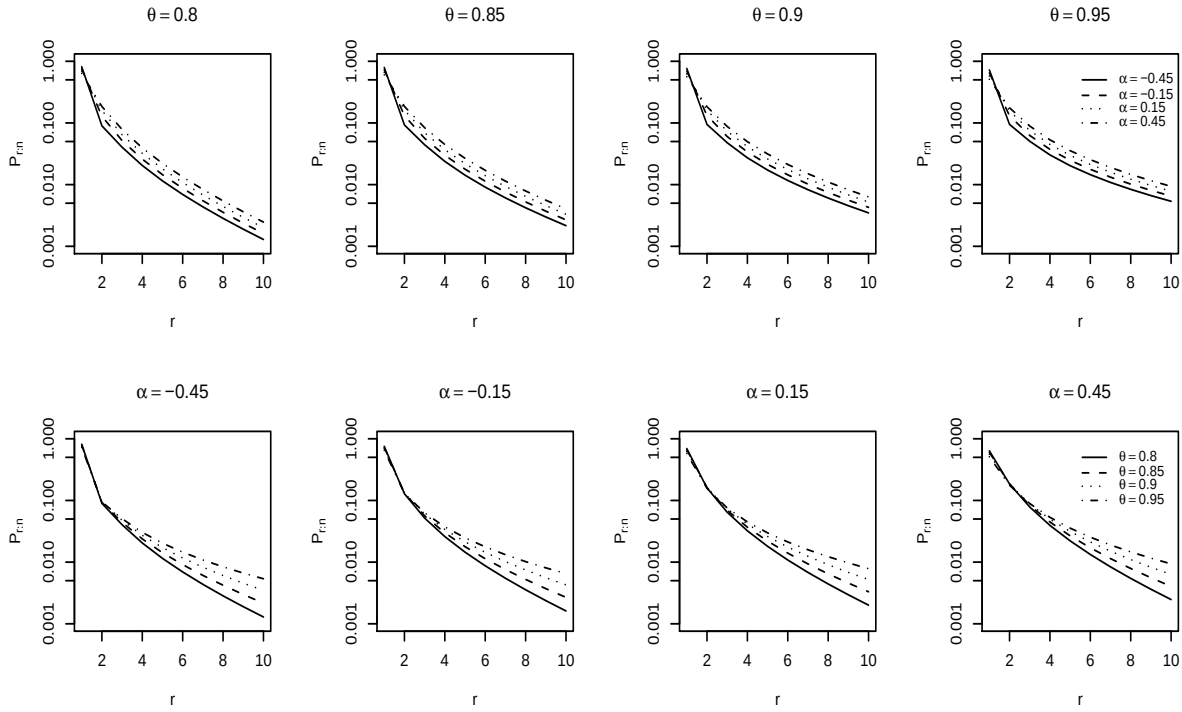

Figura 2.2: Probability mass function of the extended truncated IG-Poisson model, $p_{r:n}^{etigp}(\alpha, \theta)$, computed recursively through (2.7), (2.8) and (2.9) for sixteen pairs of values of (α, θ).

that the word frequency count data is either from a text of an author with a very rich vocabulary or from a text written in a language with a grammar leading to very small values of π_i and hence to a very large number of different words and very small word repeat rates. This phenomena will often disappear whenever one lemmatizes the text (i.e., when one replaces all words by their underlying lemma) before obtaining the word frequency count data.

Remark 2.2: When α is negative one can not interpret the $(p_{1:n}^{etigp}(\alpha,\theta), p_{2:n}^{etigp}(\alpha,\theta),\ldots)$ obtained through (2.7), (2.8) and (2.9) as the probability function of a zero truncated IG-Poisson(α,θ) mixture distribution and hence, one can not identify the corresponding mixing distribution.

2.4 The extended model in practice

To illustrate the usefulness of extending the parameter space of the truncated IG-Poisson model we fitted the extended model on the word frequency count data of each of the 425 chapters of *Tirant lo Blanc*, partially considered in Table 2.1. In order to know if the language has any effect on the gain obtained fitting the extended model we also fitted it to the data in Table 2.2, which includes texts in English, Dutch and Turkish.

If the word frequencies are independent random variables, identically distributed with an extended zero truncated IG-Poisson(α,θ) distribution, the likelihood function is such that:

$$L_{(v_{1:n},\ldots,v_{n:n})}^{etigp}(\alpha,\theta) \propto \Pi_r (p_{r:n}^{etigp}(\alpha,\theta))^{v_{r:n}}. \tag{2.12}$$

Figure 2.3 presents the contour plots of the log-likelihood functions for chapters 1 to 9 of *Tirant lo Blanc*. They are all unimodal and with a maximum attained at negative values of α and therefore in the extended part of the parameter space. Even though some of the texts considered are very short, all the corresponding likelihood surfaces have a clear absolute maximum and thus the computation of the maximum likelihood estimates by direct search is very simple.

Figure 2.4 relates the maximum likelihood estimates of α and θ with n for all the chapters considered. For 405 out of the 425 chapters the maximum likelihood estimate of α is negative and the extended truncated IG-Poisson(α,θ) model (with $\alpha > -1$) improves the fit of the unextended model (with $\alpha \geq 0$). The word frequency count data considered here was obtained without previously lemmatizing the text. This, combined with the text being written in a romanic language explains that the type token ratio, v_n/n, and

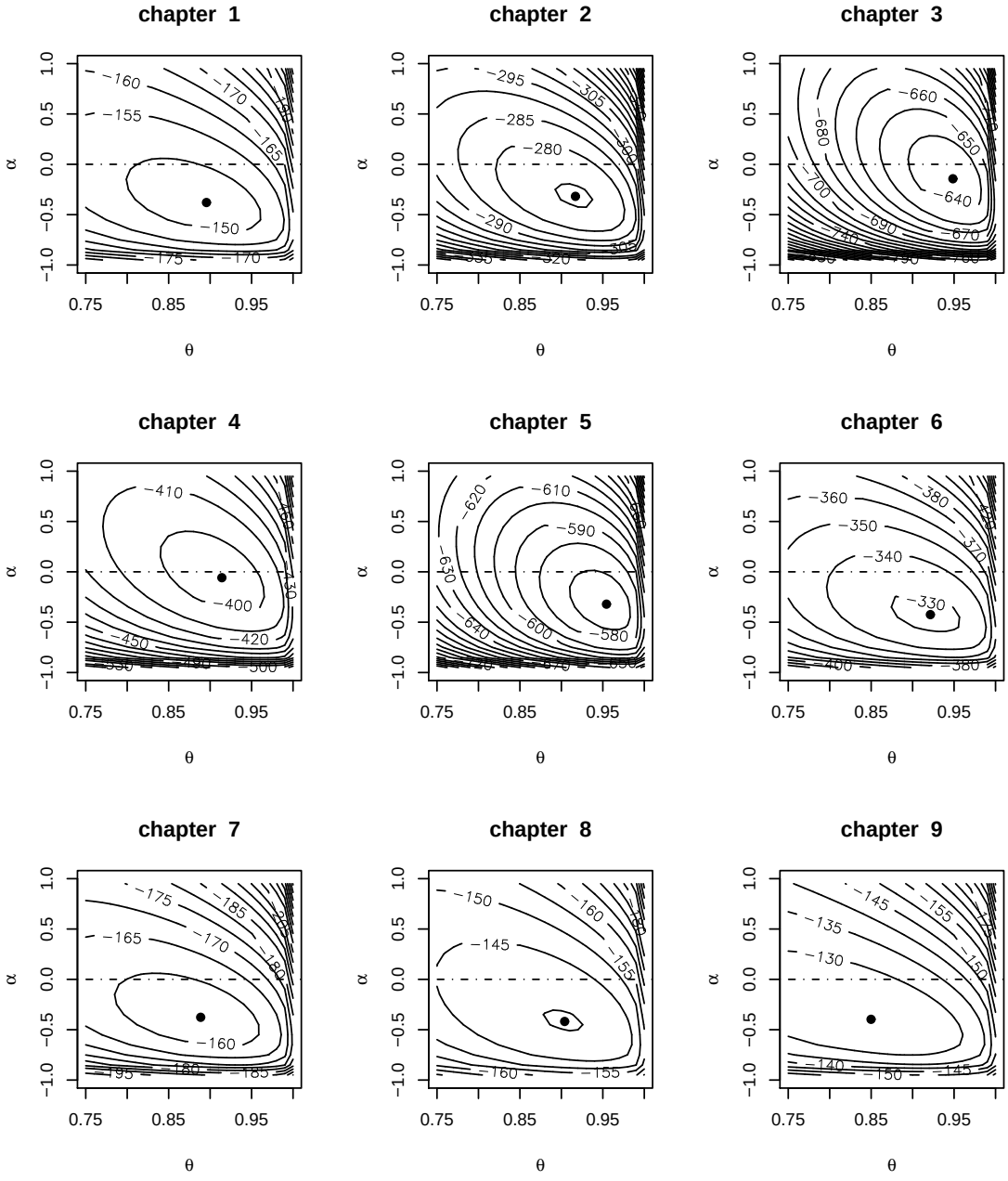

Figura 2.3: Contour plots for the log-likelihood function under the extended truncated IG-Poisson(α, θ) model, with $(\alpha, \theta) \in (-1, \infty) \times (0, 1)$, for the word frequency count data of Chapters 1 to 9 of Tirant lo Blanc partially presented in Table 2.1.

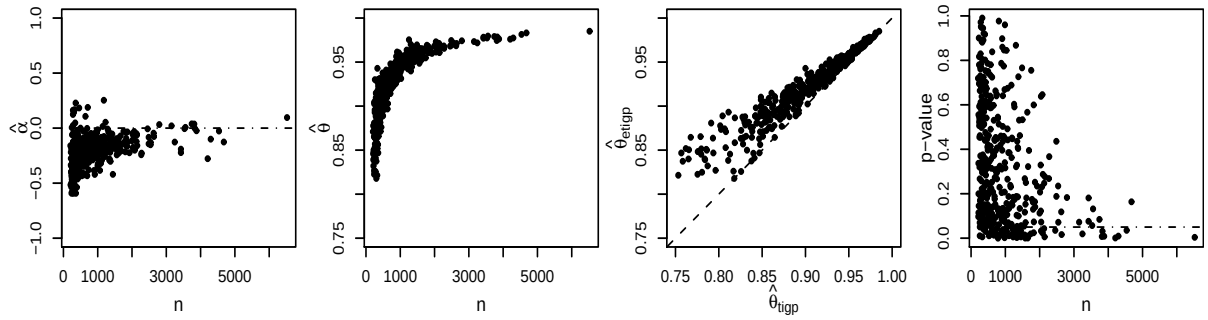

Figura 2.4: Relationship between the estimates of (α, θ) and n, between the estimate for θ under the extended model and the one under the unextended model, and between the χ^2 goodness of fit test p-value of the extended truncated IG-Poisson model and n. The data are the word frequency count sets of the 425 chapters of *Tirant lo Blanc* with more than 200 words.

the proportion of words appearing once, $\hat{p}_{1:n}$, both tend to be larger than one would find in similar texts written in English. That helps explain that, different from what tends to happen for texts written in English, here we find $\hat{\alpha}$ to be negative for more than 95% of the chapters.

Figure 2.4 also relates the maximum likelihood estimate of θ under the extended and under the unextended models, showing that the first estimate is larger than or equal to the second estimate for all the 425 chapters in Tirant lo Blanc. Hence letting α be negative always results in larger estimates for θ, which for the first nine chapters of the book is already clear from the likelihood surfaces presented in Figure 2.3.

Given that for large r the value of θ determines $p_{r:n}^{etigp}(\alpha, \theta)$, which is increasing with θ, the fits under the extended truncated IG-Poisson model always end up with heavier upper tails than the fits under the unextended model. Given that by letting α to be negative one also ends up with larger $p_{1:n}^{etigp}(\alpha, \theta)$, the extension of the parameter space helps when the data is more over-dispersed than is allowed by the unextended model.

The last panel in Figure 2.4 relates the χ^2 goodness of fit test p-value with n. The test statistic is computed after aggregating categories the least so that the expected count in each category is 4 or larger. It follows that the extended zero truncated IG-Poisson model fits the data well for the word frequency counts from most of the texts considered.

To compare the extended and the unextended zero truncated IG-Poisson model fits, the upper half part of Figure 2.5 presents the base 10 logarithm of the ratio between the maximum of the likelihood of the extended model and the maximum of the likelihood of the unextended model for the chapters under consideration. When the maximum

likelihood estimate of α for the extended model is larger than 0, the maximum of the likelihood function under both models coincide and the likelihood ratio is equal to 1, but that only happens for 20 chapters.

For the 405 chapters where the maximum likelihood estimate of α under the extended model is negative, the maximum likelihood estimate for the unextended model is equal to 0 and this likelihood ratio will be larger than 1. In particular, the likelihood ratio is larger than 10 for 127 chapters, and it is larger than 100 for 42 chapters. The downside is that for these chapters the fitted distribution can not be interpreted as a truncated IG-Poisson mixture anymore, which might be a handicap if one intended to use an estimate of the model mixing distribution as an estimate of the word frequency distribution of the vocabulary of the author.

Here one can not resort to the usual likelihood ratio test or to the AIC to compare the extended and the unextended model. Figure 2.5 presents the difference between the value of the χ^2 goodness of fit statistic under the extended and the χ^2 statistic under the unextended truncated IG-Poisson models. To compute these two statistics, both models were fitted under the initial categorization but then the categories were aggregated the least so that the expected count under the extended model was at least 4 in every category, and that re-categorization was used to compute the statistic and the p-value for both models.

Given that models are fit through the maximization of the likelihood and not through the minimization of the χ^2 statistic, there are a few instances where the χ^2 statistic for the extended model is slightly larger and the corresponding p-value is slightly smaller than for the unextended model. Most often though, the χ^2 statistic is considerably smaller and the p-value is considerably larger under the extended model.

Table 2.3 compares the observed and the expected word frequency counts under both models for Chapter 5, that is one of the many chapters where the extended truncated IG-Poisson model is not rejected with the χ^2 goodness of fit test but where the unextended model is rejected with it. Observe that the expected counts under the extended model are closer to the observed counts in all the categories which is a pattern found in almost all the chapters of that book.

In order to illustrate the role played by the language in which the texts are written we have also fitted the extended model to the word frequency count data in Table 2.2. Figure 2.6 presents the contour plots for the corresponding log-likelihood functions. For the first four texts considered, which are all written in English, the estimate of α is positive and therefore for them the extension of the parameter space is not necessary.

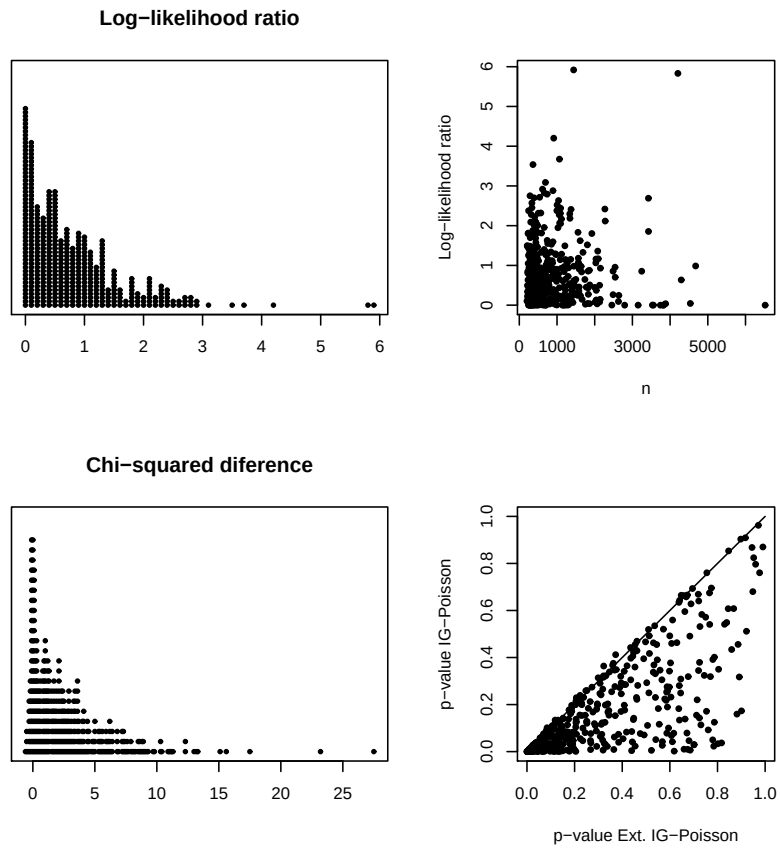

Figura 2.5: Base 10 logarithm of the ratio between the maximum of the likelihoods of the extended truncated IG-Poisson(α, θ) model (with $\alpha > -1$) and of the unextended model (with $\alpha \geq 0$), difference between the χ^2 goodness of fit test statistics of the unextended and of the extended models, and relation between the χ^2 goodness of fit test p-values of both models, all for the 425 chapters of *Tirant lo Blanc* with more than 200 words.

		Tr.IG-Poisson		Ext. Tr.IG-Poisson	
r	$v_{r:n}$	$v_n\hat{p}_{r:n}$	residual	$v_n\hat{p}_{r:n}$	residual
1	302	272.5	1.788	299.4	0.153
2	54	63.9	-1.240	48.5	0.791
3	27	30.0	-0.540	24.3	0.545
4	18	17.5	0.108	14.6	0.891
5	7	11.5	-1.330	9.8	-0.889
6	4	8.1	-1.439	7.0	-1.138
7	4	6.0	-0.803	5.3	-0.551
8-9	2	8.1	-2.140	7.3	-1.970
10-12	4	7.0	-1.130	6.6	-1.021
13-16	4	4.9	-0.401	4.9	-0.422
17-23	5	3.8	0.639	4.2	0.403
≥ 24	5	2.8	1.302	4.1	0.443
$\hat{\alpha}$		0.		-0.3211	
$\hat{\theta}$		0.9373		0.9544	
$\chi 2$		17.639		9.587	
p-value		0.034		0.385	

Taula 2.3: Observed and expected word frequency counts and the corresponding Pearson residuals, $(v_{r:n} - v_n\hat{p}_{r:n})/\sqrt{v_n\hat{p}_{r:n}}$, for Chapter 5 of *Tirant lo Blanc* under the extended and under the unextended truncated IG-Poisson models.

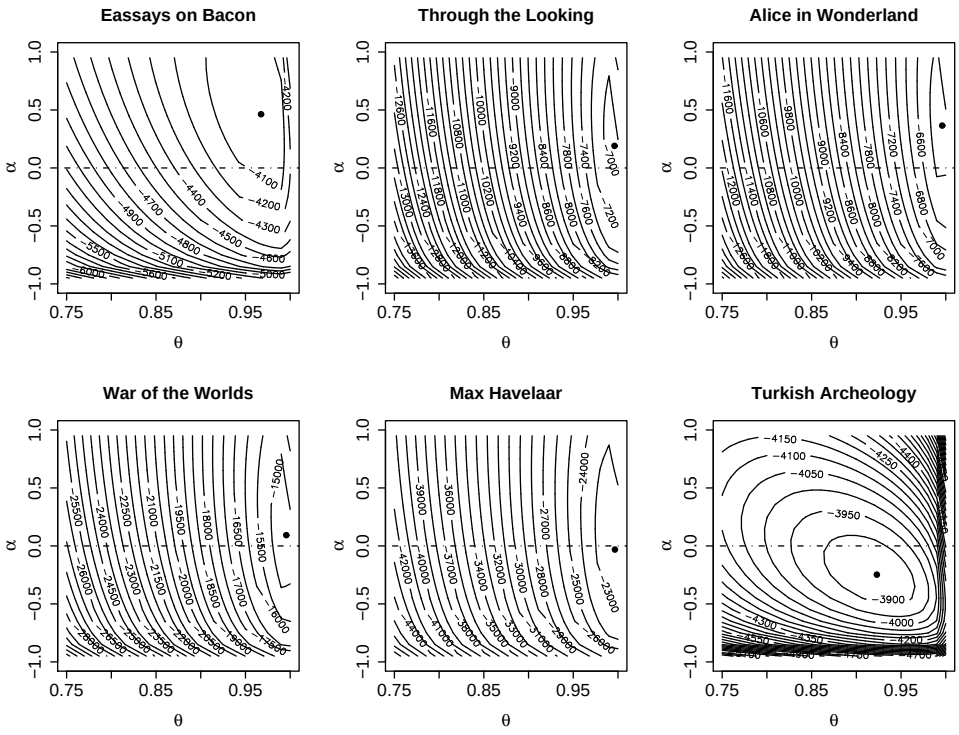

Figura 2.6: Contour plots for the log-likelihood function for the extended truncated IG-Poisson model for the word frequency count of the nouns in Macaulay's essay on Bacon, and of all the words in *Alice in Wonderland*, in *Through the Looking Glass*, in *War of the Worlds*, in *Max Havelaar* and in a Turkish archeology text, all partially presented in Table 2.2.

On the other hand the estimate of α for Max Havelaar, written in Dutch, and for the Turkish archeology text are negative, and for these two texts the extended model gives a better fit than the unextended model; these are two texts with a very large proportion of words being used only once.

Even though the texts in Table 2.2 are all considerably longer than the ones in Table 2.1, the χ^2 goodness of fit test p-value for the essays on Bacon ranges from 0.875 to 0.936, for *Alice in Wonderland* it ranges from 0.006 to 0.097 and for *Through the Looking Glass* it ranges from 0.018 to 0.039, depending on the degree of aggregation of the categories when computing that statistic. Hence the zero truncated IG-Poisson model fits the word frequency count data of these three texts reasonably well. For the texts in the second row of Figure 2.6 though, the χ^2 goodness of fit test p-values are all smaller than 0.0001.

2.5 Final comments

We have illustrated that when one truncates the sample space of the IG-Poisson model at zero it is possible to consider a parameter space that is larger than the one for the un-truncated IG-Poisson model. That extension has been proved useful in the analysis of the word frequency count data of Tirant lo Blanc, of Max Havelaar and of the turkish archeology text.

Finding the estimate of α to be negative, which is always coupled with the estimate of θ being larger, indicates that the frequency of ones is larger and the upper tail is heavier than allowed by the unextended model. Hence, the extended part of the model helps when the word frequency count data set is more over-dispersed than allowed by the unextended model.

The unextended truncated IG-Poisson model is very useful because it allows one to interpret the IG model mixing distribution as the word frequency distribution of the vocabulary of the author. This allows one to estimate the total number of words in the vocabulary of the author including both the observed and the unobserved ones, $v = v_n + v_{0:n}$, through the closest integer to the inverse of the expectation of that mixing distribution, $v = 1/E_\psi[\pi]$, as in Sichel (1986a, b), which helps one assess the richness of that vocabulary. We are also looking into the assessment of the lack of diversity of vocabulary through the variability of the IG model mixing distribution of the unextended truncated IG-Poisson model.

Unfortunately, we do not know if this could be done under the extended truncated model because when α is negative we do not know if the probability mass function obtained recursively through (2.9) can be posed as a Poisson mixture model or not. If it could be posed as that, one could also use an estimate of its model mixing distribution as an estimate of the word frequency distribution of vocabulary that could be used as a fingerprint of the style of the author in his texts. Besides, if that mixing distribution had a finite expectation one could use it to estimate the total number of words known by the author.

At this point it is natural to ask what happens when one further truncates the IG-Poisson model to exclude 0 and 1, $0, 1$ and 2 and so on. Given that the recursive relation in (2.9) that was used to extend the model to negative α's applies to any left truncated version of the IG-Poisson model, one expects that the technique presented here applies more generally even though one would have to determine the extent to which the parameter space can be expanded in each specific case. We did not follow this lead because we

do not have non-negative integer valued data requiring the truncation of the IG-Poisson model beyond zero.

One would expect that truncating the sample space of statistical models other than the beta-binomial, the negative binomial and the IG-Poisson model might allow one to expand their parameter space in ways analogous to the ones documented for these three models, and in that way help make them more flexible and useful. We hope that this chapter (Puig *et al*, 2008) will entice researchers to explore that possibility.

Capítol 3

The Sichel model and the mixing and truncation order

3.1 Introduction

The two-parameter inverse Gaussian-Poisson mixture model, denoted here as the IG-Poisson model, was introduced by Holla (1966) to model highly skewed non-negative integer data, and it has been widely used ever since in many different fields of application. This model and a three parameter generalization of it often recognized as the Sichel model, and denoted here as the GIG-Poisson model, is also used in the analysis of the frequency of word or species frequency data (Sichel, 1975, 1997), where given that one can not count unobserved words or species it is necessary to truncate these models at zero.

Even though these models are recommended just because they typically provide a good fit for this kind of data, what makes them truly useful in the analysis of word or species frequency count data is that they allow one to interpret the generalized inverse Gaussian mixing distribution as the distribution of the words or species frequencies of the vocabulary or population from which the observed count is assumed to be sampled from.

The truncated IG-Poisson model is better understood than the truncated GIG-Poisson model. The first goal is to explore the role of the parameter generalizing the IG-Poisson model and compare the fit of these two models on typical word frequency count data. The second goal is to investigate the usefulness of switching the order of the mixing and

the truncation stages.

The chapter is organized as follows. Section 3.2 describes word frequency count data. Section 3.3.1 motivates the truncated IG- and GIG-Poisson models, it explains how one can use their model mixing density estimates as fingerprints of the vocabulary of the author and it explores the role played by their parameters in the determination of their probability mass functions. Section 3.3.2 considers the IG- and GIG-Truncated Poisson mixture models. An heuristic argument makes the mixing distributions of these alternative models into the distribution of the word frequencies of an hypothetical vocabulary formed by the words actually observed; the mixing distribution estimates under these alternative models depend on text size and hence are not useful fingerprints of style of the author in his texts.

Section 3.4 compares the truncated GIG-Poisson model with the truncated IG-Poisson model, by fitting them to the word frequency counts of texts by Macaulay, Carroll, Wells, Doyle and Dekker using the maximum likelihood criteria. The three parameter model provides excellent fits for very long texts, where the two parameter model fails because the data is more overdispersed than is allowed by that model. Section 3.4 also illustrates the use of the model mixing density estimate as an estimate of the density of the word frequency of the vocabulary of the author that can help assess its diversity.

Section 3.5 shows that the IG-Truncated Poisson model performs better than the truncated IG-Poisson model because it allows for larger degrees of overdispersion. On the other hand the GIG-Truncated Poisson mixture model fares worse than the truncated GIG-Poisson model for all the sample texts on which we have tried these models.

Even though the focus is on the analysis of word frequency count data, it all trivially extends to the analysis of species frequency count data, with the understanding that in ecology the truncated Poisson mixture models that best fit the data might be different.

3.2 Description of the data

Some of the most useful tools in the analysis of literary style and in authorship attribution studies rely on the analysis of word frequency counts in sample texts with the goal of learning about the vocabulary of the author of those texts (see, e.g., Holmes, 1985, Baayen, 2001).

To characterize the style of an author through his vocabulary the basic assumption is that the author has available a list of all the words that he knows, and that the i-th

word in that list is characterized through the proportion of times that that word would be found in a text of infinite length by that author, which is denoted by π_i. The set of probabilities π_j when j ranges over all the v words known by an author, $(\pi_1, \ldots, \pi_v)$, with $\pi_i > 0$ for $i = 1, \ldots, v$ and $\sum_j^v \pi_j = 1$, constitute the probability function of his vocabulary. For mathematical convenience we will treat the word frequencies, π_j, as a continuous variable with a density function $\psi(\pi)$.

If one denotes the total number of words (tokens) in a given text by n, the number of occurrences of the i-th word by n_i, and the proportion of occurrences of that word in that text by $\hat{\pi}_i = n_i/n$, the expected value of $\hat{\pi}_i$ is π_i. Let v_n denote the number of different words (types) in a text of size n, and let $v_{r:n}$ denote the number of different words appearing exactly r times in it. The proportion of different words appearing exactly r times in a text of size n will be denoted by $\hat{p}_{r:n} = v_{r:n}/v_n$ and its expectation, which depends on n, will be denoted by $p_{r:n}$.

If one restricts attention to a fixed and given subset of words, as in Riba & Ginebra (2005) or Giron et al. (2005), one knows the total number of different words involved. Most often though one doesn't do that and as a consequence one neither knows the size v of an author's vocabulary, nor the number of words from that vocabulary that are not observed, $v_{0:n}$.

Table 3.1 presents word frequency counts that will be used later on. The first row of that table for example indicates that the Turkish archeology text has a total of $n = 6939$ words out of which there are $v_n = 3302$ different words; in that text 2326 words appear once, 477 words appear twice and so on; the most frequent word in that text appears a total of 222 times.

3.3 Description of the models

Most of the words appear only very few times and only a few words are repeated many times, and hence the distribution of the observed frequencies of word frequencies, $(v_{1:n}, v_{2:n}, \ldots, v_{n:n})$ is reverse J-shaped with an extraordinarily long upper tail. A large number of alternative models have been proposed for this type of data, sometimes under the label of large number of rare events models (see, e.g., Baayen, 2001).

Yule (1944) and Good (1953) conjecture that this strong skewness should be modelled through Poisson mixtures. Here, word frequency count distributions are first related to $\psi(\pi_i)$ and to n through the zero truncated GIG-Poisson model, and second they are

	$v_{1:n}$	$v_{2:n}$	$v_{3:n}$	$v_{4:n}$	$v_{5:n}$	$v_{6:n}$	$v_{7:n}$	$v_{8:n}$	$v_{9:n}$	$v_{10:n}$	$v_{11:n}$	...	v_n	n
Turkish A.	2326	477	178	107	53	33	22	26	7	7	12	...	3302	6939
E. Bacon	990	367	173	112	72	47	41	31	34	17	24	...	2048	8049
Alice W.	1176	402	233	154	99	57	65	52	32	36	23	...	2651	26505
Through L.	1491	460	259	148	113	78	61	47	28	26	26	...	3085	28767
Hound B.	2836	889	449	280	208	137	116	92	86	52	48	...	5741	59241
War of W.	3613	1138	567	340	250	177	135	93	72	67	44	...	7112	59938
Max H.	6004	1731	819	491	368	258	168	137	123	108	80	...	11161	99767

Taula 3.1: Part of the word frequency count sets of the nouns in Macaulay's essay on Bacon, considered in Sichel (1975), and of all the words in a Turkish archeology text, in *Alice in Wonderland* and in *Through the Looking Glass* by Lewis Carroll, in *The Hound of the Baskervilles* by Doyle, in *The War of the Worlds* by Wells, and in *Max Havelaar* by Dekker, which are all considered in Baayen (2001). Note that $v_n = \sum_{r=1}^{n} v_{r:n}$ and $n = \sum_{r=1}^{n} r v_{r:n}$.

modelled through what results from switching the order of the mixing an the truncation stages.

3.3.1 The zero truncated IG- and GIG-Poisson models

As an approximation, texts written by an author are treated as if they were random samples drawn from his vocabulary. If the specific i-th word has a probability π_i of being used each time that an author writes a word, the number of times that this word appears in a text by that author with a total of n words would be distributed as a binomial(n, π_i). Hence, if the density of the π_i's was $\psi(\pi)$ the probability that a word from the vocabulary of that author appears exactly r times in a text of size n, $p_{r:n}$, can be modelled through a $\psi(\pi)$-binomial mixture model.

Given that one can not count the words that the author knows but are not observed in the text, it is necessary to consider the zero truncated version of this mixture model. Usually n will be large and all the π_i will be very small, and hence one can approximate $p_{r:n}$ through a zero truncated $\psi(\pi)$-Poisson mixture model, with

$$p_{r:n}^{tpm}(\psi) = \frac{1}{1 - \int_{R^+} e^{-n\pi}\psi(\pi)d\pi} \int_{R^+} \frac{(n\pi)^r e^{-n\pi}}{r!}\psi(\pi)d\pi, \qquad \text{for} \quad r = 1, 2, \ldots. \tag{3.1}$$

Note that this argument provides a simple mechanistic description of the word frequency count generation process as a truncated Poisson mixture that entitles one to interpret the model mixing density $\psi(\pi)$ as the density of the word frequency of the vocabulary. That

interpretation, which is extremely helpful when characterizing the style of the author, is lacking in the empirically motivated models often considered for this kind of data.

Following a recommendation in page 249 of Good (1953), Sichel (1975, 1986a) models the mixing vocabulary distribution, $\psi(\pi)$, through a generalized inverse gaussian distribution, denoted here by GIG(b, c, g), which is defined on R^+ and has a density function

$$\psi(\pi|b, c, g) = \frac{2^{g-1}}{(bc)^g K_g(b)} \pi^{g-1} e^{-\frac{\pi}{c} - \frac{b^2 c}{4\pi}}, \tag{3.2}$$

where b is in $(0, \infty)$, c is in $(0, \infty)$, g is in $(-\infty, \infty)$, and where $K_g(\cdot)$ is the modified Bessel function of the third kind of order g, (see, e.g., Jorgensen, 1982). Even though the support of (3.2) is $(0, \infty)$, under the values of (b, c, g) that one considers in practice $\psi(\pi|b, c, g)$ is negligible for $\pi > .1$. In practice often one uses the two-parameter special case obtained when $g = -1/2$, recognized as the inverse gaussian distribution, which is denoted here as the IG(b, c) and is described in Chikara & Folks (1989) and in Seshadri (1993, 99).

By replacing (3.2) in (3.1) and solving the integral one obtains the probability function of the *zero truncated GIG-Poisson* mixture model,

$$p_{r:n}^{tsich3}(b, c, g) = \frac{1}{(1 + cn)^{g/2} K_g(b) - K_g(b\sqrt{1 + cn})} \frac{(\frac{1}{2} \frac{bcn}{\sqrt{1+cn}})^r}{r!} K_{r+g}(b\sqrt{1 + cn}), \tag{3.3}$$

for $r = 1, \ldots$. The support of these distributions is unbounded but they die out very fast with increasing r and there is no conflict with r being bounded above by n. When $g = -1/2$ the expression in (3.3) becomes the probability function of the *zero truncated IG-Poisson* model, denoted here as $p_{r:n}^{tsich2}(b, c)$.

To investigate the role of (b, c, g) in $p_{r:n}^{tsich3}(b, c, g)$, Figure 3.1 presents its contour plots for $r = 1, 2, 3, 4, 5$ when $n = 25000$ for the range of values of (b, c, g) most useful in stylometric practice. Note that if one fixes any two of the three parameters, the probability of a word appearing just once, $p_{1:n}^{tsich3}(b, c, g)$, is a decreasing function of the third parameter. Observe also that the effect of b, c and g on the distribution function is very similar across $r = 2, 3, 4, 5$. It can be checked that for values of r much larger than 5 and in the range of values of (b, c, g) found in practice, this probability mass function is rather insensitive to the value of c and it is an increasing function of b, c and g, just opposite of what happens for $r = 1$.

The left hand side panels of Figure 3.1 correspond to the contour plots of the probability function of the truncated IG-Poisson special case. Observe that $p_{1:n}^{tsich2}(b, c)$ is less than .5 for all (b, c) except for the ones very close to $(0, 0)$, which represents a very strong

practical limitation. That can be overcome by introducing the third parameter g, by extending the truncated IG-Poisson model as in Puig *et al.*(2008), or by resorting to the model presented next.

3.3.2 The IG- and GIG-Truncated Poisson models

As an alternative to the truncation of Poisson mixture models, in (3.1), one can model frequency count data without zeros by switching the order of the mixing and the truncation stages. That leads one to modelling the probability of a word being repeated exactly r times in a text of size n through a mixture of the truncated Poisson model, with

$$p_{r:n}^{mtp}(\psi^{'}) = \int_{R^{+}} \frac{(n\pi^{'})^{r} e^{-n\pi^{'}}}{(1-e^{-n\pi^{'}})r!}\psi^{'}(\pi^{'})d\pi^{'}, \qquad \text{for} \quad r = 1, \ldots, n. \tag{3.4}$$

Any model that results from truncating a Poisson mixture, as in (3.1), can be posed as a model that results from mixing the truncated Poisson, as in (3.4), but not the other way around. That is in contrast with what happens under finite mixture models, because every model that results from truncating a finite mixture of Poisson distributions can be interpreted as finite mixture of truncated Poisson models and viceversa, (see Böhning & Kuhnert, 2006).

An heuristic interpretation of (3.4) is as follows. Assume that one restricts consideration to the observed part of the vocabulary of an author, formed by the v_n different words in the text of size n, with probability distribution $(\pi_1^{'}, \ldots, \pi_{v_n}^{'})$ where $\pi_i^{'}$ is the $\pi_i / \sum_{j=1}^{v_n} \pi_j$ that results from truncating the unobserved words out from the distribution of the vocabulary of the author. This $\pi_i^{'}$ would be the probability that the i-th word is used each time a word is written in the text. Given that it is known that this i-th word is used at least once, the total number of times that it appears is distributed as a zero truncated binomial$(n, \pi_i^{'})$.

If one treats $\pi_i^{'}$, as a continuous random variable with density $\psi^{'}(\pi^{'})$, the probability that a word from the observed part of the vocabulary appears exactly r times in that text, $p_{r:n}$, can be modelled through a $\psi^{'}(\pi^{'})$-Truncated binomial mixture model. When n is large and all the $\pi_i^{'}$ are small, $p_{r:n}$ can be approximated through a $\psi^{'}(\pi^{'})$-Truncated Poisson model, as in (3.4).

According to this interpretation $\psi^{'}(\pi^{'})$ only represents the v_n words observed in the text and not all the v words in the vocabulary of the author, and hence it can not be interpreted as the density of the frequencies of the whole vocabulary of the author as the $\psi(\pi)$ in (3.1). As a consequence, $\psi^{'}(\pi^{'})$ depends on n; given that as n grows v_n will

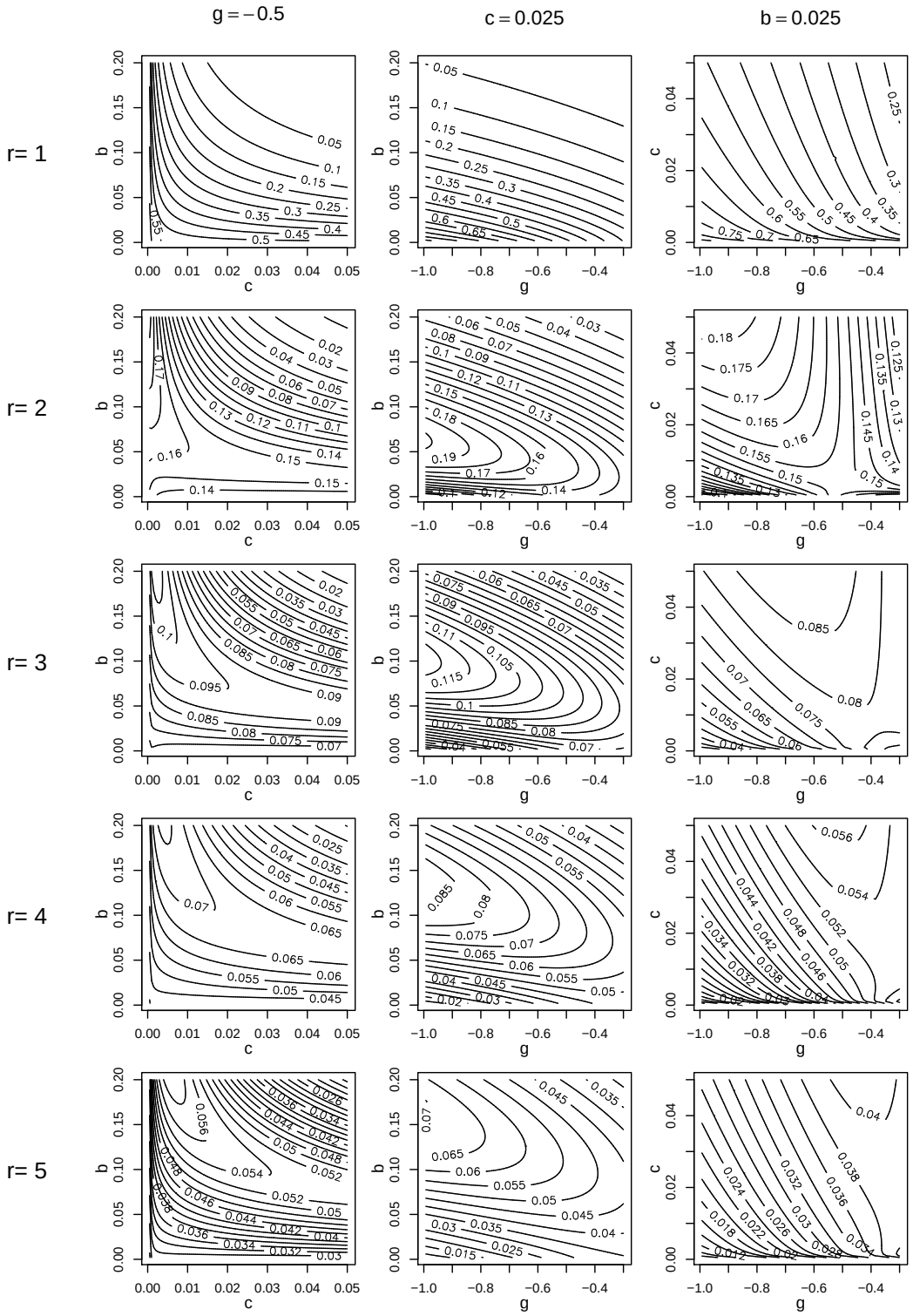

Figura 3.1: Contour plots of the probability function of the truncated GIG-Poisson model, $p_{r:n}^{tsich3}(b, c, g)$, when $r = 1, \ldots, 5$ and $n = 25000$. The left hand side panels correspond to the contour plots for the truncated IG-Poisson model, $p_{r:n}^{tsich2}(b, c)$.

tend to v and $\pi_i^{'}$ will tend to π_i, it follows that $\psi^{'}(\pi^{'})$ will tend to the $\psi(\pi)$ in (3.1) as the text length grows.

The model obtained from (3.4) when $\psi^{'}(\pi^{'})$ is a generalized inverse gaussian distribution will be denoted as the GIG-Truncated Poisson mixture model and the corresponding probability function as $p_{r:n}^{gigtp}(b,c,g)$. The version obtained when $g=-1/2$ is the IG-Truncated Poisson model, and its probability function is $p_{r:n}^{igtp}(b,c)$. The truncated GIG-Poisson mixture model in (3.3) can also be posed as in (3.4), but with a different $\psi^{'}(\pi^{'})$ distribution.

Figure 3.2 presents the contour plots of $p_{r:n}^{gigtp}(b,c,g)$ when $r=1,2,3,4,5$ for the same range of values of (b,c,g) and the same n in Figure 3.1. The role played by b, c and g is almost identical to the one played for the $p_{r:n}^{tsich3}(b,c,g)$. In particular, in the range of (b,c,g) considered the probability of a word appearing just once under the GIG-Truncated Poisson model is also a decreasing function of b, c and g while the probability of a word appearing r times with r very large is also an increasing function of each of these three parameters.

The left hand side panels of Figure 3.2 correspond to the IG-Truncated Poisson model. The main difference with the truncated IG-Poisson model in Figure 3.1 is that the model considered here can attain much higher values for $p_{1:n}^{igtp}(b,c)$ and for $p_{r:n}^{igtp}(b,c)$ with large r in that same range of values of (b,c). As a consequence, this IG-Truncated Poisson model performs better when $\hat{p}_{1:n}$ and $\hat{p}_{r:n}$ with large r are large and the word frequency count data set is too overdispersed to be properly modelled through the truncated IG-Poisson model. That helps in those instances considered in Puig et al (2008) where the maximum likelihood estimate for the truncated IG-Poisson model falls in the boundary of the parameter space.

3.4 Truncated IG- and GIG-Poisson models in practice

Sichel (1975, 1982, 1986b) proposes various estimation methods for the truncated IG-Poisson model. Pollatschek & Radday (1981), Holmes (1992), Holmes & Forsyth (1995), Baayen (2001), Riba & Ginebra (2006) and Puig *et al* (2008) fit this two parameter model to word frequency count data and find that it fits reasonably well when the text has less than 5.000 words.

The full three parameter truncated GIG-Poisson model has been rarely used in practice.

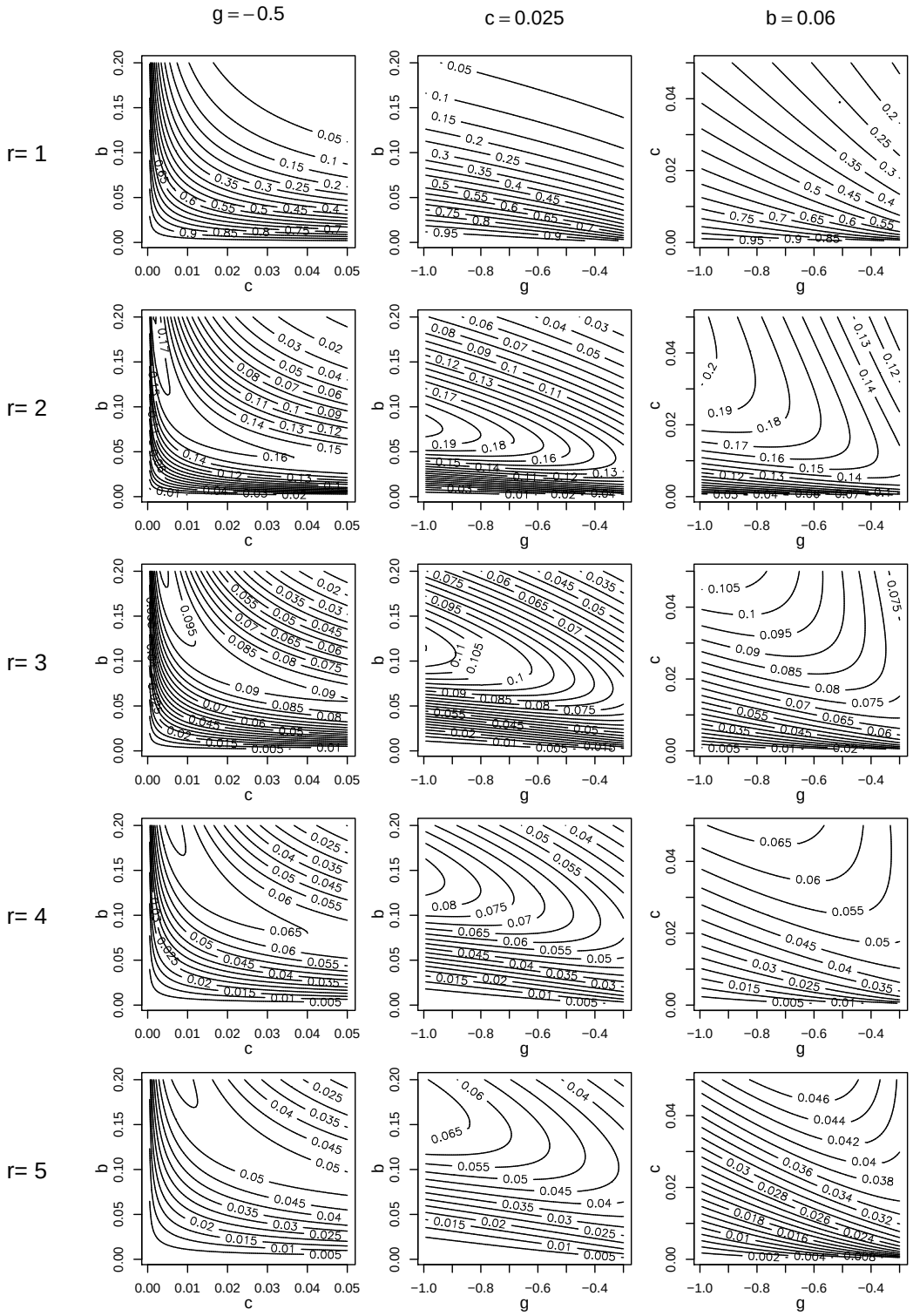

Figura 3.2: Contour plots of the probability function of the GIG-Truncated Poisson model, $p_{r:n}^{gigtp}(b, c, g)$, when $r = 1, \ldots, 5$ and $n = 25000$. The left hand side panels correspond to the contour plots for the IG-Truncated Poisson model, $p_{r:n}^{igtp}(b, c)$.

Sichel (1975, 1986a, 1997) fits it to word and species frequency count data and Sichel (1985, 1992a, b) uses it for bibliometric data where one observes the frequency of zeros but needs to model $p_{0:n}$ apart from $p_{r:n}$ for $r > 0$. In all these examples n is relatively small and one estimates b and c for a given value of g by equating $\hat{p}_{1:n}$ to $p_{1:n}^{tsich3}(b, c, g)$ and the sample mean to the population mean and then varying g and re-estimating (b, c) until the χ^2 goodness of fit statistic is minimized. Instead, here all the estimation will be done through the maximization of the likelihood function by direct search of the likelihood surface.

3.4.1 Comparison of the truncated IG- and GIG-Poisson models

The joint distribution of $(v_{1:n}, v_{2:n}, \ldots, v_{n:n})$ is very complicated but if one assumes that word frequencies are independent and identically distributed as a zero truncated GIG-Poisson distribution, the likelihood function can be assumed to be such that:

$$L_{(v_{1:n},\ldots,v_{n:n})}^{tsich3}(b, c, g) \propto \Pi_r (p_{r:n}^{tsich3}(b, c, g))^{v_{r:n}}. \tag{3.5}$$

Both this model as well as the truncated IG-Poisson model were fitted to the data in Table 3.1 by maximizing this likelihood function. To check the validity of the Poisson approximation in (3.1), the maximum likelihood fits of (3.3) were compared with the ones of the zero truncated GIG-binomial model and it was found that both fits were identical.

Figure 3.3 presents the contour plots of the log-likelihood functions near their maximum for four of the seven texts in Table 3.1. Note that these likelihood surfaces are very well behaved. As a consequence the computation of the maximum likelihood estimates through direct search of that surface does not require the use of the approximation methods in Atkinson & Yeh (1982) and the re-parametrizations proposed in Stein *et al.* (1987) for the maximum likelihood estimation of the untruncated Sichel model.

Table 3.2 presents the maximum likelihood estimate for (b, c, g), the maximum value of the log-likelihood function, the value taken by the classical Pearson goodness of fit test statistic,

$$X^2(\hat{b}, \hat{c}, \hat{g}) = \sum_r \left(\frac{v_{r:n} - v_n p_{r:n}(\hat{b}, \hat{c}, \hat{g})}{\sqrt{v_n p_{r:n}(\hat{b}, \hat{c}, \hat{g})}} \right)^2 = \sum_r e_{r:n}(\hat{b}, \hat{c}, \hat{g})^2. \tag{3.6}$$

and the corresponding p-values, all for the truncated two and three parameter Sichel models. To evaluate $X^2(\hat{b}, \hat{c}, \hat{g})$, models were fitted under the initial categorization but then the categories were aggregated the least so that the expected count under each

model was at least 5 for each category. As a consequence, models are compared across different aggregate categories.

With the exception of the word frequency count for the Essays on Bacon, under both a likelihood ratio test as well as under a χ^2 test one clearly rejects the two parameter truncated IG-Poisson model, with $g = -1/2$, in favor of the three parameter truncated GIG-Poisson model.

In fact, it is remarkable that the overall goodness of fit test based on $X^2(\hat{b}, \hat{c}, \hat{g})$ indicates that the three parameter zero truncated Sichel model provides a very good fit for all the word frequency count data sets considered, including the one for the text with almost 100.000 words.

The improvement of the three parameter model relative to the two parameter version of it is specially noticeable for the Turkish archeology text and for Max Havelaar, which are the only texts considered that are not written in english. The maximum likelihood estimates of (b, c) of the truncated IG-Poisson model for these two texts are $(0., 0.00133)$ and $(0., 0.00285)$ respectively, on the boundary of the parameter space. In these situations, allowing for negative values of b, as in Puig et al (2008), produces a better fit but that extended model is not an IG-Poisson mixture anymore, and hence it does not allow one to estimate $\psi(\pi)$.

To help compare the fit under these models, Figure 3.4 presents their Pearson residuals, $e_{r:n}(\hat{b}, \hat{c}, \hat{g})$. Except for the essays on Bacon, under the truncated IG-Poisson model the Pearson residuals for the upper tail probability and for a few small r are too small, indicating that for them the observed $v_{r:n}$ count is much smaller than the expected $v_n p_{r:n}(\hat{b}, \hat{c})$ count. To compensate for that the observed counts for the intermediate categories tend to be larger than predicted by the model. Hence this two parameter model fails because it can not capture the degree of overdispersion in typical word frequency count data of texts with more than 5000 words.

What is most remarkable from Figure 3.4 though is that the truncated GIG-Poisson model provides an excellent fit for all the data considered, in accordance with the small values of the overall goodness of fit test statistic and the corresponding large p-values in Table 3.2.

		Tr.IG-Pois (Tr.Sichel 2)	Tr.GIG-Pois (Tr.Sichel 3)	IG-TrPois	GIG-TrPois
Turkish A.	$\hat{b}$	0.	0.05139	0.14584	0.12500
$n = 6939$	$\hat{c}$	0.00133	0.01654	0.00275	0.00950
$v_n = 3302$	$\hat{g}$	$(-.5)$	-1.09165	$(-.5)$	-0.91793
	max loglik	-3882.655	-3813.247	-3831.615	-3818.615
	$X^2, (df)$	88.6799, (16)	24.8444, (17)	35.1022, (19)	28.6960, (18)
	p-value	0.00000	0.09825	0.01358	0.05222
E. Bacon	$\hat{b}$	0.08362	0.10130	0.22282	0.22276
$n = 8049$	$\hat{c}$	0.00369	0.00443	0.00385	0.00381
$v_n = 2048$	$\hat{g}$	$(-.5)$	-0.58011	$(-.5)$	-0.49542
	max loglik	-4008.887	-4008.339	-4008.861	-4008.860
	$X^2, (df)$	17.6943, (27)	19.6619, (26)	18.9094, (27)	18.8292, (26)
	p-value	0.91259	0.8073	0.87328	0.84356
Alice W.	$\hat{b}$	0.02293	0.03139	0.07343	0.06678
$n = 26505$	$\hat{c}$	0.00954	0.02175	0.00982	0.01751
$v_n = 2651$	$\hat{g}$	$(-.5)$	-0.65777	$(-.5)$	-0.64416
	max loglik	-6281.116	-6270.200	-6283.069	-6273.486
	$X^2, (df)$	85.8517, (56)	61.3366, (50)	90.5613, (56)	67.9309, (51)
	p-value	0.0063	0.13065	0.00236	0.05653
Through L.	$\hat{b}$	0.01191	0.02300	0.06354	0.05451
$n = 28767$	$\hat{c}$	0.00891	0.02136	0.00967	0.01951
$v_n = 3085$	$\hat{g}$	$(-.5)$	-0.65491	$(-.5)$	-0.64141
	max loglik	-6887.620	-6873.728	-6887.447	-6879.461
	$X^2, (df)$	82.5601, (58)	62.4665, (52)	88.7565, (58)	73.3083, (52)
	p-value	0.01875	0.15175	0.00578	0.02738
Hound B.	$\hat{b}$	0.00684	0.01638	0.05147	0.03455
$n = 59241$	$\hat{c}$	0.00573	0.02907	0.00644	0.02680
$v_n = 5741$	$\hat{g}$	$(-.5)$	-0.73097	$(-.5)$	-0.72540
	max loglik	-12445.727	-12380.668	-12437.066	-12392.897
	$X^2, (df)$	181.2624, (86)	90.0331, (71)	175.6582, (85)	108.0107, (71)
	p-value	0.00000	0.06323	0.00000	0.00306
War W.	$\hat{b}$	0.00613	0.01792	0.05980	0.03575
$n = 59938$	$\hat{c}$	0.00381	0.02887	0.00441	0.02618
$v_n = 7112$	$\hat{g}$	$(-.5)$	-0.79093	$(-.5)$	-0.78412
	max loglik	-14654.543	-14543.865	-14631.829	-14557.171
	$X^2, (df)$	216.1065, (87)	85.8329, (71)	188.8810, (87)	109.3586, (72)
	p-value	0.00000	0.11078	0.00000	0.00299
Max H.	$\hat{b}$	0.	0.01166	0.04636	0.02749
$n = 99767$	$\hat{c}$	0.00285	0.02651	0.00363	0.02177
$v_n = 11161$	$\hat{g}$	$(-.5)$	-0.77845	$(-.5)$	-0.75873
	max loglik	-22067.088	-21873.281	-22008.324	-21897.789
	$X^2, (df)$	399.2618, (113)	97.8287, (90)	324.9427, (112)	144.8910, (92)
	p-value	0.00000	0.26869	0.00000	0.00036

Taula 3.2: Maximum likelihood estimate of (b, c, g), maximum value of the log-likelihood function, value of the $X^2(\hat{b}, \hat{c}, \hat{g})$ goodness of fit test statistic and its corresponding p-value, all for the truncated IG-Poisson and GIG-Poisson models and for the IG-Truncated Poisson and GIG-Truncated Poisson models. Between brackets, the degrees of freedom associated to X^2.

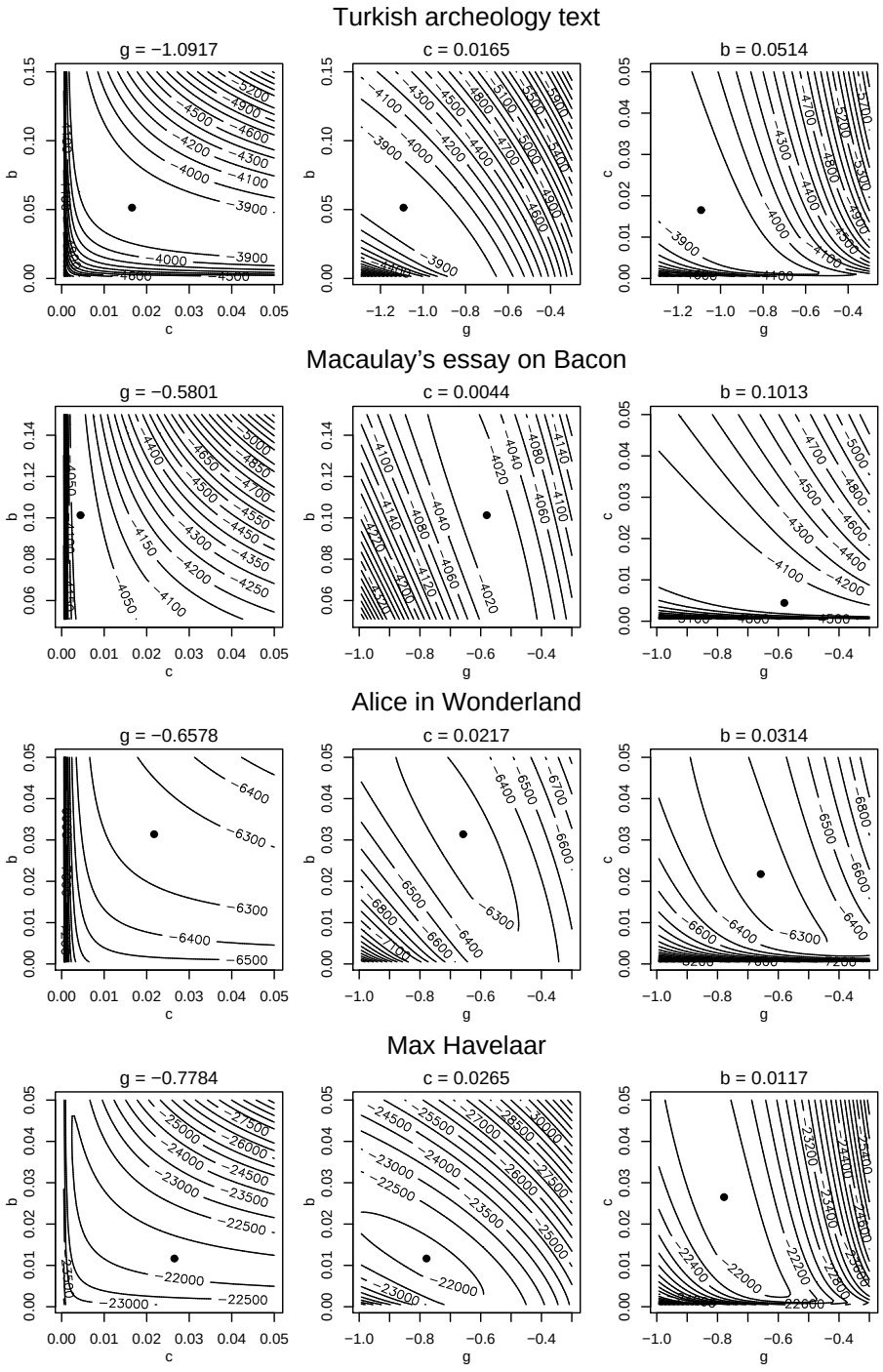

Figura 3.3: Contour plots for the log-likelihood function of the zero truncated Sichel (GIG-Poisson(b, c, g)) model, in (3.3), near its maximum for the word frequency count data sets of four of the seven texts in Table 3.1.

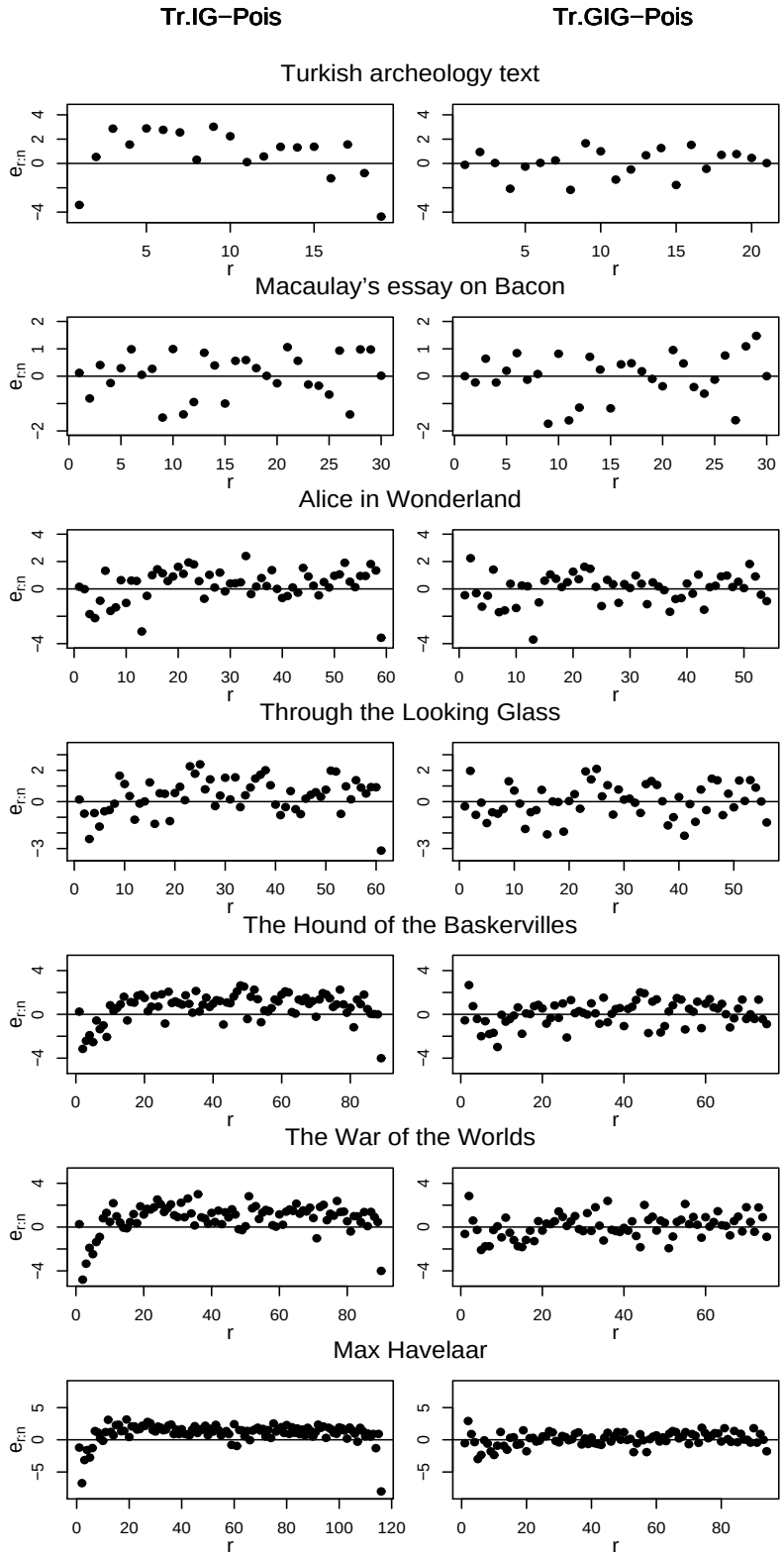

Figura 3.4: Pearson residuals, $e_{r:n}(\hat{b}, \hat{c}, \hat{g})$, under the truncated IG-Poisson model, on the left hand side, and under the truncated GIG-Poisson model, on the right hand side.

3.4.2 Estimation of the density and diversity of vocabulary

The fact that the truncated GIG-Poisson model provides excellent fits for the word frequency count of very long texts entitles one to use an estimate of the density of GIG(b, c, g) as an estimate of the density of the word frequencies of the vocabulary of the author, $\hat{\psi}(\pi)$. The left hand side panels in Figure 3.5 present the maximum likelihood estimates of these mixing densities for the data in Table 3.1, which serve as fingerprints of the style of the author.

The richness of vocabulary is usually assessed through estimates of the total number of words, v. Sichel (1986a) proposes estimating it through the closest integer to

$$v(\hat{\psi}) = \frac{1}{E_{\hat{\psi}}[\pi]}. \tag{3.7}$$

A useful measure of the diversity in a population with probability function $(\pi_1, \ldots, \pi_v)$ is the Gini-Simpson index, $D_1(\pi_1, \ldots, \pi_v) = 1 - \sum_{i=1}^{v} \pi_i^2$, which is the probability that two words picked at random from a text of infinite length would be different. Ginebra & Puig (2008) propose estimating $D_1(\pi_1, \ldots, \pi_v)$ through the expected value of this index,

$$D_1(\hat{\psi}) = E_{\hat{\psi}}[D_1(\pi_1, \ldots, \pi_v)] = 1 - \frac{E_{\hat{\psi}}[\pi^2]}{E_{\hat{\psi}}[\pi]}, \tag{3.8}$$

obtained assuming that the π_j's are identically distributed as $\hat{\psi}(\pi)$ =GIG$(\hat{b}, \hat{c}, \hat{g})$ and that $v = 1/E_{\hat{\psi}}[\pi]$. Another measure of the diversity in a population is its entropy, $D_2(\pi_1, \ldots, \pi_v) = -\sum_i \pi_i \log \pi_i$, which can be analogously estimated through,

$$D_2(\hat{\psi}) = E_{\hat{\psi}}[D_2(\pi_1, \ldots, \pi_v)] = -\frac{E_{\hat{\psi}}[\pi \log \pi]}{E_{\hat{\psi}}[\pi]}. \tag{3.9}$$

Table 3.3 presents estimates of these measures for the vocabularies of the texts in Table 3.1. According to them, the text from a richer vocabulary is Max Havelaar, followed by the Turkish archeology text and The War of the Worlds, while the smallest vocabulary is the one behind Maccaulay's essay on Bacon, which is the only case where one is only counting names. Both diversity measures agree in that the most diverse vocabularies are the ones behind the Turkish archeology text, the essay on Bacon, Max Havelaar and The War of the Worlds, in this order.

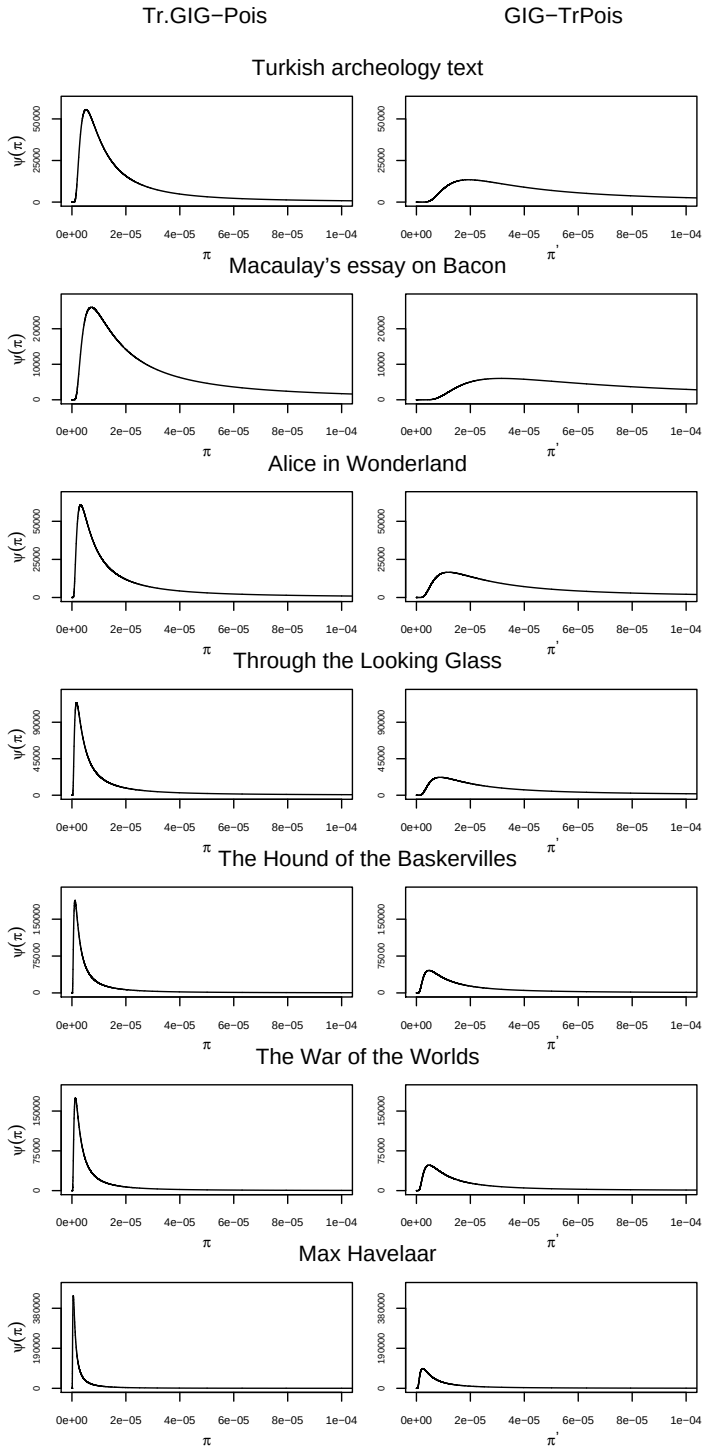

Figura 3.5: Maximum likelihood estimate of the mixing density functions, $\psi(\pi)$, under the truncated GIG-Poisson model (on the left), next to the one of $\psi'(\pi')$ under the GIG-Truncated Poisson model (on the right), for the texts in Table 3.1. The ones on the left estimate the density of the word frequencies of the theoretical vocabulary of the author; the smaller its expectation the larger v and the smaller its variability the more diverse the vocabulary. The ones on the right estimate the word frequencies of the vocabulary observed in the text.

	n	v_n	$v(\hat{\psi})$	$D_1(\hat{\psi})$	$D_2(\hat{\psi})$
Turkish A.	6939	3302	19364	0,998018	8,17899
E. Bacon	8049	2048	5651	0,997855	7,12427
Alice W.	26505	2651	6199	0,991834	6,20007
Through L.	28767	3085	9228	0,992072	6,29513
Hound B.	59241	5741	16052	0,991269	6,49482
War W.	59938	7112	19564	0,992655	6,83006
Max H.	99767	11161	37753	0,993224	7,01936

Taula 3.3: Estimates of the size and of two measures of the diversity of the vocabulary of the authors of the texts in Table 3.1, based on the maximum likelihood estimates of the mixing distribution of the truncated GIG-Poisson model.

3.5 IG- and GIG-Truncated Poisson models in practice

3.5.1 Comparison with truncated IG- and GIG-Poisson models

Figure 3.6 presents the contour plots of the logarithm of:

$$L^{gigtp}_{(v_{1:n},\ldots,v_{n:n})}(b,c,g) \propto \Pi_r(p^{gigtp}_{r:n}(b,c,g))^{v_{r:n}}, \tag{3.10}$$

near its maximum for four of the seven texts under consideration. These likelihood surfaces are not as well behaved as the ones for the truncated GIG-Poisson model, in Figure 3.3, but it is still computationally easy to locate their maximum by direct search of the surfaces.

Table 3.2 presents the maximum likelihood estimate for (b,c,g), the maximum value of the log-likelihood, the value of the $X^2_{pr}(\hat{b},\hat{c},\hat{g})$ goodness of fit test statistic and the corresponding p-value for the IG and GIG-Truncated Poisson models.

Table 3.2 indicates that the two parameter IG-Truncated Poisson model fits the data for the Turkish archeology texts and for the three longest texts better than the truncated IG-Poisson model. By comparing the Pearson residuals of the truncated IG-Poisson model, in Figure 3.3, with the ones of the IG-Truncated Poisson model, in Figure 3.7, it is clear that this later model fits better the upper tail and the small r categories and therefore the model considered here is better prepared for the large overdispersion present in typical word frequency count data.

Table 3.2 also indicates that the three parameter GIG-Truncated Poisson model fits

are worse than the truncated GIG-Poisson model fits for all the seven texts considered. The Pearson residuals indicate that the reason for the inferior performance of the GIG-Truncated Poisson model is its inability to properly model $p_{r:n}$ for small r.

3.5.2 Estimation of the density and diversity of the observed vocabulary

Here we reproduce the analysis in Section 4.2 on the GIG-Truncated Poisson model. The right hand side panels of Figure 3.3.5 present the maximum likelihood estimate of the $\psi'(\pi') = GIG(b, c, g)$ model mixing densities of this model for the seven texts in Table 3.1. The $\hat{\psi}'(\pi')$ are to the right of the estimates of the mixing density for the Sichel model in that same figure, which is to be expected because the expectation of ψ' is an estimate of the inverse of the size of the observed part of the vocabulary, while the expectation of ψ is an estimate of the inverse of the size of the whole vocabulary of the author.

Except for the Turkish archeology text the estimated sizes $v(\psi')$, in Table 3.4, are close to the observed v_n values in Table 3.1, and the ranking of texts in terms of $v(\hat{\psi}')$ and the one in terms of v_n is the same. Table 3.4 also presents the maximum likelihood estimate of the expected value of the Gini-Simpson diversity index, $D_1(\hat{\psi}')$, and of the entropy, $D_2(\hat{\psi}')$.

Note that one can also characterize the observed part of the vocabulary directly through the observed frequencies $(\hat{\pi}_1, \ldots, \hat{\pi}_{v_n})$ instead of doing it indirectly through the mixing distribution estimate of (3.4). In fact, it is simpler and more natural to assess the size of the observed vocabulary through v_n and its diversity through $D_1(\hat{\pi}) = 1 - \sum_{i=1}^{v} \hat{\pi}_i^2$ or through $D_2(\hat{\pi}) = -\sum_i \hat{\pi}_i \log \hat{\pi}_i$, instead of doing it through $v(\hat{\psi}')$, $D_1(\hat{\psi}')$ and $D_2(\hat{\psi}')$. Observe that the ranking of texts in terms of $D_2(\hat{\psi}')$ and of $D_2(\hat{\pi})$, in Table 3.4, are almost identical, while the ones in terms of $D_1(\hat{\psi}')$ and of $D_1(\hat{\pi})$ are similar (the discrepancies being due in part to the lack of fit of the GIG-Truncated Poisson model).

In practice neither v_n nor $v(\hat{\psi}')$ are good estimates of the size of the vocabulary of the author, v, and neither $D_i(\hat{\pi})$ nor $D_i(\hat{\psi}')$ are good estimates of the corresponding measure of the diversity of the vocabulary of the author, $D_i(\pi_1, \ldots, \pi_v)$, because they are all biased estimators with a bias that can be large and depends on n, v and $(\pi_1, \ldots, \pi_v)$. In order to estimate the density, size and diversity of the vocabulary of an author through the word frequency count in his texts there is no alternative to fitting models that result from truncating Poisson mixture models, in Section 3.3.1, and allow one to estimate the word frequency distribution of that vocabulary.

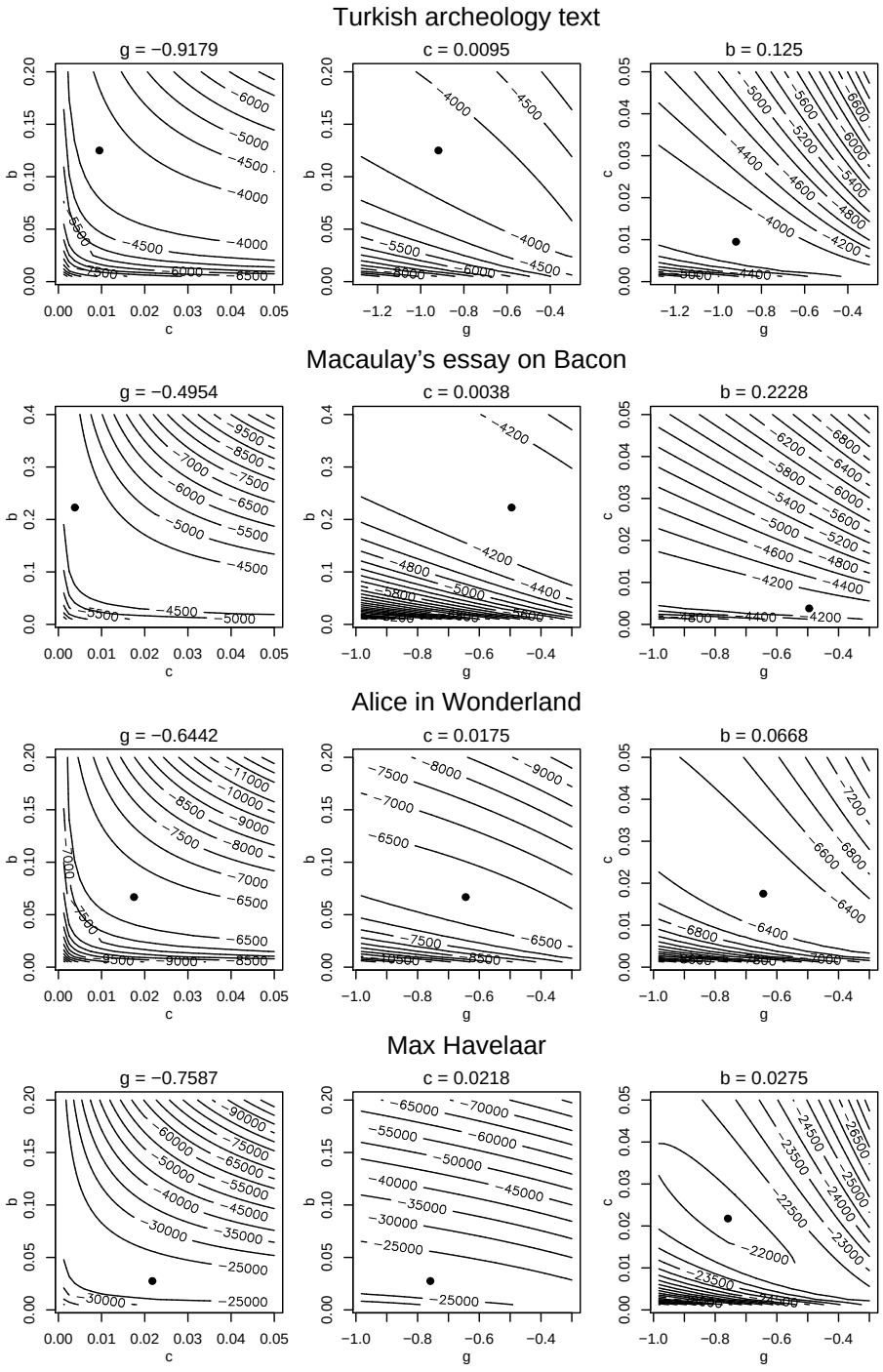

Figura 3.6: Contour plots for the log-likelihood function of the GIG-Truncated Poisson model near its maximum for the word frequency count data sets of four of the texts in Table 3.1.

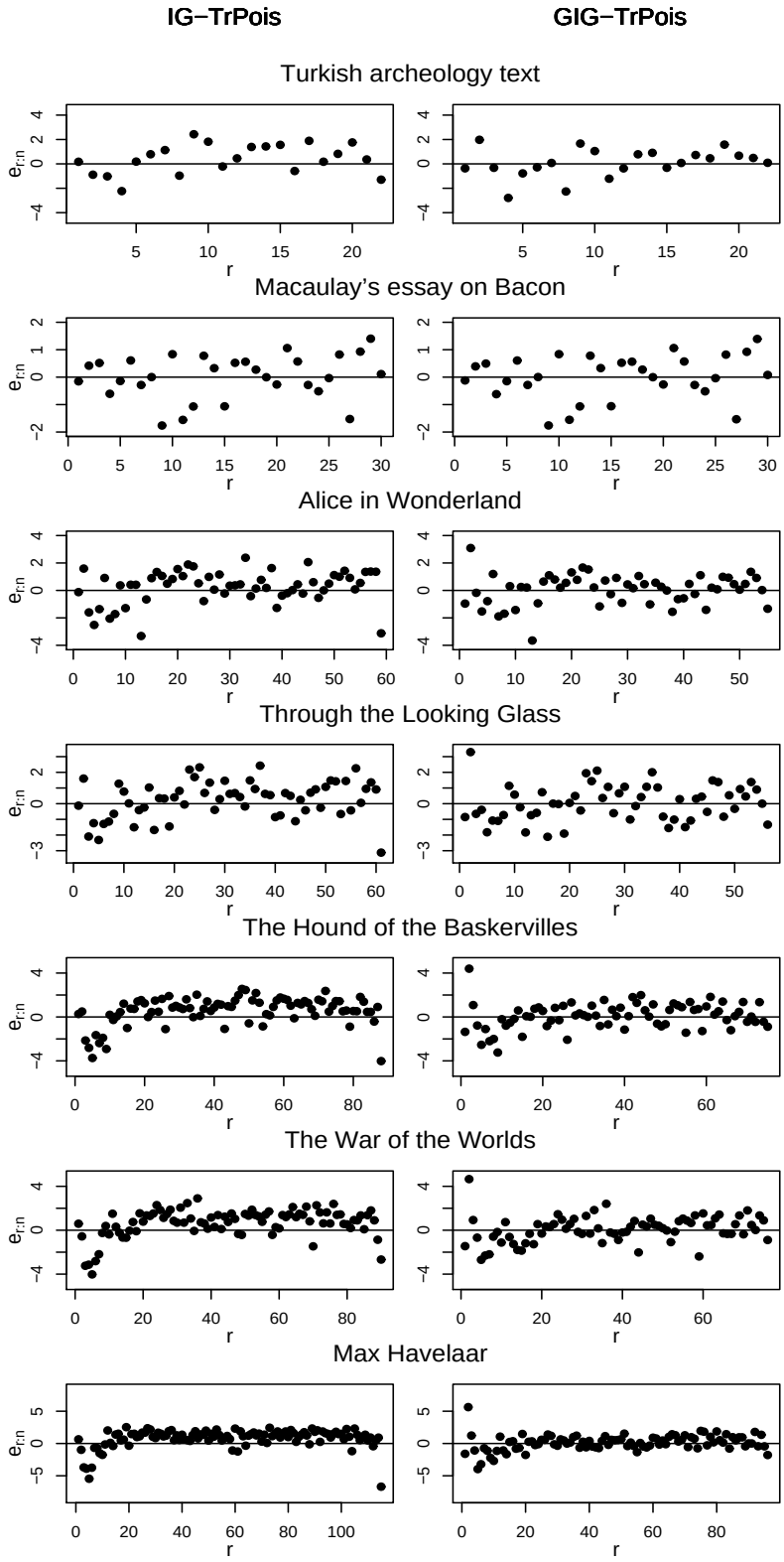

Figura 3.7: Pearson residuals, $e_{r:n}(\hat{b}, \hat{c}, \hat{b})$, under the IG-Truncated Poisson model, on the left hand side panels, and under the GIG-Truncated Poisson model, on the right hand side panels.

	n	v_n	$v(\hat{\psi}')$	$D_1(\hat{\pi})$	$D_1(\hat{\psi}')$	$D_2(\hat{\pi})$	$D_2(\hat{\psi}')$
Turkish A.	6939	3302	4940	.997108	.997479	7.39019	7.19128
E. Bacon	8049	2048	2335	.997156	.997657	6.83396	6.77050
Alice W.	26505	2651	2855	.989757	.992793	6.01525	6.10118
Through L.	28767	3085	3249	.990921	.992085	6.11352	6.05967
Hound B.	59241	5741	6020	.990529	.991350	6.29992	6.26102
War W.	59938	7112	7550	.987013	.992695	6.48687	6.54635
Max H.	99767	11161	11856	.992227	.993686	6.75465	6.73199

Taula 3.4: Estimates of $v(\psi')$, of two measures and of the diversity in the observed part of the vocabulary for the texts in Table 3.1, based on the maximum likelihood estimate of the mixing distribution of the GIG-Truncated Poisson model, and sample versions of these measures.

3.6 Concluding remarks

The truncated GIG-Poisson mixture model provides excellent fits for the word frequency counts of very long texts, when the truncated IG-Poisson model fails due to the large overdispersion of the data. That indicates that the GIG mixing distribution estimates work as estimates of the word frequency distribution of the vocabulary of the author. Instead, the mixing distribution of the GIG-Truncated Poisson model can not be interpreted as the word frequency distribution of that vocabulary. Hence, even though the IG-Truncated Poisson model adapts better than the truncated IG-Poisson model to the large overdispersion found in typical word frequency count data, these alternative models obtained by switching the order of the mixing and the truncation stages are not as helpful in the characterization of literary style.

The truncated mixture models considered in Sections 3.3.1 and 3.4, are more natural to formulate than the mixture of truncated models considered in Sections 3.3.2 and 3.5, and the mixing distribution estimates obtained from them are more helpful because they can be used as fingerprints in authorship attribution studies. And yet the mixture of truncated models might be theoretically easier to treat and might yield better fits, as is already pointed by Böhning & Kuhnert (2006) in the context of discrete mixture models.

Capítol 4

On the measure and the estimation of the evenness and diversity

4.1 Introduction

Some of the most useful tools in authorship attribution studies and in ecology rely on the analysis of word or species frequency count data. In the first case for example, texts are treated as samples from the vocabulary of their author and the word frequency counts in them are used to learn about his style and, in particular, about the size, evenness and diversity of his vocabulary, which might help distinguish his style from the style of other authors.

There has been a long lasting debate on which statistical models are most useful for word or species frequency count data, which has lead experts to consider a large number of alternative models. Given that most words (species) appear very few times and very few words (species) are repeated many times, word and species frequency count data typically have reverse J-shaped distributions with long upper tails. Yule (1944) and Good (1953) conjectured that this skewness should be modelled through Poisson mixture models, which at the time was not well accepted by everyone. Ever since, the debate on which models best suit the analysis of that type of data has been posed mainly in terms of which models best fit them.

The *first goal* of the chapter is to argue that what make Poisson mixture models truly special and useful is that they provide a simple mechanistic explanation that lets one interpret the model mixing distribution as the distribution of the word or species fre-

quencies of the vocabulary or population from which the sample was created. That interpretation is lacking in all the many purely empirical motivated models considered for this kind of data. The word frequency distribution of the vocabulary of an author characterizes his style, and the species frequency distribution of an ecosystem characterizes its population and they determine the size, evenness and diversity of that vocabulary or population. Hence the importance of identifying Poisson mixture models that fit word or species frequency count data well, and yield estimates of the word or species frequency distribution that can be used as fingerprints of the style of an author in his texts and of the population of an ecosystem in its samples.

To rank populations in terms of their evenness or diversity one needs real valued measures that capture the one aspect of evenness or diversity that one cares the most. Nevertheless, among practitioners there is a pervading feeling that the concepts of evenness and of diversity within a population can not be precisely distinguished and characterized. The *second goal* of the chapter is to propose definitions of measure of the evenness and of measure of the diversity within a population that can be used to compare populations with different total number of classes, which is the setting found most often when assessing evenness and diversity of vocabulary in stylometry, and of the population of an ecosystem in ecology.

An additional source of confusion is that the sample and the population versions of evenness and diversity measures are not always clearly distinguished. Typically the sample version is not a good estimate of the population version because it is biased with a bias that can be large. In stylometry and in ecology the estimation of diversity measures is made even more difficult due to the total number of words in the vocabulary or classes in the population being unknown. The *third goal* of the chapter is to describe how one can estimate measures of the evenness and of the diversity of a population through the expectation of these measures under the frequency distribution estimates proposed early on in the chapter. This leads to the assessment of the lack of diversity through measures of the variability of these model mixing distribution estimates.

The chapter focuses on the analysis of word frequency counts even though it all extends to species frequency counts, and it is organized as follows. Section 4.2 describes word frequency count data, it motivates the use of zero truncated Poisson mixture models on them and it illustrates how the truncated generalized inverse Gaussian-Poisson model provides excellent fits for very long texts. It also illustrates the use of the maximum likelihood estimate of that model mixing density as an estimate of the density of the word frequency of the vocabulary of the author.

Section 4.3 proposes definitions of measure of evenness and of measure of diversity, and

it explains the novelty of these two definitions when the total number of classes in the populations (words in the vocabularies) are unknown but known to be different among them. Readers not specially interested in foundational issues may want to skip the first part of this section.

Section 4.4 proposes approximating measures of the evenness and of the diversity through their expectation under the frequency distribution of the population. This combined with the main argument in Section 4.2 links lack of diversity with the ratio between a measure of the variability of the model mixing distribution of truncated Poisson mixture models, and the expected value of that mixing distribution. As an illustration, the size and various measures of the diversity of the vocabulary behind seven texts are estimated, and the behavior of various diversity measures when word frequencies have generalized inverse Gaussian distributions is explored.

4.2 Poisson mixture models and density of vocabulary

4.2.1 Vocabulary distribution and word frequency count data

To characterize the style of an author through his vocabulary, as in Holmes (1985), the basic assumption is that the author has available a list of all the words that he knows, and that the i-th word in that list is characterized through the proportion of times that that word would be found in a text of infinite length by that author, which is denoted by π_i. The set of probabilities π_j when j ranges over all the v words known by an author, $(\pi_1, \ldots, \pi_v)$, constitute the probability function of the vocabulary of that author. By identifying v with the total number of words in the vocabulary one is assuming that $\pi_i > 0$ for $i = 1, \ldots, v$.

For mathematical convenience one treats word frequencies, π_j, as a continuous variable with a density function, $\psi(\pi)$. Note that the larger the number of words in a vocabulary, v, the smaller the π_j's, and the closer the probability mass of $\psi(\pi)$ is to 0, which links a small expectation of $\psi(\pi)$ with a *rich* vocabulary. Furthermore given v, the closer the vocabulary distribution $(\pi_1, \ldots, \pi_v)$ is to the uniform distribution $(1/v, \ldots, 1/v)$, the more peaked $\psi(\pi)$ is around $1/v$, and the more evenly represented are the words in texts from that vocabulary, which links variability of $\psi(\pi)$ with lack of *evenness* and of *diversity* of vocabulary.

Texts written by an author are treated as if they were random samples drawn from his

	$v_{1:n}$	$v_{2:n}$	$v_{3:n}$	$v_{4:n}$	$v_{5:n}$	$v_{6:n}$	$v_{7:n}$	$v_{8:n}$	...	r_{max}	v_n	n
E. Bacon	990	367	173	112	72	47	41	31	...	255	2048	8049
Turkish A.	2326	477	178	107	53	33	22	26	...	222	3302	6939
Alice W.	1176	402	233	154	99	57	65	52	...	1631	2651	26505
Through L.	1491	460	259	148	113	78	61	47	...	1555	3085	28767
Hound B.	2836	889	449	280	208	137	116	92	...	3327	5741	59241
War of W.	3613	1138	567	340	250	177	135	93	...	4775	7112	59938
Max H.	6004	1731	819	491	368	258	168	137	...	4826	11161	99767

Taula 4.1: Word frequency count data for the nouns in Macaulay's essay on Bacon, considered in Sichel (1975), and for all the words in a Turkish archeology text, in *Alice in Wonderland* and in *Through the Looking Glass* by Carroll, in *The Hound of the Baskervilles* by Doyle, in *The War of the Worlds* by Wells, and in *Max Havelaar* by Dekker, all considered in Baayen (2001). Note that r_{max} denotes the frequency of the most frequent word.

vocabulary. If one denotes the total number of words (tokens) in a given text by n, the number of occurrences of the i-th word by $n_{i:n}$, and the proportion of occurrences of that word in that text by $\hat{\pi}_{i:n} = n_{i:n}/n$, the expected value of $\hat{\pi}_{i:n}$ is π_i. Let v_n denote the number of different words (types) observed in that text, and let $v_{r:n}$ denote the number of different words appearing exactly r times in it. The proportion of different words appearing exactly r times in a text of size n is $\hat{p}_{r:n} = v_{r:n}/v_n$ and its expectation, which depends on n, will be denoted by $p_{r:n}$.

In a given text most words appear only a few times and only a few words are repeated many times, and the distribution of $(v_{1:n}, v_{2:n}, \ldots, v_{n:n})$ is reverse J-shaped with an extraordinarily long upper tail. Table 4.1 presents the word frequency count of texts that will be used later on. Max Havelaar for example has a total of $n = 99767$ words out of which there are $v_n = 11161$ different words, 6004 words appear once, 1731 words appear twice and so on, with the most frequent word appearing a total of 4826 times.

4.2.2 Zero truncated Poisson mixture models

A large number of alternative models have been proposed for word frequency count data, sometimes under the label of large number of rare events models. Here Poisson mixture models are advocated for on the ground that their model mixing density estimates work as estimates of the density of the word frequencies of the vocabulary from which texts come from.

If the specific i-th word has a probability π_i of being used each time that an author

writes a word, the number of times that this word appears in a text by that author with a total of n words would be distributed as a binomial(n, π_i). Hence, if the density of the word frequencies was $\psi(\pi)$, the probability that a word appears exactly r times in a text of size n, $p_{r:n}$, follows a $\psi(\pi)$-binomial mixture model. Given that one can not count the words that the author knows but are not observed, one needs to consider the zero truncated version of this mixture.

Usually n will be large and all the π_i will be small, and one can approximate $p_{r:n}$ through a zero truncated $\psi(\pi)$-Poisson mixture model. This argument provides a simple mechanistic description of the word frequency count generating process that is lacking in the ad-hoc empirically motivated models often considered for this kind of data (see, e.g., Baayen, 2001), and it entitles one to interpret $\psi(\pi)$ as the density of the word frequencies of the vocabulary.

Sichel (1975, 1986a) models $\psi(\pi)$ through a generalized inverse gaussian distribution, described in Jorgensen (1982) and denoted here by GIG(b, c, g), which has a density function

$$\psi(\pi|b, c, g) = \frac{2^{g-1}}{(bc)^g K_g(b)} \pi^{g-1} e^{-\frac{\pi}{c} - \frac{b^2 c}{4\pi}}, \tag{4.1}$$

where b is in $(0, \infty)$, c is in $(0, \infty)$, g is in $(-\infty, \infty)$, and $K_g(\cdot)$ is the modified Bessel function of the third kind of order g. That leads to the use of the zero truncated GIG-Poisson model. A good alternative Poisson mixture model for this type of data could be the zero truncated Tweedie-Poisson model presented in Valero *et al.* (2009).

Sichel (1975, 1982, 1986b) proposes various estimation methods for the two parameter version obtained when $g = -1/2$, and Pollatschek and Radday (1981), Holmes (1992), Holmes and Forsyth (1995), Baayen (2001), Riba and Ginebra (2006) and Puig *et al.* (2009) fit it to word frequency count data. Sichel (1975, 1997) fit the full truncated GIG-Poisson model to word and species frequency count data with small n by estimating b and c for a given value of g by equating $\hat{p}_{1:n}$ to the theoretical $p_{1:n}^{tgigp}(b, c, g)$ and the sample mean to the population mean and then varying g and re-estimating (b, c) until the χ^2 goodness of fit statistic is minimized.

Table 4.2 presents the maximum likelihood estimate of the parameters of this model for the word frequency count data in Table 4.1, the value of the χ^2 goodness of fit test statistic and a lower bound of the corresponding p-value. To evaluate this statistic the model was fitted under the initial categorization but then categories were aggregated the least so that the expected count was at least 5 for each category. What is most remarkable in Table 4.2 is that the truncated GIG-Poisson model provides a very good fit for all the word frequency count data sets considered, even though some of the texts

	n	$\hat{b}$	$\hat{c}$	$\hat{g}$	$X^2, (df)$	$p - value$
E. Bacon	8049	0.1013	0.0044	-0.5801	$19.66, (26)$	0.8073
Turkish A.	6939	0.0514	0.0165	-1.0917	$24.84, (17)$	0.0982
Alice W.	26505	0.0314	0.0218	-0.6578	$61.34, (50)$	0.1306
Through L.	28767	0.0230	0.0214	-0.6549	$62.47, (52)$	0.1517
Hound B.	59241	0.0164	0.0291	-0.7300	$90.03, (71)$	0.0632
War of W.	59938	0.0179	0.0289	-0.7909	$85.83, (71)$	0.1108
Max H.	99767	0.0117	0.0265	-0.7785	$97.83, (90)$	0.2687

Taula 4.2: Maximum likelihood estimate of (b, c, g) for the truncated GIG-Poisson model, value of the χ^2 goodness of fit test statistic, and a lower bound for the corresponding p-value.

are very long.

Observe that the parameter estimates for Alice in Wonderland and Through the Looking Glass are very similar, in line with these texts being the only ones sharing the same author. Note also that the parameter estimate for the Essay on Bacon is very different from the other estimates, in line with this being the only text in which only nouns have been considered.

4.2.3 Estimation of the density of the word frequencies of vocabulary

Obtaining a good fit with a truncated Poisson mixture model indicates that the estimates of its model mixing distribution must be a good estimate of the distribution of the word frequencies of the vocabulary of the author. Hence these estimates can be viewed as fingerprints of the style of the author and can help reveal distinguishing features of his vocabulary, and hence they can be of great help in authorship attribution problems.

Figure 4.1 presents the maximum likelihood estimates of the density of the GIG-Poisson model mixing distribution for the data in Table 4.1. One could compare style through these word frequency density estimates with the help of functional data analysis tools, but it is more meaningful to summarize first these density estimates through real valued quantities that capture specific aspects of literary style like the diversity of vocabulary, which encompasses its richness and its evenness.

The richness of a vocabulary (population) is measured through an estimate of the total number of words (classes) in it, v, as reviewed in Bunge and Fitzpatrick (1993). One can

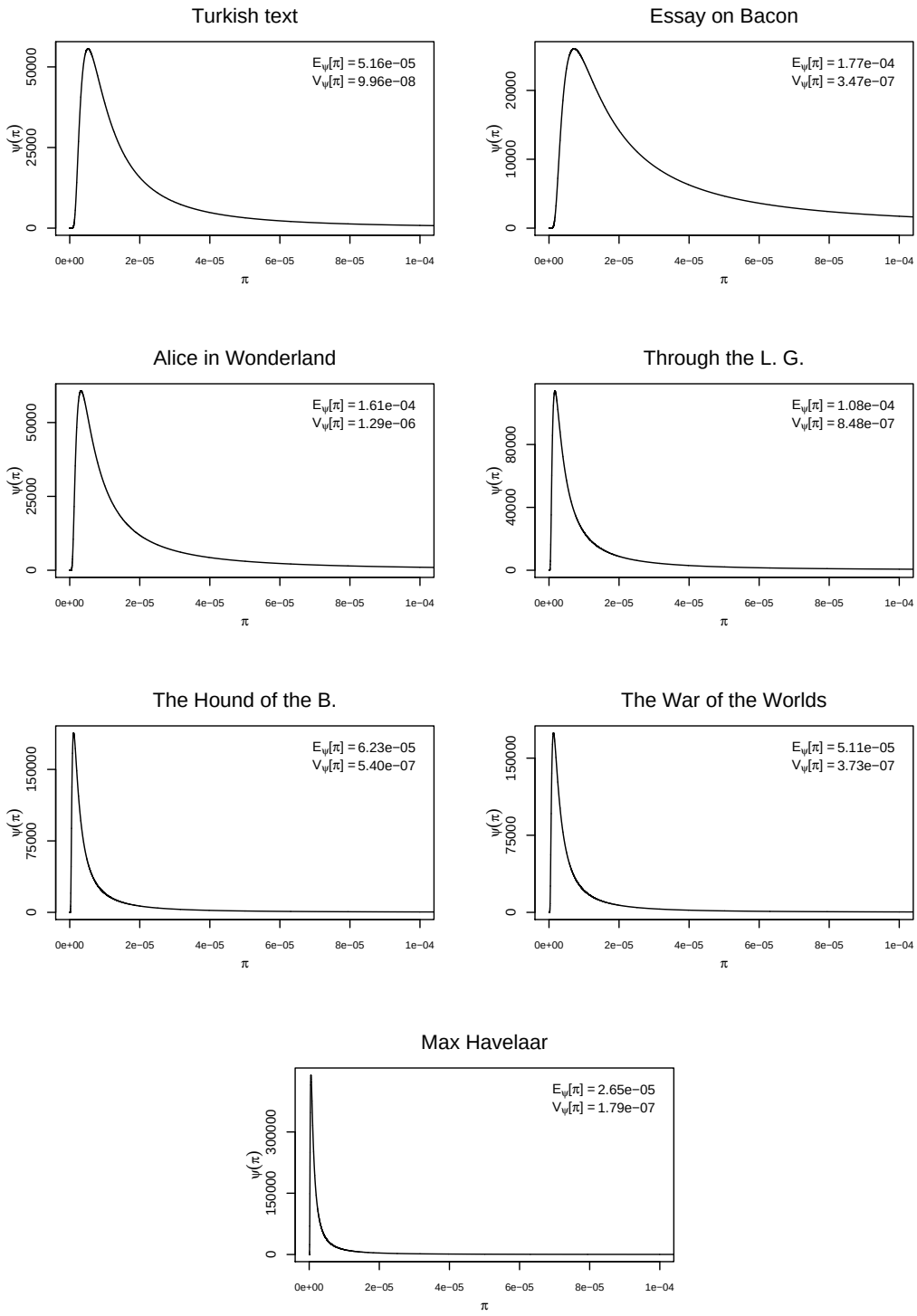

Figura 4.1: Maximum likelihood estimates of the mixing densities of the zero truncated GIG-Poisson model for the texts in Table 4.1. They serve as estimates of the density of the word frequencies of the vocabulary of the author. The smaller $E_\psi[\pi]$ the larger the total number of words in the vocabulary of the author, v. For a given v, the smaller $Var_\psi[\pi]$ the more even and diverse that vocabulary.

not estimate v through v_n because it is a biased estimator of v that underestimates it. Instead note that $\bar{\pi} = \sum_{i=1}^{v} \pi_i / v = 1/v$, and therefore v is the reciprocal of the average of the π_j's.

If one considers the π_j to be a continuous random variable with density function $\psi(\pi)$, the closer the probability mass of $\psi(\pi)$ is concentrated near zero, the smaller the π_j's and $E_\psi[\pi]$, and the richer the vocabulary. In that case, one can estimate v through the closest integer to $v(\hat{\psi}) = 1/E_{\hat{\psi}}[\pi]$ as in Sichel (1986a, 1997) and in Table 4.5. In Section 4.4.1 the variability of π under $\psi(\pi)$ is linked with lack of diversity and in Section 4.4.2 it is used to estimate it.

4.3 Measure of the evenness and of the diversity of a population

Evenness and diversity are intimately related multidimensional concepts that can not be completely captured by any pair of single real valued quantities. Nevertheless, to rank populations in terms of their diversity (evenness) one has to do it through real valued measures that capture the one aspect of diversity (evenness) that one cares the most. Articles often list various measures but they do not always identify what is common among them and when they do they do not always agree, which has left some with the feeling that evenness and diversity are concepts that can not be completely distinguished and characterized (see, e.g., Ricotta, 2005).

Here we attempt to dispel that feeling by explaining what makes into a measure of evenness, what makes into a measure of diversity and by detailing the sense in which our definition of measure of diversity differs from the definitions in the literature.

Remark 4.1: Here it is assumed that a population is characterized only through its number of classes, v, and its probability function, $(\pi_1, \ldots, \pi_v)$. Hence we do not consider the measure of the diversity within populations characterized through the pairwise distances or dissimilarities between their classes, like the ones recognized as measures of taxonomical diversity, of functional diversity or Rao's quadratic diversity (see, e.g., Rao, 1982, Solow *et al.*, 1993). Even though there exist links between distance based and distribution based diversity measures, (see, e.g., Izsak and Papp, 2000, Ricotta and Szeidl, 2006), distance based measures are objects of an intrinsically different nature and hence require to be characterized apart.

Remark 4.2: When the total number of classes (words) in the population (vocabulary) is

known to be equal to v, one has to restrict attention to probability functions $(\pi_1, \ldots, \pi_v)$ with $\pi_i > 0$ for $i = 1, \ldots, v$, and thus in the interior of the simplex of R^v. Furthermore, if the number of classes is known to be the same for all the populations under consideration, there is no essential difference between comparing them in terms of their diversity or in terms of their evenness. Nevertheless, in the practice of stylometry (ecology) the total number of words (species) in the vocabulary (population) is most often unknown and the vocabularies (populations) being compared have different v's, which complicates matters significantly because one has to estimate v and one has to distinguish evenness measures from diversity measures.

4.3.1 Measure of the evenness of populations with a given number of classes

A population with a fixed and given number of classes v is said to be completely even if its distribution, $(\pi_1, \ldots, \pi_v)$, is the uniform $(1/v, \ldots, 1/v)$; the further its distribution is away from the uniform towards the boundary of the simplex of R^v the less even that population is considered to be. The next definition of measure of evenness captures these two features.

Definition 4.3.1. *A real valued function $E^v(\cdot)$ defined on the set of probability functions of populations with a given number of classes v, $(\pi_1, \ldots, \pi_v)$ with $\pi_i > 0$ for $i = 1, \ldots, v$, is a measure of the evenness of the corresponding populations if:*

1. *$E^v(\cdot)$ is Schur concave in the interior of the simplex of R^v, and*
2. *$E^v(1, \ldots, 0) = 0$.*

Schur concave functions on the simplex of R^v are the ones such that for every probability function $(\pi_1, \ldots, \pi_v)$ in the interior of that simplex, one has that

$$E^v(\pi_1, \ldots, \pi_v) \leq E^v(\nu_1, \ldots, \nu_v) \tag{4.2}$$

where $(\nu_1, \ldots, \nu_v)$ is any convex combination of permutations of $(\pi_1, \ldots, \pi_v)$. Hence this definition states that $E^v(\pi_1, \ldots, \pi_v)$ can not be larger than the value taken by $E^v(\cdot)$ on any point in the convex hull defined by the permutations of $(\pi_1, \ldots, \pi_v)$, which is the set of points that can be considered to be "more interior"than $(\pi_1, \ldots, \pi_v)$. As a consequence $E^v(\cdot)$ takes on its smaller values over "extreme regions"and its larger values around $(1/v, \ldots, 1/v)$ as desired.

Given that $(1/v, \ldots, 1/v)$ is in the convex hull of the permutations of every point in the simplex of R^v and given that every point in that set is in the convex hull of the

permutations of $(1, \ldots, 0)$,

$$0 = E^v(1, \ldots, 0) \leq E^v(\pi_1, \ldots, \pi_v) \leq E^v(1/v, \ldots, 1/v). \tag{4.3}$$

Strictly speaking $E^v(\cdot)$ does not even have to be defined on the boundary of the simplex of R^v because that boundary represents populations with less than v classes, and therefore $E^v(1, \ldots, 0) = 0$ only needs to hold in a limiting sense and is equivalent to requiring that $E^v(\cdot)$ be non-negative. Any Schur concave function on the simplex can be made to satisfy that requirement by subtracting from it its value at $(1, \ldots, 0)$.

All Schur concave functions are permutation symmetric, but they are not always concave functions. Nevertheless, all permutation symmetric and concave functions are Schur concave.

The rationale for imposing the Schur concavity condition is that it is what guarantees that the measure of the evenness of a population increases whenever abundance is transferred from one of its v classes (words or species) to another strictly less abundant class, as long as the ranking in terms of abundance between these two classes is not modified. For a precise description of this principle of transfer and its connection to Schur concavity see Muller and Stoyan (2002, p.31-37). For details on Schur concave functions, and on the equivalence between the ordering defined inside the simplex of R^v by the condition 'being at least as even under all evenness measures,' the majorization ordering, and the ordering induced by Lorenz curves, see Marshall and Olkin (1979, ch. 3) and Shaked and Shanthikumar (1994, pp. 198-200).

Definition 4.3.1 coincides with the definition of measure of equality in economics (Sen, 1974) and if one replaced the Schur concavity condition by the "plain Jensen" concavity condition it would coincide with the definition of measure of uncertainty or of lack of information in $(\pi_1, \ldots, \pi_v)$, (DeGroot, 1962, Ginebra, 2007).

On a more intuitive level note that the closer $(\pi_1, \ldots, \pi_v)$ is to $(1/v, \ldots, 1/v)$, the more concentrated the π_j's are around $1/v$ and the smaller the variability of π_j when j varies from 1 to v. It will thus not come as a surprise that:

$$E_1^v(\pi_1, \ldots, \pi_v) = \frac{v-1}{v^2} - \frac{\sum_{i=1}^{v}(\pi_i - 1/v)^2}{v} = \frac{v-1}{v^2} - Var[\pi], \tag{4.4}$$

is a measure of evenness abiding by Definition 4.3.1. A second example is the geometric average,

$$E_2^v(\pi_1, \ldots, \pi_v) = \sqrt[v]{\pi_1 \ldots \pi_v}, \tag{4.5}$$

where remember that by definition of v one assumes that $\pi_i > 0$ for $i = 1, \ldots, v$.

4.3.2 Measure of the evenness within a population

In stylometry and in ecology, unless one restricts consideration to a given subset of words or species, the number of classes v is unknown and the vocabularies being compared have different v's, and hence Definition 4.3.1 is of a limited use. Nevertheless, any definition of measure of evenness that applies there has to collapse down to Definition 4.3.1 when v is fixed and known.

When comparing evenness across populations with different v, one has to be careful with the role of v. In particular, $E^v(1/v, \ldots, 1/v)$ will often be a function of v, as in (4.4) where it is equal to $(v-1)/v^2$, and in (4.5), where it is equal to $1/v$. If one used $E_1^v(\cdot)$ or $E_2^v(\cdot)$ to compare evenness across populations with a different v one would be wrongly assuming that the larger v the less even the corresponding uniform distribution. The next definition rules that out.

Definition 4.3.2. *A real valued function $E(\cdot)$ defined on the set of probability functions of populations of any number of classes is a measure of the evenness of the corresponding population if the restriction of $E(\cdot)$ to the interior points of the simplex of R^v, $E^v(\cdot)$, is such that:*

1. *$E^v(\cdot)$ is Schur concave,*
2. *$E^v(1, \ldots, 0) = 0$ and $E(1) = 0$, and*
3. *$E^v(1/v, \ldots, 1/v) = 1$,*

for every $v > 1$, where $E(1)$ is the evenness of a degenerate single class population.

There is no loss of generality in that this definition imposes that $0 \leq E(\cdot) \leq 1$, and in fact one can associate one such measure $E(\cdot)$ to any measure $E^v(\cdot)$ abiding by Definition 4.3.1 through

$$E(\pi_1, \ldots, \pi_v) = \frac{1}{E^v(1/v, \ldots, 1/v)} E^v(\pi_1, \ldots, \pi_v), \tag{4.6}$$

which under (4.4) leads to

$$E_{VAR}(\pi_1, \ldots, \pi_v) = 1 - \frac{v^2}{v-1} Var[\pi] = \frac{v}{v-1}(1 - \sum_{i=1}^{v} \pi_i^2), \tag{4.7}$$

and under (4.5) leads to:

$$E_{GA}(\pi_1, \ldots, \pi_v) = v \sqrt[v]{\pi_1 \ldots \pi_v}, \tag{4.8}$$

with the understanding that $\pi_i > 0$ for $i = 1, \ldots, v$.

Remark 4.3: The evenness measures (4.7) and (4.8) are discontinuous at the boundaries of the successive simplex, because

$$\lim_{\epsilon\to 0} E^{v+1}_{VAR}(\pi_1-\frac{\epsilon}{v},\ldots,\pi_v-\frac{\epsilon}{v},\epsilon) = \frac{v+1}{v}(1-\sum_{i=1}^{v}\pi_i^2) < E_{VAR}(\pi_1,\ldots,\pi_v,0) = E^v_{VAR}(\pi_1,\ldots,\pi_v), \tag{4.9}$$

and

$$\lim_{\epsilon\to 0} E^{v+1}_{GA}(\pi_1-\frac{\epsilon}{v},\ldots,\pi_v-\frac{\epsilon}{v},\epsilon) = 0 < E_{GA}(\pi_1,\ldots,\pi_v,0) = E^v_{GA}(\pi_1,\ldots,\pi_v), \tag{4.10}$$

and the same happens to all the examples of $E(\cdot)$ in Table 4.4. If instead of requiring Schur concavity in the interior of each of the successive simplex one imposed that $E(\cdot)$ be simultaneously Schur concave on the boundary and the interior points of all these simplex, (4.7), (4.8) and the examples of evenness measures in Table 4.4 would not qualify as such.

Imposing that stronger requirement would mean that whenever one added one new class and transfered abundance from the pre-existing classes to the new class while keeping the same abundance ranking among classes, all evenness measures would have to increase. That is not a desirable feature because there is nothing wrong in evenness measures assigning a larger value to a population with probability function $(.5,.5,0)$ than to a population with probability function $(.49,.49,.02)$, which would be forbidden under the overall Schur concavity requirement.

4.3.3 Measure of the diversity within a population

Loosely speaking, one typically requires that any measure of the diversity within a population combine its evenness with its number of classes, v, in a way such that:

1. among populations with the same v, the 'more even' the population the larger the number assigned by a diversity measure, where by 'more even' it is meant that all evenness measures adopt a larger number, and that
2. among populations with the 'same evenness', the larger v the larger the number assigned by that measure, (even though often it is not clear what is meant by 'same evenness').

It is also desirable that a definition of measure of diversity collapses down to Definition 4.3.1 when restricted to populations with the same known v, as in next definition.

Definition 4.3.3. *A real valued function $D(\cdot)$ defined on the set of probability functions of populations of any number of classes is a measure of the diversity within the corresponding population if the restriction of $D(\cdot)$ to the interior points of the simplex of R^v, $D^v(\cdot)$, is such that*

1. $D^v(\pi_1, \ldots, \pi_v)$ *is Schur concave,*
2. $D^v(1, \ldots, 0) = 0$ *and* $D(1) = 0$*, and*
3. $D^v(1/v, \ldots, 1/v)$ *is non-decreasing in* v*,*

for every $v > 1$*, where* $D(1)$ *is the diversity of a degenerate single class population.*

All evenness measures abiding by Definition 4.3.2 and $v - 1$ satisfy this definition. The first two conditions imply that the restriction of $D(\cdot)$ inside the simplex of R^v is such that,

$$0 \leq D^v(\pi_1, \ldots, \pi_v) \leq D^v(1/v, \ldots, 1/v). \tag{4.11}$$

The third condition, listed as a desirable property in Pielou (1975), states that among completely even populations the larger v the more diverse the population. This condition rules out the $E^v(\cdot)$'s with $E^v(1/v, \ldots, 1/v)$ decreasing with v, like (4.4) and (4.5), but their associated measures in (4.7) and (4.8) are valid diversity measures. In fact, as suggested in Pielou (1975), one can associate an evenness measure to every diversity measure through

$$E(\pi_1, \ldots, \pi_v) = \frac{1}{D(1/v, \ldots, 1/v)} D(\pi_1, \ldots, \pi_v). \tag{4.12}$$

Solomon (1979), Kempton (1979), Patil and Taillie (1982) and Tong (1983) all recognize the need to impose some type of Schur concavity condition on diversity measures. Different from the definitions in these papers though, Definition 4.3.3 requires that Schur concavity hold only one dimension v at a time and therefore only in the interior of each successive simplex of R^v. In that way Definition 4.3.3 encompasses measures like (4.7), (4.8), and $D_r(\cdot)$, $D_{SM}(\cdot)$ and $D_{RM}(\cdot)$ in Table 4.3 which would all be excluded under the alternative definitions of measure of diversity.

None of the papers listed in the above paragraph make the third requirement of Definition 4.3.3 into a part of their definitions of measure of diversity because it follows as a consequence of the "global" Schur concavity condition that they impose, which is not the case here.

If Definition 4.3.3 required, (which it doesn't), that all diversity measures be Schur concave everywhere on the simplex of R^{v^*}, where v^* is an integer larger than the largest possible v, all diversity measures would have to assign a larger value to a population with probability function $(.49, .49, .02)$ than to a population with probability function $(.5, .5, 0)$. Even though there is no problem with diversity measures behaving that way, one is also entitled to consider diversity measures under which the increase due to adding one new class might be offset by the decrease due to the loss of evenness imposed by

that addition; When deciding which animals should board Noa's ark, one prefers two cows and two horses over two cows, one horse and one pig.

Remark 4.4: Diversity measures can be discontinuous at the boundaries of the simplex, in the sense of Remark 4.3. If one required that all diversity measures be continuous everywhere on the simplex of R^{v^*}, one could drop the third requirement from Definition 4.3.3 because $(1/(v+1), \ldots, 1/(v+1))$ is in the convex hull of the permutations of $(1/v, \ldots, 1/v, 0)$ and therefore continuity plus Schur concavity implies $D(1/v, \ldots, 1/v, 0) \leq D(1/(v+1), \ldots, 1/(v+1))$ for every $v > 1$. The problem is that imposing continuity on all diversity measures would rule out all the examples of evenness measures and $D_r(\cdot)$, $D_{SM}(\cdot)$ and $D_{RM}(\cdot)$ in Table 4.3.

Instead of Schur concavity, Rao (1982) requires that all measures of diversity be (Jensen) concave because that it is what guarantees that the diversity within a mixture population be larger than or equal to the average of the diversities within the individual populations. Not all concave functions are Schur concave and not all Schur concave functions are concave and therefore neither of these two conditions is a special case of the other. We require Schur concavity and not concavity because we find it more natural to characterize diversity through the abundance transfer argument in Setion 4.3.1 than through this mixing argument.

Furthermore, increasing transformations of Schur concave functions are Schur concave and therefore if $h(\cdot)$ is increasing with $h(0) = 0$ and if $D(\cdot)$ abides by Definition 4.3.3 then $h(D(\cdot))$ also abides by it. This would not hold if one replaced Schur concavity by concavity in the definition because increasing transformations of concave functions are not always concave. In any case all the examples of measure of diversity considered below are both Schur concave as well as concave, and therefore they abide both by Definition 4.3.3 as well as by Rao's definition.

It is important to emphasize that Definition 4.3.3 lists the minimal set of requirements for functions on the set of finite probability functions to qualify as measures of the diversity of the populations that they represent. This does not mean that in specific settings, like for example when one intends to use a diversity measure in apportionment of diversity between and within populations as in Rao (1984), one is not entitled to impose additional requirements and in that way reduce the set of measures of diversity under consideration.

4.3.4 Examples of measures of diversity and of measures of evenness

By going through various permutation symmetric concave and therefore Schur concave functions Table 4.3 presents an array of measures of diversity, each bringing a different perspective on what diversity means. Most of them can be posed as

$$D_g(\pi_1, \ldots, \pi_v) = \sum_i^v g(\pi_i) - (g(1) + (v-1)g(0)), \tag{4.13}$$

where $g(\cdot)$ is concave on $[0, 1]$ and such that $vg(1/v)$ is non-decreasing in v.

A very large class of diversity measures satisfying Definition 4.3.3 is

$$D_t(\pi_1, \ldots, \pi_v) = \text{sign}(t-1)(1 - \sum_{i=1}^v \pi_i^t), \qquad \text{for } \; t > 0, t \neq 1, \tag{4.14}$$

which is equivalent to the class measures advocated for in Hill (1973) and in Patil and Taillie (1982), and is related to the ones in Hurlbert (1971), in Smith and Grassle (1977), and to the measures of the lack of information in $(\pi_1, \ldots, \pi_v)$ put forward by Renyi (1961). The advantage of using (4.14) instead of the alternative formulations in these papers will be clear when we compute its expectation in Section 4.4.1. For a given v the measures in (4.14) are graded by t according to the relative weight given to the most abundant classes, with $D_{t=0}(\pi_1, \ldots, \pi_v) = v - 1$ assigning the same weight to all classes and with $D_t(\cdot)$ approaching

$$D_M(\pi_1, \ldots, \pi_v) = 1 - \max_i \pi_i, \tag{4.15}$$

when t approaches ∞, which is only a function of the frequency of the most abundant class. The most favored measure in this class is the Gini-Simpson index,

$$D_{GS}(\pi_1, \ldots, \pi_v) = 1 - \sum_i^v \pi_i^2, \tag{4.16}$$

which is the probability that two words picked at random from a text of infinite length from that vocabulary are different. This measure induces the same ordering as $1/\sum_i \pi_i^2$, which is a Schur concave measure widely used in ecology, (see, e.g., Magurran, 2004), that is not concave and therefore would be excluded as a diversity measure if concavity was imposed as a requirement in Definition 4.3.3. The measure of evenness associated with (4.16), $E_{GS}(\pi_1, \ldots, \pi_v)$, coincides with (4.7), and its values are very similar to the ones of $D_{GS}(\pi_1, \ldots, \pi_v)$. For a description of the sense in which (4.16) is unique see Routledge (1979) and Rao (1984), and for criticism of (4.16) for being too dependent on the most abundant classes see Kempton and Wedderburn (1978). A special case of

	$D(\pi_1,\ldots,\pi_v)$	$D(\psi)$	$D(1/v,\ldots,1/v)$
$D_{t=0}$	$v-1$	$\frac{1}{E_\psi[\pi]}-1$	$v-1$
D_t	$\sum_i \pi_i^t - 1, \quad 0<t<1$	$\frac{E_\psi[\pi^t]}{E_\psi[\pi]}-1$	$\frac{1}{v^{t-1}}-1$
D_t	$1-\sum_i \pi_i^t, \quad 1<t$	$1-\frac{E_\psi[\pi^t]}{E_\psi[\pi]}$	$1-\frac{1}{v^{t-1}}$
D_E	$-\sum_i \pi_i \log \pi_i$	$-\frac{E_\psi[\pi\log\pi]}{E_\psi[\pi]}$	$\log v$
D_M	$1-\max_i \pi_i$	$1-E_\psi[\max_i \pi_i]$	$1-\frac{1}{v}$
$D_{r\geq 1}$	$1+\frac{(v-1)^r-1}{v}-\sum_i \lvert\pi_i^{1/r}-(\frac{1}{v})^{1/r}\rvert^r$	$1-E_\psi[\pi]+\frac{(1-E_\psi[\pi])^r}{E_\psi[\pi]^{r-1}}-\frac{E_\psi[\lvert\pi^{1/r}-E_\psi[\pi]^{1/r}\rvert^r]}{E_\psi[\pi]}$	$1+\frac{(v-1)^r-1}{v}$
$D_{r=1}$	$2(1-\frac{1}{v})-\sum_i \lvert\pi_i-\frac{1}{v}\rvert$	$2(1-E_\psi[\pi])-\frac{E_\psi[\lvert\pi-E_\psi[\pi]\rvert]}{E_\psi[\pi]}$	$2(1-\frac{1}{v})$
D_{SM}	$2-\frac{1}{v}-\sum_i \max\{\pi_i,\frac{1}{v}\}$	$2-E_\psi[\pi]-\frac{E_\psi[\max\{\pi,E_\psi[\pi]\}]}{E_\psi[\pi]}$	$1-\frac{1}{v}$
D_{RM}	$\sqrt[r]{v}(1-\sqrt[r]{\sum_i \pi_i^r}), \quad r\geq 2$	$\frac{1-E_\psi[\sqrt[r]{\sum_i \pi_i^r}]}{\sqrt[r]{E_\psi[\pi]}}$	$\sqrt[r]{v}-\sqrt[r]{v^{2-r}}$

Taula 4.3: Measures of the diversity within $(\pi_1,\ldots,\pi_v)$, their expected value assuming that the π_i are identically distributed as $\psi(\pi)$ and that $v=1/E_\psi[\pi]$, $D(\psi)$, and their maximum value when they are restricted for populations with a given total number of classes v. $D(\psi)$ is the ratio between a measure of the lack of variability of π distributed as $\psi(\pi)$ and $E_\psi[\pi]$.

(4.14) not so dependent on them is obtained with $t=1/2$,

$$D_{SQ}(\pi_1,\ldots,\pi_v)=\sum_{i=1}^{v}\sqrt{\pi_i}-1. \tag{4.17}$$

One popular measure abiding by Definition 4.3.3 is the entropy of $(\pi_1,\ldots,\pi_v)$, $D_E(\pi_1,\ldots,\pi_v)$ that can also be obtained as a limiting case of a variant of (4.14) when $t=1$.

Tables 4.3 and 4.4 present many other examples of evenness and diversity measures abiding by Definitions 4.3.2 and 4.3.3, including the class of measures $D_r(\pi_1,\ldots,\pi_v)$, that relate the diversity of $(\pi_1,\ldots,\pi_v)$ with the negative of a distance between $(\pi_1,\ldots,\pi_v)$ and $(1/v,\ldots,1/v)$, and are examples of diversity measures that are discontinuous on the boundaries of the simplex.

	$E(\pi_1,\ldots,\pi_v)$	$E(\psi)$
E_{GA}	$v\sqrt[v]{\pi_1\ldots\pi_v}$	$\frac{E_\psi(\pi)\sqrt{E_\psi[\pi^{E_\psi(\pi)}]}}{E_\psi[\pi]}$
E_t	$\frac{v^{t-1}}{1-v^{t-1}}(\sum_i \pi_i^t - 1), \quad 0<t<1$	$\frac{E_\psi[\pi(\pi^{t-1}-1)]}{E_\psi[\pi](E_\psi[\pi]^{t-1}-1)}$
E_t	$\frac{v^{t-1}}{v^{t-1}-1}(1-\sum_i \pi_i^t), \quad 1<t$	$\frac{E_\psi[\pi(1-\pi^{t-1})]}{E_\psi[\pi](1-E_\psi[\pi]^{t-1})}$
E_E	$-\frac{1}{\log v}\sum_i \pi_i \log \pi_i$	$\frac{E_\psi[\pi\log\pi]}{E_\psi[\pi]E_\psi[\log\pi]}$
E_M	$\frac{v}{v-1}(1-\max_i \pi_i)$	$\frac{1-E_\psi[\max_i \pi_i]}{1-E_\psi[\pi]}$
$E_{r\geq 1}$	$1-\frac{v}{(v-1)(1+(v-1)^{r-1})}\sum_i \lvert\pi_i^{1/r}-(\frac{1}{v})^{1/r}\rvert^r$	$1-\frac{E_\psi[\pi]^{r-2}}{E_\psi[\pi]^{r-1}+(1-E_\psi[\pi])^{r-1}}\frac{E_\psi[\lvert\pi^{1/r}-E_\psi[\pi]^{1/r}\rvert^r]}{1-E_\psi[\pi]}$
$E_{r=1}$	$1-\frac{v}{2(v-1)}\sum_i \lvert\pi_i-\frac{1}{v}\rvert$	$1-\frac{1}{2(1-E_\psi[\pi])}\frac{E_\psi[\lvert\pi-E_\psi[\pi]\rvert]}{E_\psi[\pi]}$
E_{SM}	$1-\frac{v}{v-1}(\sum_i \max\{\pi_i,\frac{1}{v}\}-1)$	$1-\frac{E_\psi[\max\{\pi,E_\psi[\pi]\}]-E_\psi[\pi]}{E_\psi[\pi](1-E_\psi[\pi])}$
E_{RM}	$\frac{1}{1-\sqrt[v]{v^{1-r}}}(1-\sqrt[r]{\sum_i \pi_i^r}), \quad r\geq 2$	$\frac{1-E_\psi[\sqrt[r]{\sum_i \pi_i^r}]}{1-\sqrt[r]{E_\psi[\pi]^{r-1}}}$

Taula 4.4: Measures of the evenness of $(\pi_1,\ldots,\pi_v)$ associated to the diversity measures in Table 4.3 and (4.8), and their expected value when the π_i are identically distributed as $\psi(\pi)$. To compute $E_{GA}(\psi)$ one assumes that the π_i are both independent as well as identically distributed.

4.4 Estimation of diversity measures and variability of $\psi(\pi)$

4.4.1 Diversity of $(\pi_1,\ldots,\pi_v)$ and variability of $\psi(\pi)$

When measuring the evenness and diversity of a population, most often one does neither know v nor $(\pi_1,\ldots,\pi_v)$. Estimating $D(\pi_1,\ldots,\pi_v)$ or $E(\pi_1,\ldots,\pi_v)$ through the sample versions obtained by replacing $(\pi_1,\ldots,\pi_v)$ by $(\hat{\pi}_{1:n},\ldots,\hat{\pi}_{v:n})$ is usually not convenient because $D(\hat{\pi}_{1:n},\ldots,\hat{\pi}_{v:n})$ and $E(\hat{\pi}_{1:n},\ldots,\hat{\pi}_{v:n})$ are biased estimators of $D(\pi_1,\ldots,\pi_v)$ and $E(\pi_1,\ldots,\pi_v)$ with a bias that can be large and depends on n, v and $(\pi_1,\ldots,\pi_v)$, (see Riba and Ginebra, 2006). To construct good estimators of $D(\pi_1,\ldots,\pi_v)$ through its sample versions one needs the distribution of $D(\hat{\pi}_{1:n},\ldots,\hat{\pi}_{v:n})$, which is typically unknown. This makes it very difficult to compare diversity through word frequency count of texts of different length.

As an alternative we propose estimating $D(\pi_1,\ldots,\pi_v)$ and $E(\pi_1,\ldots,\pi_v)$ by first approximating them by their expected value assuming that the π_i are identically distributed as $\psi(\pi)$,

$$D(\psi) = E_\psi[D(\pi_1,\ldots,\pi_v)], \tag{4.18}$$

and $E(\psi) = E_\psi[E(\pi_1, \ldots, \pi_v)]$, and then substituting v by $v(\psi) = 1/E_\psi[\pi]$ and ψ by a good estimate of it. For example, one could approximate (4.13) through:

$$D_g(\psi) = E_\psi[D_g(\pi_1, \ldots, \pi_v)] = \frac{E_\psi[g(\pi)]}{E_\psi[\pi]} - (g(1) + \frac{1 - E_\psi[\pi]}{E_\psi[\pi]} g(0)), \tag{4.19}$$

and estimate $D_g(\psi)$ by replacing ψ by the maximum likelihood estimate of ψ under any truncated mixed Poisson model that fits the data well.

Remark 4.5: Given that $E_\psi[g(\pi)]$ is the expectation under ψ of a concave function of π, it can be interpreted as a measure of the concentration or lack of variability of π under $\psi(\pi)$. Hence, $D_g(\psi)$ is the ratio between a measure of the lack of variability of π and $E_\psi[\pi]$. When v and hence $E_\psi[\pi]$ are fixed and known, $D_g(\psi)$ becomes a measure of the lack of variability of π under $\psi(\pi)$, in line with the fact that the more concentrated π is around $1/v$, the smaller the variability of the π_i's and the more even and diverse the corresponding population.

Good (1953, 1982) and Lyons and Hutchison (1986) present unbiased estimates of the Gini-Simpson diversity measure in (4.16) and indicate how to estimate its variance. As an alternative we propose estimating $D_{GS}(\pi_1, \ldots, \pi_v)$ by first replacing it by its expectation,

$$D_{GS}(\psi) = 1 - E_\psi[\sum_{i=1}^{v} \pi_i^2] = 1 - \frac{E_\psi[\pi^2]}{E_\psi[\pi]} = 1 - \frac{Var_\psi[\pi]}{E_\psi[\pi]} - E_\psi[\pi]. \tag{4.20}$$

To estimate $E_{GS}(\pi_1, \ldots, \pi_v)$, in (4.7), one approximates it through

$$E_{GS}(\psi) = 1 - \frac{Var_\psi[\pi]}{E_\psi[\pi](1 - E_\psi[\pi])} = \frac{E_\psi[\pi(1-\pi)]}{E_\psi[\pi]E_\psi[1-\pi]}. \tag{4.21}$$

Note that among populations with the same $v = 1/E_\psi[\pi]$, the smaller $Var_\psi[\pi]$ the more even and diverse the population is according to $E_{GS}(\psi)$ and $D_{GS}(\psi)$.

The estimation of the entropy of a population, $D_E(\pi_1, \ldots, \pi_v)$ in Table 4.3, is a difficult problem tackled for example in Blyth (1959), Basharin (1959) and Chao and Shen (2003). The alternative proposed here consists in replacing the entropy of $(\pi_1, \ldots, \pi_v)$ by

$$D_E(\psi) = E_\psi[-\sum_{i=1}^{v} \pi_i \log \pi_i] = -\frac{E_\psi[\pi \log \pi]}{E_\psi[\pi]}, \tag{4.22}$$

and then estimating $D_E(\psi)$ by substituting ψ by a good estimate of it. Tables 4.3 and 4 present the expectation assuming that the π_i are distributed as ψ, $D(\psi)$ and $E(\psi)$, for many other examples of evenness and diversity measures abiding by Definitions 4.3.2 and 4.3.3.

To compare $D(\hat{\psi})$ with other estimators of the diversity note that in our context, $D(\hat{\psi})$ can be judged both as an estimate of $D(\psi) = E_{\psi}[D(\pi_1, \ldots, \pi_v)]$ as well as an estimate of $D(\pi_1, \ldots, \pi_v)$. Hence the performance of $D(\hat{\psi})$ has to be assessed through a three stage process, that includes the study of its variation and of its precision relative to $D(\psi)$ and relative to $D(\pi_1, \ldots, \pi_v)$.

Given a specific measure, $D(\pi_1, \ldots, \pi_v)$, one first has to study the distribution of $(D(\pi_1, \ldots, \pi_v) - D(\psi))$ under $\psi(\pi)$ to help assess the quality of $D(\psi)$ as an approximation to $D(\pi_1, \ldots, \pi_v)$ for each ψ. Second one has look into the distribution of $(D(\hat{\psi}) - D(\psi))$ under ψ for a given estimate $\hat{\psi}$ of ψ; the variance of $D(\hat{\psi})$, the expectation of $(D(\hat{\psi}) - D(\psi))^2$ or any other summary of this distribution helps assess the quality of $D(\hat{\psi})$ as an estimator of $D(\psi)$ for each ψ. Finally the distribution of $(D(\hat{\psi}) - D(\pi_1, \ldots, \pi_v))$ under ψ helps assess the quality of $D(\hat{\psi})$ as an estimator of $D(\pi_1, \ldots, \pi_v)$ for each ψ. Note that in this later case, it is not completely clear how one would compare the performance of $D(\hat{\psi})$ with the one of "nonparametric"estimators of $D(\pi_1, \ldots, \pi_v)$ based on $(\hat{\pi}_{1:n}, \ldots, \hat{\pi}_{v:n})$ but not on ψ, that have to be judged in terms of their distribution given $(\pi_1, \ldots, \pi_v)$.

This assessment of the performance of $D(\hat{\psi})$ has to be made measure $D(\cdot)$ by measure $D(\cdot)$, distribution ψ by distribution ψ, estimator $\hat{\psi}$ by estimator $\hat{\psi}$ and sample size by sample size, which makes it into a complicated program that goes beyond the scope of this manuscript.

4.4.2 Diversity when word frequencies are GIG distributed

Here it is investigated how the expectation of various measures of diversity behave when word frequencies are $GIG(b, c, g)$ distributed. Figures 4.2 and 4.3 present the contour plot of the logarithm of:

$$v(\psi) = \frac{1}{E_{\psi}[\pi]} = \frac{2}{bc} \frac{K_g(b)}{K_{g+1}(b)}, \tag{4.23}$$

and of various $D(\psi)$ and $E(\psi)$ over the range of (b, c, g) most useful in stylometric practice. What is most remarkable is the differing behavior observed when one switches from one measure to another, due to the fact that different measures capture different aspects of what diversity means and weight the parts of the vocabulary distribution differently. For example,

$$D_{GS}(\psi) = 1 - c(1 + g + \frac{bK_g(b)}{2K_{g+1}(b)}) \tag{4.24}$$

is strongly influenced by c and to a lesser degree by g, but it is highly insensitive to the value of b, in contrast with what happens under $D_{r=1}(\cdot)$ or under any evenness measure other than $E_{GS}(\cdot)$, which give a lot less weight to the frequency of the most

abundant words than Gini-Simpson measures and are mostly determined by b. What is common to all diversity measures in Figure 4.2 except the last one is that they are non-increasing with b, c and g. On the other hand, all evenness measures in Figure 4.3 are non-decreasing with b and non-increasing with c and when $g < -.5$ with g.

To illustrate the use of $D(\psi)$ and $E(\psi)$ to estimate evenness and diversity, Tables 4.5 and 4.6 present the estimates of $v(\psi)$ and of various $D(\psi)$ and $E(\psi)$ obtained by replacing ψ by maximum likelihood estimates of ψ for the vocabularies of the authors of the texts in Table 4.1.

The values of $v(\hat{\psi})$ indicate that Max Havelaar is the text from a larger and thus richer vocabulary, followed by the turkish archeology text; they are the only texts considered that are not in English. Among the texts in English, the largest v corresponds to the vocabulary of the War of the Worlds while the smallest v is the one behind the Essays on Bacon, in line with that being the only text where one is only counting names instead of all words.

The vocabulary orderings obtained through different evenness measures can be inconsistent because different measures emphasize different parts of the word frequency distribution, (see, e.g., Hurlbert, 1971, Patil and Taillie, 1982). Nevertheless here most of the evenness measures consistently rank the vocabularies behind these seven texts similarly. Almost all measures agree in that the most even vocabulary corresponds to the Turkish archeology text, the second most even is the one of the essays on Bacon, and then in a close tie come the vocabulary of Alice in Wonderland and of Through the Looking Glass and the one of The War of the Worlds.

Different diversity measures weight the role of v relative to the role of evenness differently and one should expect larger discrepancies among diversity rankings than among evenness rankings. The degree of agreement between the diversity rankings here is indeed weaker than the one between evenness rankings, but almost all measures agree that the most diverse vocabulary is the one behind the turkish archeology text, while the least diverse vocabularies are the ones associated with Alice in Wonderland and Through the Looking Glass.

4.5 Final comments

This article tackled three different issues. On a conceptual level it proposes definitions of measure of the diversity and of measure of the evenness in a population that work when the total number of classes are unknown but known to be different. On an applied

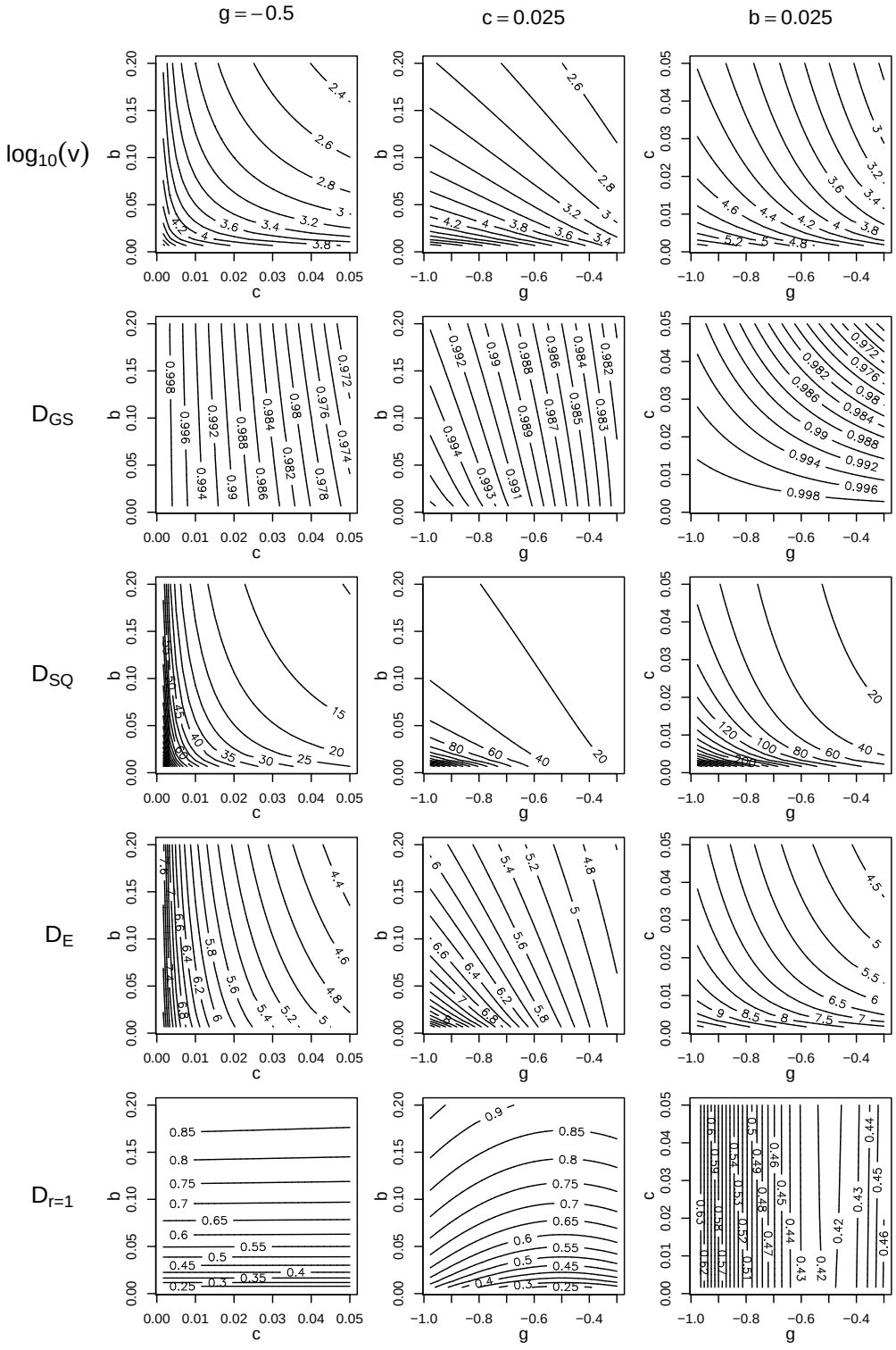

Figura 4.2: Contour plots of $\log_{10} v(\psi)$, $D_{GS}(\psi)$, $D_{SQ}(\psi)$, $D_E(\psi)$ and $D_{r=1}(\psi)$ when the distribution of word frequencies is the GIG(b, c, g). Observe that all these measures except the last one are non-increasing functions of b, c and g.

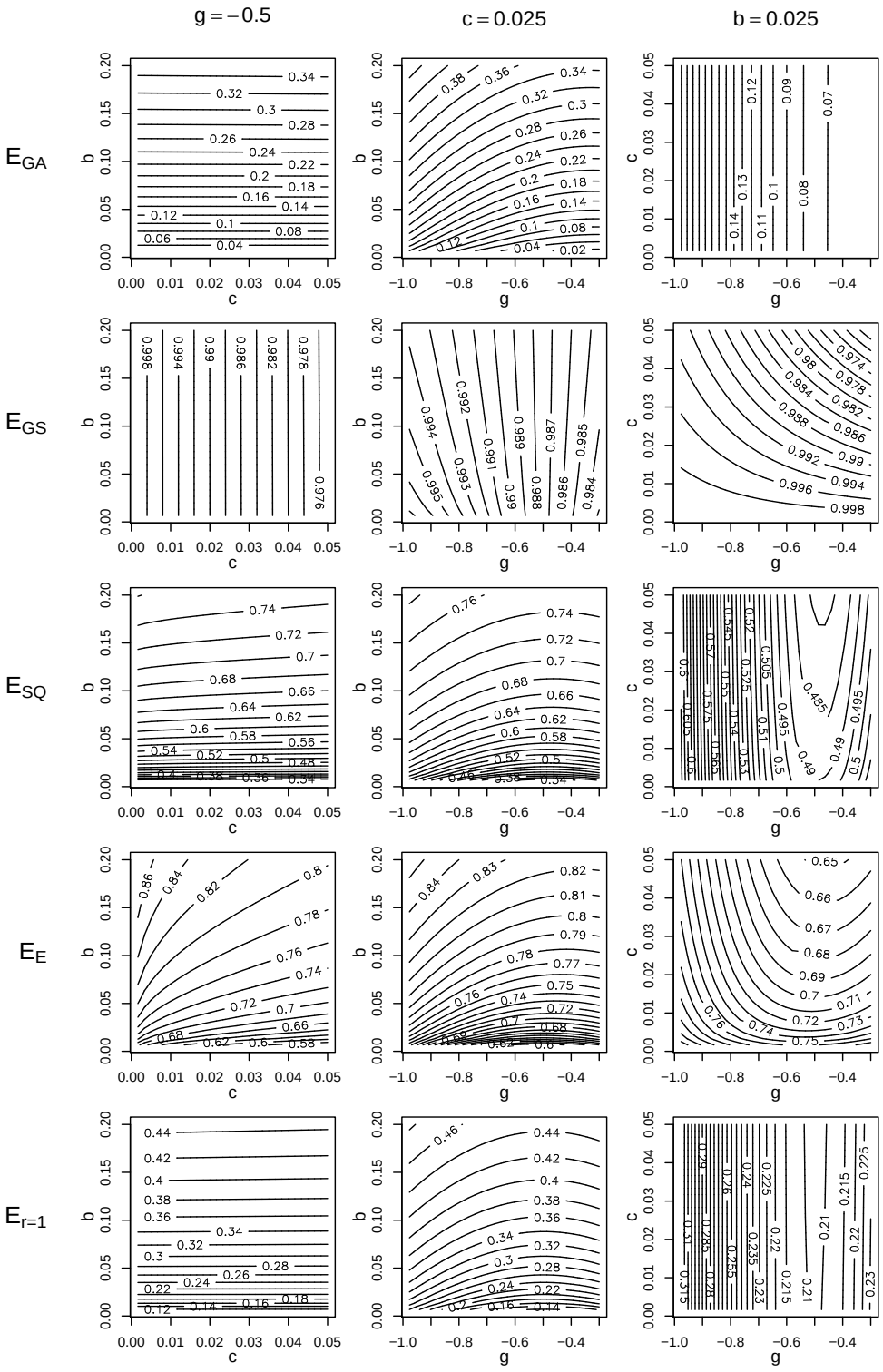

Figura 4.3: Contour plots of $E_{GA}(\psi)$, $E_{GS}(\psi)$, $E_{SQ}(\psi)$, $E_E(\psi)$ and $E_{r=1}(\psi)$ as a function of (b, c, g) when the distribution of word frequencies is the GIG(b, c, g). The left hand side panels correspond to the case where $\psi(\pi)$ is the $IG(b, c)$ distribution. Observe that all these measures are non-decreasing with b and non-increasing with c and when $g < -.5$ with g.

	n	$v(\hat{\psi})$	$D_{GS}(\hat{\psi})$	$D_{SQ}(\hat{\psi})$	$D_E(\hat{\psi})$	$D_{r=1}(\hat{\psi})$
E. Bacon	8049	5651	0.9979	50.129	7.1243	0.7191
Turkish A.	6939	19364	0.9980	97.390	8.1790	0.8058
Alice W.	26505	6199	0.9918	41.224	6.2001	0.4846
Through L.	28767	9228	0.9921	46.792	6.2951	0.4307
Hound B.	59241	16052	0.9913	59.671	6.4948	0.4117
War W.	59938	19564	0.9927	70.322	6.8301	0.4590
Max H.	99767	37753	0.9932	88.853	7.0194	0.3955

Taula 4.5: Estimates of measures of the diversity of the vocabulary of the authors of the texts in Table 4.1, based on the mle of the mixing distribution of the truncated GIG-Poisson model.

	n	$E_{GA}(\hat{\psi})$	$E_{GS}(\hat{\psi})$	$E_{SQ}(\hat{\psi})$	$E_E(\hat{\psi})$	$E_{r=1}(\hat{\psi})$
E. Bacon	8049	0.2374	0.9980	0.6758	0.8246	0.3596
Turkish A.	6939	0.3234	0.9981	0.7049	0.8286	0.4029
Alice W.	26505	0.1191	0.9920	0.5303	0.7100	0.2424
Through L.	28767	0.0959	0.9922	0.4922	0.6895	0.2154
Hound B.	59241	0.0955	0.9913	0.4747	0.6707	0.2059
War W.	59938	0.1198	0.9927	0.5064	0.6912	0.2295
Max H.	99767	0.0931	0.9933	0.4597	0.6661	0.1978

Taula 4.6: Estimates of measures of the evenness of the vocabulary of the authors of the texts in Table 4.1, based on the mle of the mixing distribution of the truncated GIG-Poisson model.

level it was first argued that the advantage in analyzing frequency count data through mixed Poisson models is that they let one estimate the frequency distribution of the population through their mixing distribution. Finally it was argued that lack of diversity in a population should be measured through the variability of these mixing frequency distribution estimates.

To help establish the performance of $D(\hat{\psi})$ one will have to first investigate the distribution $(D(\pi_1, \ldots, \pi_v) - D(\psi))|\psi$ to assess $D(\psi)$ as an approximation to $D(\pi_1, \ldots, \pi_v)$, one will have to study the distribution $(D(\hat{\psi}) - D(\psi))|\psi$ to assess $D(\hat{\psi})$ as an estimate of $D(\psi)$, and one will also have to look into the distribution $(D(\hat{\psi}) - D(\pi_1, \ldots, \pi_v))|\psi$ to assess $D(\hat{\psi})$ as an estimator of $D(\pi_1, \ldots, \pi_v)$. This three stage assessment of $D(\hat{\psi})$ will have to be made $D(\cdot)$ by $D(\cdot)$, ψ by ψ, $\hat{\psi}$ by $\hat{\psi}$ and sample size by sample size.

The truncated GIG-Poisson model was used because it fits word frequency counts extre-

mely well, but one could base this type of analysis on any other Poisson mixture model that fits the data well. In those instances where different Poisson mixture models fit the data equally well, one should investigate the sensitivity of diversity estimates relative to the choice of mixing model. In our framework the diversity of the population is only a function of the expectation and of the variability of that mixing distribution, and hence if the expectation and the variability of the mixing distribution estimates are similar the diversity estimates will be similar.

Even though the focus has been the analysis of word frequency count data, everything applies to the analysis of any type of frequency count where zeros are not observed and v is unknown. One could also adapt this type of analysis to frequency count data that include zeros, and hence where $E_\psi[\pi]$ and v are known, like when one restricts attention to a given subset of words as in Giron et al. (2005). In that case, assessing the diversity in the population through the variability of the mixing distribution estimate of any untruncated Poisson mixture model that fits the data well is tantamount to assessing the degree of overdispersion of the data.

Bibliografia

Ajiferuke I, Wolfram D, Famoye F (2006). Sample size and informetric model goodness-of-fit outcomes: a search engine log case study. *Journal of Information Science*, 32, 212-222.

Atkinson A, Yeh L (1982). Inference for Sichel's compound Poisson distribution. *Journal of the American Statistical Association*, 77, 153-158.

Baayen H (2001). *Word frequency distributions.* Dordretch: Kluwer.

Basharin GP (1959). On a statistical estimate for the entropy of a sequence of indepenent random variables. *heory of Probability and its Applications*, 4, 333-336.

Blyth CB (1959). Note on estimating information. *Ann. Math. Statist.*, 30, 71-79.

Bunge J, Fitzpatrick M (1993). Estimating the number of species: A review. *J. Am. Statist. Ass.*, 88, 364-373.

Böhning D, Kuhnert R (2006). Equivalence of truncated count mixture distributions and mixtures of truncated count distributions. *Biometrics*, 62, 1207-1215.

Bunge J, Fitzpatrick M, (1993). Estimating the number of species: A review. *J. Am. Statist. Ass.*, 88, 364-373.

Burrell QL, Fenton MR (1993). Yes, the GIGP really does work–and is workable!. *Journal of the American Society for Information Science*, 44, 61-69.

Carlson M (2002). Assessing microdata disclosure risk using the Poisson-inverse Gaussian distribution. *Statistics in Transition*, 5, 901-925.

Chao A, Shen TJ (2003). Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environmental and Ecological Statistics*, 10, 429-443.

Chikkara RS, Folks JL (1989). *The inverse Gaussian distribution: theory, methodology and applications*. New York: Marcel Dekker.

DeGroot MH (1962). Uncertainty, information and sequential experiments. *Ann. Math. Statist.*, 33, 404-419.

Engen S (1974). On species frequency models. *Biometrika*, 61, 263-270.

Ginebra J (2007). On the measure of the information in a statistical experiment. *Bayesian Analysis*, 2, 167-212.

Ginebra J, Puig X (2009). On the measure and the estimation of the evenness and diversity of vocabulary. *Submitted for publication.*

Giron J, Ginebra J, Riba A (2005). Bayesian analysis of a multinomial sequence and homogeneity of literary style. *The American Statistician*, 32, 61-74.

Good IJ (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, 40, 237-264.

Good IJ (1982). Comment to "Diversity as a concept and its measurement". *Journal of the American Statistical Association*, 77, 561-563.

Griffiths DA (1973). Maximum likelihood estimation for the beta-binomial distribuiton and an application to the household distribution of the total of cases of a disease. *Biometrics*, 29, 637-648.

Heller G (1997). Estimation of the number of classes. *South African Statistical Journal*, 31, 65-90.

Herdan G (1961). A critical examination of Simon's model of certain distribution functions in linguistics. *Applied Statistics*, 10, 65-76.

Herdan G (1964). *Quantitative linguistics*. London: Butterworth.

Hill MO (1973). Diversity and evenness: A unifying notation and its consequences. *Ecology*, 54, 427-432.

Holla MS (1966). On a Poisson-inverse Gaussian distribution. *Metrika*, 11, 115-121.

Holmes DI (1985). The analysis of literary style:A review, *J. R. Statist. Soc. A*, 148, 328-341.

Holmes DI (1992). A stylometric analysis of mormon scripture and related texts. *Journal of the Royal Statistical Society, Series A*, 155, 91-120.

Holmes DI, Forsyth RS (1995). The Federalist revisited. New directions in authorship attribution. *Literary and Linguistics Computing*, 10, 111-127.

Hoshino N (2005). Engen's extended negative binomial model revisited. *Annals of the Institute of Statistical Mathematics*, 57, 369-387.

Hurlbert SH (1971). The nonconcept of species diversity: a critique and alternative parameters. *Ecology*, 52, 577-586.

Izsak J, Papp L (2000). A link between ecological diversity indices and measures of biodiversity. *Ecological Modelling*, 130, 151-156.

Johnson N, Kotz S (1994). *Univariate continuous distributions.* New York: Wiley.

Johnson N, Kotz S, Kemp A (1993). *Univariate discrete distributions.* New York: Wiley.

Jorgensen B (1982). *Statistical properties of the generalized inverse Gaussian distribution* New York: Wiley.

Karlis D (2001). A general EM approach for maximum likelihood estimation in mixed Poisson regression models. *Statistical Modelling*, 1, 305-318.

Kempton RA (1979). The structure of species abundance and measurement of diversity. *Biometrics*, 35, 307-321.

Kempton RA, Wedderburn RWM (1978). A comparison of three measures of species diversity. *Biometrics*, 34, 25-37.

Klugman SA, Panjer HH, Willmot GE (1998). *Loss models. From data to decisions.* New York: Wiley.

Lyons NI, Hutcheson K (1986). Estimation of Simpson's diversity when counts follow a Poisson distribution. *Biometrics*, 42, 171-176.

Magurran AE (2004). *Measuring Biological Diversity.* Blackwell, New York.

Marshall AW, Olkin I (1979). *Inequalities: Theory of Majorization and its Applications.* Academic Press, New York.

Muller A, Stoyan D (2002). *Comparison Methods for Stochastic Models and Risk.* Wiley, New York.

Ord JK, Whitmore G (1986). The Poisson-inverse Gaussian distribution as a model for species abundance. *Communications in Statistics, Theory and Methods*, 15, 853-871.

Patil GP, Taillie C (1982). Diversity as a concept and its measurement (with discussion). *J. Am. Statist. Ass.*, 77, 548-567.

Pielou EC (1975). *Ecological Diversity*. New York: Wiley.

Pollatschek M, Radday YT (1981). Vocabulary richness and concentration in Hebrew biblical literature. *Association for Literary and Linguistical Computing Bulletin*, 8, 217-231.

Puig X, Ginebra J, Perez-Casany M (2009). Extended truncated inverse Gaussian-Poisson model. *Statistical Modelling*, 9,151-171.

Puig X, Ginebra J, Font M (2009). The Sichel Model and the Mixing and Truncation Order. *Journal of Applied Statistics.*, to appear.

Rao CR (1982). Diversity and dissimilarity coefficients: A unified approach. *Theoretical Population Biology*, 21, 24-43.

Rao CR (1984). Rao's axiomatization of diversity measures. In: *Encyclopedia of Statistics*, Vol 7 (Eds. Johnson NL, Kotz S, Read CB), 614-617. Wiley, New York.

Rényi A (1961). On measures of entropy and information. In: *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (Ed. Neyman J), 547-561. University of California Press, Berkeley.

Riba A (2002). Estadística i Homogeneïtat d'estil al Tirant lo Blanc (in catalan). *Unpublished PhD Thesis*, Technical University of Catalonia.

Riba A, Ginebra J (2005). Change-point estimation in a multinomial sequence and homogeneity of literary style. *Journal of Applied Statistics*, 32, 61-74.

Riba A, Ginebra J (2006). Diversity of vocabulary and homogeneity of literary style. *Journal of Applied Statistics*, 33, 729-741.

Ricotta C (2005). Through the jungle of biological diversity. *Acta Biotheoretica*, 53, 29-38.

Ricotta C, Szeidl L (2006). Towards a unifying approach to diversity mesures: Bridging the gap between Shannon entropy and Rao's quadratic index. *Theoretical Population Biology*, 70, 237-243.

Routledge RD (1979). Diversity indices: Which ones are admissible? *Journal of Theoretical Biology*, 76, 503-515.

Sankaran M (1968). Mixtures by the inverse Gaussian distribution. *Sankhya, Series A*, 30, 455-458.

Sen A (1974). Poverty, inequality and unemployment: Some conceptual issues in measurement. *Sankhya C*, 36, 67-82.

Seshadri V (1993) *The inverse Gaussian distribution: A case study in exponential families.* Oxford: Clarendon Press.

Seshadri V (1998.) *The inverse Gaussian distribution: Statistical theory and applications.* New York: Springer Verlag.

Shaked M, Shanthikumar JG (1994). *Stochastic orders and their applications.* Boston: Academic Press.

Sichel HS (1971). On a family of discrete distributions particularly suited to represent long-tailed frequency data. In Laubscher, N.F. ed. *Proceedings of the third symposium on mathematical statistics.* Pretoria: C.S.I.R, 51-97.

Sichel HS (1973). The density and size distribution of diamonds. *Bulletin of the International Statistical Institute*, 45, 420-427.

Sichel HS (1974). On a distribution representing sentence-length in written prose. *Journal of the Royal Statistical Society, Series A*, 137, 25-34.

Sichel HS (1975). On a distribution law for words frequencies. *Journal of the American Statistical Association*, 70, 542-547.

Sichel HS (1982a). Asymptotic efficiencies of three methods of estimation for the inverse Gaussian-Poisson distribution. *Biometrika*, 69, 467-472.

Sichel HS (1982b). Repeat-buying and the generalized inverse Gaussian-Poisson distribution. *Applied Statistics*, 31, 193-204.

Sichel HS (1985). A bibliometric distribution that really works. *Journal of the American Society for Information Science*, 36, 314-321.

Sichel HS (1986a). Word frequency distributions and type-token characteristics. *Mathematical Scientist*, 11, 45-72.

Sichel HS (1986b). Parameter estimation for a word frequency distribution based on occupancy theory. *Communications in Statistics, Theory and Methods*, 15, 935-949.

Sichel HS (1992a). Anatomy of the generalized inverse gaussian-Poisson distribution with special applications to bibliometric studies. *Information Processing and Management*, 28, 5-17.

Sichel HS (1992b). Note on a strongly unimodal bibliometric size frequency distribution. *Journal of the American Society for Information Science*, 43, 299-303.

Sichel HS (1997). Modelling species-abundance frequencies and species-individual functions with the generalized inverse Gaussian-Poisson distribution. *South African Statistical Journal*, 31, 13-37.

Smith W, Grassle JF (1977). Sampling properties of a family of diversity measures. *Biometrics*, 33, 283-292.

Solomon DL, (1979). A comparative approach to species diversity. In: *Ecological Diversity in Theory and Practice* (Eds. Grassle, J.F., Patil, G.P., Smith, W., Taillie, C.), 29-35. International Cooperative Publishing House, Maryland.

Solow A, Polasky S, Broadus J (1993). On the measurement of biological diversity. *Journal of Environmental Economics and Management*, 24, 60-68.

Shoukri MM, Asyali MH, VanDorp R and Kelton D (2004). The Poisson inverse Gaussian regression model in the analysis of clustered counts data. *Journal of Data Science*, 2, 17-32.

Stein GZ, Juritz JM (1988). Linear models with an inverse Gaussian-Poisson error distribution. *Communications in Statistics: Theory and Methods*, 17, 557-571.

Stein GZ, Zucchini W, Juritz JM (1987). Parameter estimation for the Sichel distribution and its multivariate extension. *Journal of the American Statistical Association*, 82, 938-944.

Tong YL (1983). Some distribution properties of the sample species-diversity indices and their applications. *Biometrics*, 39, 999-1008.

Tremblay L (1992). Using the Poisson inverse Gaussian in bonus-malus systems. *ASTIN Bulletin*, 22, 97-106.

Valero J, Ginebra J, Perez-Casany M (2009). Extended truncated Tweedie-Poisson model. Manuscript submitted for publication.

Willmot GE (1986). Mixed compound Poisson distributions. *ASTIN Bulletin*, 16, 59-79.

Willmot GE (1987). The Poisson-inverse Gaussian distribution as an alternative to the negative binomial. *Scandinavian Actuarial Journal*, 2, 113-127.

Willmot GE (1988a). Parameter orthogonality for a family of discrete distributions. *Journal of the American Statistical Association*, 83, 517-521.

Willmot GE (1988b). A remark on the Poisson Pascal and some other contagious distributions. *Statistics and Probability Letters*, 7, 217-220.

Willmot GE (1988c). Sundt and Jewell's family of discrete distributions. *ASTIN Bulletin*, 18, 17-29.

Yule GU (1944). *The statistical study of literary vocabulary*. London: Cambridge University Press.

Zipf GK (1932). *Selected studies of the principle of relative frequency in language*. Cambridge: Harvard University Press

Part II

Anàlisi Cluster Multinomial Bayesià i Dades Electorals

Capítol 5

Introducció

L'anàlisi estadística basada en models bayesians passa per: a) formular el model estadístic i escollir la distribució a priori, b) càlcular o simular de la distribució a posteriori o predictiva a posteriori, c) validar els models, detectar anomalies i així permetre seleccionar models de forma indirecta, i d) presentar les distribucions a posteriori o predictiva a posteriori de forma clara, concisa i informativa, sovint a través de gràfics o taules.

La segona part de la Tesi aborda aquests problemes en el context de l'anàlisi cluster sobre variables categòriques espacials. En aquesta part farem servir els resultats de les darreres cinc eleccions al Parlament de Catalunya com a fil conductor. El fons metodològic però, és extensible a molts altres problemes.

La diversitat de partits polítics que obtenen representació a Catalunya donen lloc a un mapa polític d'una gran riquesa. Al capítol 6 presentarem les dades i posarem de manifest aquesta riquesa mitjançant una anàlisi descriptiva.

Al capítol 7 presentarem els models cluster multinomial Bayesians que utilitzarem per realitzar una anàlisi que permeti identificar àrees amb comportament homogeni, descriure patrons de vot, estimar la probabilitat de cada àrea de pertànyer a cada un dels patrons de vot, i eventualment monitoritzar l'evolució de tots aquest paràmetres al llarg de les diferents eleccions. Al Capítol 8 presentarem eines per validar i per comparar els diferents models considerats, tant jeràrquics com no jeràrquics. Les eines de validació ens permetran descartar els models que no reprodueixen prou bé característiques rellevants de les dades, i per tant permetran seleccionar els models de forma indirecta.

Al Capítol 9 presentarem el procés de comparació i validació dels models per els resultats

de les eleccions del 2003 a Barcelona ciutat i al Capítol 11 compararem i validarem aquests models per els resultats de les eleccions de 1992, 1995, 1999, 2003 i 2006. En totes les eleccions s'han descartat els models no jeràrquics perquè no capturen prou bé la variabilitat dels resultats, i s'ha acabat seleccionant com a models més adequats els models jeràrquics de tres i quatre clusters. El Capítol 10 presenta la interpretació del model de tres i quatre clusters a les eleccions del 2003 a Barcelona ciutat, i el Capítol 12 repeteix el mateix per a totes les eleccions considerades.

El dia després d'unes eleccions és habitual especular sobre el transvasament de vots d'un partit a un altre. Per modelar aquest fenomen cal separar primer les àrees en grups homogenis en quant a comportaments de vot, i això és justament el que hem fet en aquesta tesi. Extendre els nostres models, de forma que permetin modelar dues eleccions consecutives alhora que permetin estimar les matrius de transició entre les diferents opcions electorals és una de les continuacions de la tesi que descriurem al Capítol 13. Finalment, a l'Apèndix C s'apunta quines són les dificultats d'extendre aquest tipus d'anàlisi als resultats electorals de tot Catalunya.

Capítol 6

Descripció de les dades

6.1 Introducció

Aquesta segona part de la tesi gira al voltant de l'anàlisi de dades electorals, i més concretament, dels resultats de les eleccions al parlament de Catalunya dels anys 1992, 1995, 1999, 2003 i 2006. En aquest capítol presentarem les dades, així com una anàlisi descriptiva a nivell primer de Barcelona ciutat i a continuació a nivell de tot Catalunya.

El Parlament és la institució que representa el poble de Catalunya. Aquest Parlament està format per una sola cambra que exerceix la potestat legislativa, aprova els pressupostos, impulsa i controla l'acció política i de Govern i exerceix les restants competències que li són atribuïdes per l'ordenament jurídic i, en especial, per l'Estatut d'autonomia. Fins a l'actualitat s'han celebrat 9 eleccions al Parlament de Catalunya; les primeres van ser l'any 1932, però les segones no es van poder celebrar fins l'any 1980.

A les eleccions al parlament de Catalunya s'elegeixen per sufragi universal els 135 diputats que l'integren. Cadascuna de les quatre circumscripcions elegeix els seus diputats que li corresponen segons l'Estatut d'autonomia:

1. Barcelona: 85 diputats,
2. Girona: 17 diputats,
3. Lleida: 15 diputats, i
4. Tarragona: 18 diputats

La fórmula electoral utilitzada és la Llei d'Hondt i s'aplica un cop superada la barrera

legal del 3% dels vots vàlids a cada circumscripció necessària per tenir dret a tenir al primer diputat.

6.2 Estructura dels partits

Des de les primeres eleccions a Catalunya després de la reinstauració de la democràcia, a l'any 1980, les llistes electorals que s'han presentat han anat canviant. Aquest és un dels motius pels que hem acotat l'anàlisi des del 1992 fins a l'actualitat amb la finalitat de treballar en un context més homogeni. Així, en comicis anteriors al 1992 trobem representats al parlament de Catalunya partits extingits com la *Unión de Centro Democrático*, el *Partit Socialista Unificat de Catalunya* (PSUC), *Centro Democrático y Social*, *Alianza Popular* i el *Partido Socialista Andaluz*. Un altre motiu important per començar l'anàlisi al 1992, i no abans, han estat els canvis que han sofert les unitats geogràfiques amb les que treballarem.

El període estudiat és molt més estable pel que fa a l'estructura dels partits tot i que també hi trobem algunes petites variacions pel que fa a les llistes que han obtingut representació parlamentària.

Així, l'any 1992 *Iniciativa per Catalunya* es presenta sol, sense formar l'actual coalició amb els *Verds* ni amb el *Partit Comunista de Catalunya* (PCC). L'any 1995 *Iniciativa per Catalunya*, *Verds* i el PCC es presenten conjuntament formant la coalició amb sigles ICV. Al mateix temps els anys 2003 i 2006 ICV també presenta llistes conjuntes amb *Esquerra Unida i Alternativa* (EUiA), partit fundat l'any 1998, els membres del qual provenien principalment del PCC i del "PSUC-viu". L'any 1999 ICV només es presenta com a llista pròpia a la circumscripció de Barcelona i separats dels seus actuals socis EUiA, i ho fa conjuntament amb el PSC a la resta de circumscripcions. En aquest cas hem optat per imputar els vots d'ICV fora de Barcelona, vots que s'han restat del PSC, en base a la informació de Barcelona de l'any 1999, així com de la informació de Barcelona i Catalunya dels anys anteriors i posteriors, a la vegada que hem sumat els seus vots amb els d'EUiA.

Un altre fet rellevant és l'aparició l'any 2006 d'un nou partit, *Ciutadans per Catalunya* (Cs) que obté representació política, si bé només a la circumscripció de Barcelona.

Tots els partits amb menys d'un 1% del global de vots els hem agregat en la categoria d'*altres*. Aquests partits també són tots els que no han obtingut cap diputat en cap de les quatre circumscripcions, excepte EUiA que l'any 1999 treu l'1,4% dels vots i no obté

representació parlamentària i que hem sumat a la categoria ICV.

Per a les anàlisis hem optat per treballar amb l'agregació dels vots nuls i dels vots en blanc, ja que representaven pocs casos i estan molt correlacionats, cosa que no passa amb l'abstenció, tal i com hem pogut comprovar en anàlisis preliminars.

Així les categories que considerarem en les nostres anàlisis seran:

1. CIU: Convergència i Unió,
2. PSC: Partit Socialista de Catalunya,
3. PPC: Partit Popular de Catalunya,
4. ICV: Iniciativa per Catalunya els verds (+ EUIA els anys 1999, 2003 i 2006),
5. ERC: Esquerra Republicana de Catalunya,
6. Cs: Ciutadans per Catalunya (només l'any 2006),
7. *altres*: Agregació dels partits amb menys d'un 1% dels vots,
8. *b+n*: blancs i nuls, i
9. *abs*: abstenció.

Els sis partits de la llista és poden classificar en el bloc sobiranista, el bloc unionista i ICV. Considerarem com a bloc sobiranista ERC i CIU, i com a bloc unionista PSC, PPC, i Cs. Una altra classificació natural seria separar els partits en centre-esquerra i centre-dreta, considerant que el bloc d'esquerres estaria format per ICV, ERC i PSC i el bloc de dretes per CIU i PPC. Som conscients del simplisme d'aquestes classificacions que no sempre responen a una realitat complexa i plena de matisos.

6.3 Les unitats geogràfiques

Les unitats geogràfiques seleccionades per a l'anàlisi han estat els districtes, excepte a Barcelona ciutat, on degut a les seves característiques i volum de població, hem cregut que en comptes d'utilitzar els 10 districtes seria millor treballar amb una divisió territorial més fina com són les *zones de recerca petites* (zrp).

Les zrp són una divisió territorial creada per l'ajuntament de Barcelona l'any 1984 amb finalitats estadístiques, i que no s'han modificat des de la seva creació. Es poden trobar més detalls sobre les mateixes a: *http://www.bcn.cat/estadistica*. No obstant a partir de l'any 2009 l'ajuntament de Barcelona té previst utilitzar una nova divisió territorial

construïda a partir dels 73 barris de Barcelona que consisteix en 233 àrees estadístiques bàsiques i que substituiran les 248 zrp.

Per simplificar el llenguatge parlarem sovint de districtes però hem de tenir present que a Barcelona ciutat la unitat d'anàlisi sempre seran les zrp. Malgrat que als comicis anteriors a l'any 1984 les zrp encara no estaven definides, es podria disposar de les dades amb força fiabilitat mitjançant un convertor en base a les seccions censals, però al truncar les nostres anàlisis a l'any 1992 no hem hagut de recórrer a aquestes aproximacions.

Les dades electorals a nivell de zrp que facilita i que publica el departament d'estadística de l'ajuntament de Barcelona es tracten de resultats oficiosos que poden tenir petites discrepàncies amb els resultats oficials que a posteriori publica l'IDESCAT a nivell de districte.

Inicialment, havíem plantejat l'anàlisi a nivell de secció censal, ja que a priori hom la identifica com la unitat natural, més fina i, per tant, més homogènia que els districtes. El problema és que les seccions censals estan regulades pels criteris de manteniment marcats a la Llei Orgànica sobre Règim Electoral General sobre la distribució dels electors d'un municipi per seccions censals, i aquest seccionat és variable en el temps, depenent dels canvis de població. Des de l'any 1991 les seccions censals han sofert diverses modificacions, com a resultat de les revisions del cens electoral. Per exemple, a Barcelona ciutat l'any 1992 hi havien 1811 seccions censals, mentre que a l'any 2006 tant sols n'hi havia 1483 i en la darrera revisió que entra en vigor aquest any 2009 el nombre de seccions censals ha disminuït a 1061. Aquesta variabilitat ens ha portat a escollir els districtes com a la unitat d'anàlisi, ja que ens ofereix una divisió territorial relativament estable i alhora raonablement detallada, i més si tenim en compte que a Barcelona, que és on hi ha els districtes més grans, baixem el nivell de desagregació fins a les zrp. Això ens permetrà estudiar molt més còmodament els canvis que hi ha d'unes eleccions a les següents.

No obstant això, durant els anys subjectes que analitzem també hi ha hagut alguns canvis de districtes i s'ha segregat algun municipi. Amb la finalitat de tenir un divisió comuna a tots els anys de cara a modelar l'evolució temporal dels resultats, hem optat per agregar totes les parelles de municipis segregats en algun moment entre 1992 i 2006. Aquestes parelles són: Riu de Cerdanya i Bellver de Cerdanya segregats l'any 1998, La Palma de Cervelló i Cervelló segregats també l'any 1998, i Sant Julià de Cerdanyola i Guardiola de Berguedà segregats l'any 1996. Hem tractat diferent el cas de Badia del Vallès que es va segregar l'any 1994 de Barberà del Vallès i de Cerdanyola del Vallès, extingint d'aquesta manera la mancomunitat entre aquestes dues poblacions. Seguint el mateix criteri hauríem d'agregar en una les tres poblacions, però donades les seves

grandàries i característiques hem cregut oportú imputar els seus valors per a l'any 1992 i tractar-los sempre com a tres municipis independents.

Els canvis de districtes entre 1992 i 2006 s'han generat tots per municipis en els que s'han creat nous districtes: Cardedeu passa de 2 a 3 districtes, Esplugues de Llobregat de 5 a 10, Sant Joan Despí de 4 a 5, Torelló de 2 a 3, Martorell de 3 a 4, Sant Feliu de Llobregat de 6 a 7 i Calafell de 2 a 3. En aquests 7 municipis hem agregat tots els districtes i hem tractat cadascun dels municipis com un sol districte.

A la Figura 6.1 es mostra la divisió territorial utilitzada per a l'anàlisi, a nivell de districtes per Catalunya i de zrp per Barcelona ciutat. A la figura s'han marcat també els districtes de Barcelona i s'aprecia l'estructura aniuada, en el sentit que cada una de les zrp pertany a un i només un districte. Treballar amb les zrp ens permet augmentar notablement el nombre de particions a Barcelona, passant de 10 districtes a 248 zrp. Pel que fa al total de Catalunya, excloent Barcelona, tenim 1199 districtes, la majoria dels quals representen la totalitat d'un municipi, mentre que les ciutats més grans estan integrades per varis districtes.

6.4 Anàlisi descriptiva a nivell de Barcelona ciutat

Si la circumscripció de Barcelona elegeix a 85 dels 135 diputats que integren el parlament de Catalunya, la ciutat de Barcelona actualment representa el 30% dels electors del total de la província, si bé al 1992 aquest percentatge era del 37%. Aquest fet, juntament amb el fet que sigui la ciutat més gran de Catalunya, i que disposem d'un nivell de desagregació notable de les dades fan que sigui interessant la seva anàlisi de forma separada al total de Catalunya. A la Taula 6.1 hi trobem la llista de districtes així com el nombre de zrp pertanyents a cada districte.

Amb la finalitat de visualitzar les dades a la Taula 6.2 presentem parcialment els resultats a les eleccions al Parlament de Catalunya de la ciutat de Barcelona per a l'any 2003, que donen lloc a una taula de contingència de 248 files i 8 columnes. Per exemple, la primera fila de la taula ens diu que la zrp 1, que pertany al districte de Ciutat Vella i que correspon a un sector de la Barceloneta, té un total de 1510 electors, dels quals 195 han votat CIU, 375 han votat PSC, i així successivament per a la resta de formacions polítiques; en aquesta zrp de la Barceloneta 4 electors han votat *altres*, que són els partits que han obtingut menys d'1% dels vots, 15 electors han votat en blanc o bé el seu vot ha estat considerat nul, i finalment 701 s'han abstingut.

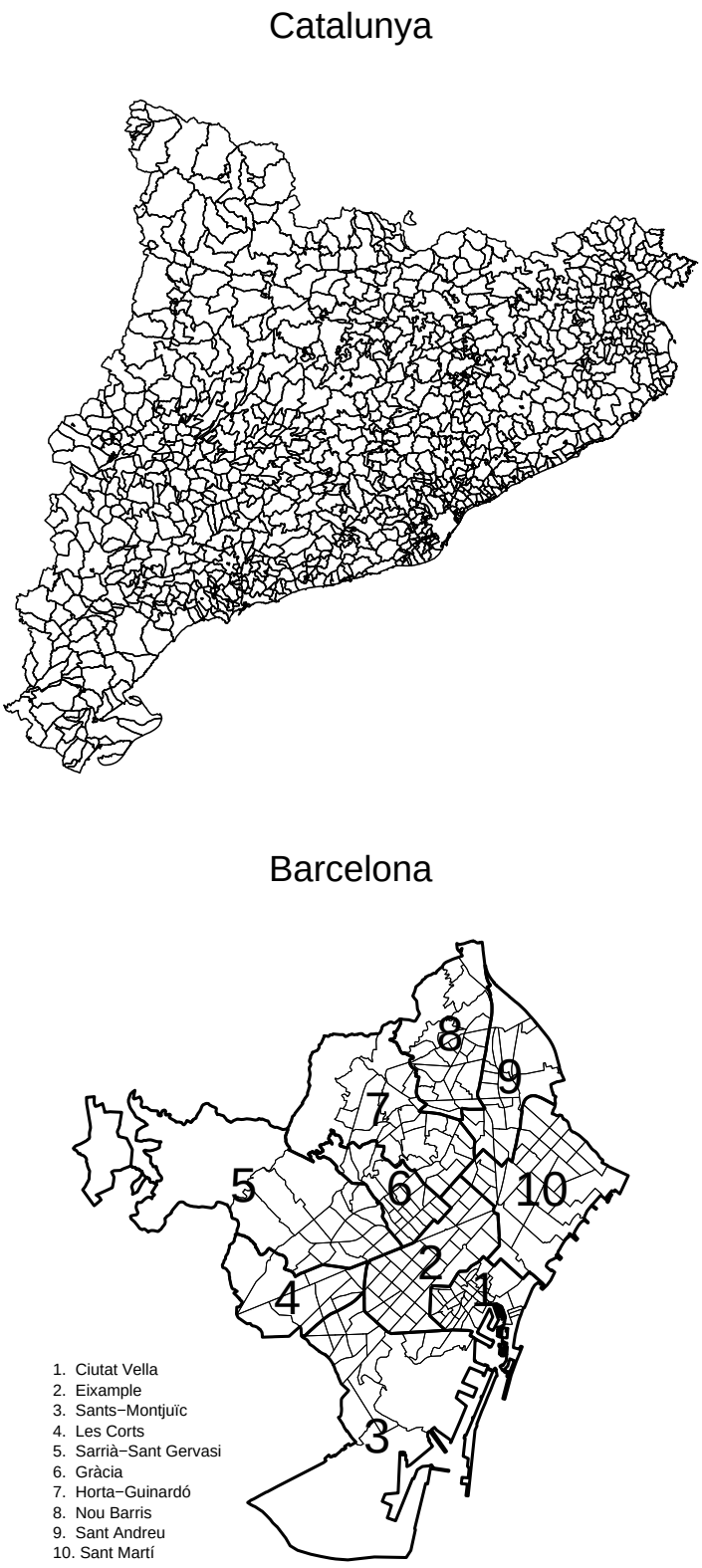

Figura 6.1: Divisió territorial utilitzada per a les anàlisi. A dalt, Catalunya a nivell de districte. A baix, Barcelona a nivell de zrp i, en un contorn més gruixut, de districte.

Codi Districte	Nom Districte	zrp	nº de zrp
1	Ciutat Vella	1-37	37
2	Eixample	38-72	35
3	Sants-Montjuïc	73-90	18
4	Les Corts	91-101	11
5	Sarrià-Sant Gervasi	102-121	20
6	Gràcia	121-144	23
7	Horta-Guinardó	145-171	27
8	Nou Barris	172-197	26
9	Sant Andreu	198-218	21
10	Sant Martí	219-248	30

Taula 6.1: Districtes als que pertanyen les 248 zrp de Barcelona ciutat i nombre de zrp pertanyents a cadascun dels districtes.

Districte	zrp	CIU	PSC	PPC	ICV	ERC	*altres*	*b+n*	*abs*	*N*
1	1	195	375	76	58	86	4	15	701	1510
1	2	208	333	75	70	97	20	6	790	1599
1	3	226	307	70	54	133	8	11	635	1444
1	4	484	675	166	127	263	39	16	1384	3154
...	...	...	...	...	...	...	...	...	...	...
1	36	346	329	148	168	157	30	16	715	1909
1	37	94	89	43	48	53	8	5	240	580
2	38	2095	1541	773	545	1157	72	49	3118	9350
...	...	...	...	...	...	...	...	...	...	...
10	247	471	1489	555	195	235	50	48	2185	5228
10	248	441	1535	592	229	245	48	34	2202	5326

Taula 6.2: Part de la taula dels resultats a les eleccions al Parlament de Catalunya de l'any 2003 per a la ciutat de Barcelona. *N* és el nombre total d'electors, *abs* és l'abstenció, *b+n* són el total de vots blancs i nuls, *altres* és l'agregació de tots els partits amb menys d'1% del global de vots, i la resta de categories respon a les sigles dels respectius partits.

Any	CIU	PSC	PPC	ICV	ERC	Cs	*altres*	*b+n*	*abs*	*N*
1992	360537	178355	65203	53454	57300	-	33244	12906	610489	1371488
1995	352797	191187	156026	88340	82531	-	8047	12186	461028	1352142
1999	281171	304928	103583	36385	66770	-	9962	10796	490046	1303641
2003	227783	249020	123163	69234	126626	-	9813	9482	407294	1222415
2006	208753	172447	93932	84713	86050	31951	15607	17864	462276	1173593

Taula 6.3: Nombre de vots a la ciutat de Barcelona per a cadascuna de les opcions i el nombre d'electors (N) per a cadascuna de les darreres cinc eleccions al parlament de Catalunya. Les dades presentades a la taula han estat prèviament processades de manera que a l'any 1999 als vots de ICV se li han sumat els de EUiA.

Any	CIU	PSC	PPC	ICV	ERC	Cs	*altres*	*b+n*	*abs*
1992	0.263	0.130	0.048	0.039	0.042	-	0.024	0.009	0.445
1995	0.261	0.141	0.115	0.065	0.061	-	0.006	0.009	0.341
1999	0.216	0.234	0.079	0.028	0.051	-	0.008	0.008	0.376
2003	0.186	0.204	0.101	0.057	0.104	-	0.008	0.008	0.333
2006	0.178	0.147	0.080	0.072	0.073	0.027	0.013	0.015	0.394

Taula 6.4: Percentatge de vots a la ciutat de Barcelona per a cadascuna de les opcions per a cadascuna de les darreres cinc eleccions al parlament de Catalunya. Les dades presentades a la taula han estat prèviament processades de manera que a l'any 1999 als vots de ICV se li han sumat els de EUiA.

A tota anàlisi estadística és important una primera anàlisi de caire descriptiu. A la Taula 6.3 hi presentem els vots totals per a cada categoria en cada any; això ens permet copsar la disminució del nombre d'electors en els successius anys. La disminució del nombre d'electors respon a la despoblació que ha sofert Barcelona en les dues darreres dècades, principalment degut a que la gent jove n'ha hagut de marxar per la dificultat que han suposat els preus a l'hora d'accedir a un habitatge. I a la taula 6.4 hi presentem el percentatge de vots de cada categoria en cada any.

A la Figura 6.2 presentem la distribució dels percentatges a nivell de zrp entre les diferents categories per a cadascun dels anys mitjançant diagrames de caixes. En aquesta figura hi destaca que en les dues primeres eleccions estudiades el partit més votat amb diferència era CIU, mentre que a les darreres tres eleccions s'iguala amb el PSC. Crida l'atenció el comportament de PPC, ICV i ERC que al 1992 presentaven percentatges molt similars i, després d'oscilacions a les últimes eleccions tornen a prendre percentatges similars, però amb percentatges més alts que el 1992.

La Figura 6.3 és una reordenació de la Figura 6.2 amb l'objectiu d'observar l'evolució de la distribució percentual per a cadascuna de les opcions de vot; hi on destaca la contínua però constant davallada de CIU, la pujada del PSC l'any 1999, que coincideix amb la candidatura d'en Pasqual Maragall, i la notable pujada d'ERC l'any 2003. Els valors més alts del PPC que s'observen en totes les eleccions corresponen majoritàriament a zrp dels districtes de Les Corts i de Sarrià-Sant Gervasi. Pel que fa a l'abstenció, destaca l'alta abstenció d'una zrp del districte de Sants-Montjuïc ubicada al sector de la Zona Franca-Port, que l'any 1999 va arribar a ser del 88%.

Tant a la mateixa Figura 6.2 com a la Figura 6.3 hi trobem representat amb un punt el percentatge marginal el qual, apareix centrat en els diagrames de caixes degut a la relativa homogeneïtat de grandàries a nivell d'electors de les zrp; en el cas de Catalunya aquest punt apereixerà descentrat, indicant que el patró de vot a nivell de tot el principat està

molt relacionat amb el nombre d'electors del districte, fet que hi complica molt l'anàlisi.

Per tenir idea de la distribució de les grandàries de les zrp, la Figura 6.4 presenta l'-histograma per la distribució del nombre d'electors per a l'any 2006. El 95% de les zrp tenen entre 700 i 9500 electors i destaca una zrp del districte de Sants-Montjuïc amb més de 15000 electors. Els diagrames bivariants de la figura 6.5 relacionen gràficament el percentatge de vot de cadascuna de les opcions en front la grandària de les zrp per a l'any 2006, manifestant molt poca dependència entre aquests percentatges i N.

A la Figura 6.6 representem la distribució espacial del percentatge de vot per a les quatre llistes més votades, CIU, PSC, PPC i ERC, així com per a l'abstenció per el 1992 i el 2006. Per a la representació s'han categoritzat els percentatges de cada partit segons les quartiles del 1992. D'aquesta manera visualitzem conjuntament la distribució espacial i els canvis ocorreguts entre la primera i la última elecció considerada. En aquesta Figura s'observa com CIU obté els percentatges més alts a la part central de Barcelona, als districtes de Les Cort, Sarrià-Sant Gervasi, l'Eixample i Gràcia, i el PPC també obté els seus percentatges més alts en aquests districtes, a excepció de Gràcia. En quant al patró de l'*abstenció*, és complementari al del CIU, i en quan al PSC, obté els percentatges més alts als districtes del nord, especialment a Nou Barris i Sant Andreu. Finalment, les zrp amb percentatges alts d'ERC és distribueixen força homogèniament per a tots els districtes a excepció de Nou Barris.

Donat que les dades formen una taula de contingència, una anàlisi exploratoria força informativa consisteix en realitzar una anàlisi de correspondències, que és una eina gràfica que captura part de l'estructura subjacent a les dades i que identifica en què són diferents i com s'agrupen les files (o columnes) d'una taula de contingència com la Taula 6.2. El que volem és esbrinar quins perfils fila (o columna) s'assemblen, però aquests estan situats en un espai de dimensió 248 (8, o 9 al 2006) on és difícil veure què queda a la vora de què. L'anàlisi de correspondències simple identifica un subespai de dimensió dos que captura el màxim d'informació sobre tots els perfils fila (248) i perfils columna (8 o 9), i hi projecta aquests perfils per veure com s'hi agrupen. Una mesura de la qualitat de la projecció és el percentatge d'inèrcia representat per a cada una de les dues dimensions del subespai.

Els resultats es presenten via gràfic simètric en el que les distàncies entre fila i fila i les distàncies entre columna i columna aproximen les distàncies χ^2 entre files i entre columnes, però les distàncies entre files i columnes no indiquen el grau d'associació entre fila i columna i, per tant, no són directament interpretables; en canvi, punts fila i columna apareixen igualment repartits per tota l'àrea del gràfic. El que sempre val és interpretar una dimensió cada vegada, fent servir les posicions relatives dels punts columna (fila) per

atribuir els factors que determinen cada un des eixos i observant les posicions dels punts fila (columna) relatives al mateix eix. Greenacre (2007) és una molt bona referència sobre l'anàlisi de correspondències.

La Figura 6.7 presenta els gràfics de les dues primeres components resultat de l'anàlisi de correspondències per a les nostres dades per a cada any, on els punts representen les zrp. La primera component explica al voltant del 70% de la inèrcia, concretament, i per ordre cronològic, 84.2%, 73.4%, 72.5%, 67.2% i 66.5%. S'observa que la primera component discrimina PSC de la resta de partits, sobretot de CIU, mentre que la segona component separa principalment per una banda ERC i per l'altra el PPC. La segona component ja només explica al voltant del 15% de la inèrcia.

La figura 6.8 presenta l'anàlisi de correspondències per al 2006 representant cada districte per separat. Així, les zrp dels districtes de Ciutat Vella, Sants Montjüic, Horta-Guinardó i Sant Martí es troben majoritàriament al subespai que podríem anomenar d'esquerres, on hi trobem els partits PSC, ICV i ERC. Les zrp dels districtes de l'Eixample i Gràcia els trobem al subespai nacionalista a cavall d'ERC i CIU. I les zrp de Sarrià-Sant Gervasi i Les Corts, els dos districtes més benestants de Barcelona, se situen entre CIU i PPC, passant per Cs. El districte de Nou Barris crida l'atenció per la concentració de les seves zrp al voltant del PSC i de l'abstenció.

Per completar l'anàlisi descriptiva estudiem el comportament d'estadístics que capturen diferents característiques de les dades. Inspirats en l'anàlsisi de correspondències definim tres estadístics que recullen les principals fonts de variabilitat i que són:

$$log(\frac{y_{CIU}}{y_{PSC}}),$$

$$log(\frac{y_{CIU+PPC}}{y_{PSC+ERC+ICV}})$$

i

$$log(\frac{y_{CIU+ERC}}{y_{PSC+PPC+ICV}}),$$

així com un quart estadístic referent al percentatge d'abstenció,

$$log(\frac{y_{abs}}{N}),$$

que és l'opció de vot més escollida.

La representació d'aquests estadístics estratificada per districtes es presenten a la Figura 6.9. Allà s'hi observa com les zrp d'un mateix districte tendeixen a obtenir valors similars. Aquesta estructura subjacent a les dades l'aprofitarem més endavant a l'hora de validar els models Bayesians utilitzats.

A la Figura 6.10 es presenta conjuntament la distribució de vots a nivell de zrp entre CIU, PSC i PPC, així com de CIU, PSC i ERC mitjançant diagrames ternaris. Per a la representació ternària cal utilitzar dades composicionals (Aitchison, 1986), que sumen 1, per això els percentatges de vot s'han reescalat. Al diagrama ternari de l'esquerra, s'observa com les zrp en que el PPC obté perecentatges alts també ho fa CIU i en canvi és on el PSC obté els percentatges més baixos, mentre que el diagrama ternari de la dreta mostra com els percentatges alts d'ERC es situen en valors mitjans del PSC i de CIU. Aquestes representacions ternàries també ens seràn d'utilitat en la validació dels models, al Capítol 9.

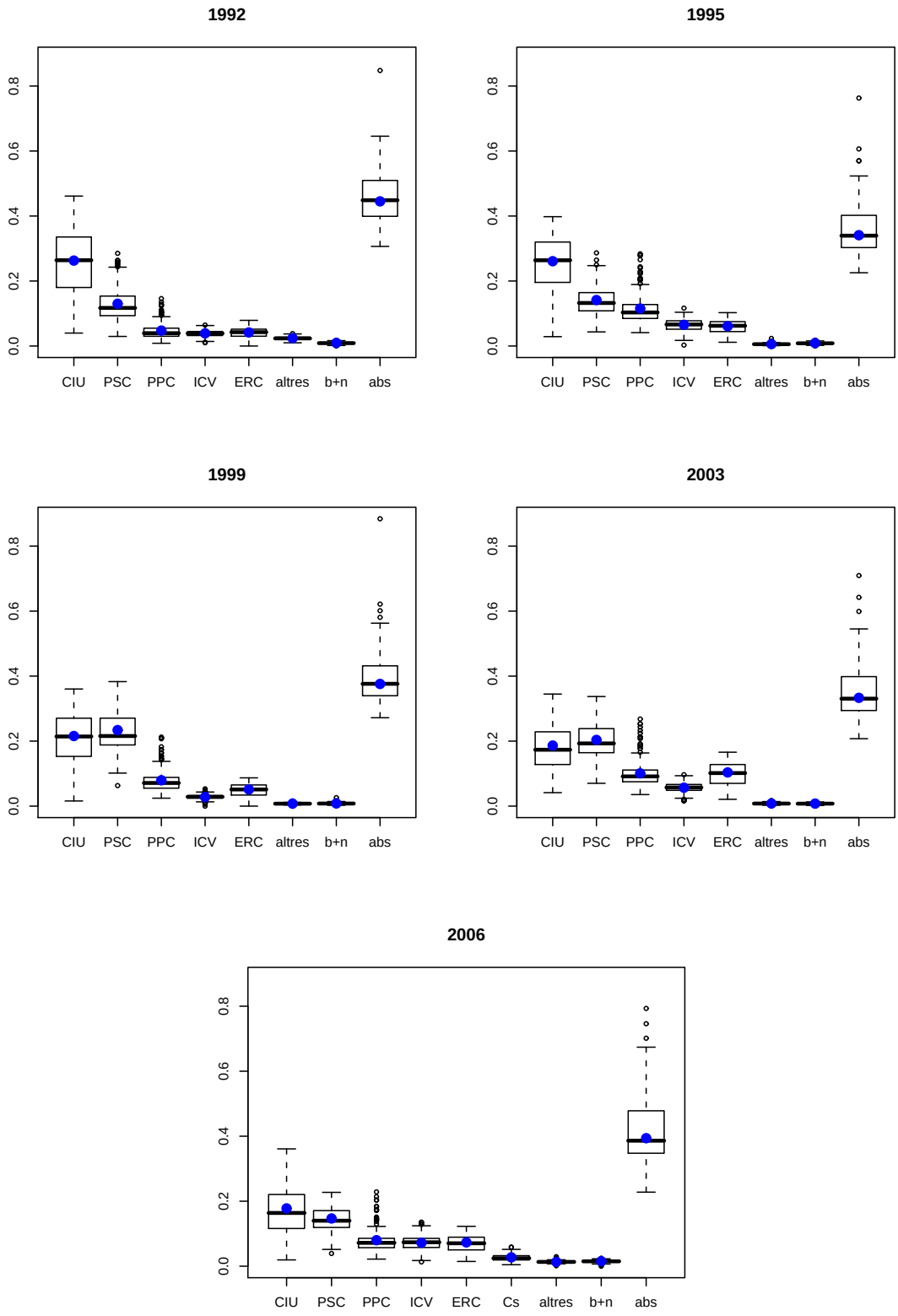

Figura 6.2: Distribució percentual dels vots a les eleccions al Parlament de Catalunya a nivell de les zrp de la ciutat de Barcelona entre les diferents categories per a cadascun dels anys. El punt blau és el percentatge marginal de cada categoria.

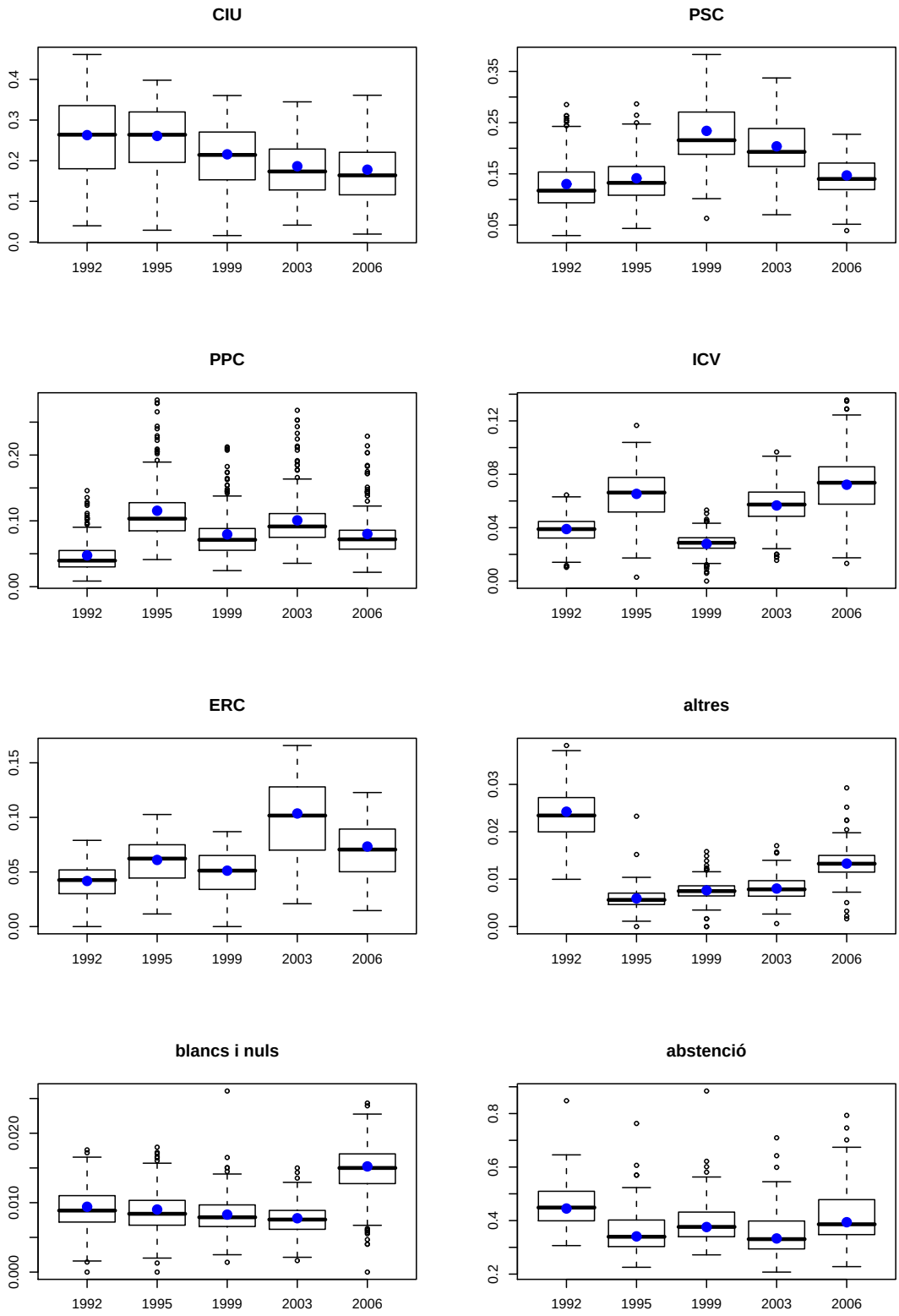

Figura 6.3: Evolució de la distribució percentual de vot a les eleccions al Parlament de Catalunya a nivell de zrp de la ciutat de Barcelona per a cadascuna de les opcions de vot. El punt blau és el percentatge marginal de cada categoria.

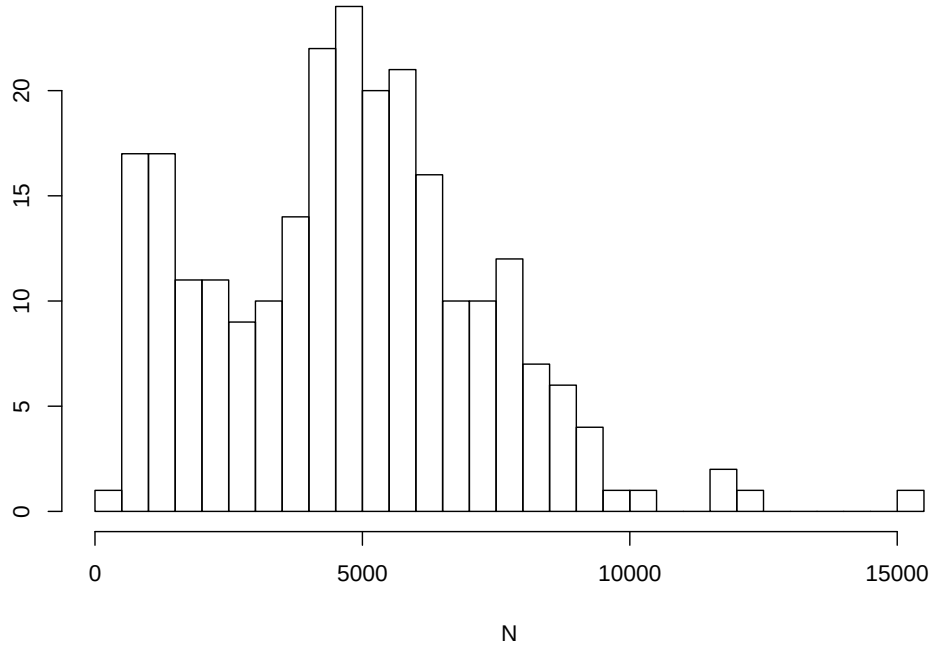

Figura 6.4: Distribució del nombre d'electors de les zrp de la ciutat de Barcelona el 2006.

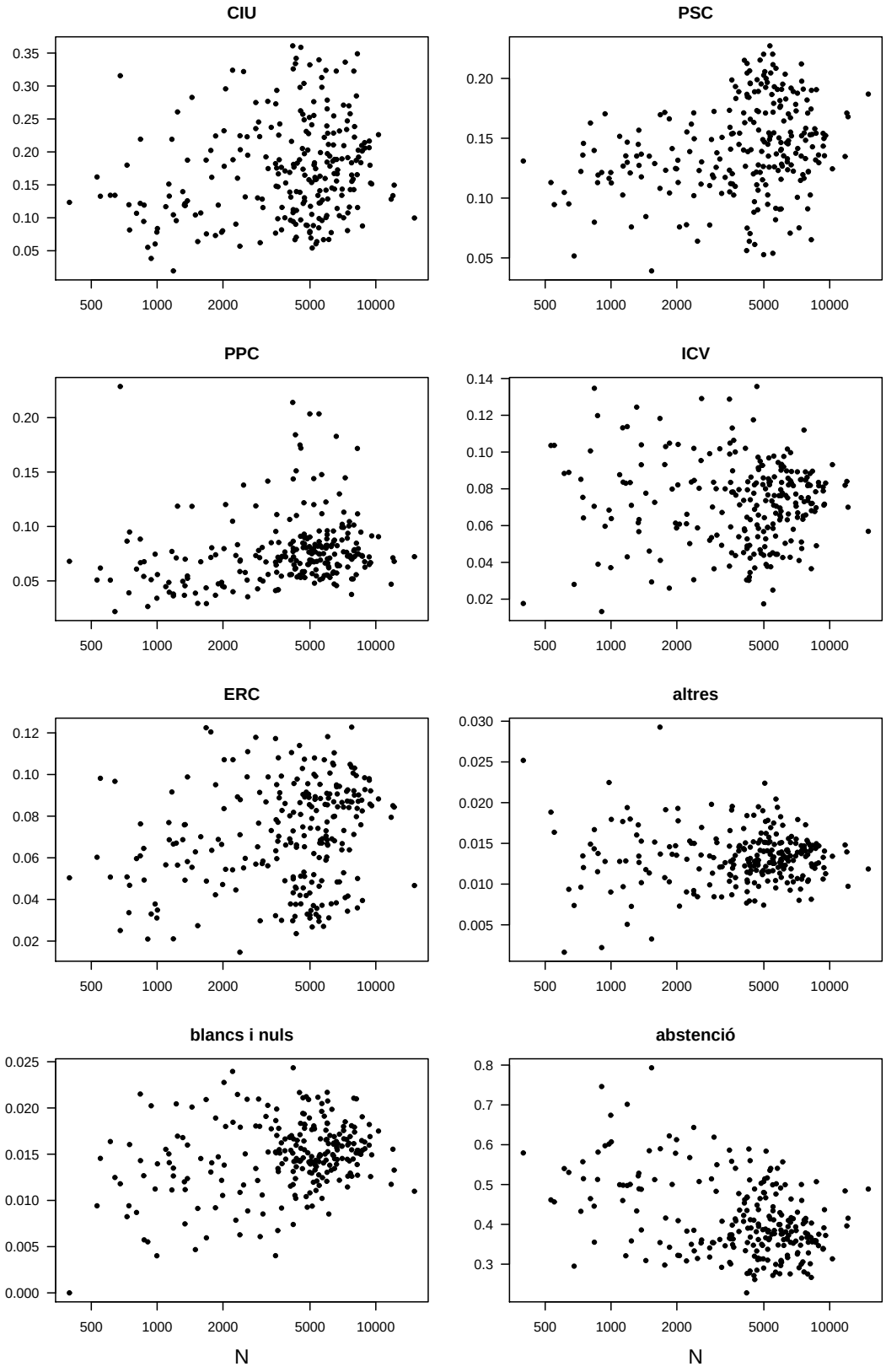

Figura 6.5: Percentatge de vot en funció del nombre d'electors (N) de les zrp de la ciutat de Barcelona per a cadascuna de les opcions de vot a les eleccions al Parlament de Catalunya del 2006. El nombre d'electors s'ha representat en escala logarítmica.

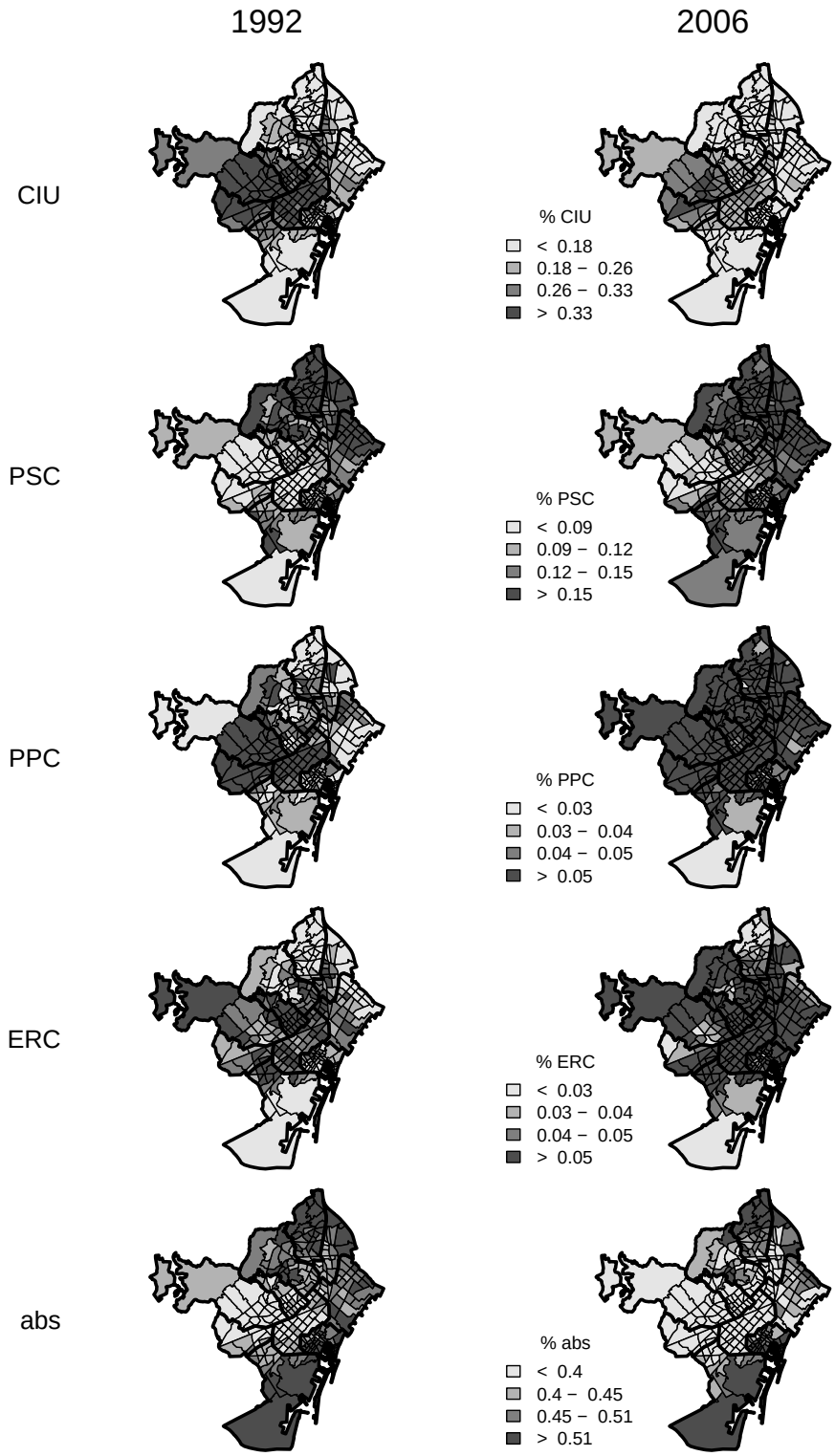

Figura 6.6: Distribució espacial del percentatge de vot a les Eleccions al Parlament de Catalunya per CIU, PSC, PPC, ERC i l'abstenció a nivell de zrp a Barcelona ciutat per als anys 1992 i 2006 . S'han categoritzat els percentatges de cada partit segons les quartiles de 1992.

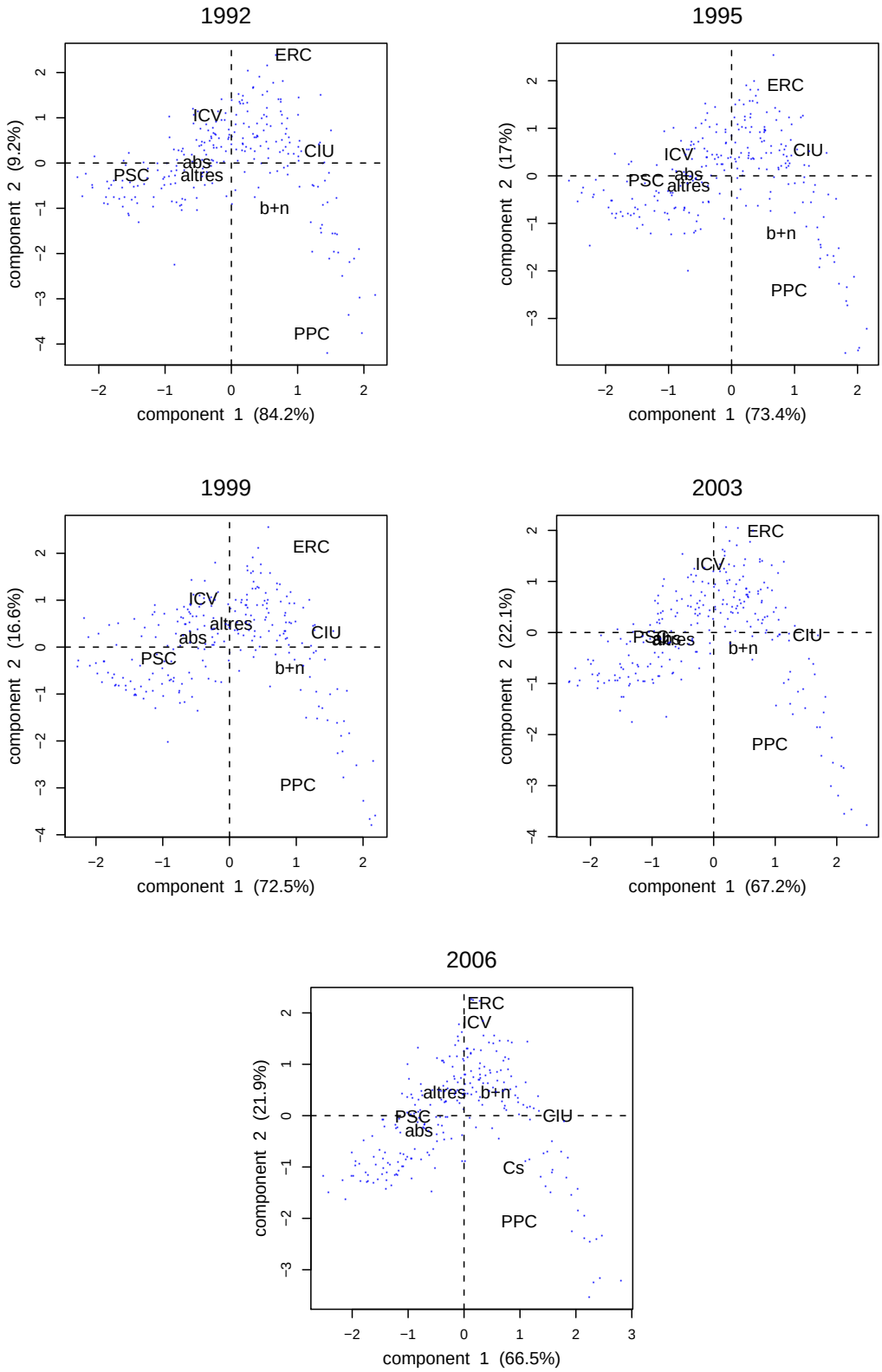

Figura 6.7: Anàlisi de correspondències per a les dades de Barcelona ciutat a les cinc darreres eleccions del Parlament de Catalunya. En les etiquetes dels eixos, entre parèntesi, hi trobem la inèrcia explicada per les respectives components.

Figura 6.8: Anàlisi de correspondències per Barcelona ciutat a les eleccions del Parlament de Catalunya del 2006, presentant per separat les zrp de cada districte.

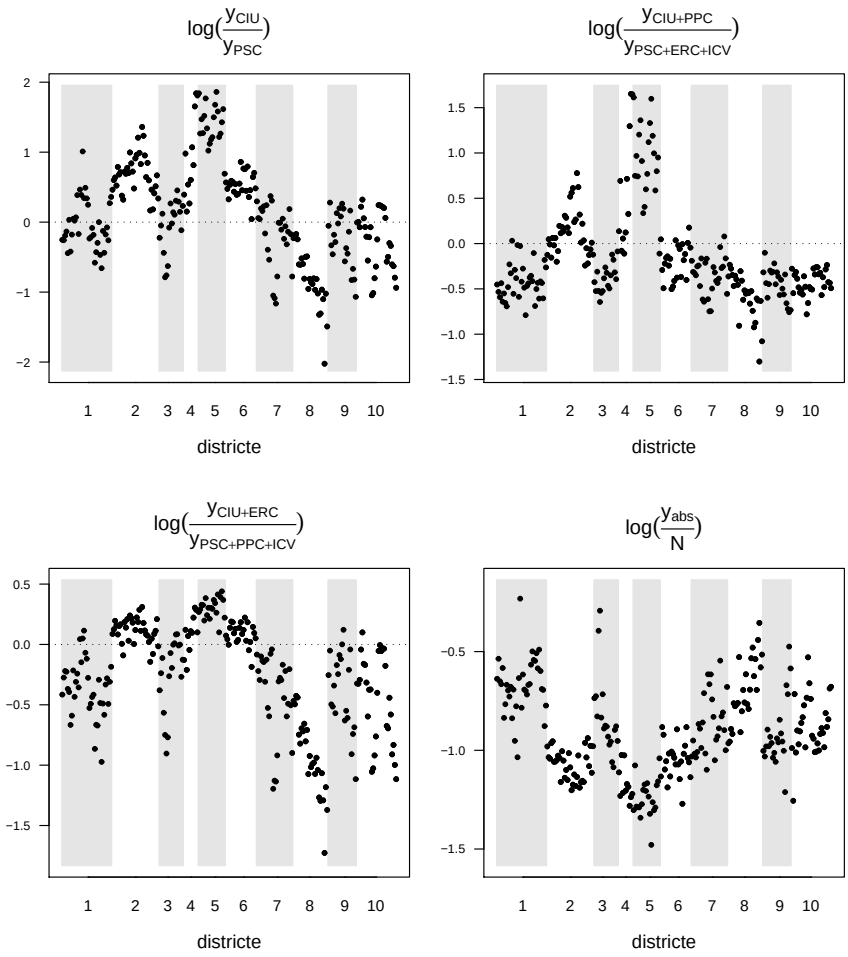

Figura 6.9: Valor que prenen quatre estadístics a nivell de zrp de Barcelona ciutat a les eleccions al Parlament de Catalunya del 2006.

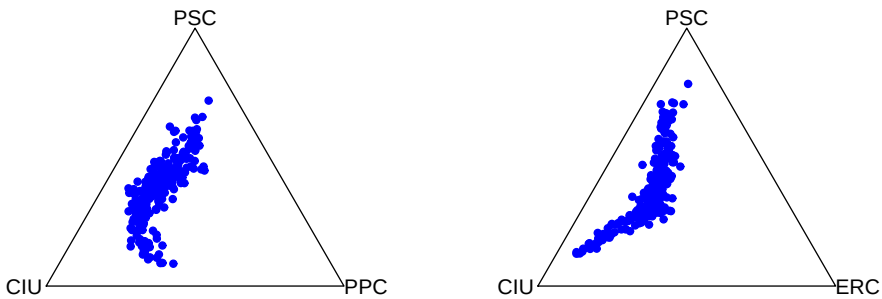

Figura 6.10: Representació ternària de la distribució de vot de (PSC, CIU, PPC), i (PSC, CIU, ERC) a nivell de zrp de Barcelona ciutat a les eleccions al Parlament de Catalunya del 2006.

6.5 Anàlisi descriptiva a nivell de Catalunya

En aquesta secció presentem una anàlisi descriptiva per a tot Catalunya incloent Barcelona. Els resultats a les eleccions al Parlament de Catalunya tal com els analitzarem formen una taula de contingència de 1447 files i 8 columnes. A la Taula 6.5 presentem parcialment els resultats de les eleccions de l'any 2003. Per exemple, la primera fila de la taula ens diu que el municipi d'Abrera, que només té un districte, té un total de 7140 electors, dels quals 2931 s'han abstingut, 883 han votat CIU, 1746 han votat PSC, 585 han votat PPC, 397 han votat ICV, 467 han votat ERC, 77 electors han votat altres partits amb menys de l'1% dels vots, i 54 han votat en blanc o be el seu vot ha estat considerat nul.

Municipi	Districte	zrp	CIU	PSC	PPC	ICV	ERC	altres	b+n	abs	N
Abrera	1	-	883	1746	585	397	467	77	54	2931	7140
Aguilar de Segarra	1	-	73	9	6	12	22	4	1	52	179
Alella	1	-	1781	813	659	238	1036	49	51	2105	6732
Alpens	1	-	94	8	0	11	71	0	4	38	226
L'Ametlla del Vallès	1	-	1308	685	374	181	752	29	36	1626	4991
Arenys de Mar	1	-	1304	604	354	138	722	27	37	1353	4539
Arenys de Mar	2	-	1404	1033	408	195	747	26	50	2048	5911
...	...	...	...	...	...	...	...	...	...	...	...
Barcelona	1	36	9346	329	148	168	157	30	16	715	1909
Barcelona	1	37	94	89	43	48	53	8	5	240	580
Barcelona	2	38	2095	1541	773	545	1157	72	49	3118	9350
...	...	...	...	...	...	...	...	...	...	...	...

Taula 6.5: Part de la taula dels resultats a les eleccions al Parlament de Catalunya de l'any 2003. N és el nombre total d'electors, *abs*: és l'abstenció, *b+n*: són el total de vots blancs i nuls, *altres* és l'agregació de tots els partits amb menys d'1% del global de vots, i la resta de categories respon a les sigles dels respectius partits.

A la Taula 6.6 hi presentem la marginal columna de la Taula 6.5, i les marginals corresponents a les altres quatre eleccions, que ens permet copsar l'augment del nombre d'electors en els successius anys, contràriament al que passa a la ciutat de Barcelona, i com es reparteixen entre els diferents partits. I a la taula 6.7 hi presentem el percentatge de vots de cada categoria en cada any.

En la mateixa linea la Figura 6.11 presenta la distribució a nivell percentual dels electors en els districtes per a cadascuna de les opcions de vot en cadascun dels anys. Aquesta figura no mostra grans canvis de patró a nivell global. En aquesta figura hi trobem

Any	CIU	PSC	PPC	ICV	ERC	Cs	altres	b+n	abs	N
1992	1218662	726211	157347	171451	209909	-	122313	42213	2168272	4816378
1995	1314222	797251	420135	312462	304796	-	28908	40527	1811416	5029717
1999	1172799	1159553	295759	140467	270149	-	42898	36862	2086496	5204984
2003	1018165	1026227	390748	240228	542059	-	44049	39058	1906886	5207420
2006	928926	789950	313388	281449	414144	89557	68690	73503	2252816	5212423

Taula 6.6: Nombre de vots per a cadascuna de les opcions i el nombre d'electors (N) per a cadascuna de les darreres cinc eleccions al Parlament de Catalunya. Les dades presentades a la taula ja han estat processades de manera que a l'any 1999 els vots de ICV se li han sumat els de EUiA així com la part imputada de vots fora de la circumscripció de Barcelona que han estat restats del PSC.

Any	CIU	PSC	PPC	ICV	ERC	Cs	*altres*	*b+n*	*abs*
1992	0.253	0.151	0.033	0.036	0.044	-	0.025	0.009	0.450
1995	0.261	0.159	0.084	0.062	0.061	-	0.006	0.008	0.360
1999	0.225	0.223	0.057	0.027	0.052	-	0.008	0.007	0.401
2003	0.196	0.197	0.075	0.046	0.104	-	0.008	0.008	0.366
2006	0.178	0.152	0.060	0.054	0.079	0.017	0.013	0.014	0.432

Taula 6.7: Percentatge de vots a per a cadascuna de les opcions per a cadascuna de les darreres cinc eleccions al parlament de Catalunya. Les dades presentades a la taula han estat prèviament processades de manera que a l'any 1999 als vots de ICV se li han sumat els de EUiA així com la part imputada de vots fora de la circumscripció de Barcelona que han estat restats del PSC.

representat amb un punt el percentatge marginal de cada categoria. A diferència del que passava a l'analitzar només la ciutat de Barcelona, aquest punt no coincideix amb el centre dels diagrames de caixes degut a la diferent grandària, a nivell d'electors, dels districtes i al fet que la proporció de vot a cada partit depèn en gran mesura de la grandària del districte. Així quan la caixa està per sobre del punt indica una major representació d'aquella categoria en districtes petits mentre que quan la caixa està per sota del punt indica una major representació en districtes grans. El vot per PSC, PPC, ICV i l'abstenció és més gran en les grans ciutats que en els pobles, a l'inversa del que succeeix amb el vot per CIU i ERC. Això fa que mentre no es controli per la grandària de districte, aquests no es podran suposar intercanviables. Aquest fet caldrà tenir-lo molt present al modelar les dades ja que la grandària dels districtes, o en altres paraules al fet de viure en grans ciutats, està molt associat a la distribució de vots del districte o ciutat.

La Figura 6.12 es una reordenació de la Figura 6.11 amb l'objectiu d'observar l'evolució

de la distribució percentual per a cadascuna de les opcions de vot. En aquesta figura destaca la continua davallada de CIU, la pujada del PSC l'any 1999 que coincideix amb la primera candidatura d'en Pasqual Maragall com a cap de llista i la notable pujada d'ERC l'any 2003.

Els diagrames bivariants del percentatge de vot en front la grandària dels districtes per a cadascuna de les opcions de vot i per a cada any, a la Figura 6.13 hi presentem l'any 2006, posen de manifest que els perfils fila depenen molt més de N del que havíem observat a Barcelona ciutat. En el cas del percentatge de vot a CIU els valors més alts corresponen a municipis amb menys de 500 electors, quelcom similar al que passa amb ERC encara que aquí trobem municipis fins a 1000 electors. Pel que fa al PSC part dels districtes amb valors més alts es corresponen a grans ciutats del cinturó de Barcelona, com per exemple districtes de Molins de rei, Sabadell, Mataró i Cornellà de Llobregat. Els valors més alts del PPC es concentren sobretot en zrp dels districtes de Sarrià-Sant Gervasi i Les Corts de la ciutat de Barcelona. ICV és el partit que presenta més homogeneïtat entre els valors més alts, barrejant-se municipis petits amb zrp dels districtes de Gràcia i Ciutat Vella. Referent a la categoria d'altres ressalten valors alts de l'any 1999, molts dels quals corresponen a una àrea geogràfica determinada on destaca el partit *Unió d'Independents Conca de Barberà*. També és peculiar el patró dels percentatges alts de blancs i nuls, ja que encara que es corresponen a municipis petits, en cadascun dels comicis han estat diferents. Finalment els percentatges alts d'abstenció es corresponen majoritàriament a districtes de ciutats grans així com zrp concretes de la ciutat de Barcelona, principalment dels districtes de Sants-Montjuïc, Ciutat Vella i Nou Barris.

La Figura 6.14 ens presenta en forma d'histograma la distribució de la grandària dels districtes per a l'any 2006, quan el 96% dels districtes tenen menys de 15000 electors tot i que en trobem dos amb més de 40.000, que corresponen a un districte de Badalona i al municipi de Cerdanyola del Vallès el qual només té un districte. Al generalitzar l'anàlisi a tot Catalunya, les grandàries presenten una distribució menys homogènia i més asimètrica que la que trobàvem per les zrp a l'anàlisi descriptiva de les dades de Barcelona a l'anterior secció.

Per tal de conèixer com es reparteix la població pel territori, a la Figura 6.15 mostrem el mapa dels electors per municipi per als anys 1992 i 2006. Aquí s'observa com globalment el patró de concentració d'electors no presenta grans canvis i el més destacat és l'augment de la concentració d'electors a l'àrea d'influència de Barcelona en detriment del nombre d'electors a la ciutat de Barcelona, si bé aquest fet no s'aprecia en els mapes de la figura per la categorització utilitzada.

I a la Figura 6.16 representem la distribució espacial del percentatge de vot per a les

quatre llistes més votades, CIU, PSC, PPC i ERC, així com per a l'abstenció per el 1992 i el 2006. Per a la representació s'han categoritzat els percentatges de cada partit segons les quartiles del 1992, d'aquesta manera visualitzem conjuntament la distribució espacial i els canvis ocorreguts entre la primera i la última elecció considerada. En aquesta figura s'observa com CIU obté percentatges més alts a l'interior mentre que PSC i l'abstenció obtenen percentatges més alts al litoral i al nord-oest, aquesta dualitat de patrons s'associa a la distribució del nombre d'electors, fet que es comprova observant la figura 6.15, és a dir PSC i l'abstenció tendeixen a obtenir percentatges de vot més alt en les zones més poblades. Els mapes de la Figura 6.16 també posen de manifest l'augment de vots de l'any 1992 al 2006 per part d'ERC, i PPC, la disminució de vots de CIU i els valors similars de l'abstenció.

Donat que les dades formen una taula de contingència, una anàlisi exploratoria força informativa consisteix en realitzar una anàlisi de correspondències. La Figura 6.17 presenta els gràfics de les dues primeres components resultat de l'anàlisi de correspondències per a les nostres dades per a cada any, on els punts representen els districtes. La primera component explica al voltant del 70% de la inèrica (concretament, i per ordre cronològic, 70.7%, 67.9%, 71.6%, 70.4% i 61.5%). S'observa que la primera component separa sobretot per una banda CIU i ERC, i per l'altre el PSC, ICV i l'abstenció, mentre que la segona, que ja només explica al voltant del 15% de la inèrcia, caracteritza el PPC i Cs, aquest darrer partit després de la seva aparició l'any 2006 acompanya el PPC situant-se en el pla com el partit més allunyat de CIU i sobretot d'ERC.

En el Capítol 7 formularem un model bayesià per aquest tipus de dades, i en els Capítols 10 i 12 presentarem els resultats de l'anàlisi bayesià no jeràrquic i jeràrquic per a les dades de la ciutat de Barcelona ciutat. A l'Apèndix A es presenten els resultats de l'anàlisi preliminar per a les dades de tot Catalunya.

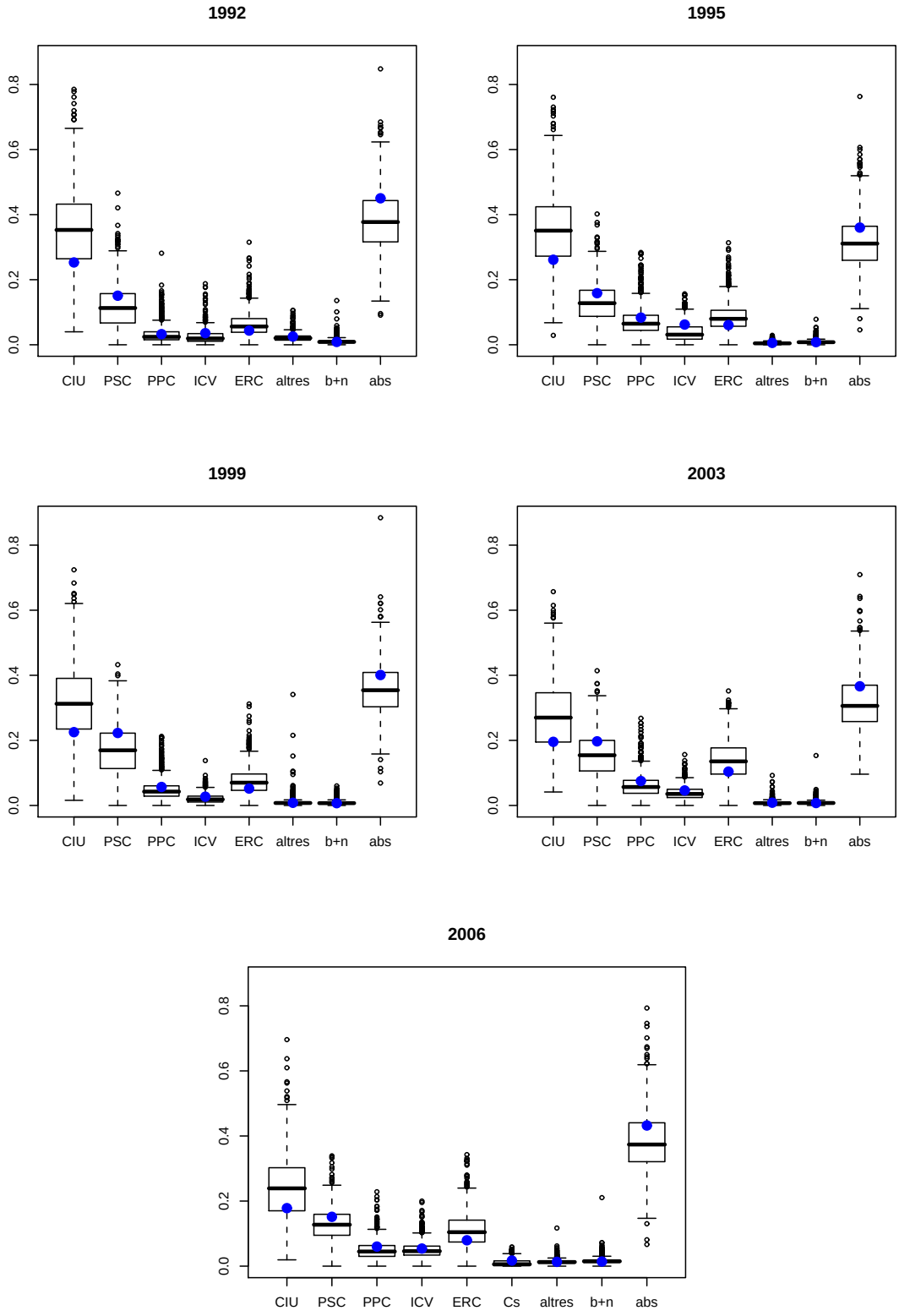

Figura 6.11: Distribució percentual dels vots a les eleccions al Parlament de Catalunya a nivell de districte, i en el cas de Barcelona de zrp, entre les diferents categories per a cadascun dels anys. El punt blau és el percentatge marginal de cada categoria.

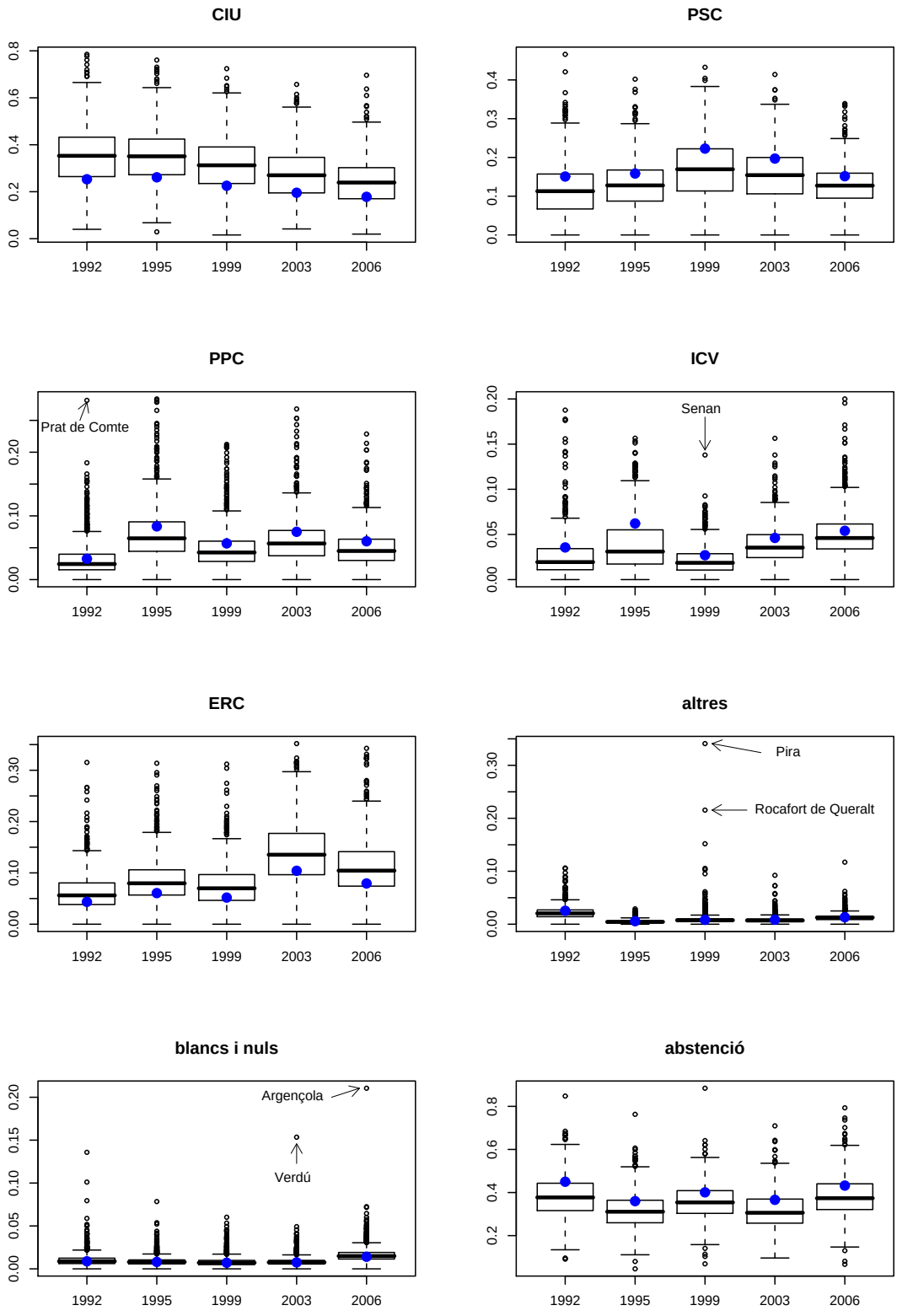

Figura 6.12: Evolució de la distribució percentual de vot a les eleccions al Parlament de Catalunya a nivell de districte, i en el cas de Barcelona de zrp, per a cadascuna de les opcions de vot. El punt blau és el percentatge marginal de cada categoria.

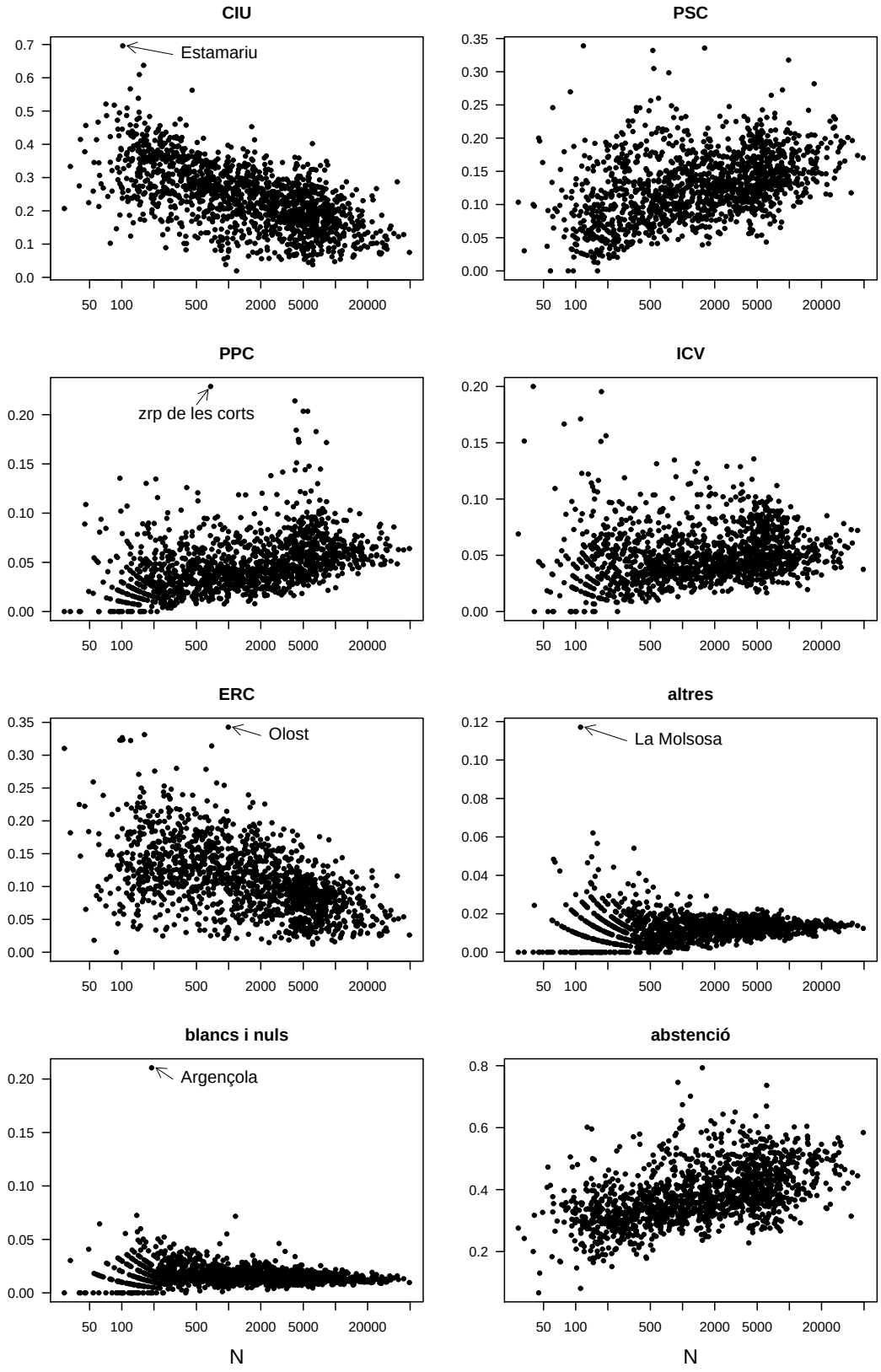

Figura 6.13: Percentatge de vot en funció del nombre d'electors (N) del districte, i en el cas de Barcelona zrp, per a cadascuna de les opcions de vot per a les eleccions al Parlament de Catalunya del 2006. El nombre d'electors s'ha representat en escala logarítmica.

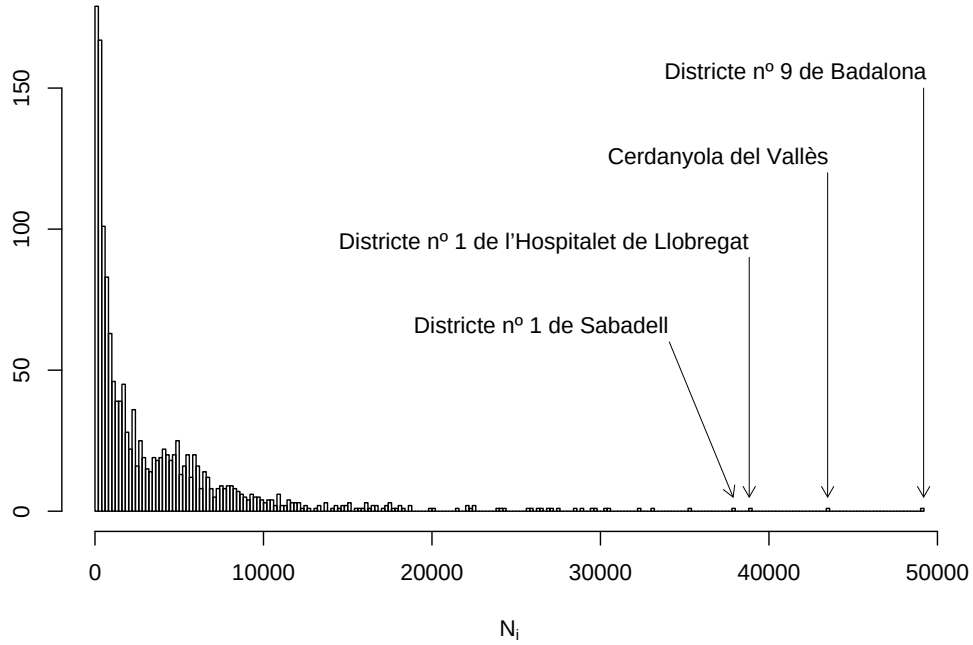

Figura 6.14: Distribució del nombre d'electors per districtes, i en el cas de Barcelona zrp, pel 2006 a Catalunya.

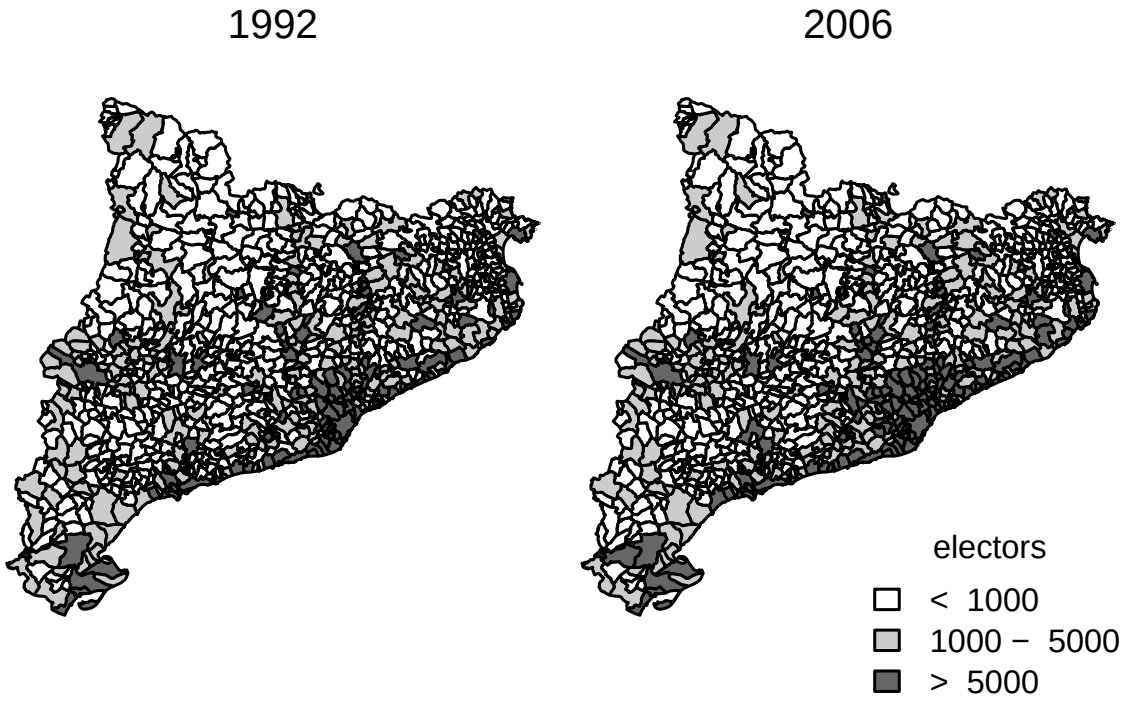

Figura 6.15: Nombre d'electors per municipis a Catalunya els anys 1992 i 2006.

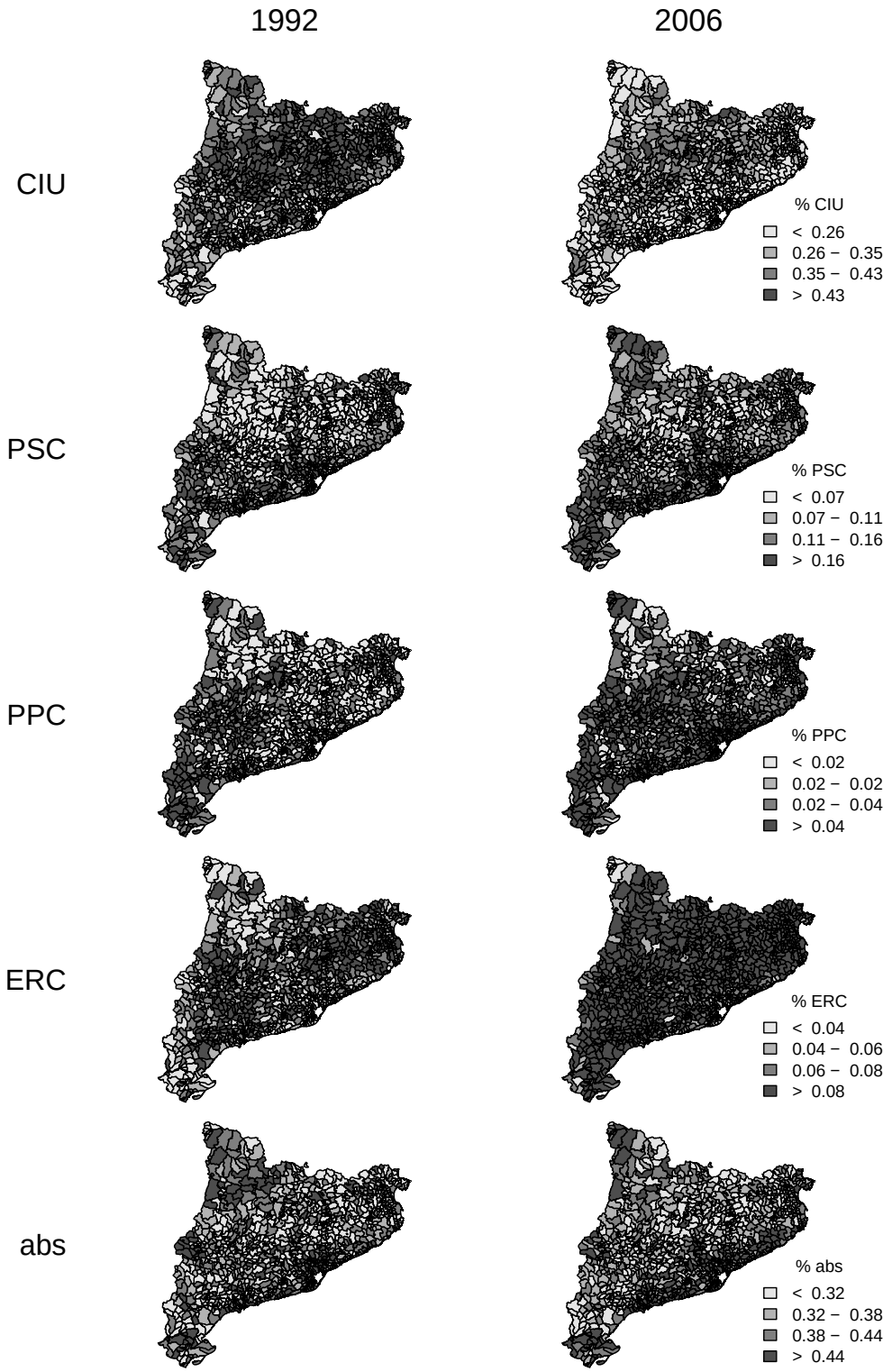

Figura 6.16: Distribució espacial del percentatge de vot a les Eleccions al Parlament de Catalunya per CIU, PSC, PPC, ERC i l'abstenció a nivell de districtes per als anys 1992 i 2006. Barcelona està representada a la Figura 6.6. Els percentatges de cada partit s'han categoritzat segons les quartiles de 1992.

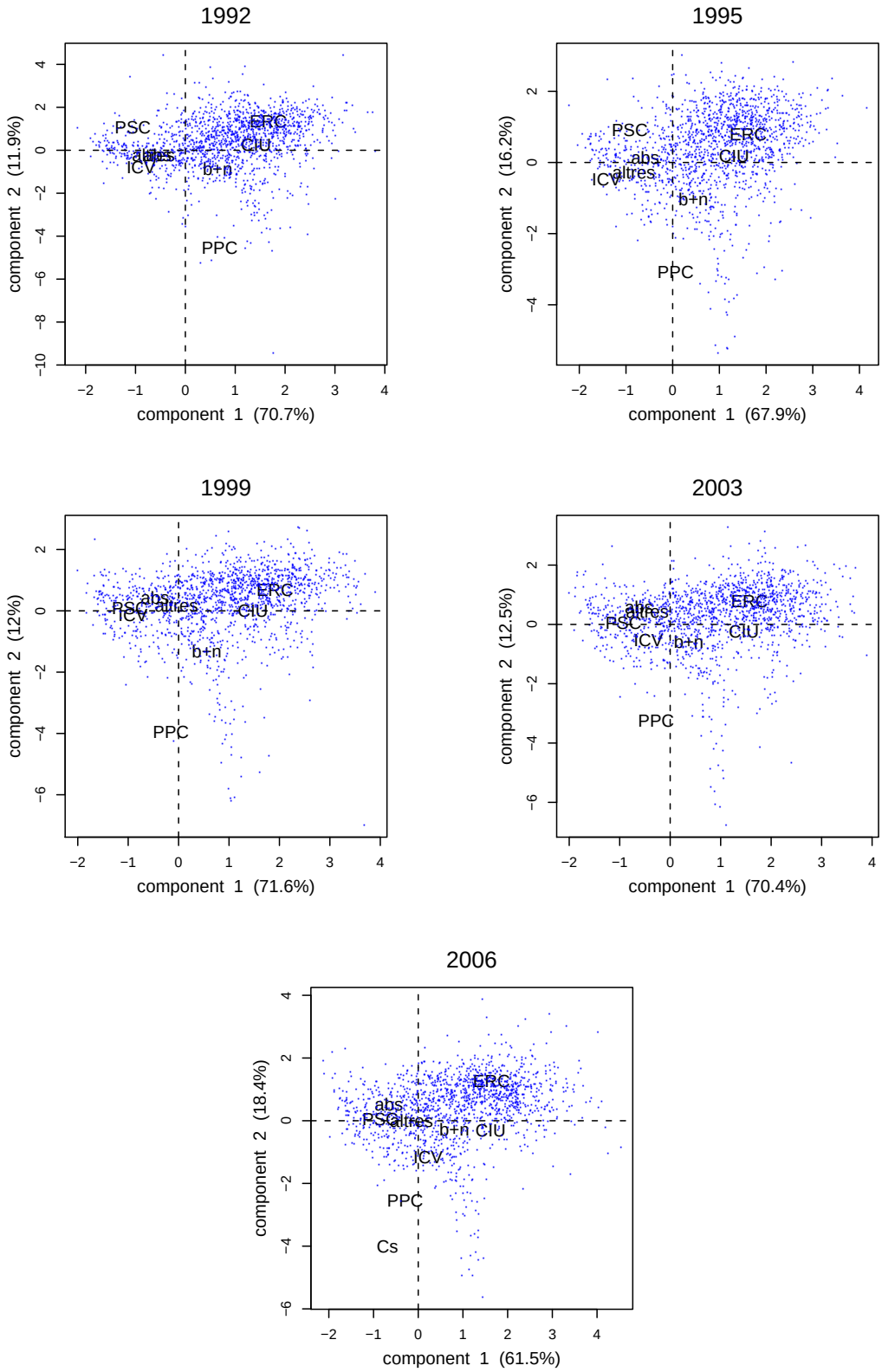

Figura 6.17: Anàlisi de correspondències dels diferents comicis al Parlament de Catalunya. Els punts representen els districtes i, en el cas de Barcelona, les zrp. En les etiquetes dels eixos entre parèntesis hi trobem el percentatge d'inèrcia explicada per les respectives components.

Capítol 7

Anàlisi Cluster Multinomial Bayesià

En aquest capítol presentem els models Bayesians que farem servir per a realitzar una anàlisi cluster per a les dades presentades al capítol 6, i que ha de servir de base per a més endavant acabar modelant l'evolució temporal de les dades. Per arribar-hi presentarem el paradigma bayesià, el model Multinomial bayesià no jeràrquic i jeràrquic, i el model que permet realitzar una anàlisi de cluster Multinomial bayesiana pels casos no jeràrquic i jeràrquic. Al capítol 8 presentarem maneres de validar aquests models.

7.1 El paradigma Bayesià

Entenem com a model estadístic una llista de distribucions de probabilitat que comparteixen el mateix espai mostral, i que escriurem com:

$$M = \{p(y|\theta); \theta \in \Omega\},$$

on y és la variable aleatòria observable que se suposa que té una distribució $p(y|\theta^*)$ de la llista. El problema de la inferència estadística consisteix en, havent observat una realització d'un dels models de probabilitat $p(y|\theta^*)$ que pertanyen a M intentar endevinar el valor del paràmetre $\theta^* \in \Omega$ que ha generat les dades.

El model bayesià parteix d'un model estadístic, M, però tracta el paràmetre, θ, com una variable aleatòria i està disposat a triar una distribució a priori sobre l'espai de paràmetres Ω, $\pi(\theta)$, que representi la nostra incertesa sobre θ abans d'observar les dades. Així podem escriure el model bayesià com:

$$(M; \pi(\theta)) = (p(y|\theta); \theta \in \Omega; \pi(\theta))$$

que esdevé una llista de distribucions sobre un únic espai mostral, ordenades de més a menys creïbles en base a $\pi(\theta)$. Al tractar el paràmetre com a variable aleatòria, el paradigma Bayesià utilitza dos tipus d'informació:

1. el model Bayesià a priori, que inclou el model estadístic que hem suposat per les dades i la distribució que especifiquem sobre el paràmetre, $\pi(\theta)$, i
2. les dades, la informació de les quals es canalitza a través de la funció de versemblança que notem per $L_y(\theta)$ i que és una funció de θ proporcional a $p(y|\theta)$.

El teorema de Bayes ens ofereix una manera formal d'integrar els dos tipus d'informació, a través de la distribució a posteriori, $\pi(\theta|y)$,

$$\pi(\theta|y) = \frac{p(y|\theta)\pi(\theta)}{\int_\Omega p(y|\theta)\pi(\theta)d\theta} \propto L_y(\theta)\pi(\theta). \tag{7.1}$$

El Teorema de Bayes ens diu que la distribució a posteriori és proporcional al producte de la distribució a priori per la funció de versemblança, $L_y(\theta)\pi(\theta)$. La informació que suposem a priori sobre el paràmetre, $\pi(\theta)$, s'actualitza a través de la informació que aporten les dades a través de $L_y(\theta)$, i tot combinat porta a $\pi(\theta|y)$.

Una vegada observades les dades, y, seguim tenint el mateix objecte com a model però actualitzat l'ordre de credibilitat de les distribucions de la llista,

$$(M, \pi(\theta|y)) = (p(y|\theta); \theta \in \Omega; \pi(\theta|y)),$$

de manera que el Bayesià no ajusta un model si no que l'actualitza.

Quan treballem sota la perspectiva bayesiana la inferència sobre els paràmetres es fa tota a partir de la distribució a posteriori, que neix d'un compromís entre les nostres creences a priori i la informació que n'aporten les dades. El punt més compromès és l'elecció de la distribució a priori dels paràmetres i és molt important pensar bé quina és la $\pi(\theta)$ que captura millor el que sabem sobre θ.

Una família de distribucions a priori molt útil és la formada per les anomenades distribucions conjugades. Una distribució a priori és conjugada per un model donat si la distribució a posteriori és de la mateixa família que la distribució a priori. Aquestes distribucions a priori tenen l'avantatge que faciliten molt l'anàlisi Bayesià, perquè per calcular la distribució a posteriori només cal actualitzar el valor dels paràmetres de la distribució a priori.

Un altre element important per a la selecció i la validació de models Bayesians són les distribucions predictiva a priori i predictiva a posteriori. La distribució predictiva a

priori té per densitat de probabilitat:

$$p_\pi(\widetilde{y}) = E_{\pi(\theta)}[p(\widetilde{y}|\theta)] = \int_\Omega p(\widetilde{y}|\theta)\pi(\theta)d\theta, \tag{7.2}$$

on $\widetilde{y}$ simbolitza una observació futura i on $p_\pi(\widetilde{y})$ és un promig ponderat de totes les $p(\widetilde{y}|\theta)$ fent servir com a ponderació $\pi(\theta)$. Anàlogament definim la distribució predictiva a posteriori com:

$$p_\pi(\widetilde{y}|y) = E_{\pi(\theta|y)}[p(\widetilde{y}|\theta)] = \int_\Omega p(\widetilde{y}|\theta)\pi(\theta|y)d\theta, \tag{7.3}$$

que no és més que un promig ponderat de $p(\widetilde{y}|\theta)$, ara ponderat segons $\pi(\theta|y)$, i que representa el que saps sobre una observació futura a la llum del que has observat.

Per introduir-se i aprofundir en l'estadística Bayesiana es pot consultar Lee (2004), Leonard i Hsu (2001), Carlin i Louis (2000) i un llibre que ens agrada especialment és Gelman *et al* (2004). A nivell pràctic són interessants els llibres d'en Cogndon (2003, 2005 i 2006) en el que s'utilitza el programari de lliure distribució WinBugs. Bones referències a nivell de fonaments són els llibres de Bernardo i Smith (1994), Berger (1985) i Robert (2001).

Els mètodes bayesians han experimentat una gran expansió en les dues darreres dècades. El mèrit d'aquest increment sobtat de l'ús d'aquest paradigma sens dubte es pot atribuir als mètodes de simulació de Montecarlo basats en Cadenes de Markov (MCMC), que es van desenvolupar a finals dels 80 i principis dels 90 i que eviten haver de resoldre integrals analíticament intractables per calcular la distribució a posteriori i predictiva a posteriori quan es parteix de distribucions a priori no conjugades. Això ha fet que els esforços computacionals dels bayesians es concentrin a simular mostres de $\pi(\theta|y)$ que permeten aproximar la distribució $\pi(\theta|y)$ tant bé com es vulgui com a alternativa a haver de calcular $\pi(\theta|y)$ de forma explícita.

Els mètodes de *gibbs sampling*, que són un cas particular dels mètodes MCMC, són especialment atractius perquè tot el que es necessita és saber simular de les distribucions condicionals completes. Tot i així altres mètodes més generals basats en l'algorisme de *Metropolis-Hasting* poden resultar més eficients en determinats casos. Per una descripció d'aquests algorismes es pot consultar Gelfand (2000), Smith i Gelfand (1992), Robert i Casella (1999), Gilks, Best i Tan (1995) i Brooks, Giudici i Roberts (2003). Per simular de les distribucions a posteriori dels paràmetres dels diferents models utilitzarem el programari de lliure distribució WinBugs.

Els mètodes MCMC parteixen d'uns valors inicials arbitraris. Per aquest motiu les simulacions d'una cadena es divideixen en dues parts, una primera part anomenada

d'escalfament i que és descartada, i una segona part en que se suposa que la cadena ha convergit i és amb la que es farà la inferència, d'aquesta manera es pretén eliminar l'efecte dels valors inicials tirats de forma arbitrària. És diu que una cadena de Markov ha convergit a la seva distribució estacionària si està mostrejant de la vertadera distribució a posteriori. Per això és important avaluar la convergència de les cadenes.

La part més problemàtica de la computació MCMC és primer decidir quan la cadena o cadenes han convergit, és a dir ha arribat a la seva veritable distribució estacionària, i després decidir quantes simulacions utilitzar per a la inferència. Una bona revisió d'aquests aspectes la trobem a Cowles i Carlin (1996), i a Kass (1998) hi trobem la transcripció d'un diàleg interessant sl respecte entre el mateix Kass i Carlin, Gelman i Neal.

7.2 Model Estadístic Multinomial

El model estadístic de partida a l'hora de modelar dades categòriques és el model Multinomial, que suposa que les dades provenen de la realització de N experiments independents tals que cada experiment pot prendre una de les k possibles categories amb una probabilitat associada a cada una d'elles. Les k categories de la Multinomial se suposen mútuament excloents i exhaustives, en el sentit que cobreixen tots els possibles resultats d'un experiment.

Un exemple de dades que poden seguir aquest tipus de model serien les files de la Taula 6.5, suposant que cadascun dels electors d'un districte ha escollit de forma independent als altres electors del districte una i només una de les 8 opcions (CIU, PSC, PPC, ICV, ERC, *altres*, *b+n* i abstenció) d'acord amb un perfil de probabilitats $\theta = (\theta_1, \ldots, \theta_8)$.

Denotem el model estadístic Multinomial per una de les files de la Taula 6.5 com:

$$M = \{\text{Mult}(y|N, \theta = (\theta_1, \ldots, \theta_k)); \theta_j \in [0,1] \text{ per } j = 1, \ldots, k, \text{ amb } \textstyle\sum_{j=1}^{k} \theta_j = 1\},$$

on θ_j és la probabilitat de la j-èssima categoria, k és el nombre de categories, N és el nombre d'elements de la fila de la taula i on $y = (y_1, y_2, \ldots, y_k)$ són els comptetjos de les k categories de la fila. $\text{Mult}(y|N, \theta)$ representa el model de probabilitat amb funció de probabilitat igual a:

$$p(y|N, \theta) = \frac{N!}{\Pi_{j=1}^{k} y_j!} \Pi_{j=1}^{k} \theta_j^{y_j}, \tag{7.4}$$

on l'esperança d'aquesta distribució és:

$$E[y|N,\theta] = (N\theta_1, \ldots, N\theta_k), \tag{7.5}$$

i la matriu de variàncies i covariances d'aquesta distribució és:

$$V(y|N,\theta) = \begin{pmatrix} N\theta_1(1-\theta_1) & -N\theta_1\theta_2 & \ldots & -N\theta_1\theta_k \\ -N\theta_2\theta_1 & N\theta_2(1-\theta_2) & \ldots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ -N\theta_k\theta_1 & \ldots & \ldots & N\theta_k(1-\theta_k) \end{pmatrix}.$$

En aquest cas per tant, el model estadístic de partida és una llista de distribucions Multinomials indexada per $\theta = (\theta_1, \ldots, \theta_k)$ que pot prendre qualsevol valor del símplex de $\mathbb{R}^k$ ($\theta_j \in [0,1]$ per $j = 1, \ldots, k$, amb $\sum_{j=1}^k \theta_j = 1$), que suposarem que és l'espai de paràmetres.

7.3 Model Multinomial Bayesià: cas no jeràrquic

El model Multinomial no jeràrquic suposa que les files de la taula, $y_i = (y_{i1}, y_{i2}, \ldots, y_{ik})$ per $i = 1, 2, \ldots, n$ és una mostra d'observacions condicionalment independents i idènticament distribuides d'una Mult(N_i, θ), on el valor de θ és el mateix per totes les files.

Per definir el model bayesià cal especificar la distribució a priori dels paràmetres, que en el cas de la distribució Multinomial tenen com a espai de paràmetres el simplex de $\mathbb{R}^k$. El més habitual és fer servir com a distribució a priori la conjugada de la Multinomial. Aquesta és la distribució Dirichlet amb k categories, representada per Dir$(\alpha_1, \alpha_2, \ldots, \alpha_k)$, que té per suport el simplex de $\mathbb{R}^k$ i per densitat de probabilitat

$$\pi(\theta|\alpha) = \frac{\Gamma(\sum_{j=1}^k \alpha_j)}{\Pi_{j=1}^k \Gamma(\alpha_j)} \Pi_{j=1}^k \theta_j^{\alpha_j - 1}, \tag{7.6}$$

on $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_k)$ és tal que $\alpha_j > 0$ per $j = 1, \ldots, k$.

Per tenir completament especificat el model resta triar els paràmetres de la Dirichlet. En tot el capítol denotarem els paràmetres desconeguts, no observables i dels quals farem

inferència, amb lletres gregues i els paràmetres coneguts de les distribucions a priori amb lletres llatines. Per tant el model Multinomial Bayesià no jeràrquic conjugat per a una de les files de la taula l'escriurem com:

$$y_i|\theta \sim \text{Multinomial}(N_i, \theta)$$
$$\theta \sim \text{Dirichlet}(a_1, \ldots, a_k)$$

Model 1. Multinomial Bayesià no jeràrquic.

La distribució predictiva a priori per al Model 1 és la distribució Multinomial-Dirichlet, que denotem per Mult-Dir$(N, a_1, \ldots, a_k)$, i que té per densitat de probabilitat:

$$p_\pi(\widetilde{y}) = p(\widetilde{y}|\widetilde{N}, a_1, a_2, \ldots, a_k) = \frac{\widetilde{N}!\Gamma(\sum_{j=1}^k a_j)}{\Gamma(\widetilde{N} + \sum_{j=1}^k a_j)} \prod_{j=1}^k \frac{\Gamma(\widetilde{y}_j + a_j)}{\widetilde{y}_j!\Gamma(a_j)},$$

on $\widetilde{N}$ seria el compteig total per l'observació futura $\widetilde{y}$.

A les aplicacions cal triar els paràmetres de la distribució a priori de manera que aquesta modeli la nostra incertesa a priori sobre θ. A l'hora de triar $(a_1, \ldots, a_k)$ sol ser útil tenir en compte que si $\theta \sim \text{Dir}(a_1, a_2, \ldots, a_k)$, i $t = \sum_{j=1}^k a_j$, aleshores,

$$E[\theta] = (\frac{a_1}{t}, \ldots, \frac{a_k}{t}), \tag{7.7}$$

i per tant si suposem que els valors de les θ_j són similars hauríem de triar una distribució Dirichlet simètrica amb $a_1 = \ldots = a_k$. Suposar que $\theta \sim \text{Dir}(1, .., 1)$ correspon a assumir una distribució uniforme en el símplex. En canvi si per exemple suposem que els valors de les θ_j estan ordenats de gran a petit, llavors és natural assumir $a_1 \geq a_2 \geq \ldots \geq a_k$.

A més a més:

$$V(\theta) = \begin{pmatrix} \frac{\frac{a_1}{t}(1-\frac{a_1}{t})}{t+1} & \frac{-\frac{a_1 a_2}{t^2}}{t+1} & \cdots & \frac{-\frac{a_1 a_k}{t^2}}{t+1} \\ \frac{\frac{-a_2 a_1}{t^2}}{t+1} & \frac{\frac{a_2}{t}(1-\frac{a_2}{t})}{t+1} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\frac{-a_k a_1}{t^2}}{t+1} & \cdots & \cdots & \frac{\frac{a_k}{t}(1-\frac{a_k}{t})}{t+1} \end{pmatrix},$$

i per tant t té una relació inversa amb les variàncies i covariances a priori dels components de θ, fet que ens ajudarà a reflexar el nostre grau d'incertesa sobre θ a través de t. Per un

perfil donat, $(\frac{a_1}{t}, \ldots, \frac{a_k}{t})$, com més gran és t més petites són les variàncies i les covariances a priori de θ_j i més informativa és la distribució a priori sobre θ_j. Podem interpretar t com una grandària de mostra equivalent aportada per la distribució a priori.

L'esperança predictiva a priori de $\widetilde{y}$ és:

$$E[\widetilde{y}] = \widetilde{N}(\frac{a_1}{t}, \ldots, \frac{a_k}{t}) = \widetilde{N}E[\theta], \tag{7.8}$$

i la matriu de variàncies i covariances d'aquesta distribució val:

$$V(\widetilde{y}) = (\widetilde{N}^2 + t\widetilde{N}) \begin{pmatrix} \frac{\frac{a_1}{t}(1-\frac{a_1}{t})}{t+1} & \frac{-\frac{a_1 a_2}{t^2}}{t+1} & \cdots & \frac{-\frac{a_1 a_k}{t^2}}{t+1} \\ \frac{\frac{-a_2 a_1}{t^2}}{t+1} & \frac{\frac{a_2}{t}(1-\frac{a_2}{t})}{t+1} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\frac{-a_k a_1}{t^2}}{t+1} & \cdots & \cdots & \frac{\frac{a_k}{t}(1-\frac{a_k}{t})}{t+1} \end{pmatrix} = (\widetilde{N}^2 + t\widetilde{N})V(\theta),$$

on s'observa que t també té una relació inversa amb les variàncies i les covarianes a priori de les components de $\widetilde{y}$, al igual que succeïa amb la variància a priori de θ, de manera que per un perfil donat, $(\frac{a_1}{t}, \ldots, \frac{a_k}{t})$, com més gran és t més petites són les variàncies i les covariances a priori de $\widetilde{y}$ i més informativa és la distribució predictiva a priori.

Una reparametrització equivalent per les distribucions Dirichlet que facilita la interpretació és definir-la com a $\text{Dir}(t(m_1, \ldots, m_k))$ on $t = \sum_{j=1}^{k} a_j \in (0, \infty)$ i on $m = (m_1, \ldots, m_k) = (\frac{a_1}{t}, \ldots, \frac{a_k}{t}) \in$ Simplex de $\mathbb{R}^k$. D'aquesta manera m captura la informació a priori sobre on està centrat el vector dels perfils de probabilitat θ, i t reflexa el grau d'incertesa a priori sobre el nostre coneixement de θ.

Si el model Bayesià per y_i és el Model 1, la distribució a posteriori una vegada observat $y_i = (y_{i1}, \ldots, y_{ik})$ és

$$\pi(\theta|y_i) = Dir(a_1 + y_{i1}, \ldots, a_k + y_{ik}) \tag{7.9}$$

i per (7.7) tenim que l'esperança a posteriori de cada component de θ serà

$$E[\theta_j|y_i] = \frac{y_{ij}+a_j}{N+t} = \frac{N}{N+t}\frac{y_{ij}}{N} + \frac{t}{N+t}\frac{a_j}{t}, \qquad j = 1, \ldots, k,$$

i per tant es pot interpretar com un promig ponderat de l'estimador màxim versemblant

de θ_j i de l'esperança a priori de θ_j. La predictiva a posteriori és:

$$p_\pi(\widetilde{y}|y_i) = \text{Mult-Dir}(\widetilde{N}, a_1 + y_{i1}, \ldots, a_k + y_{ik}), \tag{7.10}$$

i per tant

$$E[\widetilde{y}|y_i] = \widetilde{N}E[\theta|y_i],$$

on $\widetilde{N}$ seria el compteig total per l'observació futura $\widetilde{y}$.

7.4 Model Bayesià jeràrquic

Quan hom analitza dades discretes fent servir el model Multinomial és molt freqüent que la variabilitat present entre les diferents files de la taula sigui major que l'esperada pel model. Aquest fenomen, anomenat de sobredispersió, sovint és degut a no incloure al model variables explicatives rellevants; quan aquestes variables explicatives són categòriques donen lloc a alguna mena d'agregació en les dades originals.

El fenomen de la sobredispersió és més la norma que no pas l'excepció. Quan la sobredispersió no pot ser corregida incorporant les variables que falten és convenient utilitzar models que tinguin en compte aquest fenomen, ja que del contrari la inferència sobre els paràmetres d'interès no serà correcta.

Una possible solució per tenir en compte la sobredispersió passa per modelar el procés de generació de dades a traves de dos nivells. En un primer nivell la distribució de cada fila de la taula, y_i, es suposa $\text{Mult}(N_i, \theta_i)$ però amb un perfil de probabilitats θ_i que canvia de fila a fila i que és aleatori. Modelant la distribució d'aquest valor esperat, un arriba de forma natural als models de barreja o d'efectes aleatoris, i en el cas Bayesià als models jeràrquics Bayesians.

Si bé el model jeràrquic pot fer augmentar la complexitat de la inferència, representa una situació més realista que permet modelar les dades a través de variables latents o variables explicatives no mesurades i que incorpora de forma natural els diferents nivells d'incertesa que hi ha.

Les components d'un model jeràrquic es poden expresar amb els tres nivells següents:

1. la distribució de probabilitat de les dades en funció dels paràmetres θ_i, $y_i|\theta_i \sim p(y_i|\theta_i)$ per $i = 1, \ldots, n$, que suposem condicionalment independents i per tant tals que $p(y|\theta) = \prod_{i=1}^{n} p(y_i|\theta_i)$, on $y = (y_1, \ldots, y_n)$ i on $\theta = (\theta_1, \ldots, \theta_n)$,

2. la distribució a priori dels paràmetres en funció dels hiperparàmetres α, $\theta_i|\alpha \sim \pi(\theta_i|\alpha)$ per $i = 1, \ldots, n$, que suposem condicionalment independents i idènticament distribuïts per tant tals que $\pi(\theta|\alpha) = \prod_{i=1}^{n} \pi(\theta_i|\alpha)$, i

3. la distribució hiperpriori dels hiperparàmetres que se suposa coneguda, $\alpha \sim \psi(\alpha)$

La diferència entre el model no jeràrquic i el jeràrquic és que el primer suposa que les dades són condicionalment independents i idènticament distribuïdes mentre que el segon només suposa que són condicionalment independents, però no idènticament distribuïdes.

Escriurem un model Bayesià jeràrquic per a una observació com:

$$\begin{gathered} y_i|\theta_i \sim p(y_i|\theta_i) \\ \theta_i|\alpha \sim \pi(\theta_i|\alpha) \\ \alpha \sim \psi(\alpha) \end{gathered}$$

Model 2. Bayesià jeràrquic per a una observació.

Quan modelem totes les observacions a l'hora suposem que les observacions $y = (y_1, \ldots, y_n)$ són condicionalment independents un cop conegut $\theta = (\theta_1, \ldots, \theta_n)$ i que $\theta = (\theta_1, \ldots, \theta_n)$ també ho són un cop conegut α i per tant podem escriure de forma equivalent el mateix model de forma compacta per a totes les observacions com:

$$\begin{gathered} y = (y_1, \ldots, y_n)|\theta = (\theta_1, \ldots, \theta_n) \sim \prod_{i=1}^{n} p(y_i|\theta_i) \\ \theta = (\theta_1, \ldots, \theta_n)|\alpha \sim \prod_{i=1}^{n} \pi(\theta_i|\alpha) \\ \alpha \sim \psi(\alpha) \end{gathered}$$

Model 2. Bayesià jeràrquic per a totes les observacions.

És habitual escollir distribucions hiperpriori molt poc informatives. D'aquesta manera es relativitza l'efecte de la informació subjectiva que s'aporta sobre la distribució a posteriori, i per tant de la inferència que se'n derivi.

Podem expressar el model de forma compacta com $(M, \pi, \psi) = (p(y_i|\theta_i), \pi(\theta_i|\alpha), \psi(\alpha))$. Per tant tenim que la densitat de l'observació i-èssima y_i donat θ_i és $p(y_i|\theta_i)$, per tot $i = 1..., n$, on cada θ_i és distribueix d'acord a una densitat de barreja o distribució a priori, governada per l'hiperparàmetre α que controla la distribució dels $\theta = (\theta_1, \ldots, \theta_n)$

i en particular la seva variabilitat. De forma indirecta, la distribucio d'α governa el grau de sobresipersió de les observacions $y = (y_1, \ldots, y_n)$.

El model jeràrquic Bayesià permet fer inferència tant sobre el primer nivell de $\theta = (\theta_1, \ldots, \theta_n)$ com sobre el nivell d'α. Això no és així a les aproximacions clàssiques que inclouen coeficients aleatoris per a un subconjunt de paràmetres i fan inferència sobre els hiperàmetres α sense permetre inferir sobre els paràmetres θ_i del primer nivell.

Partint del model jeràrquic a tres nivells, i aplicant el teorema de Bayes, tenim que la distribució conjunta de y, θ i α serà:

$$p(y, \theta, \alpha) = \prod_{i=1}^{n} p(y_i|\theta_i) \prod_{i=1}^{n} \pi(\theta_i|\alpha)\psi(\alpha)$$

i per tant la posteriori de θ i α condicionada a les dades observades és tal que:

$$\pi(\theta_1, \ldots, \theta_n, \alpha|y_1, \ldots\ y_n) \propto \prod_{i=1}^{n} p(y_i|\theta_i)\pi(\theta_i|\alpha)\psi(\alpha).$$

A la pràctica les estimacions sota aquest model Bayesià jeràrquic bàsic dels paràmetres del primer nivell, θ_i, acaben sent un compromís entre una mesura de tendència central global a priori de tots els θ_i i una funció dels valors observats y_i. És possible però sofisticar encara més el model introduint algun tipus d'estructura de manera que si, per exemple, s'assumeix una estructura espacial llavors el suavitzat es veu atret cap a una mesura de tendència central de les àrees adjacents. En aquesta tesi no els implementarem perquè no són necessaris pel nostre problema.

En lloc de definir el model jeràrquic a través de $(p(y_i|\theta_i), \pi(\theta_i|\alpha), \psi(\alpha))$ com el requadre del Model 2, podem formular dos versions alternatives d'aquest model per y. Aquestes formulacions alternatives sorgeixen de forma natural al colapsar nivells del model jeràrquic, al preu de tenir o bé el model estadístic o bé la distribució a priori més complexa.

Així si colapsem els dos primers nivells, el model estadístic i la a priori, tenim que el Model 2 per a totes observacions també és pot formular com:

$$\begin{gathered} y|\alpha \sim \int \ldots \int \textstyle\prod_{i=1}^{n} p(y_i|\theta_i)\pi(\theta_i|\alpha)d\theta_1 \ldots d\theta_n \\ \alpha \sim \psi(\alpha), \end{gathered}$$

i donat que $\int \ldots \int \prod_{i=1}^{n} p(y_i|\theta_i)\pi(\theta_i|\alpha)d\theta_1 \ldots d\theta_n = \prod_{i=1}^{n} \int p(y_i|\theta_i)\pi(\theta_i|\alpha)d\theta_i$, el model per a una observació serà:

$$y_i|\alpha \sim p(y_i|\alpha) = \int p(y_i|\theta_i)\pi(\theta_i|\alpha)d\theta_i$$
$$\alpha \sim \psi(\alpha).$$

No obstant al treballar amb la marginal desapareix l'estructura jeràrquica, cosa que no permet estimar les θ_i, ni permet aprofitar la flexibilitat dels mètodes de simulació MCMC. Si en canvi colapsem la distribucions a priori i la hiperpriori, tenim una segona versió del Model 2 en la que

$$y|\theta \sim \prod_{i=1}^{n} p(y_i|\theta_i)$$
$$\theta \sim \int \prod_{i=1}^{n} \pi(\theta_i|\alpha)\psi(\alpha)d\alpha.$$

A l'hora de triar la formulació a utilitzar caldrà tenir en compte l'eficiència computacional, així com els paràmetres sobre els quals volem fer inferència.

7.5 Model Multinomial Bayesià: cas jeràrquic

Tornant a les dades del capítol 6, assumir que la distribució dels vots per a cada districte o zrp, i per tant una fila de la Taula 6.5, són condicionalment independents i idènticament distribuïdes com a Mult(N_i, θ) on el perfil de probabilitats és el mateix per tots els districtes o zrp és poc realistic. És més raonable assumir que existeix certa heterogeneïtat en aquests perfils.

Una opció per modelar aquesta heterogeneïtat, és adoptar un model jeràrquic bayesià, suposant una distribució per als paràmetres per la observació y_i, $\theta_i = (\theta_1, \ldots, \theta_k)$, assumint que els valors de θ_i per $i = 1, \ldots, n$ són intercanviables. Els paràmetres $\theta_1, \ldots, \theta_n$ es diu que són intercanviables si la seva distribució conjunta $\pi(\theta_1, \ldots, \theta_n)$ és invariant davant de permutacions dels índexs $(1, \ldots, n)$, i per tant no incorpora informació a priori que permeti distingir les θ_i entre elles. Per tant la distribució hiperpriori assumeix simetria entre els paràmetres de diferents files (observacions). D'aquesta manera la distribució hiperpriori ens permetrà fer inferència sobre el grau de sobredispersió present a les dades. Si els models no incorporen correctament aquesta heterogeneïtat, i aquesta està present a les dades, la inferència sobre els paràmetres d'interès serà incorrecta.

A continuació llistem diferents models Multinomials jeràrquics que classificarem d'entrada entre els basats en el model Multinomial-Dirichlet, i els basats en el model Multinomial-logistic.

7.5.1 Models basats en la Multinomial-Dirichlet

Tots els models considerats en aquesta subsecció parteixen del model $\text{Mult}(N_i, \theta_i)$ per $y_i|\theta_i$ amb distribució a priori $\text{Dir}(\alpha_1, \ldots, \alpha_k)$ per θ_i:

$$
\begin{aligned}
y_i|\theta_i &\sim \text{Multinomial}(N_i, \theta_i) \\
\theta_i|\alpha &\sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_k) \\
\alpha &\sim \psi(\alpha)
\end{aligned}
$$

Model 3. Multinomial jeràrquic amb priori Dirichlet.

Aquí el vector $\alpha = (\alpha_1, \ldots, \alpha_k)$ és una variable aleatòria i no un vector conegut. Per acabar de definir el model caldrà assignar una distribució coneguda $\psi(\alpha)$ per α, que serà lo únic que distingira els models d'aquesta subsecció. Com ja s'ha dit, una parametrització alternativa de la Dirichlet que pot ser més útil a l'hora d'incorporar informació a priori és $\text{Dir}(\tau(\mu_1, \ldots, \mu_k))$ on:

$$\mu = \left(\frac{\alpha_1}{\sum_{j=1}^{k} \alpha_j}, \ldots, \frac{\alpha_k}{\sum_{j=1}^{k} \alpha_j}\right) = E[\theta_i|\alpha]$$

representa el perfil de probabilitats i on

$$\tau = \sum_{j=1}^{k} \alpha_j$$

representa el grau d'heterogeneïtat, que regula la variabilitat de la distribució de les θ_i donat que

$$Var(\theta_i|\alpha) = \frac{\frac{\alpha_i}{\tau}(1 - \frac{\alpha_i}{\tau})}{\tau + 1}$$

i que

$$Cov(\theta_i, \theta_j|\alpha) = \frac{-\frac{\alpha_i \alpha_j}{\tau^2}}{\tau + 1},$$

de manera que com més grans els valors de τ més homogènies seran les θ_i, i en el cas límit en el que τ tendeix a infinit les θ_i seran totes iguals, $\theta = \theta_i$ per $i = 1, \ldots, n$, i per tant el model jeràrquic esdevindrà un model no jeràrquic.

Tal i com s'ha indicat a la secció 7.4 podríem formular el Model 3 de dues formes alternatives, com el model no jeràrquic:

$$y|\alpha \sim \prod_{i=1}^{n} \text{Mult-Dir}(N_i, \alpha_1, \ldots, \alpha_k)$$
$$\alpha \sim \psi(\alpha),$$

i com a:

$$y|\theta = (\theta_1, \ldots, \theta_n) \sim \prod_{i=1}^{n} p(y_i|\theta_i)$$
$$\theta = (\theta_1, \ldots, \theta_n) \sim \int \prod_{i=1}^{n} \pi(\theta_i|\alpha)\psi(\alpha)d\alpha.$$

A l'hora de triar la formulació a utilitzar caldrà tenir en que la formulació no jeràrquica no permet fer inferència sobre els paràmetres intermitjos, θ_i. Quan formulem el model cluster discutirem la formulació d'acord als nostres objectius i/o necessitats. A continuació llistem diferents models jeràrquics obtinguts a l'anar variant $\psi(\alpha)$:

1) Una opció és utilitzar com a ditribució $\psi(\alpha)$ la distribució a priori conjugada de la distribució $\text{Dir}(\alpha_1, \ldots, \alpha_k)$. No obstant, tot i que la distribució $\text{Dir}(\alpha_1, \ldots, \alpha_k)$ pertany a la familia exponencial es tracta d'un cas especial en el que la distribució conjugada no té cap avantatge respecte a les altres prioris i que no té ni tant sols forma tancada (Robert 2004, Lefkimmiatis 2009). Això fa que aquesta opció sigui pràcticament inexplorada a la literatura.

2) Una altra opció sorgeix com a extensió natural de la proposta feta per a la Beta-Binomial en un exemple de Gelman et al. (1995), que consisteix a fer servir com a hiperpriori per els paràmetres de la $\text{Dir}(\alpha_1, \ldots, \alpha_k)$, $\psi(\alpha)$ una distribució proporcional a $(\alpha_1 + \ldots + \alpha_k)^{-(k+1/2)}$:

$$y_i|\theta_i \sim \text{Multinomial}(N_i, \theta_i)$$
$$\theta_i|\alpha \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_k)$$
$$\psi(\alpha) \propto (\alpha_1 + \ldots + \alpha_k)^{-(k+1/2)}$$

Model 3.2. Multinomial jeràrquic amb priori Dirichlet i hiperpriori de Gelman.

Schuckers (1998) implementen aquest model per als resultats del referèndum per l'independència d'Eslovènia del 1991. Més recentment Giron, Ginebra i Riba (2005), proposen aquesta hiperpriori com a extensió del seu treball, però no l'arriben a implementar per la dificultat que representa implementar algoritmes MCMC per mostrejar de les distribucions a posteriori corresponents.

3) Good (1976) proposa reparametritzar la distribució Dirichlet a través de $\alpha = (\alpha_1, \ldots, \alpha_k) = \tau(\mu_1, \ldots, \mu_k)$ on $\sum_{j=1}^{k} \mu_j = 1$, i suposa que $(\mu_1, \ldots, \mu_k)$ és igual a un vector conegut, $(m_1, \ldots, m_k)$, i que la distribució de probabilitat de τ, $\psi(\tau)$, es una *log-Cauchy*:

$$
\begin{aligned}
y_i|\theta_i &\sim \text{Multinomial}(N_i, \theta_i) \\
\theta_i|\alpha &\sim \text{Dirichlet}(\alpha = \tau(m_1, \ldots, m_k)) \\
\tau &\sim \textit{log}\text{-Cauchy}
\end{aligned}
$$

Model 3.3. Multinomial jeràrquic amb priori Dirichlet i hiperpriori *log*-Cauchy.

En aquest model les dades actualitzen la informació a priori de τ peró no de $\mu = m = (m_1, \ldots, m_k)$ que ve fixat d'entrada. Això fa que si aquest vector perfil m no és el correcte ens trobem amb que quan més lluny estàn els vertaders valors θ_i de m, més petita és la τ per tal de que el model tingui prou flexibilitat per fer compatibles les θ_i amb m. Aquest fet provoca que la distribució a posteriori de la τ tendeixi a subestimar el verdader valor de τ.

4) Una quarta opció, utilitzada per exemple a Congdon (2005), consisteix a suposar que les k components de $\alpha = (\alpha_1, \ldots, \alpha_k)$ són independents i que cada α_j té una distribució Gamma(a_j, b_j):

$$
\begin{aligned}
y_i|\theta_i &\sim \text{Multinomial}(N_i, \theta_i) \\
\theta_i|\alpha &\sim \text{Dirichlet}(\alpha = (\alpha_1, \ldots, \alpha_k)) \\
\alpha &\sim \textstyle\prod_{j=1}^{k} \text{Gamma}(a_j, b_j)
\end{aligned}
$$

Model 3.4. Multinomial jeràrquic amb priori Dirichlet i hiperpriori Gamma.

5) Nandram (1998) proposa reparametritzar la Dirichlet a través de $(\alpha_1, \ldots, \alpha_k) = \tau(\mu_1, \ldots, \mu_k)$ on $\sum_{j=1}^{k} \mu_j = 1$, i utilitzar com a hiperpriori de μ una Dir$(m_1, \ldots, m_k)$ i com a hiperpriori de τ una Gamma(c, d):

$$
\begin{aligned}
y_i|\theta_i &\sim \text{Multinomial}(N_i, \theta_i) \\
\theta_i|\alpha &\sim \text{Dirichlet}(\alpha = \tau(\mu_1, \ldots, \mu_k)) \\
\mu = (\mu_1, \ldots, \mu_k) &\sim \text{Dirichlet}(m_1, \ldots, m_k) \\
\tau &\sim \text{Gamma}(c, d)
\end{aligned}
$$

Model 3.5. Multinomial jeràrquic amb priori Dirichlet i hiperpriori Dirichlet-Gamma.

Així μ modela l'esperança del perfil de probabilitats θ_i, i τ determina el nivell d'heterogeneitat de les θ_i, i per tant el grau de sobredispersió de les dades. Valors de τ petits equival a una major variabilitat dels perfils de probabilitat θ_i, i per tant a més sobredispersió de les files de la taula.

Es pot comprobar que si al Model 3.4 suposem que $b_1 = \ldots = b_k = b$, aquest es converteix en un cas particular del Model 3.5 amb $(m_1, \ldots, m_k) = (a_1, \ldots, a_k)$, $c = \sum_{j=1}^{k} a_j$ i $d = b$.

Hem implementat els Models 3.4 i 3.5 mitjançant el programa WinBugs. Tot i que la inferència resultant d'ambdós models és força similar, el Model 3.5 presenta l'avantatge de que quan simules de la distribució a posteriori fent servir el mètode MCMC les seves cadenes convergeixen més ràpidament cap al procés estacionari, alhora que les simulacions presenten menor autocorrelació. Per aquest motiu hem acabat utilitzant el Model 3.5.

7.5.2 Model basat en el model logístic

Una parametrització alternativa de la distribució Multinomial, neix del fet que la distribució Multinomial és la distribució conjunta de k variables independents, y_{ij}, amb distribució $y_{ij} \sim \text{Poisson}(\lambda_j)$ condicionada a que $\sum_{j=1}^{k} y_{ij} = N_i$, ja que en aquest cas:

$$p(y_i|N_i, \lambda) = \frac{e^{-\sum \lambda_j} \prod_{j=1}^{k} \frac{\lambda_j^{y_{ij}}}{y_{ij}!}}{\frac{e^{-\sum \lambda_j} (\sum \lambda_j)^{\sum_{j=1}^{k} y_{ij}}}{N_i!}} = \frac{N_i!}{\prod_{j=1}^{k} y_{ij}!} \prod_{j=1}^{k} (\frac{\lambda_j}{\sum \lambda_j})^{y_{ij}} = \frac{N_i!}{\Pi_{j=1}^{k} y_{ij}!} \Pi_{j=1}^{k} \theta_j^{y_{ij}}, \quad (7.11)$$

on $\lambda = (\lambda_1, \ldots, \lambda_k)$ és una reparametrització de $\theta = (\theta_1, \ldots, \theta_k)$ tal que, $\theta_j = \frac{\lambda_j}{\sum \lambda_j}$.

El model resultant d'assumir una distribució a priori $\lambda_j \sim \text{Gamma}(a_j, b)$, per $j = 1, \ldots, k$, és equivalent al Model 1 (model Multinomial Bayesià no jeràrquic amb distribució a priori $\text{Dir}(a_1, \ldots, a_k)$.

Aquesta parametrització permet definir models Multinomials Bayesians alternatius. El més utilitzat és el model Multinomial-logístic (Leonard i Hsu 1994), que suposa una distribució a priori Normal multivariant per $(log\lambda_1, \ldots, log\lambda_k)$ amb mitjana $(\mu_1, \ldots, \mu_k)$ i matriu de covariances Σ de dimensió k, i utilitzar com a distribució hiperpriori de μ_j una Normal, per tenir la identificabilitat és necessari imposar a més a més una restricció a la suma del vector de mitjanes, i com a hiperpriori de $1/\Sigma$ una Wishart amb k graus de llibertat. Escrivim aquest model com:

$$
\begin{gathered}
y_i|\theta_i \sim \text{Multinomial}(N_i, \theta_i = (\theta_{i1}, \ldots, \theta_{ik})) \\
\theta_i = (\theta_{i1}, \ldots, \theta_{ik}) = (\tfrac{\lambda_{i1}}{\sum_{j=1}^{k}\lambda_{ij}}, \ldots, \tfrac{\lambda_{ik}}{\sum_{j=1}^{k}\lambda_{ij}}) \\
log\lambda_i = (log\lambda_1, \ldots, log\lambda_k) \sim \text{Normal}((\mu_1, \ldots, \mu_k), \Sigma) \\
\mu_j = \rho_j - \overline{\rho} \ \text{ t.q. } \textstyle\sum_{j=1}^{k}\mu_j = 0 \\
\rho_j \sim \text{Normal}(m_j, s_j) \\
1/\Sigma \sim \text{Wishart}(Id_{k\times k}, k)
\end{gathered}
$$

Model 4. Multinomial-logístic.

Aquesta aproximació és útil en el cas de voler incoroporar covariables en el model, però aquest no és el nostre cas.

7.6 Model per a s clusters multinomials: cas no jeràrquic

La majoria d'anàlisi cluster es fan partint de dades contínues i utilitzen algoritmes heurístics de partició, que porten a una metodologia que dificulta fer inferència sobre els paràmetres de les poblacions i sobre la probabilitat de pertànyer a un cluster. No obstant, l'anàlisi cluster també pot ser formulat a partir de models paramètrics de mixtura discreta, on dues observacions pertanyen al mateix cluster si tenen la mateixa distribució (McLachlan i Basford, 1988; Banfield i Raftery, 1993).

Un model de mixtura discreta suposa que la resposta y ha sigut generada per una de les s poblacions possibles. Una distribució de mixtura de s components té associats uns pesos $\omega_1, \omega_2, \ldots, \omega_s$ tals que $\omega_j > 0$, i $\sum_{j=1}^{s}\omega_j = 1$, i la seva densitat de probabilitat serà:

$$p(y_i|\omega_1, \ldots, \omega_s, \theta_1, \ldots, \theta_s) = \sum_{r=1}^{s}\omega_r p_r(y_i|\theta_r), \tag{7.12}$$

on $p_r(y_i|\theta_r)$ és la densitat de probabilitat de les observacions del cluster r.

Aquest plantejament és apropiat quan existeixen diferents subgrups el tamany del qual es proporcional a ω_r, i desconeixem a quin dels s subgrups pertany cada observació y_i. L'enfoc bayesià permet formular tests sobre l'existència de dos o més clusters, i permet fer inferència sobre la probabilitat de que una observació pertanyi a un dels clusters, sobre les ω_r's i sobre els paràmetres que caracteritzen les s poblacions, θ_r.

Donades les dades descrites al capítol 6, estem interessats a dividir el conjunt de districtes i zrp en dos o més subconjunts lo més homogenis possibles amb l'objectiu d'identificar els patrons dels grups, així com identificar els districtes que pertanyen a cada grup. L'anàlisi cluster que formularem particiona el conjunt de totes les dades en grups més homogenis que el total, però sense imposar cap restricció d'ordre espacial per formar els grups. Un dels objectius és comporvar si el model cluster captura tota l'estructura espacial de les dades o bé si encara queda estructura espacial als errors del model.

A continuació es descriu el model de mixtura de s components Multinomial en el cas que s'utilitza com a distribució a priori la conjugada, la Dirichlet. El cas particular de 2 clusters s'ha utilitzat per exemple a Giron, Ginebra i Riba (2005). Sota aquest model, cada una de les observacions, $y_i = (y_{i1}, \ldots, y_{ik})$, corresponents a una de les files de la Taula 6.5, es considera que prové d'una de les s possibles distribucions $\text{Mult}(N_i, \theta_r)$ per $r = 1, \ldots, s$ amb probabilitat ω_r, i per tant seguint la formulació de (7.12) escrivim:

$$p(y_i|N_i, \omega, \theta_1, \ldots, \theta_s) = \sum_{r=1}^{s} \omega_r \text{Mult}(y_i|N_i, \theta_r) \tag{7.13}$$

on $\theta_r = (\theta_{r1}, \ldots, \theta_{rk})$ representa el perfil de probabilitat del cluster r, i on ω_r representa la proporció d'elements de la mostra que pertanyen al cluster r i que en el nostre cas es tradueix a la proporció de districtes (zrp) del cluster r.

El conjunt de les files de la taula, $y_1, \ldots, y_n$, les considerem com n observacions condicionalment independents i distribuïdes d'acord a una mixtura de multinomials, i per tant amb funció de versemblança:

$$L_y(\theta) \propto \prod_{i=1}^{n} \sum_{r=1}^{s} \omega_r \text{Mult}(y_i|N_i, \theta_r). \tag{7.14}$$

Aquesta expressió de la versemblança és molt difícil de tractar, ja que és una expressió que conté s^n sumands. I tanmateix l'assignació dels districtes a cadascun dels grups, que és el problema central de l'anàlisi de cluster, no es dedueix directament de 7.14.

L'assignació de cada districte (zrp) a cadascun dels s clusters es pot aconseguir introduint al model un nou vector de variables categòriques latents no observables, $\zeta = (\zeta_1, ..., \zeta_n)$,

tal que $\zeta_i = r$ si l'observació y_i pertany al cluster r i per tant és tal que

$$\begin{aligned} p(y_i|N_i, \theta_1, \ldots, \theta_s, \zeta_i = 1) &= Mult(N_i, \theta_1) \\ p(y_i|N_i, \theta_1, \ldots, \theta_s, \zeta_i = 2) &= \text{Mult}(N_i, \theta_2) \\ &\vdots \\ p(y_i|N_i, \theta_1, \ldots, \theta_s, \zeta_i = s) &= \text{Mult}(N_i, \theta_s), \end{aligned}$$

que és una idea de Dempster (1977). D'aquesta manera podem re-escriure la funció de versemblança com:

$$p(y|N, \theta_1, \ldots, \theta_s, \zeta) = \prod_{i=1,..n} \text{Mult}(N_i, \theta_{\zeta_i}). \tag{7.15}$$

Considerant la distribució conjunta de (y, ζ),

$$p(y, \zeta|N, \omega, \theta_1, \ldots, \theta_s) = p(y|N, \omega, \theta_1, \ldots, \theta_s, \zeta)\pi(\zeta|\omega) = \prod_{i=1}^{n}\prod_{r=1}^{s} \omega_{\zeta_i}\text{Mult}(y_i|N_i, \theta_{\zeta_i}),$$

la mixtura de la versemblança desapareix i això simplifica molt els càlculs. A més a més la introducció de les variables latents, ζ_i, permet simplificar no només el model de mixtura sinó també el càlcul de la distribució a posteriori dels paràmetres, $\omega, \theta_1, \ldots, \theta_s$ i l'assignació dels districtes a cada grup. De fet ζ esdevé el principal paràmetre d'interès perquè la distribució a posteriori de cada component, $\pi(\zeta_i|y)$, representa la probabilitat a posteriori de que l'observació i-èssima pertanyi a cadascun dels clusters, i això és justament el que hom necessita per assignar les observacions a cada cluster.

Sota el paradigma bayesià hem d'escollir una distribució a priori per $(\omega, \theta_1, \ldots, \theta_s, \zeta)$, que s'adapti a les nostres necessitats. En el cas no jeràrquic suposarem que:

$$\begin{aligned} \pi(\theta_1) &= \text{Dir}(a_{11}, \ldots, a_{1k}) \propto \theta_{11}^{a_{11}-1} \ldots \theta_{1k}^{a_{1k}-1} \\ \pi(\theta_2) &= \text{Dir}(a_{21}, \ldots, a_{2k}) \propto \theta_{21}^{a_{21}-1} \ldots \theta_{2k}^{a_{2k}-1} \\ &\vdots \\ \pi(\theta_s) &= \text{Dir}(a_{s1}, \ldots, a_{sk}) \propto \theta_{s1}^{a_{s1}-1} \ldots \theta_{sk}^{a_{sk}-1} \end{aligned}$$

on els s vectors $a_r = (a_{r1}, \ldots, a_{rk})$ són coneguts. Al tractar $\zeta = (\zeta_1, \ldots, \zeta_n)$ com una variable latent, no observable, s'ha de tractar com un paràmetre, i per tant li hem d'especificar una distribució a priori. Per la naturalesa de la variable ζ, una distribució a priori sobre cada una de les seves components ζ_i serà del tipus:

$$\pi(\zeta_i = r|\omega) = \begin{cases} \omega_1 & r = 1 \\ \omega_2 & r = 2 \\ \vdots & \vdots \\ \omega_s & r = s \end{cases}$$

i per tant

$$\pi(\zeta = (\zeta_1, \ldots, \zeta_n)|\omega) = \prod_{i=1}^{n} \omega_{\zeta_i}.$$

Per al paràmetre ω, al tractar-se d'un vector de probabilitats lo natural és escollir una distribució hiperpriori $\text{Dir}(b_1, \ldots, b_s)$. Finalment s'assumeix que $\omega, \theta_1, \ldots, \theta_s$ són independents. A mode de recapitulació al capítol 9 farem servir el model del següent requadre amb $s = 1, \ldots, 5$.

$$\begin{aligned}
y_i|\theta_1, \ldots, \theta_s &\sim \text{Multinomial}(N_i, \theta_{\zeta_i}) \\
\theta_1 &\sim \text{Dirichlet}(a_{11}, \ldots, a_{1k}) \\
&\vdots \\
\theta_s &\sim \text{Dirichlet}(a_{s1}, \ldots, a_{sk}) \\
\pi(\zeta_i = r|\omega) &= \begin{cases} \omega_1 & r = 1 \\ \omega_2 & r = 2 \\ \vdots & \vdots \\ \omega_s & r = s \end{cases} \\
(\omega_1, \ldots, \omega_s) &\sim \text{Dirichlet}(b_1, \ldots, b_s),
\end{aligned}$$

Model 5. Model Bayesià no jeràrquic per a s clusters Multinomials.

Observem que per aquest model de s clusters hem definit un vector de variables latents, $\zeta = (\zeta_1, \ldots, \zeta_n)$, on cada component ζ_i pot prendre s valors diferents. El mateix model es pot re-escriure de forma equivalent definint les variables latents ζ_i com a un vector de s components que prenen valors 0 o 1 tals que $\zeta_i = (\zeta_{1i}, \ldots \zeta_{si})|w \sim \text{Mult}(1, (\omega_1, \ldots, \omega_s))$. Tot i l'equivalència d'aquestes dues parametritzacions, aquesta segona versió obliga a escriure el model de manera menys compacta, i per això hem triat definir les variables latents com un únic vector de variables ζ_i que pren valors de 1 a s. Per la parametrització que considera ζ_i com a un vector de s components *dummy* hauríem d'escriure la versemblança com a $\Pi_{i=1}^{n}\Pi_{r=1}^{s} p(y_i, |\theta_r)^{\zeta_{ri}}$ mentre que amb la primera parametrització escrivim $\Pi_{i=1}^{n} p(y_i, |\theta_{\zeta_i})$.

El model de mixtura presenta un problema d'identificabilitat ja que per exemple en el cas de dos clusters, $s = 2$, tenim que $p(y|\omega_1, \theta_1, \theta_2) = p(y|\omega_2 = 1 - \omega_1, \theta_2, \theta_1)$, i la funció de versemblança avaluada a $(\omega_1, \theta_1, \theta_2)$ val igual que a $(\omega_2, \theta_2, \theta_1)$. Davant d'aquest problema cal una curosa inspecció visual de les cadenes simulades de les distribucions a posteriori per tal de no cometre errors en la inferència. Altres autors proposen proposen afegir constriccions d'odre als paràmetres, per exemple en el model de dos clusters $\theta_2 > \theta_1$, per

tal d'eliminar el problema de la identificabilitat. No obstant Jasra *et al* (2005) il·lustren com imposar aquest tipus de constriccions no es garantia d'eliminar el problema.

La distribució a posteriori per $(\omega, \theta_1, \ldots, \theta_s)$ condicionat a y i ζ la podem obtenir de

$$\pi(\omega, \theta_1, \ldots, \theta_s | y, \zeta) \propto p(y, \zeta | N, \omega, \theta_1 \ldots, \theta_s)\pi(\omega, \theta_1, \ldots, \theta_s) =$$

$$p(y|N, \omega, \theta_1, \ldots, \theta_s, \zeta)\pi(\theta_1) \ldots \pi(\theta_s)\pi(\zeta|\omega)\pi(\omega).$$

Amb la introducció de les variables latents, ζ, condicionat respecte a y i ζ, es manté l'estructura conjugada perquè la distribució a posteriori per $(\omega, \theta_1, \ldots, \theta_s)$ donat y i ζ torna a ser el producte de s distribucions Dirichlet actualitzades per $\theta_1, \ldots, \theta_s$ i el producte d'una distribució Dirichlet actualitzada per ω, de manera que $\theta_1, \ldots, \theta_s$ i ω són encara condicionalment independents, ja que

$$\pi(\omega, \theta_1 \ldots, \theta_s | y, \zeta) \propto \prod_{r=1}^{s} \prod_{j=1}^{k} \theta_{rj}^{a_{rj}(\zeta)-1} \prod_{r=1}^{s} \omega_r^{b_r(\zeta)-1}, \tag{7.16}$$

on $b_r(\zeta) = b_r + \sum_{i=1}^{n} I_{\{\zeta_i = r\}}$ i $a_{rj}(\zeta) = a_{rj} + \sum_{i=1}^{n} y_{ij} I_{\{\zeta_i = r\}}$, i on $I_{\{A\}}$ és la funció indicadora del succés A. No obstant això, la distribució a posteriori de $(\omega, \theta_1, \ldots, \theta_s)$ donat y, és una mixtura intractable de s^n termes.

En lloc de càlcular les distribucions marginals a posteriori dels paràmetres de forma analítica, $\pi(\zeta|y)$, $\pi(\omega|y)$ i $\pi(\theta_1, \ldots, \theta_s|y)$, hom pot recórrer a mètodes de mostreig MCMC, com els descrits a Lavine i West (1992) i Diebolt i Robert (1994). En aquesta tesi hem utilitzat el WinBugs per simular de les distribucions a posteriori.

La probabilitat a posteriori que y_i pertanyi al cluster r-èssim serà $\pi(\zeta_i = r|y)$, i això és fàcilment estimable a través de les cadenes que proporcionen els mètodes MCMC. Una vegada estimada aquesta probabilitat el més natural és adjudicar cada observació al cluster amb la distribució a posteriori més gran, i per tant assignant y_i al cluster r si la moda a posteriori de $\pi(\zeta_i|y)$ val r.

7.7 Model per a s clusters multinomials: cas jeràrquic

El mateix plantejament d'anàlisi cluster fet per al cas Multinomial no jeràrquic es pot estendre al cas jeràrquic adaptant el model proposat per Nandram (1998) pel cas homogeni d'un sol cluster (Model 3.5). Escrivim el model Multinomial per a s clusters jeràrquic com:

$$
\begin{aligned}
y_i|\theta_i &\sim \text{Multinomial}(N_i, \theta_i) \\
\theta_i|\tau, \mu, \zeta &\sim \text{Dirichlet}(\tau_{\zeta_i}(\mu_{\zeta_i 1}, \ldots, \mu_{\zeta_i k})) \\
\mu_1 = (\mu_{11}, \ldots, \mu_{1k}) &\sim \text{Dirichlet}(m_{11}, \ldots, m_{1k}) \\
&\vdots \\
\mu_s = (\mu_{s1}, \ldots, \mu_{sk}) &\sim \text{Dirichlet}(m_{s1}, \ldots, m_{sk}) \\
\tau_1 &\sim \text{Gamma}(c_1, d_1) \\
&\vdots \\
\tau_s &\sim \text{Gamma}(c_s, d_s) \\
\pi(\zeta_i = r|\omega) &= \begin{cases} \omega_1 & r = 1 \\ \omega_2 & r = 2 \\ \vdots & \vdots \\ \omega_s & r = s \end{cases} \\
(\omega_1, \ldots, \omega_s) &\sim \text{Dirichlet}(b_1, \ldots, b_s),
\end{aligned}
$$

Model 6.1. Model Multinomial Bayesià jeràrquic per a s clusters.

Ara $(\mu_{r1}, \ldots, \mu_{rk})$ representa el perfil de probabilitat del cluster r, τ_r representa el grau d'heterogeneïtat del cluster r, i ω_r representa la probabilitat de que una fila de la taula y_i donada pertanyi al cluster r, i per tant acaba representant la proporció de districtes assignats al cluster r.

Aquest mateix model es pot plantejar de manera equivalent en forma de model no jeràrquic com a un model de s clusters de Multinomial-Dirichlet on cada una de les observacions, $y_i = (y_{i1}, \ldots, y_{ik})$, té com a distribució de probabilitat condicionada:

$$p(y_i|N_i, \omega, \alpha_1, \ldots, \alpha_s) = \sum_{r=1}^{s} \omega_r \text{Mult-Dir}(y_i|N_i, \alpha_r),$$

és a dir,

$$
\begin{aligned}
y_i|\tau_1,\dots\tau_s,\mu_1,\dots,\mu_s &\sim \text{Multinomial-Dirichlet}(N_i,\tau_{\zeta_i}\mu_{\zeta_i})\\
\mu_1=(\mu_{11},\dots,\mu_{1k}) &\sim \text{Dirichlet}(m_{11},\dots,m_{1k})\\
&\vdots\\
\mu_s=(\mu_{s1},\dots,\mu_{sk}) &\sim \text{Dirichlet}(m_{s1},\dots,m_{sk})\\
\tau_1 &\sim \text{Gamma}(c_1,d_1)\\
&\vdots\\
\tau_s &\sim \text{Gamma}(c_s,d_s)\\
\pi(\zeta_i=r|\omega)&=\begin{cases}\omega_1 & r=1\\ \omega_2 & r=2\\ \vdots & \vdots\\ \omega_s & r=s\end{cases}\\
\omega=(\omega_1,\dots,\omega_s) &\sim \text{Dirichlet}(b_1,\dots,b_s),
\end{aligned}
$$

Model 6.2. Model Multinomial-Dirichlet Bayesià per a s clusters.

El nostre objectiu a l'hora de formular el model per a s clusters multinomials jeràrquic serà fer inferència sobre:

a) ζ_i, és a dir de la pertinença de cada districte a cadascun dels clusters,
b) sobre μ_r, que descriu els perfils de cada un dels s cluster i les diferències entre ells, i
c) sobre τ_r que mesura el grau d'heterogeneïtat de cada cluster.

Com que en el nostre cas no ens fa falta fer inferència sobre $\theta_1,\dots,\theta_n$, la segona formulació, denotada com a Model 6.2, és prou adequada. Si bé formular el model jeràrquic a tres nivells, tal i com ho fem al Model 6.1, permetria afrontar objectius secundaris relatius al segon nivell que fan referència al perfil de cadascun dels districtes com avaluar probabilitats del tipus $p(\theta_{ij}>\theta_{i'j}|y)$ on $i\neq i'$ per comparar perfils entre districtes, aquests per nosaltres no són l'objectiu principal. No obstant per raons de computació i d'implementació és avantatjós tractar el model a través de la formulació del Model 6.1.

7.8 Presentació de resultats

La inferència Bayesiana passa per avaluar i presentar la distribució conjunta a posteriori del vector de paràmetres, i no té perquè limitar-se a donar estimacions puntuals ni regions

de confiança. A l'hora de la veritat però, la dimensió de l'espai de paràmetres acostuma a ser massa gran per representar tota la distribució a posteriori, i cal dedicar esforços a pensar com presentar de forma creativa, eficient i entenedora el màxim d'informació en el mínim de gràfics i taules. La presentació dels resultats no només s'ha de pensar a mida per cada model, sinó que també s'ha d'adaptar al context del problema a analitzar i a les preguntes que es vol respondre.

Així per exemple per a l'anàlisi de les dades electorals de Barcelona basats en el model Multinomial no jeràrquic amb s clusters, el vector de paràmetres d'interès $(\theta_1, \ldots, \theta_s, \zeta_i)$ pertany al $\{\text{Símplex de } \mathbb{R}^k\}^s \times \{1, \ldots, s\}^{248}$, on $k = 8$ o 9 i on $s = 1, \ldots, 5$. Per resumir la informació de forma compacta en el capítol 9 representarem gràfics que resumiran les distribucions a posteriori de:

$$\pi(\theta_{r1}, \ldots, \theta_{rk}|y),$$

per $r = 1, \ldots, s$,

$$\pi(log\frac{\theta_{i1}}{\theta_{j1}}|y), \ldots, \pi(log\frac{\theta_{ik}}{\theta_{jk}}|y),$$

per $j = 1, \ldots, s$ i $i = j + 1, \ldots, s$, i

$$\pi(\omega_r|y),$$

per $r = 1, \ldots, s$, i per al model Multinomial jeràrquic amb s clusters representarem gràfics que resumiran les distribucions a posteriori de:

$$\pi(\mu_{r1}, \ldots, \mu_{rk}|y),$$

per $r = 1, \ldots, s$,

$$\pi(log\frac{\mu_{i1}}{\mu_{j1}}|y), \ldots, \pi(log\frac{\mu_{ik}}{\mu_{jk}}|y),$$

per $j = 1, \ldots, s$ i $i = j + 1, \ldots, s$, i

$$\pi(\omega_r|y),$$

així com de

$$\pi(\tau_r|y),$$

per $r = 1, \ldots, s$.

A mode de resum també representarem una taula amb el valor esperat a posteriori d'algunes d'aquestes distribucions i, mapes de classificació dels districtes a cada cluster fruit de categoritzar-los en base a la moda a posteriori de $\pi(\zeta_i|y)$.

En el context de la validació del model, el qual descriurem de forma detallada al Capítol 8, presentarem, majoritàriament de forma gràfica, les distribucions a posteriori de mesures de discrepància global i l'esperança a posteriori de mesures de discrepància a nivell de districte i de zrp, algunes de les quals representarem en mapes, així com distribucions

predictives a posteriori d'estadístics creats ad-hoc. Totes les representacions gràfiques han estat realitzades amb el programari de lliure distribució R.

Capítol 8

Validació i millora del model

La validació d'un model és una part crucial de qualsevol anàlisi estadística. Abans d'emetre conclusions d'una anàlisi l'investigador hauria d'estar segur de que les característiques importants de les dades han estat adequadament capturades pel model. Quan això falla cal millorar el model. Avaluar la qualitat de l'ajust del model escollit també serveix per conèixer les seves limitacions. L'objectiu no és tant determinar si el model és cert o fals, ja que un model rarament és perfecte, si no determinar si les limitacions del model tenen un efecte substancial en les conclusions de l'anàlisi.

Part de la dificultat de validar un model Bayesià neix de la necessitat d'estudiar simultàniament les limitacions tant del model estadístic com de la distribució a priori. Donat que el model estadístic i la distribució a priori queden integrades en la distribució predictiva, l'estudi d'aquesta serà crucial en la validació dels models.

La validació de models Bayesians és una línea de recerca que encara està molt oberta. Les principals estratègies que es fan servir a la literatura són les proposades a Box (1980), Petit (1986), Allenby i Rossi (1999), Carlin i Louis (2000), Bayarri i Berger (2000), Gelman et. al. (2003), Stern i Sinharay (2005) i Bayarri i Castellanos (2007). Aquí les descrivim parcialment, insistint sobretot en la manera de validar i millorar el model que hem adoptat pel nostre problema.

8.1 Validació i selecció de models

En aquesta tesi, el problema de la selecció de models apareix a l'hora de:

1. triar el nombre de clusters, i
2. triar entre el model jeràrquic i el model no jeràrquic.

La validació de models és el motor del procés seqüencial de construcció de models. La validació i la selecció de models són problemes complementaris ja que la validació d'un model sempre suggereix un conjunt de possibles models alternatius i un cop seleccionat un model aquest s'ha de tornar a validar, ja que el millor model d'entre la llista de models seleccionats sovint encara té algunes limitacions que es poden corregir simplificant o estenent el model.

En aquesta tesi utilitzarem la validació de models com a principal eina de selecció de models. La idea serà escollir el model més senzill que capturi les característiques més importants de les dades. No obstant existeixen altres alternatives més pròpies de selecció de models i que en aquest capítol explicarem les raons per les quals no les hem utilitzat. Les estratègies per construir/seleccionar models des de la perspectiva Bayesiana es poden classificar en les següents tres grans families:

1. assignar probabilitat a priori a cada un dels models considerats i calcular la seva probabilitat a posteriori. En aquesta linea hom generalment acaba triant model en base als factors de Bayes (FB), que és el quocient entre el promig ponderat de la funció de versemblança sota cada model, fent servir com a factor de ponderació les distribucions a priori sobre els paràmetres de cada un dels models (Kass 1995),

2. comparar models a través d'estadístics que mesurin la qualitat de l'ajust penalitzant per la complexitat del model. L'estadístic més popular en el camp Bayesià és el *deviance information criterion* (DIC) (Spiegelhalter et al. 2002), i

3. comparar les dades o resums d'interés de les dades, amb simulacions de dades o de resums de les dades fent servir la predictiva a priori o la predictiva a posteriori (Gelman, Meng i Stern 1996). Aquesta estratègia serà la que utilitzarem.

El FB no hi ha dubte que és el mètode Bayesià de selecció de models pròpiament dit. En canvi la tercera estratègia, que compara simulacions de la predictiva a priori o a posteriori amb les dades, estrictament parlant és més un mètode de validació que de selecció de models, però que veurem que és molt útil a l'hora d'ajudar a escollir els models.

La selecció de models mitjançant el FB tracta la selecció del model com un cas especial de la teoria de decisió on la funció de pèrdua val 0 si el model triat és el correcte i 1 en altre cas. En aquest cas l'acció òptima és triar el model amb probabilitat a posteriori més alta, i en cas de que la probabilitat a priori dels models siguin iguals això porta a

triar el model amb FB més gran. Seleccionar models d'aquesta manera és elegant però sovint només és viable quan les distribucions a priori són pròpies i els models presenten un nombre reduït de paràmetres. En altres casos resulta difícil o impossible calcular el FB (Rossi et. al 2005, Congdon 2006).

D'altra banda el factor de Bayes és bastant sensible a les distribucions a priori escollides, i això fa que l'investigador hagi de tenir molta cura al seleccionar les distribucions a priori de cada model. Si aquestes són impròpies el FB no es pot calcular i si comparem un model amb una distribució a priori molt més difusa que un altre model, el factor de bayes tendirà a afavorir el model amb la distribució a priori menys difusa. Aquest fet limita si no invalida el FB per comparar models no jeràrquics amb models jeràrquics, ja que llavors el FB serà sensible tant al nivell d'incertesa de la distribució hiperpriori com al nombre de d'observacions, en les nostre dades al nombre de districtes, tal i com assenyalen Gelman et al (2003). Gelman també és un detractor del FB en el sentit que ell considera que és irrellevant des del moment en que es basa en les probabilitats relatives dels models condicionats a que un dels models és el bo.

En l'anàlisi Bayesiana, treballs recents han buscat una mesura que ponderi l'ajust i la complexitat d'un model per tal de poder comparar models amb estructures arbitràries. És en aquest context que Spiegelhalter et al. (2002) proposa el DIC, que utilitza com a mesura de bondat d'ajust la deviança, i la penalització per la complexitat es fa a través del nombre efectiu de paràmetres del model. Alguns autors motiven el DIC com a la versió bayesiana del AIC (Akaike 1973). Tot i que el DIC és àmpliament utilitzat com a criteri de selecció per la seva senzillesa de càlcul a partir de MCMC, presenta limitacions. En particular quan s'aplica en els models cluster com el nostre no resulta un bon criteri de selecció per la dificultat d'estimar el nombre efectiu de paràmetres que en general porta a aquest criteri a sobreestimar el nombre de clusters (Deloiro i Robert 2002, Richardson 2002 i Ceuleux et. al. 2006). I tampoc està clar com calcular el DIC en el cas dels models jeràrquics que farem servir en les nostres anàlisis.

La idea essencial al voltant de la tercera estrategia de construcció/selecció de models a partir de la validació dels mateixos, que serà la que adoptarem, és avaluar la compatibilitat de les dades amb el model assumit. Si el model és adequat, llavors rèpliques generades pel model haurien d'assemblar-se a les dades observades. El model que capturi millor les característiques principals de les dades per al nostre objectiu i que sigui més plausible com a model generador de les dades serà el model escollit. A la pràctica cal ser curós i enginyós a l'hora de triar els aspectes a contrastar, dissenyant gràfics i formulant estadístics i mesures de discrepància en consonància amb els objectius fonamentals de l'anàlisi.

8.2 Validació amb la predictiva a posteriori

Tal i com assenyalen Bayarri i Castellanos (2007) la majoria de mètodes de validació de models, tant estadístics com Bayesians, segueixen els tres passos següents:

1. triar un estadístic per quantificar el grau d'incompatibilitat del model amb les observacions,
2. triar una distribució de referència per a l'estadístic sota la hipòtesis de que el model és correcte, i
3. triar una mesura de conflicte entre l'estadístic observat i la distribució de referència.

Aquest plantejament que cobreix tant gran part de la metodologia freqüentista per validar models estadístics com metodologia per validar models Bayesians, permet validar diferents aspectes del model perquè variant l'estadístic un va canviant el punt de vista sota el que està jutjant el model (Petit 1986, Gelman et al. 1995). A continuació abordem cada un dels tres passos per separat.

8.2.1 Elecció de l'estadístic

Escollir l'estadístic de forma inteligent és molt important però no hi ha pautes clares i sovint s'escull de forma intuïtiva. A diferència del que passa sota l'estadística freqüentista, la perspectiva bayesiana permet fer servir estadístics $D(y, \theta)$ que depenen tant del paràmetre com de les dades. És per això que en el nostre context l'estadístic pot ser des de funcions d'observacions individuals $D(y_i)$, fins a funcions complexes de totes les dades i dels paràmetres $D(y, \theta)$, passant per funcions de totes les dades i prou $D(y)$.

Un cop definit l'estadístic aquest l'avaluarem amb les dades observades, simbolitzades per y_{obs}, i ho denotarem per $D(y_{obs}, \theta)$ o $D(y_{obs})$. Si l'estadístic $D(y)$ només depèn de les dades, llavors $D(y_{obs})$ serà un número o un vector de números, i si l'estadístic $D(y, \theta)$ depèn de les dades i dels paràmetres llavors $D(y_{obs}, \theta)$ és una funció de θ.

L'objectiu final serà comparar $D(y_{obs}, \theta)$ o $D(y_{obs})$ amb la respectiva distribució de referència $h(D(\widetilde{y}, \theta))$ o $h(D(\widetilde{y}))$ on $\widetilde{y}$ simbolitza una observació futura i $h(\cdot)$ representa la distribució de referència.

La idea és escollir $D(y)$ o $D(y, \theta)$ de forma que capturi el tipus de discrepància que més

preocupi entre les dades el model estadístic i la distribució a priori. A continuació llistem els estadístics que utilitzarem per a la validació/selecció dels nostres models:

1) Un primer conjunt d'estadístics que estudiarem, definits sota la inspiració de l'anàlisis de correspondències fet al capítol 6 són:

$$D_{ai}(y_i) = log(\tfrac{y_{CIU}}{y_{PSC}}),$$

$$D_{bi}(y_i) = log(\tfrac{y_{CIU+PPC}}{y_{PSC+ERC+ICV}}),$$

$$D_{ci}(y_i) = log(\tfrac{y_{i,CIU+ERC}}{y_{i,PSC+PPC+ICV}}),$$

i

$$D_{di}(y_i) = log(\tfrac{y_{i,abs}}{N_i}),$$

per $i = 1, \ldots, n$, que resumeixen cadascuna de le files de 8 (o 9) components de les taules 6.2 i 6.5 a través de 4 dels aspectes més rellevants pels analistes polítics. Aquest primer conjunt d'estadístics els calcularem per a cada una de les zrp o districtes i per tant és important remarcar que cada un d'aquests estadístics pren un valor a $\mathbb{R}^{248}$ o $\mathbb{R}^{1447}$. Degut a l'elevat nombre d'observacions caldrà pensar en representacions gràfiques per representar de forma útil i compacta aquests estadístics, i rèpliques simulades dels mateixos.

2) Un altre estadístic que utilitzarem esta basat en l'estadístic de la deviança $D_1(y, \theta)$, que mesura la bondat d'ajust i que per al cas general del model Multinomial no jeràrquic amb s clusters, que al capítol 7 hem denotat com a Model 5, és:

$$D_1(y, \theta_1, \ldots, \theta_s, \zeta) =$$

$$-2 \sum_{i=1}^{n} [\log(N_i!) + \sum_{j=1}^{k} (y_{ij} \log(\theta_{\zeta_i j}) - \log(y_{ij}!))] = \sum_{i=1}^{n} D_{1i},$$

on

$$D_{1i}(y_i, \theta_1, \ldots, \theta_s, \zeta) = -2[\log(N_i!) + \sum_{j=1}^{k} (y_{ij} \log(\theta_{\zeta_i j}) - \log(y_{ij}!))].$$

Per al cas jeràrquic que hem denotat com a Model 6, aquesta mesura presenta dues possibilitats segons a quin nivell es vulgui fer la inferència. Si volem fer inferència del segon nivell calcularem aquest estadístic com:

$$D_1(y, \theta_1, \ldots, \theta_n) =$$

$$-2\sum_{i=1}^{n}[\log(N_i!)+\sum_{j=1}^{k}(y_{ij}\log(\theta_{ij})-\log(y_{ij}!))]=\sum_{i=1}^{n}D_{1i},$$

on

$$D_{1i}(y_i,\theta_i)=-2[\log(N_i!)+\sum_{j=1}^{k}(y_{ij}\log(\theta_{ij})-\log(y_{ij}!))],$$

mentre que si volem fer inferència del tercer nivell aquesta mesura es converteix en:

$$D_1(y,\tau,\mu_1,\ldots,\mu_s,\zeta)=-2\sum_{i=1}^{n}[\log(N_i!)+log(\Gamma(\tau_{\zeta_i}))-log(\Gamma(N_i+\tau_{\zeta_i}))$$

$$+\sum_{j=1}^{k}[log(\Gamma(y_{ij}+\tau_{\zeta_i}\mu_{\zeta_i j})-log(y_{ij}!)-log(\Gamma(\tau_{\zeta_i}\mu_{\zeta_i j})))]]=\sum_{i=1}^{n}D_{1i},$$

on

$$D_{1i}(y_i,\tau,\mu_1,\ldots,\mu_s,\zeta)=-2[\log(N_i!)+log(\Gamma(\tau_{\zeta_i}))-log(\Gamma(N_i+\tau_{\zeta_i}))$$

$$+\sum_{j=1}^{k}[log(\Gamma(y_{ij}+\tau_{\zeta_i}\mu_{\zeta_i j})-log(y_{ij}!)-log(\Gamma(\tau_{\zeta_i}\mu_{\zeta_i j})))]].$$

El nostre objectiu principal no és fer inferència del perfil de vot de cada districte concret, $(\theta_1,\ldots,\theta_n)$, si no del perfil de vot de cada cluster, $(\mu_1,\ldots,\mu_s)$, per aquesta raó per al cas jeràrquic utilitzarem aquesta segona expressió per D_1.

3) Una altra mesura de discrepància que utilitzarem s'inspira en el test χ^2 de bondat d'ajust que per al model Multinomial no jeràrquic de s clusters, Model 5, consisteix en:

$$D_2(y,\theta_1,\ldots,\theta_s,\zeta)=\sum_{i=1}^{n}\sum_{j=1}^{k}\frac{(y_{ij}-N_i\theta_{\zeta_i j})^2}{N_i\theta_{\zeta_i j}}=\sum_{i=1}^{n}D_{2i}^2,$$

on

$$D_{2i}^2(y,\theta_1,\ldots,\theta_s,\zeta)=\sum_{j=1}^{k}\frac{(y_{ij}-N_i\theta_{\zeta_i j})^2}{N_i\theta_{\zeta_i j}}$$

es pot interpretar com una mesura de discrepància entre la fila i-èssima de la taula i el seu valor esperat segons el model estadístic.

Per al cas jeràrquic, igual que succeïa amb l'estadístic D_1, tindrem dues possibilitats segons a quin nivell es vulgui fer la inferència. Si volem fer inferència del segon nivell calcularem aquest estadístic com:

$$D_2(y,\theta_1,\ldots,\theta_n)=\sum_{i=1}^{n}\sum_{j=1}^{k}\frac{(y_{ij}-N_i\theta_{ij})^2}{N_i\theta_{ij}}=\sum_{i=1}^{n}D_{2i}^2,$$

on

$$D_{2i}^2(y,\theta_i)=\sum_{j=1}^{k}\frac{(y_{ij}-N_i\theta_{ij})^2}{N_i\theta_{ij}},$$

mentre que si volem fer inferència del tercer nivell aquesta mesura es converteix en:

$$D_2(y,\tau,\mu_1,\ldots,\mu_s,\zeta)=\sum_{i=1}^{n}\sum_{j=1}^{k}\frac{(y_{ij}-N_i\mu_{\zeta_i j})^2}{N_i\mu_{\zeta_i j}}=\sum_{i=1}^{n}D_{2i}^2,$$

on

$$D_{2i}^2(y,\tau,\mu_1,\ldots,\mu_s,\zeta)=\sum_{j=1}^{k}\frac{(y_{ij}-N_i\mu_{\zeta_i j})^2}{N_i\mu_{\zeta_i j}}.$$

Donat que el nostre objectiu serà fer inferència del tercer nivell, pel perfil de vot per cada un dels s clusters, $(\mu_1,\ldots,\mu_s)$, calcularem D_2 utilitzant la segona versió.

Aquestes mesures de discrepància s'estudiaran tant a nivell global, fent servir D_1 i D_2, com a nivell de cada observació, fent servir D_{1i} i D_{2i} per $i=1,\ldots,n$. Al mateix temps donat que les nostres observacions, y_i, són un vector de dades discretes i D_{2i}^2 és una mesura d'ajust per a cada observació resultant de sumar el grau de desavinença per cada component de y_i, també seria possible analitzar l'ajust component a component de y_i a través de:

$$D_{2ij}=\frac{y_{ij}-N_i\mu_{\zeta_i j}}{\sqrt{N_i\mu_{\zeta_i j}}}.$$

Aquest últim estadístic permet interpretar el signe de la discrepància i així identificar per a cada observació quines són les components específiques que el model no és capaç de predir correctament.

Al capítol 9 presentarem gràfics de les distribucions a posteriori dels estadístics D_1 i D_2 per a cada model jeràrquic i no jeràrquic considerats per els resultats a les eleccions al Parlament de Catalunya del 2003 a Barcelona ciutat. També presentarem per cada model gràfics de l'esperança a posteriori de D_{2i}^2 per a cada observació (zrp) en funció del seu nombre d'electors, així com mapes per reflexar la distribució espacial de l'esperança a posteriori de D_{2i}^2. I també una bateria de gràfics relacionats amb els estadístics $(D_{a1},\ldots,D_{an}),\ldots,(D_{d1},\ldots,D_{dn})$ que permetran entre altres comparar els valors observats d'aquests amb rèpliques obtingudes de les respectives distribucions predictives a posteriori per a cada model jeràrquic i no jeraràrquic considerats.

8.2.2 Elecció de la distribució de referència

Triar la distribució de referència $h(\cdot)$ per un estadístic $D(\widetilde{y}, \theta)$ o $D(\widetilde{y})$ amb la que compararem els valors observats, $D(y_{obs}, \theta)$ o $D(y_{obs})$, passa per triar la manera de considerar la incertesa sobre les dades futures i els paràmetres del model. Les diferents estratègies donaràn lloc a diferents propostes de validació. A continuació llistem les tres principals distribucions de referència utilitzades sota el paradigma Bayesià.

1. Quan l'estadístic només és funció de les dades observades, Box (1980) proposa fer servir com a $h(\cdot)$ la distribució predictiva a priori. Si l'estadístic depèn alhora de dades observables i de paràmetres aleshores es farà servir la distribució conjunta a priori de $(\widetilde{y}, \theta)$. Avaluar el model o aspectes del model mitjançant la predictiva a priori es limita al cas dels models amb a prioris pròpies i no acostuma a ser viable en els problemes amb molts paràmetres com és el nostre cas.

2. Quan l'estadístic només depèn de les dades, Rubin et al (1984) proposa fer servir com a $h(\cdot)$ la predictiva a posteriori. Si l'estadístic depèn de les dades i de paràmetres aleshores es farà servir la distribució conjunta a posteriori de $(\widetilde{y}, \theta)$.

3. Gelfand, Dey i Chang (1992) proposen comparar l'estadístic avaluat per a cada observació amb la distribució de referència $h(\cdot)$ calculada a partir de la predictiva a posteriori o la conjunta a posteriori de $(\widetilde{y}, \theta)$ condicionada a totes les observacions menys la que s'està comparant. Quan es disposa d'un nombre suficient d'observacions, com és el nostre cas, aquesta proposta inspirada en la idea de la validació creuada no difereix gaire de la basada en la predictiva a posteriori, i d'altra banda per al nostre conjunt de dades requereix d'un temps de computació inassumible.

Per raons pràctiques utilitzarem com a distribució de referència les predictives a posteriori dels estadístics tal i com es descriu a la segona alternativa. Val a dir que aquest tipus de plantejament ha estat criticat (Bayarri i Berger, 2000, Bayarri i Castellanos 2007) pel fet d'utilitzar les dades dues vegades, una per a calcular la distribució a posteriori i l'altre per calcular la distribució de referència, i això té un efecte conservador que pot impedir detectar incompatibilitats del model. No obstant, per la grandària de les mostres que considerarem és d'esperar que aquest efecte serà molt petit.

En el context bayesià, un cop elegida la mesura de discrepància i la distribució de referència basada en la predictiva a posteriori o la conjunta de $(\widetilde{y}, \theta)$ a posteriori es pot simular de la distribució de referència automàticament a partir de les simulacions de la predictiva a posteriori pel cas $D(\widetilde{y})$ o de la predictiva conjunta a posteriori pel cas $D(\widetilde{y}, \theta)$.

Així, en el nostre cas, si tenim L simulacions de la distribució a posteriori de θ, $\theta^{(1)}, \ldots, \theta^{(L)}$, simularem una taula de contingència $\widetilde{y} = (\widetilde{y}_1, \ldots, \widetilde{y}_n)$ per cada $\theta^{(l)}$ i per el mateix nombre total d'electors, $N = (N_1, \ldots, N_n)$ que per la taula y_{obs} observada, de manera que tindrem L simulacions $\widetilde{y}^{(l)}$, per $l = 1, \ldots, L$, generades com a rèpliques de y_{obs}; les parelles $(\widetilde{y}^{(l)}, \theta^{(l)})$ seran simulacions de la distribució a posteriori $p(\widetilde{y}, \theta|y)$. Finalment $D(\widetilde{y}^{(l)})$ i $D(\widetilde{y}^{(l)}, \theta^{(l)})$ per $l = 1, \ldots, L$ seran L simulacions de la distribució de referència que farem servir per $D(y_{obs})$ o per $D(y_{obs}, \theta)$.

8.2.3 Elecció de la manera de mesurar el conflicte

En el cas de poder avaluar per a cada observació, y_i per $i = 1, \ldots, n$, el valor que pren la densitat de la distribució de referència de $D(\widetilde{y}_i)$ a $D(y_{i,obs})$, $h(D(y_{i,obs}))$, els valors per als quals $h(D(y_{i,obs}))$ és petit són poc coherents amb el model, i es poden considerar valors anormals sota l'assumpció del model. Massa valors petits entre els valors de $h(D(y_{1,obs})), \ldots, h(D(y_{n,obs}))$ indiquen que el model no és adequat i hauria de ser modificat. Un problema d'aquesta manera de procedir és la dificultat de decidir què vol dir petit i què vol dir massa.

De la mateixa manera si hem definit un estadístic global $D(y_{obs}) = D(y_{1,obs}, \ldots, y_{n,obs})$, un valor massa petit de $h(D(y_{obs}))$ indica que el model no és adequat perquè fa massa poc probable el que hem observat i haurà de ser modificat. Altra vegada caldrà decidir que vol dir massa petit. La dificultat augmenta en el cas en que l'estadístic de prova depengui tant de y com de θ, $D(y, \theta)$. En tots els casos haurem de comparar els valors observats, $D(y_{obs})$ o $D(y_{obs}, \theta)$, amb les respectives distribucions de referència $h(D(\widetilde{y}))$ o $h(D(\widetilde{y}, \theta))$. A continuació presentem tres alternatives a l'hora de jutjar tot això.

a) Box (1980), suggereix fer-ho a partir de

$$p(h(\widetilde{y}) \geq h(y)),$$

on $h(\cdot)$ és la densitat de la distribució predictiva a priori de $\widetilde{y}$ i on $p(\cdot)$ es calcula respecte de la mateixa distribució, de manera que com més petit sigui aquest valor millor serà el model. En el nostre cas podríem fer servir la mateixa mesura però amb $h(\cdot)$ sent la densitat de la predicitiva a posteriori i $p(\cdot)$ la probabilitat respecte a la mateixa distribució. En compte d'això s'acostuma a fer servir l'àrea de la cua:

$$p_B = p(D(\widetilde{y}) \geq D(y_{obs})|y_{obs}),$$

o be si $D(y, \theta)$ depèn de y i de θ, es fa servir:

$$p_B = p(D(\widetilde{y}, \theta) \geq D(y_{obs}, \theta)|y_{obs}),$$

on la probabilitat $p(\cdot)$ es calcula respecte a la distribució conjunta a posteriori per (y, θ). Com més a prop siguin aquests valors p_B a 0 o a 1 pitjor serà el model, en el sentit que més inversemblant serà que les dades hagin estat generades pel mateix model que ha generat les rèpliques. El valor d'aquesta àrea de cua és coneix com a *p valor* Bayesià. Aquesta mesura de conflicte serà útil per posar en dubte un model però no per rebutjar-lo excepte quan trobem un model alternatiu millor. En altres paraules valors extrems poden suggerir millores en el model tot i que si es considera que les mancances no afecten a la inferència sobre els principals objectius es pot donar el model com a prou bo.

b) Roberts (1965) proposa fer servir l'índex de sorpresa relativa (RPS de l'anglès *relative predictive surprise*),

$$RPS = \frac{h(D(y_{obs}))}{Max\{h(D(\widetilde{y}))\}},$$

que sempre estarà entre 0 i 1 i és de forma que com més gran és més creïble serà que y_{obs} provingui del model considerat. En el cas en el que l'estadístic $D(y, \theta)$ depengués de y i de θ, treballaríem amb

$$RPS = \frac{h(D(y_{obs,\theta}))}{Max\{h(D(\widetilde{y}, \theta))\}}.$$

c) Seguint les recomanacions de Gelman (2003), nosaltres utilitzarem sobretot eines gràfiques per manifestar conflicte, presentant en un gràfic un estadístic $D(y)$ avaluat per les observacions i gràfics successius amb rèpliques d'aquest estadístic generades a partir de la seva distribució de referència que en el nostre cas serà la predictiva a posteriori. Diferències sistemàtiques entre les simulacions i les dades indicaran mancances del model.

Les mesures RPS i p_B es poden calcular per cada observació y_i, si em definit un estadístic $D(y_i)$ o $D(y_i, \theta)$, o bé globalment per totes les observacions de cop, $y = (y_1, \ldots, y_n)$, si em definit un estadístic $D(y_{obs})$ o $D(y_{obs}, \theta)$. Amb la validació individual observació a observació s'avalua la compatibilitat de cada observació amb el model Bayesià. El nivell de discrepància entre cadascuna de les observacions i la distribució predictiva corresponent permetrà veure si aquestes discrepàncies responen a algun patró que suggereixi un model millor.

RPS i p_B són dues bones alternatives de mesura de conflicte. Nosaltres utilitzarem el p_B per tractar-se d'una probabilitat, i per tant està afitat entre 0 i 1, tot i que cal tenir molt present que $p_B^{(D)}$ no és la probabilitat que el model sigui correcte. En general triat

un estadístic $D(y,\theta)$ i com a distribució de referència la seva predictiva a posteriori, el *p valor* Bayesià associat a aquest estadístic, $p_B^{(D)}$ serà

$$p_B^{(D)} = p(D(\widetilde{y},\theta) \geq D(y_{obs},\theta)|y_{obs}) =$$

$$\int\int I_{D(\widetilde{y},\theta)\geq D(y_{obs},\theta)} p(\widetilde{y}|\theta)\pi(\theta|y_{obs})d\widetilde{y}d\theta.$$

Així en el cas de $D(y)$ si tenim L simulacions de $\pi(\theta|y_{obs})$, $\theta^1,\ldots,\theta^L$ i per tant L rèpliques simulades de $p_\pi(\widetilde{y}|y_{obs})$, $\widetilde{y}^1,\ldots,\widetilde{y}^L$, es podrà estimar el *p valor* com,

$$\widehat{p_B}^{(D)} = \frac{\sum_{l=1}^{L} I_{\{D(\widetilde{y}^l)\geq D(y_{obs})\}}}{L},$$

i en el cas de $D(y,\theta)$ el podrem estimar com,

$$\widehat{p_B}^{(D)} = \frac{\sum_{l=1}^{L} I_{\{D(\widetilde{y}^l,\theta^l)\geq D(y_{obs},\theta^l)\}}}{L}.$$

Una característica important del *p valor* és que quan es considera com una variable aleatòria i s'utilitzen com a distribució de referència les predictives a priori llavors té una distribució uniforme quan el model del que simules és correcte (Bayarri i Berger 2000; Cook i Gelman 2006). Aquesta propietat dels *p valors* no es compleix quan s'utilitzen com a distribució de referència les predictives a posteriori (Bayarri i Berger 2000; Robins et al. 2000; Hjort et al. 2006, SteinBakk, 2009), i empíricament s'observa que a menor grandària de mostra major desviació de la uniformitat.

Tot i així el *p valor* és vàlid com a eina de validació en tant que *p valors* extrems, propers a 0 o a 1, indiquen mancances del model, però obtenir *p valors* entre 0.05 i 0.95 no serà suficient per estar segur de que el model sigui bo.

8.3 Validació i dependència espacial

A sèries temporals un model no es dóna per bo fins que els seus residus no siguin independents i idènticament distribuïts. Aquesta idea és extensible al cas de les dades espacials, on es poden validar els models en base a tests que verifiquin si els residus del model es poden considerar com a independents i idènticament distribuïts o encara hi queda dependència espacial. Una manera d'avaluar l'existència de correlació podria ser calculant algun tipus de coeficient de correlació espacial sobre alguna mesura de discrepància del model, i d'aquesta manera avaluar si les desviacions del model estan estructurades espacialment o no.

Escollim com a mesura de discrepància l'esperança a posteriori de D_{2i}^2, $E[D_{2i}^2|y]$. Un cop definida la mesura caldrà triar l'estadístic per verificar la dependència, la distribució de referència d'aquest estadístic en cas d'independència i la mesura de conflicte entre el valor observat de l'estadístic i la seva distribució de referència.

Per analitzar la correlació espacial de qualsevol mesura prèviament s'ha de definir el terme veïns, i el grau de dependència o proximitat entre aquests mitjançant pesos. La nostra manera de definir els veïns és mitjançant la seva posició relativa en el mapa, de manera que dues àrees contigües, amb una aresta o vèrtex en comú, es consideraran veïnes mentre que la resta no ho seran. Utilitzarem pesos binaris, que assignen un pes d'una unitat als veïns i de zero altrament.

Un cop definida l'estructura de veïnatge cal especificar un estadístic que mesuri la correlació. El que utilitzarem és l'anomenat índex de Moran. Una possible alternativa hauria estat fer servir el coeficient de Geary. Per aprofundir en l'anàlisi estadística espacial i en l'ús d'aquests estadístics es pot consultar Ripley (1981), Cressie (1993), Bivand (2008) i Gelfand (2004). L'índex de Moran i el coeficient de Geary representen a nivell espacial el mateix que en sèries temporals representen el coeficient d'autocorrelació i l'estadístic de Durbin Watson respectivament.

L'Índex de Moran es defineix com:

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_{ij}(z_i-\overline{z})(z_j-\overline{z})}{\sum_{i=1}^{n}(z_i-\overline{z})^2},$$

on n és el nombre d'observacions, z_i és l'observació i-èssima que en el nostre cas serà $E[D_{2i}^2|y]$, $\overline{z}$ és la mitjana de la variable d'interès i λ_{ij} és el pes espacial de les zrp, i i j.

Definit l'estadístic cal especificar una distribució de referència. Nosaltres la construirem

com a un test de permutació clàssic. Si les dades són independents espacialment, llavors intercanviant les etiquetes de les àrees s'haurien d'obtenir valors de l'índex de Moran comparables amb l'observat. La idea consisteix en anar permutant aleatòriament les etiquetes de les àrees i avaluar l'índex de Moran per a cada permutació.

Un cop construïda la distribució de referència a partir de P permutacions, utilitzarem com a mesura de conflicte l'àrea de la cua calculada a partir de la posició de la I observada a dins de la mostra de P permutacions incloent la observada. És a dir farem servir un *p valor* que determinarà quan extrem és el valor observat respecte els valors que s'obtindrien permutant les etiquetes de les àrees sense permutar els seus valors.

Aprofundint en l'anàlisi de la correlació es pot crear un correlograma espacial, a base de calcular l'Índex de Moran, I, reemplaçant els pesos λ_{ij} per tal de definir els veins de primer ordre, els de segon ordre i així successivament. Dues àrees són veïnes de primer ordre si són contigües; dues àrees son veïnes de segon ordre si no ho són de primer ordre i tenen un veí comú, i així successivament. Així si prenem $\lambda_{ij}^{(1)}$=1 si i i j són veins de primer ordre i $\lambda_{ij}^{(1)}$=0 en altre cas obtenim $I^{(1)}$, que coincideix amb I, $I^{(1)} = I$. De forma similar definim la matriu de veïnatge de segon ordre definint els pesos com $\lambda_{ij}^{(2)}$=1 si i i j són veins de segon ordre i $\lambda_{ij}^{(2)}$=0 en altre cas i obtenim $I^{(2)}$, i així successivament. El gràfic de $I^{(t)}$ versus t s'anomena correlograma. Si existeix un patró espacial és d'esperar que inicialment $I^{(t)}$ decreixi i que després potser fluctuï al voltant del 0. En canvi si no hi ha dependència espacial es d'esperar que $I^{(t)}$ prengui valors petits al voltant del zero per tot t.

Al procediment plantejat per analitzar la correlació espacial no explicada pel model se'n podrien fer dues objeccions. La primera és que en lloc d'utilitzar $E[D_{2i}^2|y]$ es podria utilitzar directament D_{2i}^2 entenent-la com una variable aleatòria i d'aquesta manera incorporar la incertesa que tenim dels paràmetres en el càlcul de la distribució de referència. Una segona objecció és l'haver construït la distribució de referència en base a un procediment no paramètric i no tal i com hem suggerit a la secció 8.2.2. Bayesianitzar aquest procediment en aquestes dues direccions serà un repte per al futur. De totes maneres el procediment utilitzat és prou robust per suposar com a vàlides les conclusions que se'n derivin, i és d'esperar que les conclusions a les que s'arribaria fent servir aquestes altre dues vies seran molt semblants.

Capítol 9

Comparació i validació dels models per al 2003 a Barcelona

En aquest capítol presentarem, per a les eleccions al Parlament de Catalunya de l'any 2003 a Barcelona ciutat, el procés de construcció del model que ens permetrà identificar l'existència de patrons de vot i veure per a cada una de les 248 zones de recerca petites (zrp) de la ciutat de Barcelona a quin grup pertanyen. En aquest procés de construcció de models seran molt importants les representacions gràfiques. Al Capítol 10 interpretarem aquests models.

Les dades utilitzades en aquest capítol es presenten parcialment a la Taula 6.2; les columnes corresponen a les opcions de vot que constitueixen les categories de la Multinomial, i que són:

1. CIU: Convergència i Unió,
2. PSC: Partit Socialista de Catalunya,
3. PPC: Partit Popular de Catalunya,
4. ICV: Iniciativa per Catalunya els verds (i EUIA els anys 1999, 2003 i 2006),
5. ERC: Esquerra Republicana de Catalunya,
6. *altres*: agregació dels partits amb menys d'un 1% dels vots,
7. *b+n*: blancs i nuls, i
8. *abs*: abstenció,

i cada fila de la Taula 6.2 correspon a una zrp. La divisió administrativa dels districtes de Barcelona coincideix amb agregacions de les zrp. Tot i que la variable *districte* no la contemplem en el model sí que ens serà de gran utilitat en la validació, ja que

Codi Districte	Nom Districte
1	Ciutat Vella
2	Eixample
3	Sants-Montjuïc
4	Les Corts
5	Sarrià-Sant Gervasi
6	Gràcia
7	Horta-Guinardó
8	Nou Barris
9	Sant Andreu
10	Sant Martí

Taula 9.1: Districtes de Barcelona ciutat.

presentarem estadístics estratificant per districte per avaluar per a cada model la seva capacitat d'inferir a nivell de districte. A a la Taula 9.1 es mostra la relació de districtes. El detall de les dades s'ha presentat al Capítol 6.

A la primera secció d'aquest capítol compararem els diferents models cluster no jeràrquics, etiquetats com a Model 5 al Capíol 7, utilitzats sobre els resultats de les eleccions al parlament de Catalunya a la ciutat de Barcelona per a les 248 zrp a l'any 2003. A la segona secció hi compararem els models cluster jeràrquics, etiquetats com a Model 6, per als mateixos resultats.

9.1 Comparació dels models no jeràrquics

En aquesta secció compararem models no jeràrquics amb diferent número de clusters, és a dir models resultants d'anar variant s del Model 5.

En aquest punt cal especificar completament els models especificant totes les distribucions a priori i hiperpriori, donant valors als paràmetres que se suposen coneguts i que a través dels quals es poden introduir les creences a priori sobre la incertesa dels paràmetres desconeguts. Donada la complexitat del model i lo compromès que podria resultar introduir informació subjectiva, hem optat per utilitzar distribucions poc informatives. El model Bayesià a priori que hem escollit és:

$$
\begin{aligned}
y_i|\theta_1,\ldots,\theta_s &\sim \text{Multinomial}(N_i,\theta_{\zeta_i})\\
\theta_1 &\sim \text{Dirichlet}(1,\ldots,1)\\
&\vdots\\
\theta_s &\sim \text{Dirichlet}(1,\ldots,1)\\
\pi(\zeta_i = r|\omega) &= \begin{cases} \omega_1 & r=1\\ \omega_2 & r=2\\ \vdots & \vdots\\ \omega_s & r=s \end{cases}\\
\omega = (\omega_1,\ldots,\omega_s) &\sim \text{Dirichlet}(1,\ldots,1),
\end{aligned}
$$

Model 5. Model Bayesià a priori no jeràrquic per a s clusters Multinomials.

i per tant suposem que el perfil de probabilitat per a cada cluster, θ_r, té una distribució uniforme sobre el símplex de $\mathbb{R}^8$ i que el vector de probabilitats a priori ω de que l'observació i-èssima pertanyi a cada un dels clusters també té una distribució uniforme.

Actualitzarem aquest model per $s = 1,\ldots,5$, donant lloc als cinc models utilitzats, i que denotem per:

M_1: model multinomial homogeni d'un sol cluster,

M_2: model multinomial de dos clusters,

M_3: model multinomial de tres clusters,

M_4: model multinomial de quatre clusters, i

M_5: model multinomial de cinc clusters.

Per simular de les distribucions a posteriori dels paràmetres dels diferents models hem utilitzat el programari WinBugs. Per a cadascun dels models hem fet córrer dues cadenes MCMC en paral·lel amb diferents valors inicials. Hem avaluat la convergència de les cadenes utilitzant mesures d'autocorrelació mostral dintre de cada cadena, gràfics de les traces mostrals i la mesura diagnòstic proposada per Gelman i Rubin (1992).

S'han descartat les primeres 1000 iteracions de cada cadena com a escalfament previ a la convergència. De les iteracions posteriors a l'escalfament se n'han retingut una de cada 10, i la mostra final per a les anàlisis de la distribució a posteriori ha estat de

3000 simulacions, 1500 simulacions de cada cadena. Quan s'han observat problemes d'identificabilitat en els que en una mateixa cadena es permutaven les etiquetes dels clusters s'han descartat totes les simulacions i repetit el procés.

Compararem els models analitzant diferents aspectes en funció de l'estadístic escollit, i utilitzant com a distribució de referència la predictiva a posteriori de cada estadístic. Començarem avaluant gràficament l'habilitat dels models considerats a l'hora de generar dades com les observades a partir d'una bateria de quatre estadístics definits al capítol anterior.

La primera característica analitzada, presentada a la Figura 9.1, és el vector de logaritmes del quocient entre els dos partits més votats,

$$D_{ai}(y_i) = log(\frac{y_{i,CIU}}{y_{i,PSC}}),$$

per $i = 1, \ldots, 248$, que a jutjar per les anàlisis de correspondències presentades al Capítol 6 és la principal característica a l'hora de resumir l'estructura de les dades. Amb l'objectiu de visualitzar possibles patrons inherents a les dades a la Figura 9.1 s'han ordenat els districtes en funció del valor de les mitjanes d'aquest estadístic a totes les *zrp* de cada districte; les *zrp* dins de cada districtes no s'han reordenat de cap manera i apareixen per ordre del seu codi. A la Figura es presenten 4 rèpliques sota cadascun dels 5 models no jeràrquics, M_1, M_2, M_3, M_4 i M_5, per observar la capacitat d'aquests models de generar dades com les observades. Observis que les dades presenten molta més variabilitat de la que és capaç de capturar cada model.

Les Figures 9.2, 9.3 i 9.4 analitzen la qualitat del model a través de

$$D_{bi}(y_i) = log(\frac{y_{CIU+PPC}}{y_{PSC+ERC+ICV}}),$$

$$D_{ci}(y_i) = log(\frac{y_{i,CIU+ERC}}{y_{i,PSC+PPC+ICV}})$$

i

$$D_{di}(y_i) = log(\frac{y_{i,abs}}{N_i}),$$

per $i = 1, \ldots, 248$, amb l'objectiu de comprovar si aquests models són capaços de capturar els trets més rellevants per als politòlegs. Per cadascun dels tres estadístics s'observa

que les dades observades no són compatibles amb la distribució predictiva a posteriori corresponent a cap dels 5 models considerats.

La Figura 9.5 resumeix les 4 figures anteriors. Les dades es presenten també en funció del districte però en aquest cas ordenats pel codi de districte i la distribució de les *zrp* dintre d'un districte és presenta mitjançant un diagrama de caixa. Per a cada model i estadístic s'ha construït el gràfic a partir d'una rèplica. S'observa que els models M_1, M_2 i M_3 estàn lluny de capturar el comportament de les dades mentre que els models M_4 i M_5 si be capturen be els nivells, no capturen be les variabilitats.

La part superior de les Figures 9.6, 9.7 i 9.8 representen simultàniament tres columnes, o tres agregacions de columnes en el cas de la Figura 9.8, d'entre les vuit columnes de les dades mitjançant diagrames ternaris i a la part inferior de les Figures es presenten tres rèpliques simulades utilitzant la predictiva a posteriori de cada model. Aquestes Figures tornen a posar de manifest la incapacitat dels models no jeràrquics estudiats per a generar dades com les observades. Observis que també en aquest subespai ternari les dades presenten més variabilitat de la que és capaç de capturar el model.

A la Taula 9.2 es mostra l'esperança a posteriori de les mesures de bondat d'ajust

$$D_1(y, \theta_1, \ldots, \theta_s, \zeta) = -2 \sum_{i=1}^{n} (\log(N_i!) + \sum_{j=1}^{k} (y_{ij} \log(\theta_{\zeta_i j}) - \log(y_{ij}!))),$$

i

$$D_2(y, \theta_1, \ldots, \theta_s, \zeta) = \sum_{i=1}^{n} \sum_{j=1}^{k} \frac{(y_{ij} - N_i \theta_{\zeta_i j})^2}{N_i \theta_{\zeta_i j}},$$

per a cadascun dels models considerats i a la Figura 9.9 s'hi representen les respectives distribucions a posteriori, $\pi(D_1|y)$ i $\pi(D_2|y)$. Al passar d'un a dos clusters reduïm dràsticament el valor esperat a posteriori de les mesures de discrepància. Al augmentar el nombre de clusters a més de dos reduïm encara més aquest valor esperat, tot i que menys dràsticament.

També hem calculat el *p valor* Bayesià, p_B, associat a les mesures de discrepància D_1 i D_2, $p_B^{D_1}$ i $p_B^{D_2}$, i per a tots els anys el seu valor és 0 posant de manifest que els models no ajusten bé. Les dades presenten una variabilitat molt superior a la que és capaç de capturar el model ajustat, tal i com s'intueix en les Figures 9.1-9.4. En el model Multinomial la variància depèn fortament de l'esperança. Sovint les dades presenten major variabilitat de la que permet el model, aquest fenomen conegut com a sobredispersió és el que estem observant en les nostres dades.

Tot sembla indicar que per trobar un model no jeràrquic capaç de reproduir fidelment

Model	$E[D_1\|y]$	$p_B^{D_1}$	$E[D_2\|y]$	$p_B^{D_2}$
M_1	108089.7	0	96169.9	0
M_2	62829.7	0	51604.7	0
M_3	43000.9	0	30627.6	0
M_4	35553.3	0	23217.4	0
M_5	32164.0	0	19891.2	0

Taula 9.2: Esparança a posteriori de les mesures de discrepància D_1 i D_2, i els *p valors* bayesians associats, $p_B^{D_1}$ i $p_B^{D_2}$ per els resultats a les eleccions al Parlament de Catalunya del 2003 a la ciutat de Barcelona.

les nostres dades caldria un model amb molts més de cinc clusters. El fet que la deviança es redueixi tant per cada cluster afegit, essent aquesta reducció molt superior al nombre de paràmetres afegits, fa que qualsevol mesura inspirada en el DIC ens portaria a triar un model no jeràrquic amb un nombre de clusters molt més gran de 5. Aquest model amb tants clusters el faria inservible com a síntesi de les dades i no permetria extreure'n conclusions útils.

Un cop posat de manifest la manca de bondat d'ajust global pot ser interessant examinar la contribució a l'estadístic $D_2(y, \theta)$ de cada *zrp*. El valor esperat de la mesura de discrepància D_{2i}^2, on

$$D_{2i}^2(y, \theta_1, \ldots, \theta_s, \zeta) = \sum_{j=1}^{k} \frac{(y_{ij} - N_i\theta_{\zeta_i j})^2}{N_i\theta_{\zeta_i j}},$$

per a cada observació, $E[D_{2i}^2|y]$, es representa en funció del nombre d'electors de cada *zrp* a la Figura 9.10. Aquesta Figura mostra una dèbil associació de $E[D_{2i}^2|y]$ amb el nombre d'electors. Al mateix temps s'observa que la magnitud de discrepància és molt gran per als models d'un i dos clusters, però es redueix notablement al passar de 2 a 3 clusters i es redueix de forma més moderada al passar de 3 a 4 i de 4 a 5 clusters.

Un cop observat que D_{2i}^2 no presenta cap patró gaire fort associat al nombre d'electors de les zrp, s'haurà d'avaluar l'existència d'una possible correlació espacial de D_{2i}^2 indicativa de que la manca d'ajust respon a característiques de la població estructurades a l'espai i no capturades pel model de clusters. Si aquest fos el cas s'hauria de sospesar la possible utilitat d'un model amb correlació espacial.

Per avaluar l'existència de correlació espacial la Figura 9.11 representa el valor esperat de D_{2i}^2 en mapes. Visualment es percep una clara agrupació espacial en el cas dels models M_1 i M_2, que sembla atenuar-se molt en els altres tres models. Per eliminar subjectivitat es podria calcular un índex com el de Moran (Moran 1950) i veure si existeix correlació espacial en D_{2i}^2 i com varia al canviar d'un model a un altre.

El fet que les dades presentin més variabilitat que l'esperada de tots els models considerats aquí indica que els models utilitzats no són bons. Si estéssim treballant amb un model de regressió clàssic, això ens portaria a pensar que hi falten covariables. En el nostre cas, on l'objectiu d'entrada no és identificar covariables responsables de l'estructura de les dades si no identificar possibles patrons de vot i descriure'n els perfils, hem de treballar per arribar a un model que contempli aquesta sobredispersió sense introduir variables explicatives, per tal que la inferència que fem sigui correcta.

El model s'ha de reformular per capturar aquesta sobredispersió, ja sigui treballant amb models amb més variabilitat intrínseca, via models jeràrquics i/o be modelant la dependència espacial. Nosaltres optarem per la primera opció.

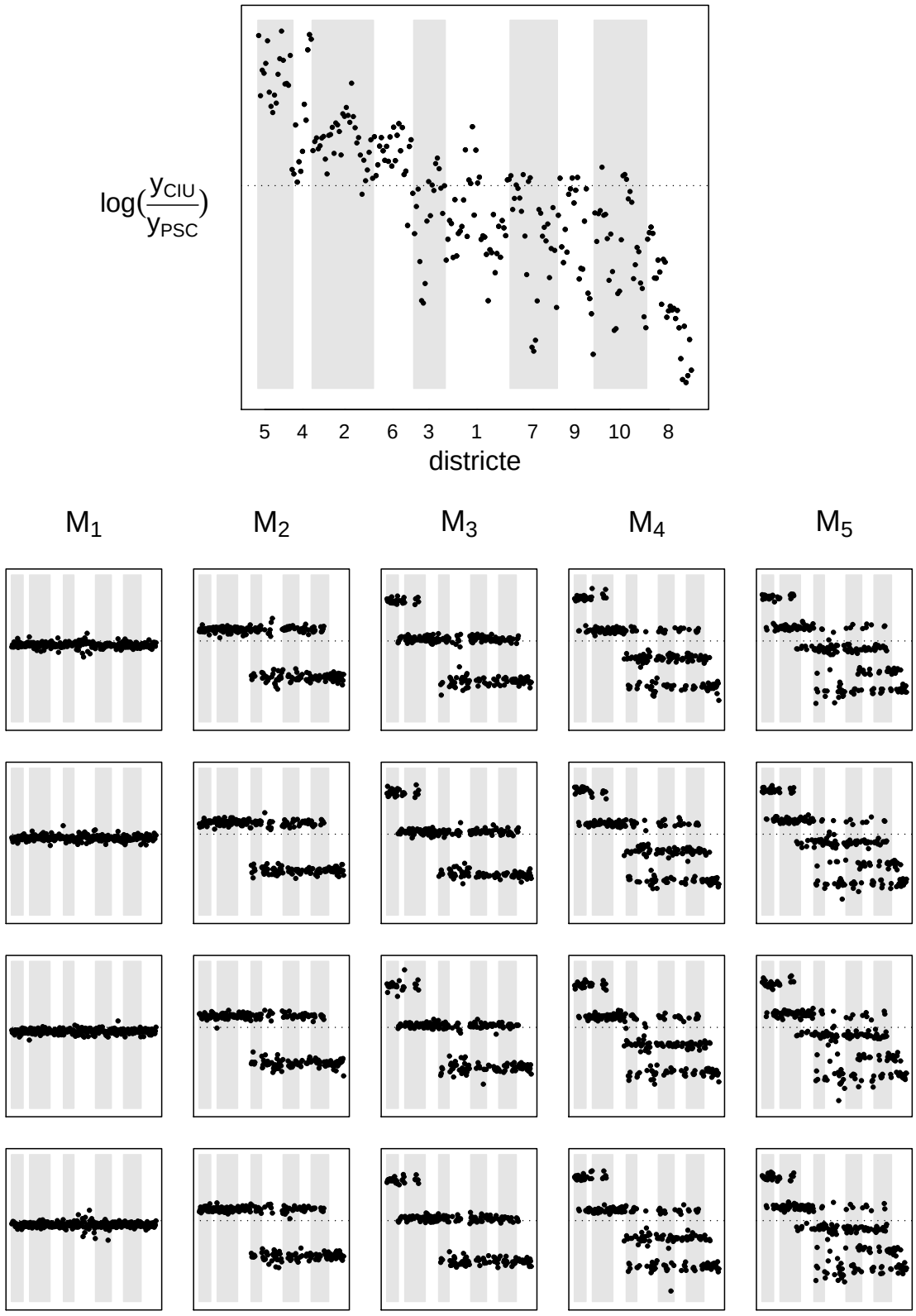

Figura 9.1: El primer gràfic representa els valors observats per $D_{ai}(y_i)$ a les zrp de cada districte a les eleccions al Parlament de Catalunya del 2003 a Barcelona ciutat i les quatre files següents corresponen a quatre rèpliques d'aquest estadístic simulades a partir de la predictiva a posteriori de cada model no jeràrquic considerat.

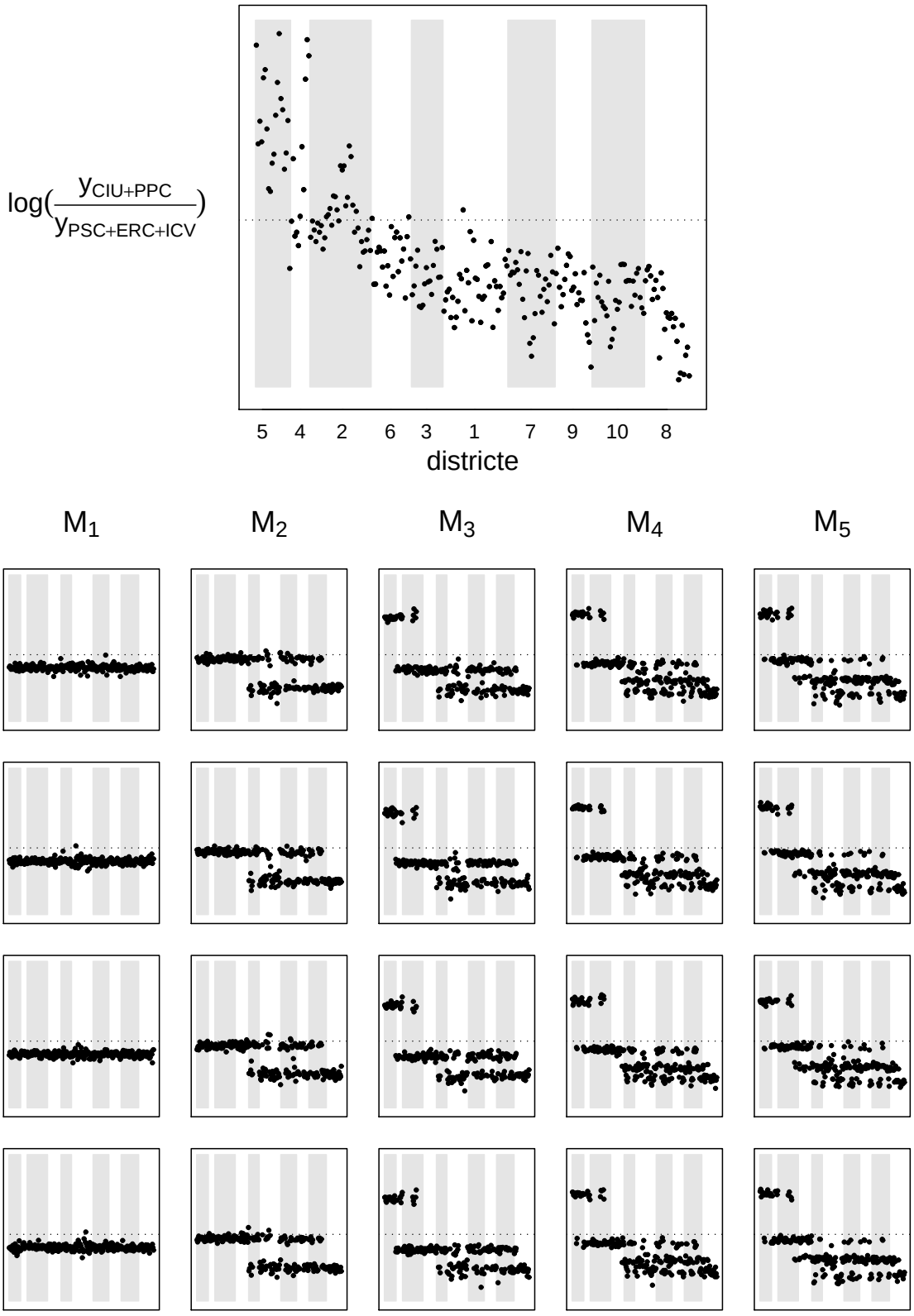

Figura 9.2: El primer gràfic representa els valors observats per $D_{bi}(y_i)$ a les zrp de cada districte a les eleccions al Parlament de Catalunya del 2003 a Barcelona ciutat i les quatre files següents corresponen a quatre rèpliques d'aquest estadístic simulades a partir de la predictiva a posteriori de cada model no jeràrquic considerat.

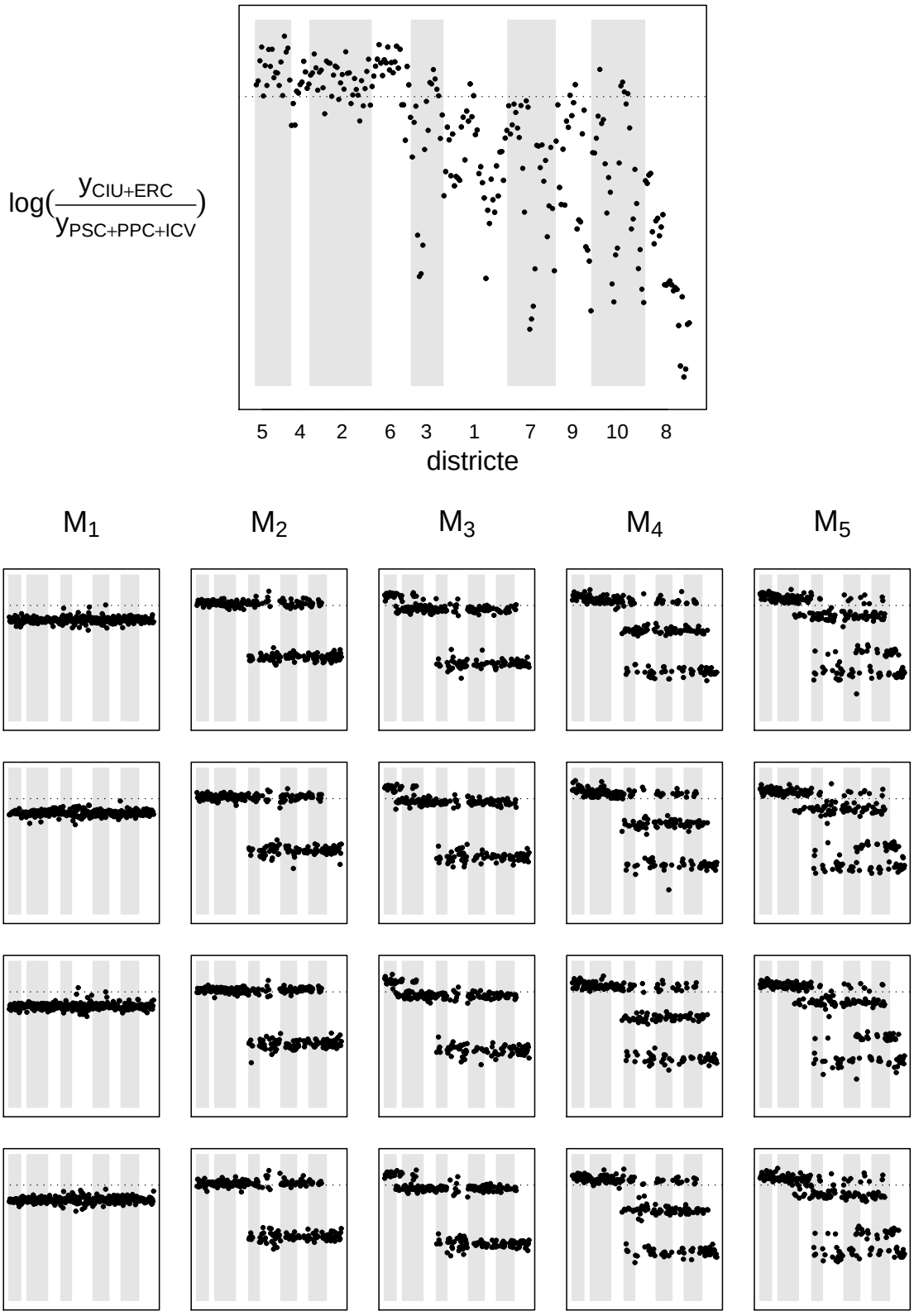

Figura 9.3: El primer gràfic representa els valors observats per $D_{ci}(y_i)$ a les zrp de cada districte a les eleccions al Parlament de Catalunya del 2003 a Barcelona ciutat i les quatre files següents corresponen a quatre rèpliques d'aquest estadístic simulades a partir de la predictiva a posteriori de cada model no jeràrquic considerat.

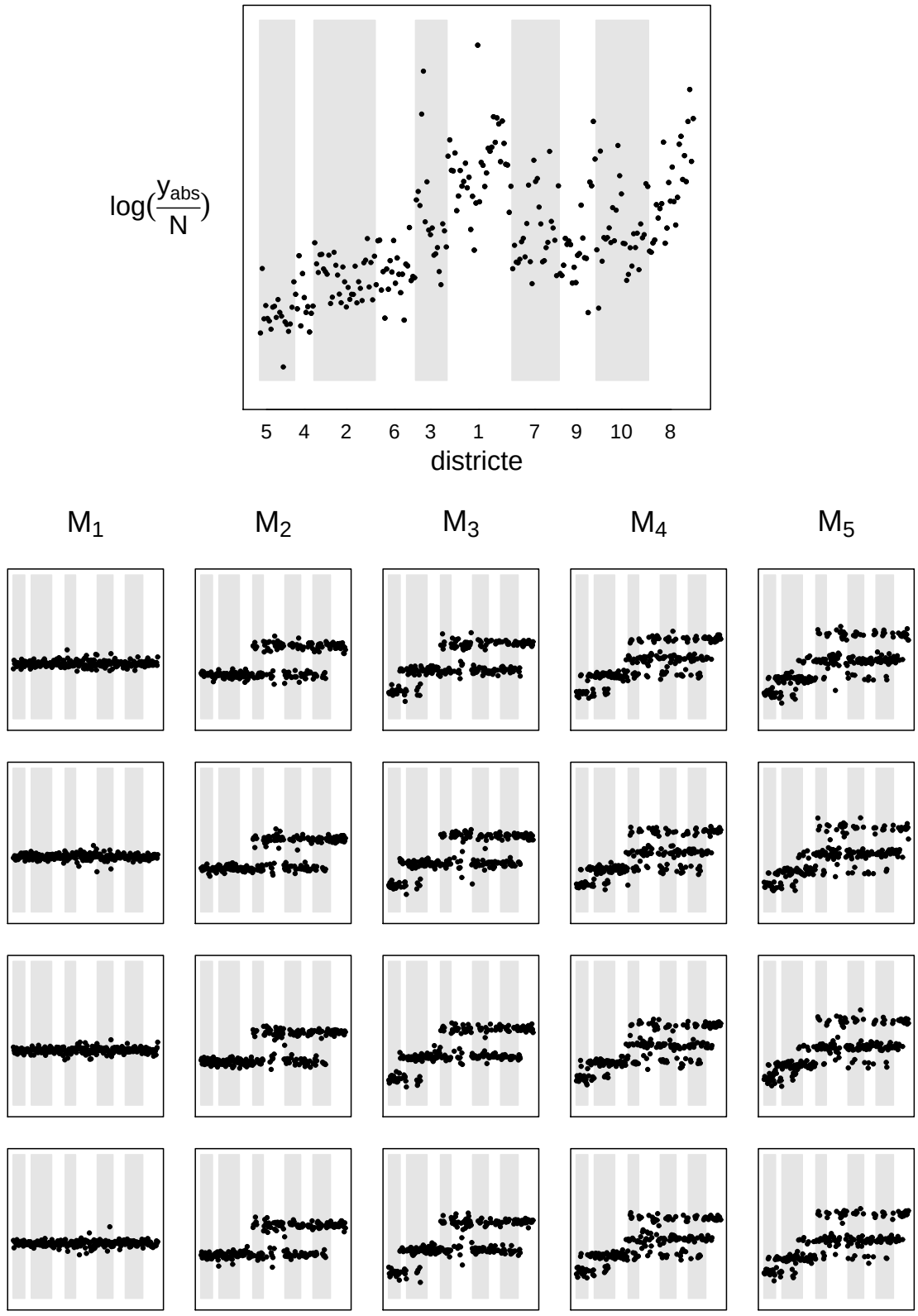

Figura 9.4: El primer gràfic representa els valors observats per $D_{di}(y_i)$ a les zrp de cada districte a les eleccions al Parlament de Catalunya del 2003 a Barcelona ciutat i les quatre files següents corresponen a quatre rèpliques d'aquest estadístic simulades a partir de la predictiva a posteriori de cada model no jeràrquic considerat.

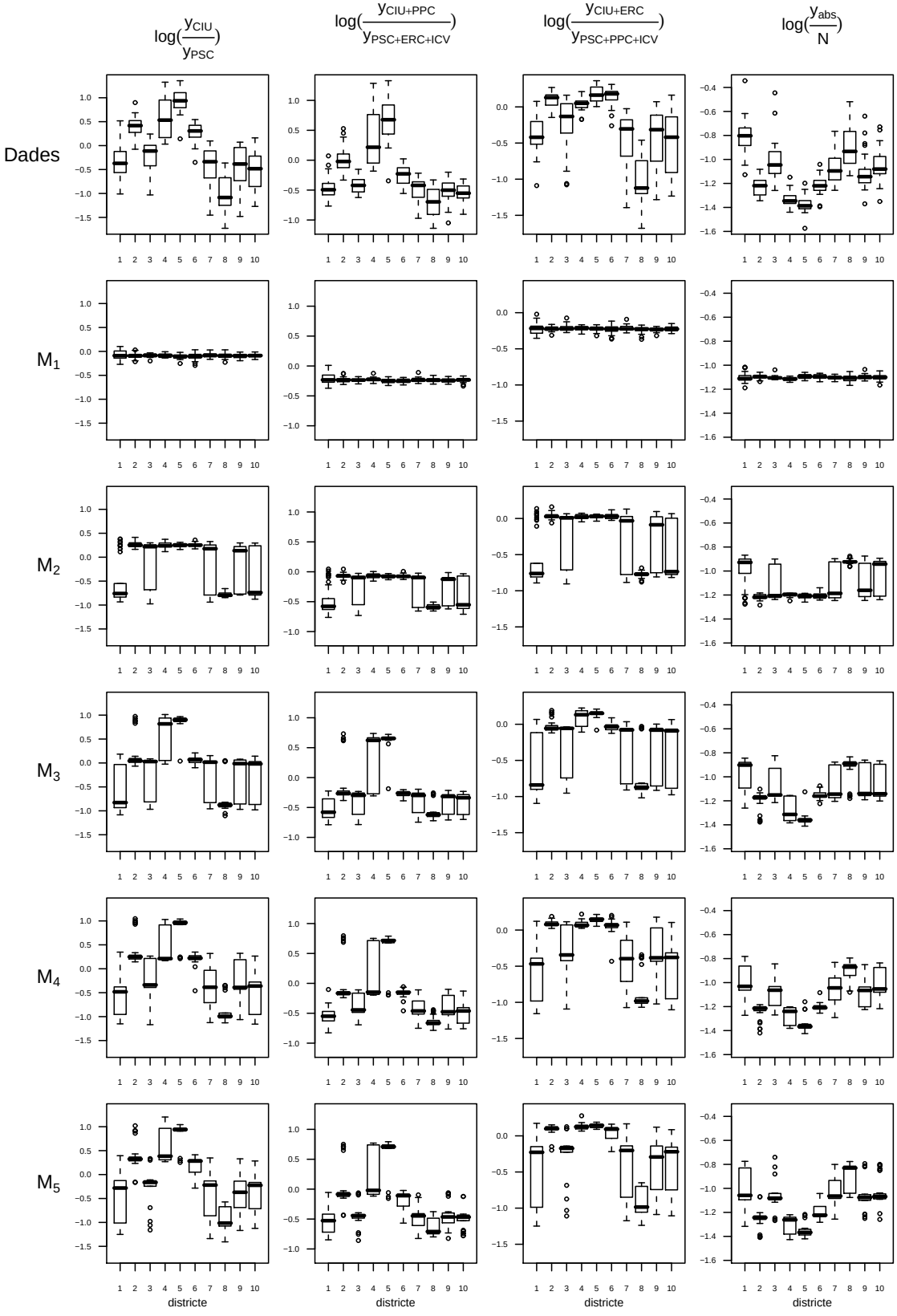

Figura 9.5: La fila superior presenta els valors observats per $D_{ai}(y_i)$, $D_{bi}(y_i)$, $D_{ci}(y_i)$ i $D_{di}(y_i)$ a les eleccions al Parlament de Catalunya del 2003 a les zrp de Barcelona per districtes, i les altres files presenten una rèplica de les dades obtingudes a partir de la prdictiva a posteriori de cada model no jeràrquic considerat.

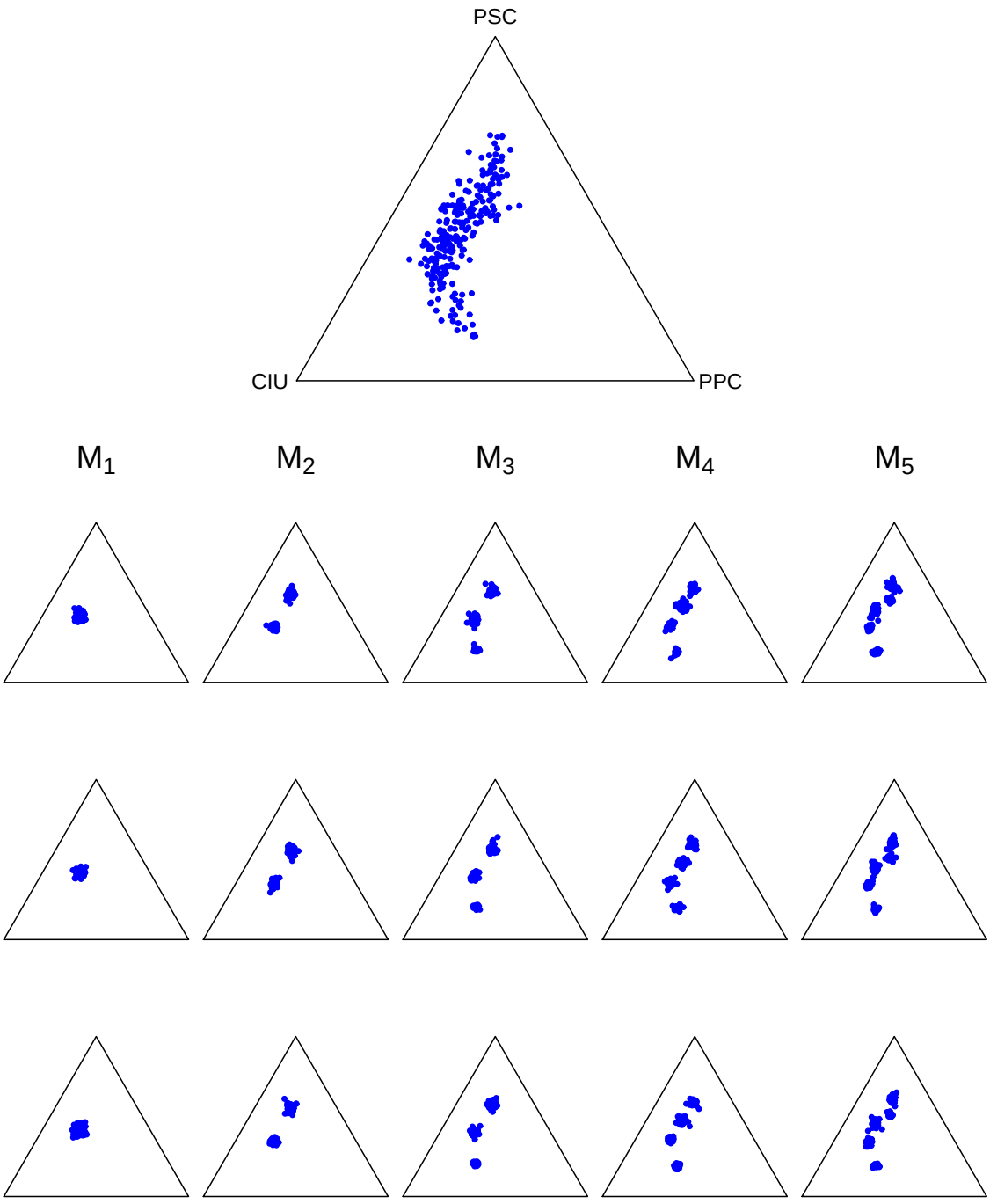

Figura 9.6: Validació gràfica dels diferents models a partir de representacions ternàries dels perfils de tres columnes de la Taula 6.2. El primer gràfic correspon als perfils observats a les eleccions al Parlament de Catalunya del 2003 a Barcelona ciutat i les altres files de corresponen a rèpliques de les dades obtingudes simulant de la predictiva a posteriori de cada un dels cinc models no jeràrquics considerats.

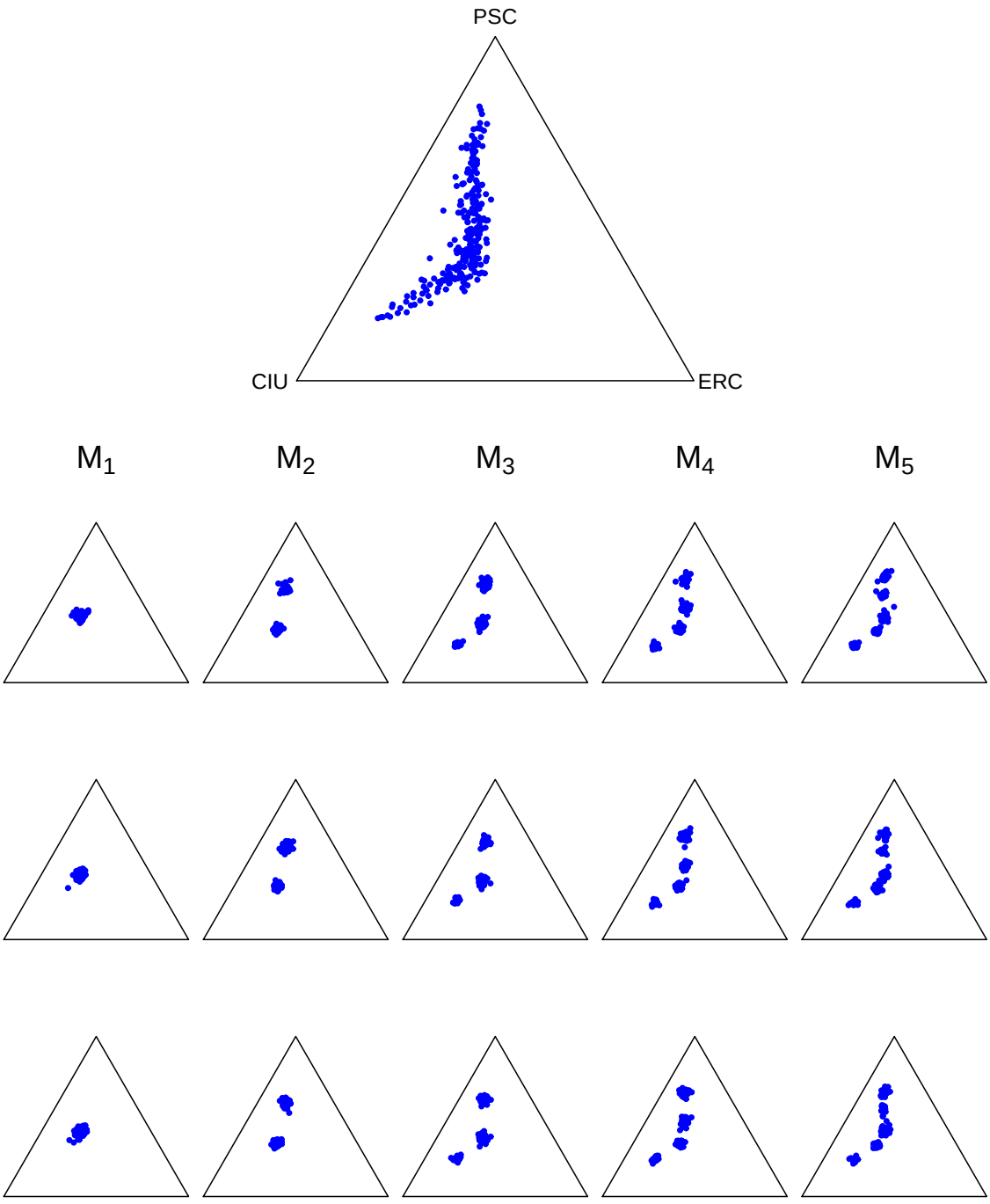

Figura 9.7: Validació gràfica dels diferents models a partir de representacions ternàries dels perfils de tres columnes de la Taula 6.2. El primer gràfic correspon als perfils observats a les eleccions al Parlament de Catalunya del 2003 a Barcelona ciutat i les altres files de corresponen a rèpliques de les dades obtingudes simulant de la predictiva a posteriori de cada un dels cinc models no jeràrquics considerats.

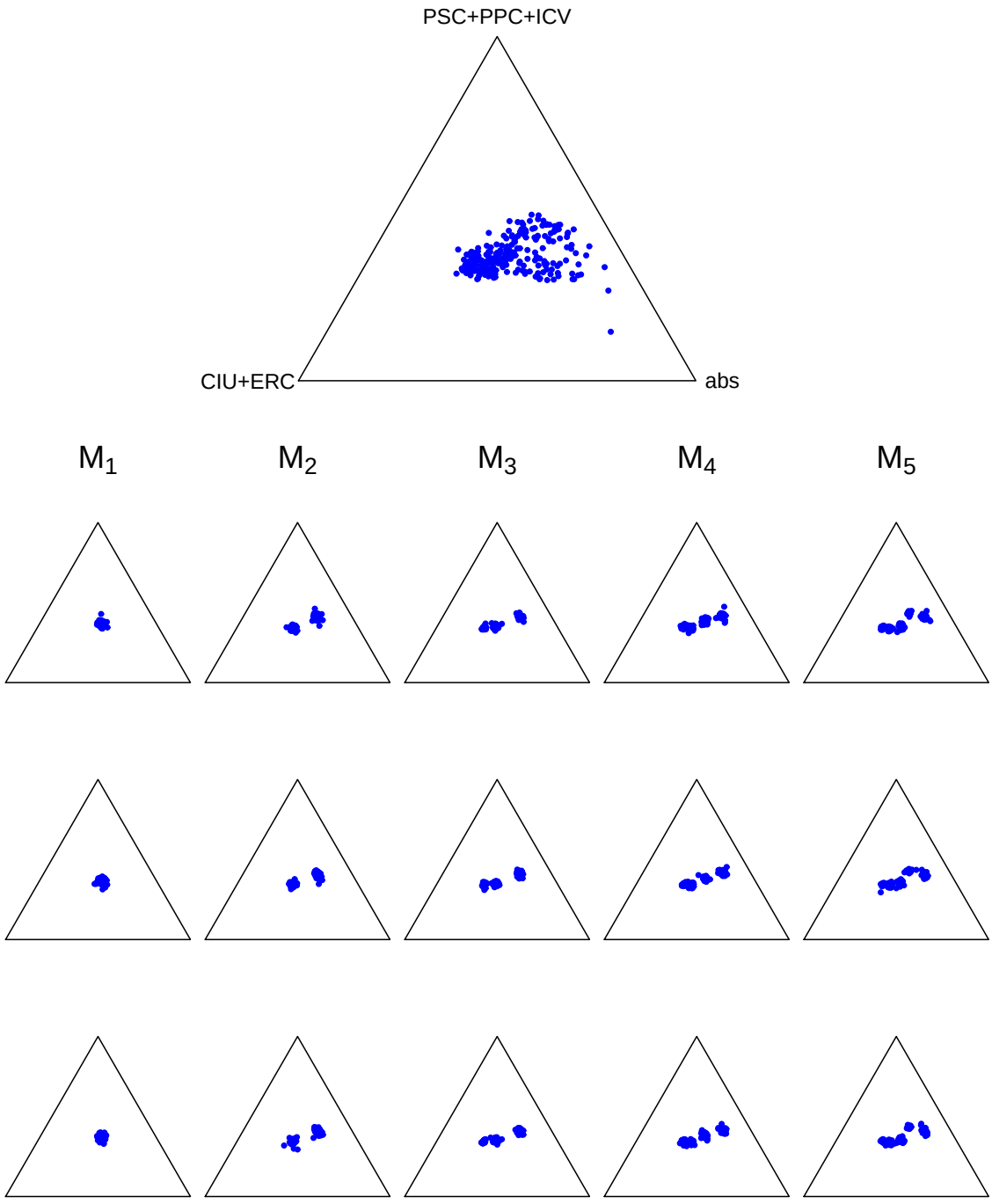

Figura 9.8: Validació gràfica dels diferents models a partir de representacions ternàries dels perfils resultants d'agregar columnes de la Taula 6.2. El primer gràfic correspon als perfils observats a les eleccions al Parlament de Catalunya del 2003 a Barcelona ciutat i les altres files corresponen a rèpliques de les dades obtingudes simulant de la predictiva a posteriori de cada un dels cinc models no jeràrquics considerats.

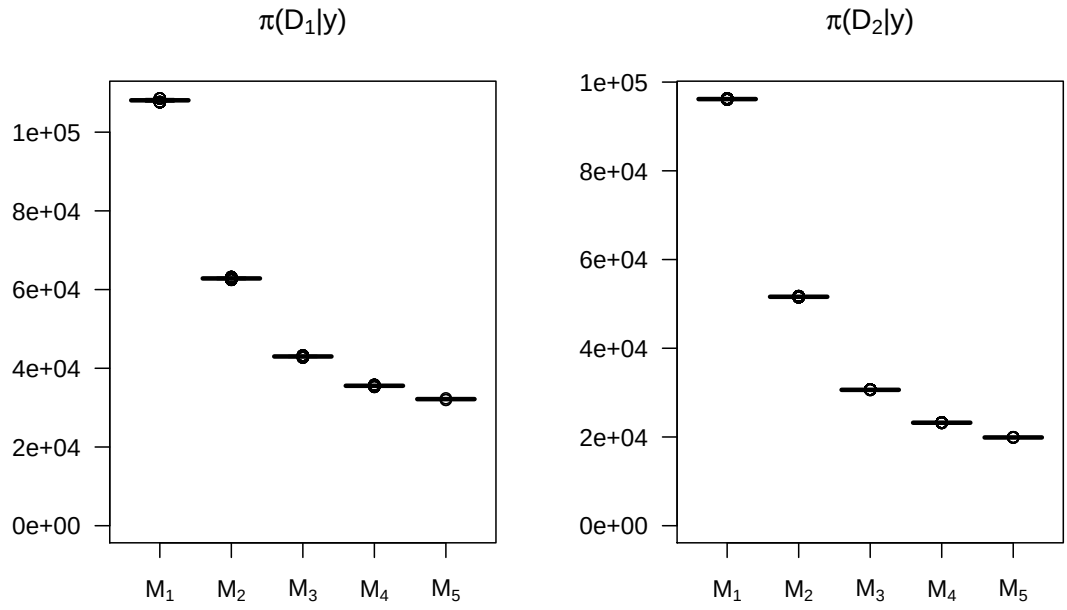

Figura 9.9: Distribucions a posteriori de les mesures de discrepància D_1 i D_2 per als cinc models no jeràrquics considerats.

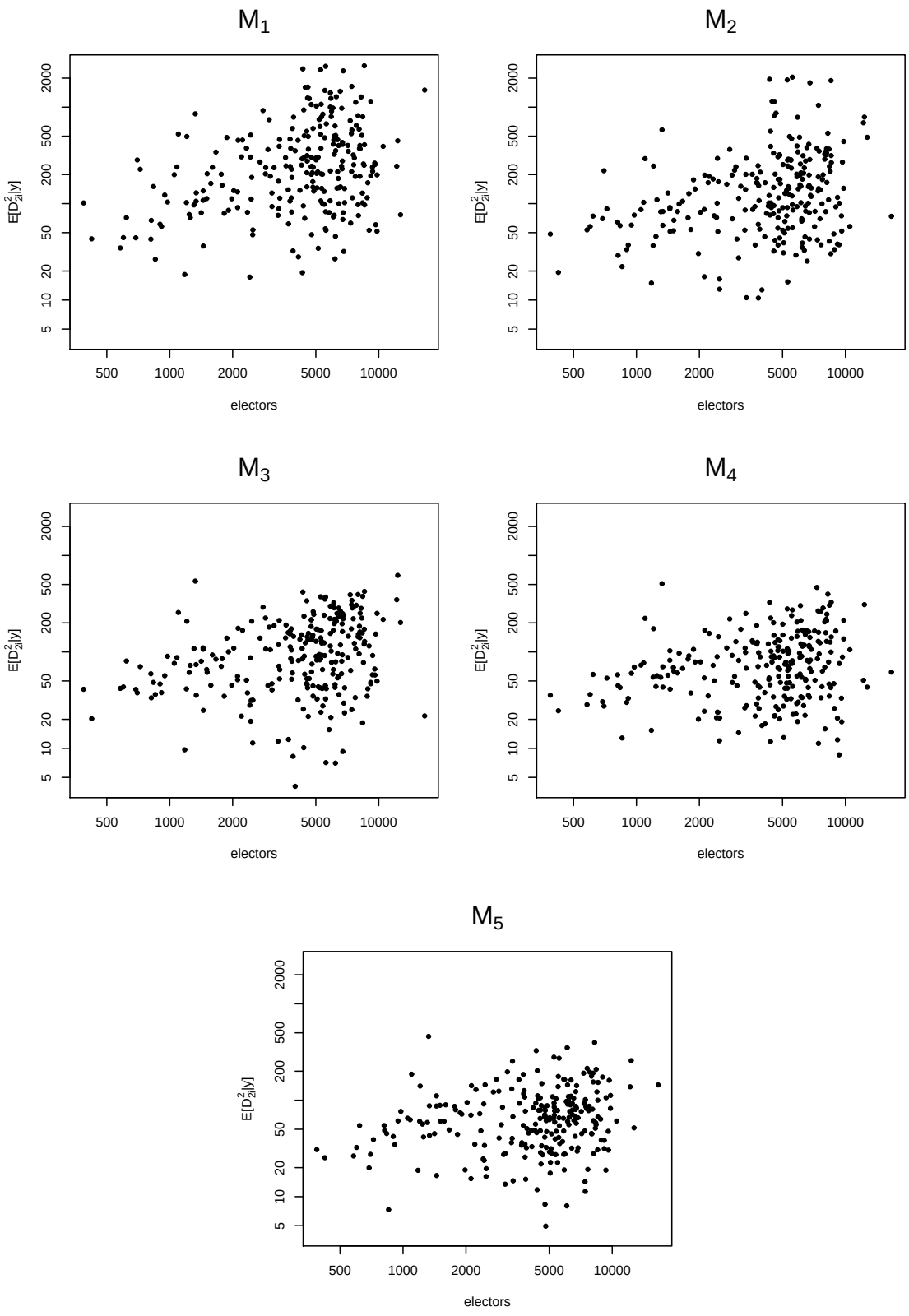

Figura 9.10: Esperança a posteriori de D_{2i}^2, $E[D_{2i}^2|y]$, per a cada zrp en funció del seu nombre d'electors per cada model no jeràrquic considerat.

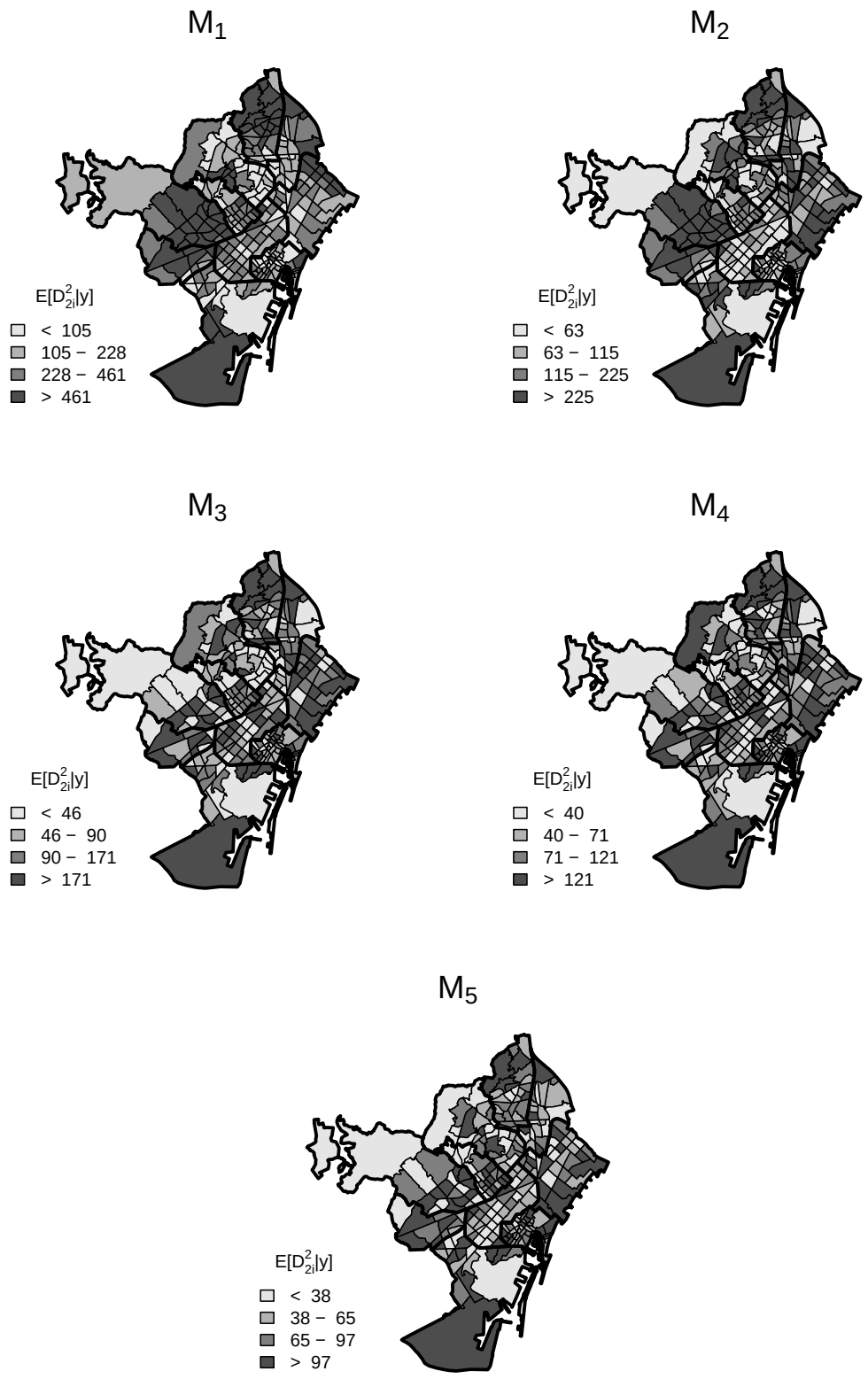

Figura 9.11: Distribució espacial de l'esperança a posteriori de D_{2i}^2 per a cada un dels cinc models no jeràrquics considerats per els resultats a les eleccions al Parlament de Catalunya del 2003 a Barcelona ciutat.

9.2 Comparació dels models jeràrquics

En aquesta secció compararem els models jeràrquics, utilitzats sobre les dades del 2003. De la mateixa menera que hem fet amb el model no jeràrquic, comencem especificant completament el model a priori utilitzat:

$$
\begin{aligned}
y_i|\theta_i &\sim \text{Multinomial}(N_i, \theta_i) \\
\theta_i|\tau\mu\zeta &\sim \text{Dirichlet}(\tau_{\zeta_i}(\mu_{\zeta_i 1}, \ldots, \mu_{\zeta_i k})) \\
\mu_1 = (\mu_{11}, \ldots, \mu_{1k}) &\sim \text{Dirichlet}(1, \ldots, 1) \\
&\vdots \\
\mu_s = (\mu_{s1}, \ldots, \mu_{sk}) &\sim \text{Dirichlet}(1, \ldots, 1) \\
\tau_1 &\sim \text{Gamma}(1, 0.001) \\
&\vdots \\
\tau_s &\sim \text{Gamma}(1, 0.001) \\
\pi(\zeta_i = r|\omega) &= \begin{cases} \omega_1 & r = 1 \\ \omega_2 & r = 2 \\ \vdots & \vdots \\ \omega_s & r = s \end{cases} \\
\omega = (\omega_1, \ldots, \omega_s) &\sim \text{Dirichlet}(1, \ldots, 1),
\end{aligned}
$$

Model 6.1. Model Bayesià a priori jeràrquic per a s clusters Multinomial.

i per tant suposem que el perfil de probabilitat a priori per a cada cluster, μ_r, té una distribució uniforme sobre el símplex de $\mathbb{R}^8$, que el grau d'heterogeneïtat a priori de cada cluster té una distribució gamma amb variància molt gran pert tal de reflexar la nostra incertesa del grau d'heterogeneïtat de cadascun dels clusters. I el vector de probabilitats a priori ω de que l'observació i-èssima pertanyi a cada un dels clusters té una distribució uniforme.

Els models jeràrquics utilitzats en aquesta secció, resultants de variar el nombre de clusters, són:

M_{1J}: model multinomial jeràrquic homogeni d'un sol cluster,

M_{2J}: model multinomial jeràrquic de dos clusters,

M_{3J}: model multinomial jeràrquic de tres clusters,

M_{4J}: model multinomial jeràrquic de quatre clusters, i

M_{5J}: model multinomial jeràrquic de cinc clusters.

Per simular de les distribucions a posteriori dels paràmetres dels diferents models hem utilitzat el programari WinBugs. Per a cadascun dels models hem fet córrer dues cadenes MCMC en paral·lel amb diferents valors inicials. Hem avaluat la convergència de les cadenes utilitzant mesures d'autocorrelació mostral dintre de cada cadena, gràfics de les traces mostrals i la mesura diagnòstic proposada per Gelman i Rubin (1992). En els models jeràrquics la convergència ha estat molt més lenta que la que havíem observat amb els models no jeràrquics.

Aquí les nostres eines de diagnòstic suggereixen descartar les primeres 50000 iteracions de cada cadena com a escalfament previ a la convergència. Donada la dificultat d'assolir la convergència s'han donat com a valors inicials de les cadenes tant valors empírics com valors obtinguts del model no jeràrquic, i els valors inicials per τ han estat valors molt grans per una cadena i petits per l'altre, de manera que s'ha establert el període d'escalfament fins que les respectives cadenes no han mostrejat en el mateix espai. De les iteracions posteriors a l'escalfament se n'han retingut una de cada 100, i la mostra final per a les anàlisis de la distribució a posteriori ha estat de 3000 simulacions. En tots els models s'han portat a terme un inspecció visual de les cadenes simulades de les distribucions a posteriori per una banda validar la convergència i per altre identificar possibles problemes d'identificabilitat. A l'apèndix A es donen més detalls en relació a l'estudi de la convergència.

En aquesta secció compararem els diferents models jeràrquics de la mateixa manera que a la secció anterior ho hem fet pels no jeràrquics. Simultàniament també compararem els models no jeràrquics amb els jeràrquics.

El primer gràfic de la Figura 9.12 presenta l'estadístic $D_{ai}(y_i) = log(y_{i,CIU}/y_{i,PSC})$ per a cada zrp. La representació també es fa estratificada per districtes, ordenats en funció de la mitjana de l'estadístic en cada districte, tal i com havíem fet amb els gràfics per validar els models no jeràrquics. Les quatre files de gràfics següents corresponen a quatre rèpliques de la predictiva a posteriori per a cadascun dels 5 models, M_{1J}, M_{2J}, M_{3J}, M_{4J} i M_{5J}. Comparant aquesta figura amb l'obtinguda pels models no jeràrquics, a la Figura 9.1, encoratja a pensar que els models jeràrquics seran una bona eina per capturar la sobredispersió que no capturen els models no jeràrquics de la secció 9.1. Això permetra determinar el nombre de clusters, assignar les zrp als clusters i fer inferència sobre els paràmetres dels models per cada cluster de forma més correcte.

El model d'un sol cluster M_{1J} no és capaç de generar dades tant extremes com les observades, i el model M_{2J} no és capaç de genrar valors de l'estadístic com els observats en les zrp dels districtes 5 i 4, Sarrià-Sant Gervasi i Les Corts. Segons el model M_{2J} el valor d'aquest estadístic per aquests dos districtes infraestimaria sistemàticament els valors observats, i per tant el perfil de probabilitat del cluster assignat a ells no és compatible amb les dades observades, fet que indica que cal un model més flexible i general. En canvi els models M_{3J}, M_{4J} i M_{5J} presenten patrons en les rèpliques similars als observats.

Òbviament els models més complexos tendiran a generar dades més similars a les observades. Es tracta doncs d'escollir el model més parsimoniós que capturi les principals característiques de les dades. Aquest primer gràfic ens apunta els models de 3 i 4 clusters com a possibles models candidats a ser escollits. Caldrà decidir si el model de 3 clusters captura de forma raonable les característiques que considerem importants de les dades a l'hora de fer la inferència d'acord als nostres objectius o bé presenta mancances que justifiquen un model més complet de 4 clusters.

Les conclusions a les que s'arriba a partir de la Figura 9.14 són variacions amb matissos de les descrites al paràgraf anterior. A la Figura 9.13 observem que els models M_{3J}, M_{4J} i M_{5J} presenten patrons de rèplica similar entre ells però en tots ells s'observa un petit allunyament del patró de les dades pels districtes de Sarrià-Sant Gervasi, Les Corts i l'Eixample. I a la Figura 9.15, que presenta l'estadístic $D_{di}(y_i) = log(y_{abs,i}/N_i)$, també s'observa com a partir del model de tres clusters un ja és capaç de generar rèpliques raonablement semblants a les dades observades, si bé en les dades s'observa que una zrp del districte de Ciutat Vella, amb una alta abstenció, podria tractar-se d'un *outlier*.

A la Figura 9.16 visualitzem de forma compacta els estadístics $D_{ai}(y_i)$, $D_{bi}(y_i)$, $D_{ci}(y_i)$ i $D_{di}(y_i)$. A la primera fila es representa la distribució de l'estadístic observat a nivell de zrp estratificat per districte i a la resta de files hi trobem la distribució fruit de generar una rèplica per a cada model i estadístic. S'observa com els models M_{1J} i M_{2J} estan lluny de capturar el comportament de les dades mentre que el model de tres clusters, M_{3J}, captura els matisos a nivell de districtes i ho fan encara més curosament a l'augmentar el nombre de clusters. La qüestió ara és decidir quin és el nombre mínim de clusters que captura les característiques principals de les dades.

A la Figura 9.17 es presenta la distribució dels *p valors*, p_B, de les zrp estratificant per districte associats a aquests quatre estadístics. Hom a priori desitjaria que els diagrames de caixes estessin centrats en 0.5 i fossin simètrics, però cal tenir present que els diagrames de caixa estan construïts per districtes amb un nombre molt diferent de zrp a cada un; la grandària dels districtes oscil·la entre el districte 4 de Les Corts amb 11 zrp i el districte 1

de Ciutat Vella amb 37 zrp. Per a tots els estadístics els diagrames de caixes dels *p valors* pels diferents districtes sota el model M_{3J} presenten un aspecte més desitjable que no pas els del model M_{2J}. El fet que les zrp d'un mateix districte tendeixin a tenir *p valors* semblants suggereix un possible patrò espacial, en el sentit que zrp properes tendeixen a tenir comportaments similars que el model no és capaç de capturar correctament, i sembla que a l'augmentar el nombre de clusters s'aconsegueix capturar l'estructura espacial ja que com veurem a la Figura 10.4, les zrp pertanyents a un mateix cluster apareixen molt agrupades en el mapa.

Aprofundint més en la distribució dels estadístics $D_{ai}(y_i)$, $D_{bi}(y_i)$, $D_{ci}(y_i)$ i $D_{di}(y_i)$ dins de cada districte, hem estudiat de forma gràfica mesures de tendència central com la mediana i la mitjana, mesures de la variabilitat com el rang interquartílic i la variància i mesures de posició com el primer i tercer quartil, així com el mínim i el màxim de la distribució d'aquests estadístics per cada districte. Hem comparat la distribució predictiva a posteriori d'aquestes mesures sota cada un dels models jeràrquics considerats, amb el respectiu valor observat.

A la Figura 9.18, a la columna de l'esquerra, hem representat per a cada districte la distribució predictiva a posteriori de la mediana dels $D_{ai}(y_i)$ de les zrp d'aquell districte, i el valor de la mediana observat. En el model d'un cluster en molts dels districtes el valor observat s'allunya dels valors simulats a partir del model, en el model de dos clusters millora la compatibilitat dels valors simulats amb els valors observats però encara trobem districtes en que és poc versemblant que els valors observats hagin estat generats pel model, especialment el districte de Sarrià -Sant Gervasi. En el model de tres clusters trobem majors coincidències entre els valors observats i els predits pel model tot i que en la majoria de casos trobem els valors observats a les cues de la distribució de manera que si calculéssim els *pvalor* associats obtindríem valors propers a 0 i a 1. I en el model de quatre clusters ja no trobem valors observats marcadament allunyats dels valors simulats a partir del model.

A nivell de conclusió general per al conjunt de Figures 9.18-9.25 podem dir que els models d'un i dos clusters estan lluny de captura el comportament dels estadístics, el model de tres clusters captura a grans trets el comportament de les dades, el model de quatre clusters presenta una compatibilitat força curosa entre les dades i el model, i que la millora del model de cinc clusters respecte el de quatre és molt petita.

L'estadístic $D_{ci}(y_i)$, a le Figures 9.22-9.23, que registra la relació entre el vot en clau catlana respecte el vot en clau espanyola, és el que els models tenen més dificultat de modelar, especialment en els districtes de Ciutat Vella i Nou Barris. Per a l'estadístic $D_{di}(y_i)$, que registra l'abstenció, els models de tres i quatre clusters no capturen be el

comportament del màxim per als districtes de Ciutat Vella i Sants-Montjuïc on els nivells d'abstenció van ser més alts dels que assumeix el model. En general però els models de tres i més clusters modelen de forma raonable els diferents aspectes analitzats.

La part superior de les Figures 9.26, 9.27 i 9.28 representen simultàniament tres columnes, o tres agregacions de columnes en el cas de la Figura 9.28, de les vuit columnes de les dades mitjançant diagrames ternaris i a la part inferior de les Figures es presenten tres rèpliques simulades utilitzant la predictiva a posteriori de cada model. Comparant aquestes figures amb les obtingudes pels models no jeràrquics, Figures 9.6, 9.7 i 9.8, reflexen la capacitat dels models jeràrquics de modelar la sobredispersió, i a l'augmentar el nombre de clusters, especialment a partir de tres clusters, recullen millor el patró de les dades.

A la Taula 9.3 es mostra l'esperança a posteriori de la mesura de bondat d'ajust

$$D_2(y, \mu_1, \ldots, \mu_s, \zeta) = \sum_{i=1}^{n} \sum_{j=1}^{k} \frac{(y_{ij} - N_i \mu_{\zeta_i j})^2}{N_i \mu_{\zeta_i j}},$$

per a cadascun dels models considerats i a la Figura 9.29 s'hi representen les respectives distribucions a posteriori, $\pi(D_2|y)$. Al passar d'un a dos clusters reduïm dràsticament el valor esperat a posteriori de D_2, i al augmentar el nombre de clusters a més de dos reduïm encara més aquest valor esperat, tot i que menys dràsticament. Les distribucions a posteriori de D_2 són similars a les observades pels models no jeràrquics de la secció 9.1, però cal tenir present que les distribucions de referència pel model jeràrquic i pel model no jeràrquic són molt diferents, de manera que per als models jeràrquics la magnitud de les discrepàncies estan dintre les esperades i això fa que obtinguem *p valors* acceptables per a tots els models. Ressalta el fet de que els *p valors* pels models M_1 i M_2 són més aviat baixos, mentre que pels de més clusters són més aviat alts, indicant que, malgrat per aquest criteri concret tots els models són compatibles amb les dades, pels models M_1 i M_2 les dades tendeixen a obtenir valors de l'estadístic lleugerament més alts que els simulats, mentre que pels models M_3, M_4 i M_5 aquests valors són més petits i en conseqüencia de ben segur que al passar de 2 a 3 clusters quelcom canviarà de la inferència.

De la mateixa forma que hem fet per al cas no jeràrquic examinarem la contribució a l'estadístic D_2 de cada zrp. El valor esperat de la mesura de discrepància D_{2i}^2, on

$$D_{2i}^2(y, \mu_1, \ldots, \mu_s, \zeta) = \sum_{j=1}^{k} \frac{(y_{ij} - N_i \mu_{\zeta_i j})^2}{N_i \mu_{\zeta_i j}}.$$

per a cada observació, $E[D_{2i}^2|y]$, graficada a la Figura 9.30 en escala logarítmica, mostra una associació dèbil amb el nombre d'electors de cada zrp. Al mateix temps s'observa

Model	$E[D_2\|y]$	$p_B^{D_2}$
M_1	107350.3	0.10125
M_2	56405.1	0.18545
M_3	33254.7	0.9283
M_4	25800.3	0.8090
M_5	22234.5	0.9208

Taula 9.3: Esparança a posteriori de la mesura de discrepància D_2, i el *p valor* bayesià associats, $p_B^{D_2}$, per els resultats a les eleccions al Parlament de Catalunya del 2003 a la ciutat de Barcelona.

que la magnitud de discrepància és molt gran per als models d'un i dos clusters, però es redueix notablement al passar de 2 a 3 clusters i es redueix de forma més moderada al passar de 3 a 4 i de 4 a 5 clusters.

Per avaluar l'existència d'una possible correlació espacial de D_{2i}^2 amb la finalitat d'observar si les desviacions del model estan estructurades espacialment la Figura 9.31 representa el valor esperat de D_{2i}^2 en mapes. Visualment es percep una agrupació espacial en els model M_{1J} i M_{2J} a diferència de la resta de models.

Per eliminar subjectivitat sobre l'existència o no de correlació espacial hem calculat l'índex de Moran, I, a partir de $E[D_{2i}^2|y]$, i l'hem contrastat amb una distribució de referència obtinguda a partir de 5000 permutacions. Això ens permetrà observar si el grau de dependència espacial en D_{2i}^2 varia al canviar d'un model a un altre.

A la columna esquerra de la Figura 9.32 s'hi representa la distribució de referència i el valor observat per I per a cadascun dels models, i a la columna dreta s'hi representa el correlograma fins al retard de seté ordre, també per a cadascun dels models. En els models M_{1J} i M_{2J} els valors obtinguts per I es troben extremadament allunyats de la distribució de referència i els respectius correlogrames presenten la típica forma d'estructura espacial, en els que les correlacions de primer ordre presenten valors elevats per després anar disminuint fins a valors al voltant de zero a partir d'un cert retard. Els models M_{3J}, M_{4J} i M_{5J} presenten valors per I semblants entre ells i alhora totalment diferents dels models d'un i dos clusters.

A partir del model de tres clusters el valor observat per I és compatible amb els valors de la distribució de referència tot i que sempre son part de la cua dreta de la distribució. Per mesurar el nivell de conflicte entre el valor observat i la distribució de referència hem calculat l'àrea de cua. Aquests *p valors* es mostren a la Taula 9.4. S'observa que pels models de tres i més clusters s'obtenen *p valors* de magnituds semblants. Els corelogrames dels models de tres i més clusters presenten valors de $I^{(t)}$ relativament

	I	$p\ valor$
M_{1J}	0.594	0.000
M_{2J}	0.558	0.000
M_{3J}	0.066	0.031
M_{4J}	0.069	0.024
M_{5J}	0.071	0.023

Taula 9.4: Índex de Moran i *p valor* associat obtingut fent servir el test de permutació partint de $E[D^2_{2i}|y]$ per a cadascun dels cinc models considerats.

petits per a tots els retards.

Tot i que l'estructura fruit de modelar tres o més clusters ha reduït considerablement l'existència de correlació espacial el fet d'observar *p valors* petits revelen l'existència de certa dependència espacial en tots els models. El fet de que a partir de tres clusters els *p valors* es mantinguin relativament estables fa pensar que en cas de voler abordar aquesta correlació convindria seguir una altra estratègia diferent a la d'anar augmentant el número de clusters.

A la vista dels gràfics de validació i d'acord als nostres objectius cal escollir entre el model de 3 clusters jeràrquic, M_{3J}, o el model de 4 clusters jeràrquic, M_{4J}. Amb criteris merament estadístics escolliríem el model de 4 clusters jeràrquic en el sentit que recull més adequadament les principals característiques de les dades. Ara bé, cal valorar si la complicació que representa afegir un cluster, a nivell de presentar i interpretar de forma comprensible els resultats, compensa les mancances que pot tenir assumir el model M_{3J}.

En el capítol següent analitzarem els resultats basant-nos tant en el model de tres com en el de quatre clusters, amb la finalitat d'explorar l'efecte que la decisió d'escollir un model o l'altre pot tenir sobre les conclusions finals.

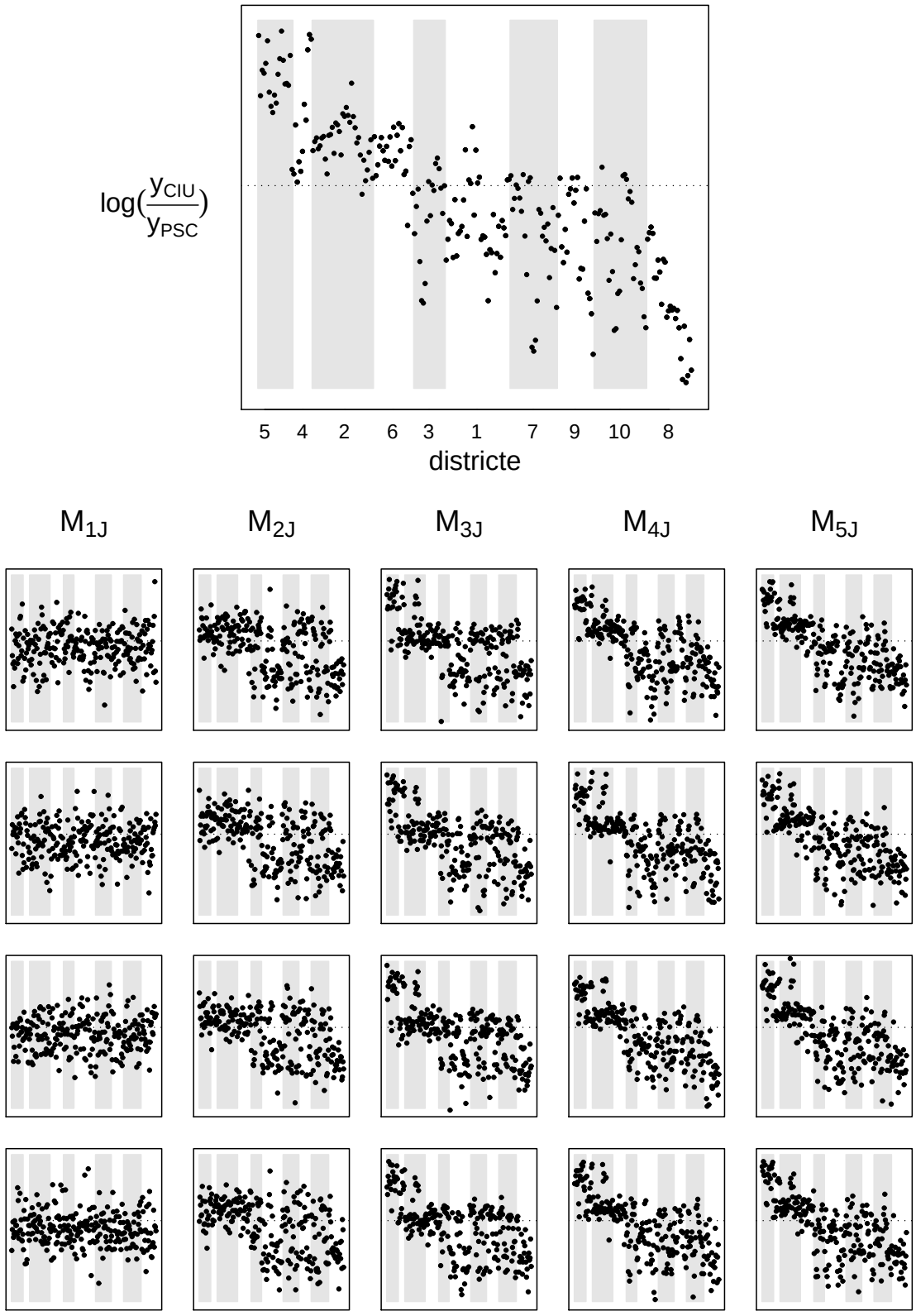

Figura 9.12: El primer gràfic representa els valors observats per $D_{ai}(y_i)$ a les zrp de cada districte a les eleccions al Parlament de Catalunya del 2003 a Barcelona ciutat i les quatre files següents corresponen a quatre rèpliques d'aquest estadístic simulades a partir de la predictiva a posteriori de cada model jeràrquic considerat.

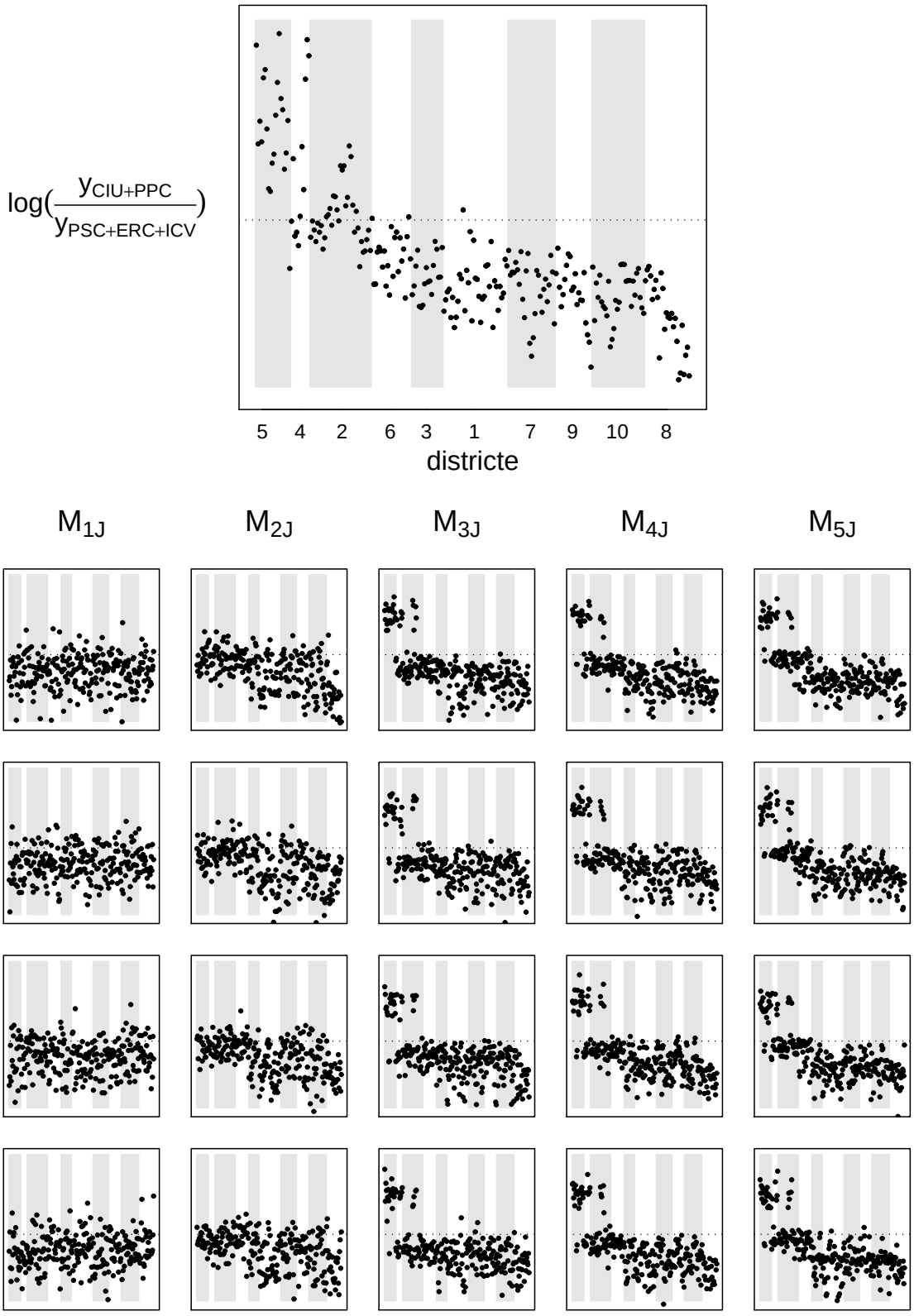

Figura 9.13: El primer gràfic representa els valors observats per $D_{bi}(y_i)$ a les zrp de cada districte a les eleccions al Parlament de Catalunya del 2003 a Barcelona ciutat i les quatre files següents corresponen a quatre rèpliques d'aquest estadístic simulades a partir de la predictiva a posteriori de cada model jeràrquic considerat.

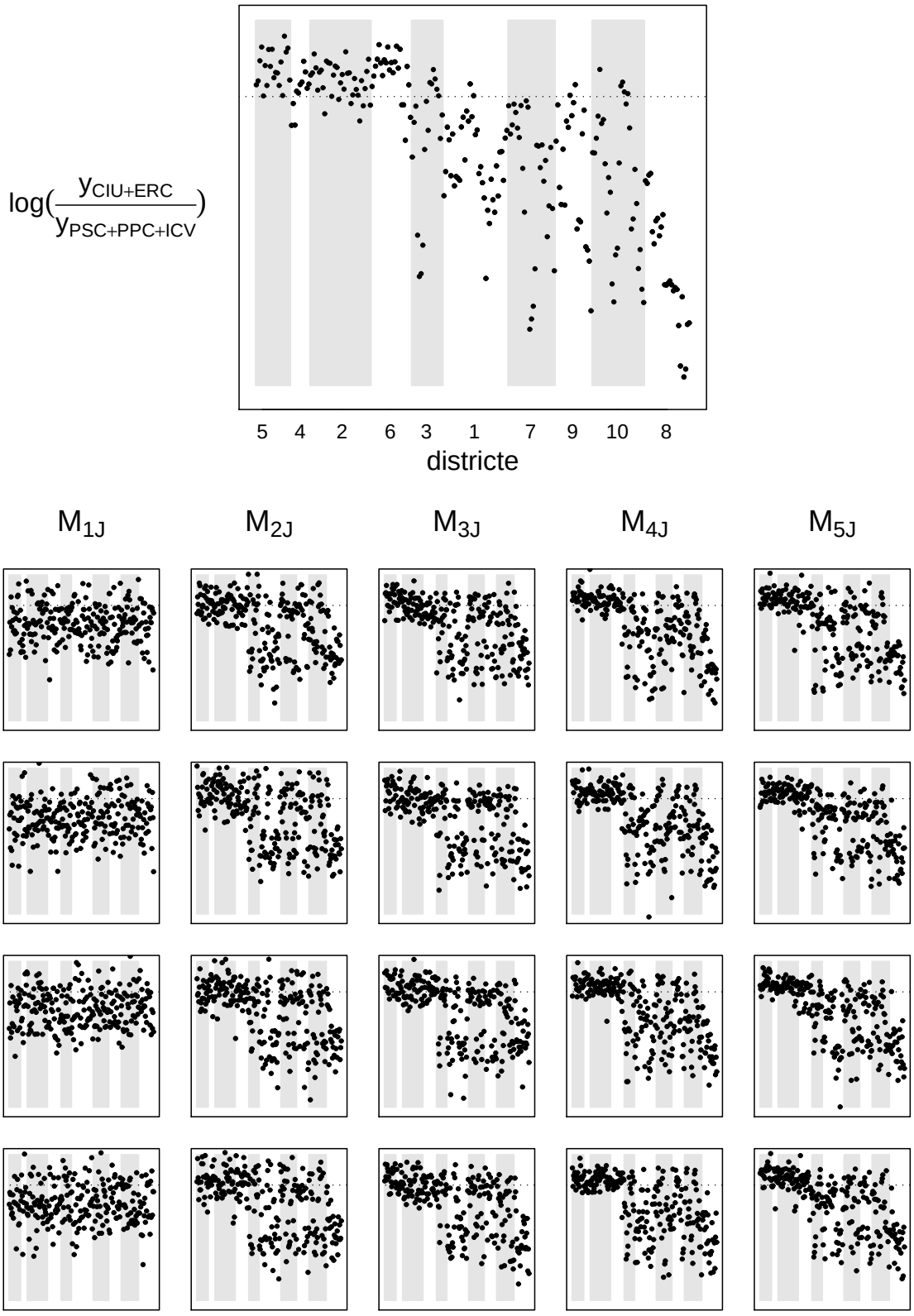

Figura 9.14: El primer gràfic representa els valors observats per $D_{ci}(y_i)$ a les zrp de cada districte a les eleccions al Parlament de Catalunya del 2003 a Barcelona ciutat i les quatre files següents corresponen a quatre rèpliques d'aquest estadístic simulades a partir de la predictiva a posteriori de cada model jeràrquic considerat.

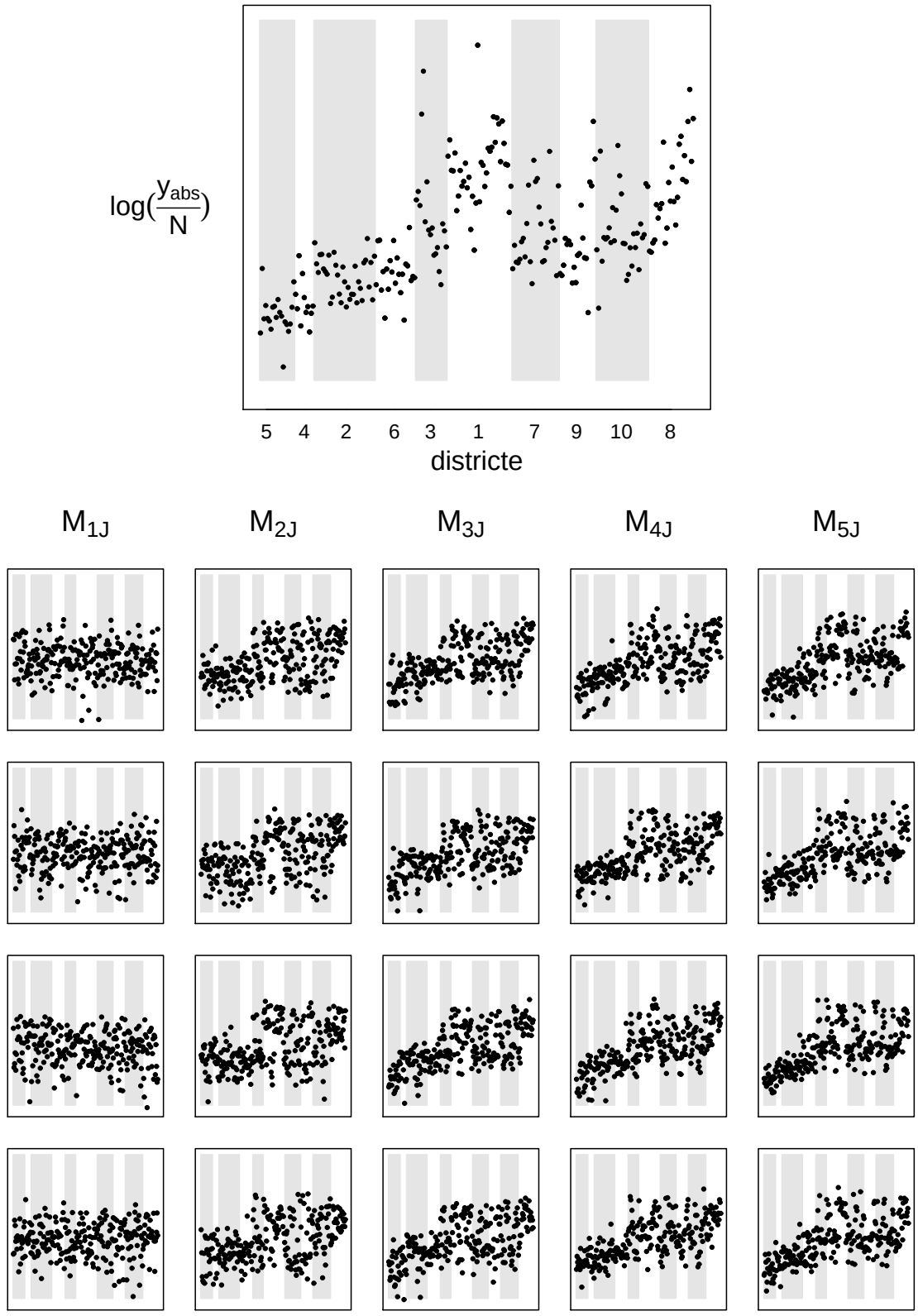

Figura 9.15: El primer gràfic representa els valors observats per $D_{di}(y_i)$ a les zrp de cada districte a les eleccions al Parlament de Catalunya del 2003 a Barcelona ciutat i les quatre files següents corresponen a quatre rèpliques d'aquest estadístic simulades a partir de la predictiva a posteriori de cada model jeràrquic considerat.

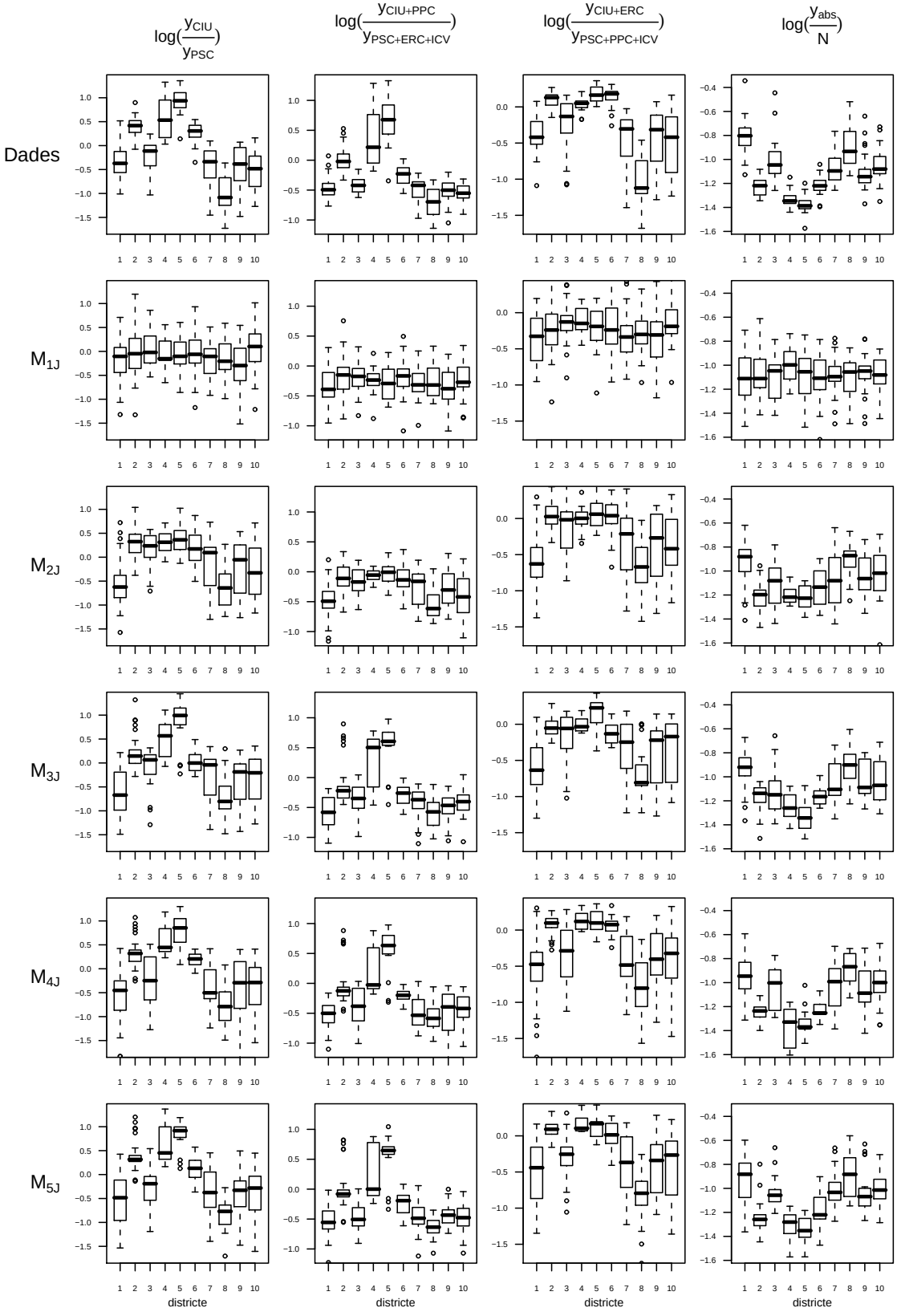

Figura 9.16: La fila superior presenta els valors observats per $D_{ai}(y_i)$, $D_{bi}(y_i)$, $D_{ci}(y_i)$ i $D_{di}(y_i)$ a les eleccions al Parlament de Catalunya del 2003 a les zrp de Barcelona per districtes, i les altres files presenten una rèplica de les dades obtingudes a partir de la predictiva a posteriori de cada model jeràrquic considerat.

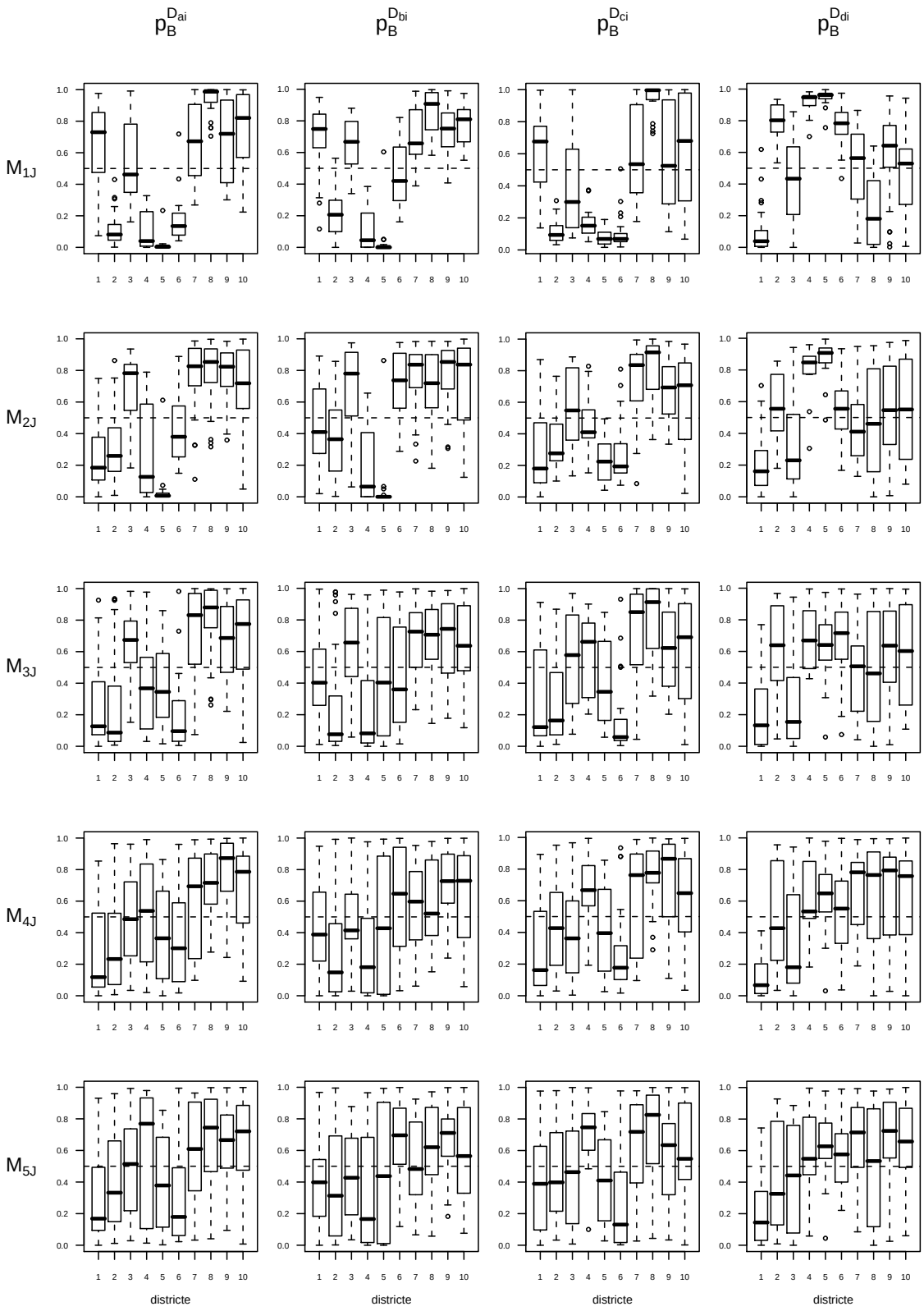

Figura 9.17: Distribució dels *p valors*, p_B, de les zrp estratificant per districte associats als estadístics $D_{ai}(y_i)$, $D_{bi}(y_i)$, $D_{ci}(y_i)$ i $D_{di}(y_i)$ a les eleccions al parlament de Catalunya del 2003 de Barcelona.

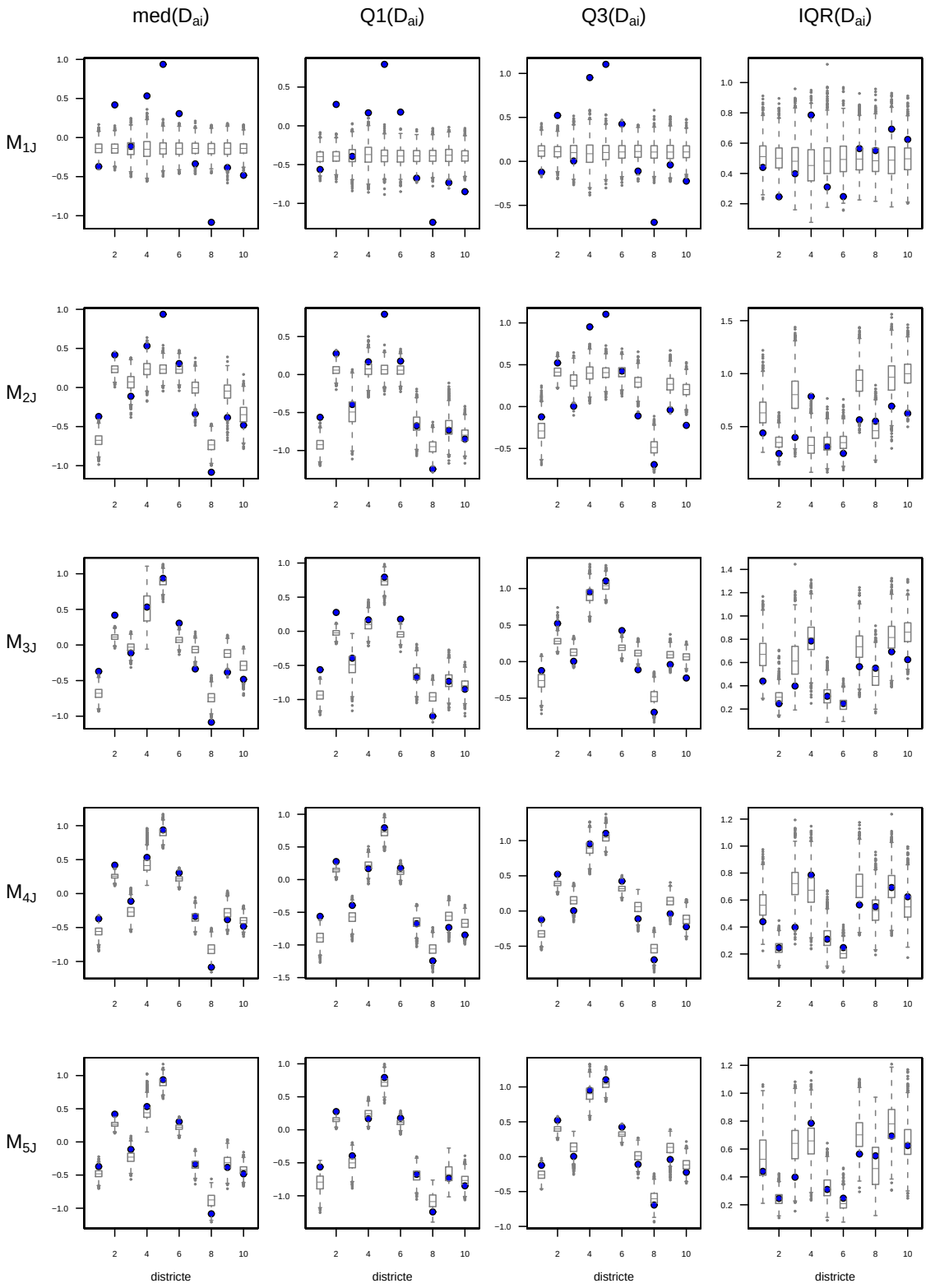

Figura 9.18: Els punts representen els valors observats per la mediana (med), el primer quartil (Q1), el tercer quartil (Q3) i el rang interquartílic (IQR) de la distribució de l'estadístic $D_{ai}(y_i) = log(y_{i,CIU}/y_{i,PSC})$ a cada un dels districtes, i els diagrames de caixa representen la distribució predictiva a posteriori d'aquest estadístic per cada districte sota cada model jeràrquic considerat.

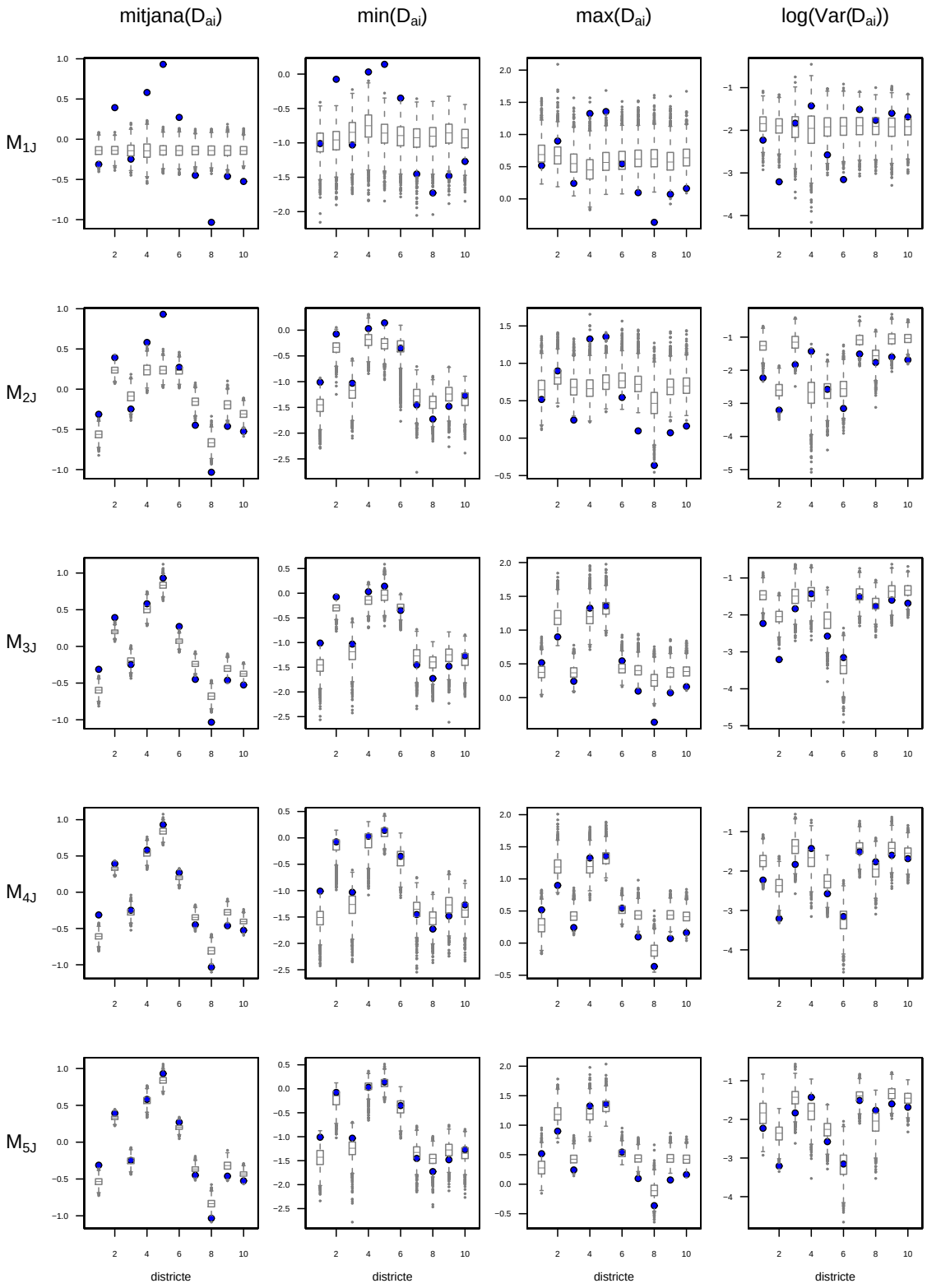

Figura 9.19: Els punts representen els valors observats per la mitjana, el mínim (min), el màxim (max) i el logaritme de la variància de la distribució de l'estadístic $D_{ai}(y_i) = log(y_{i,CIU}/y_{i,PSC})$ a cada un dels districtes, i els diagrames de caixa representen la distribució predictiva a posteriori d'aquest estadístic per cada districte sota cada model jeràrquic considerat.

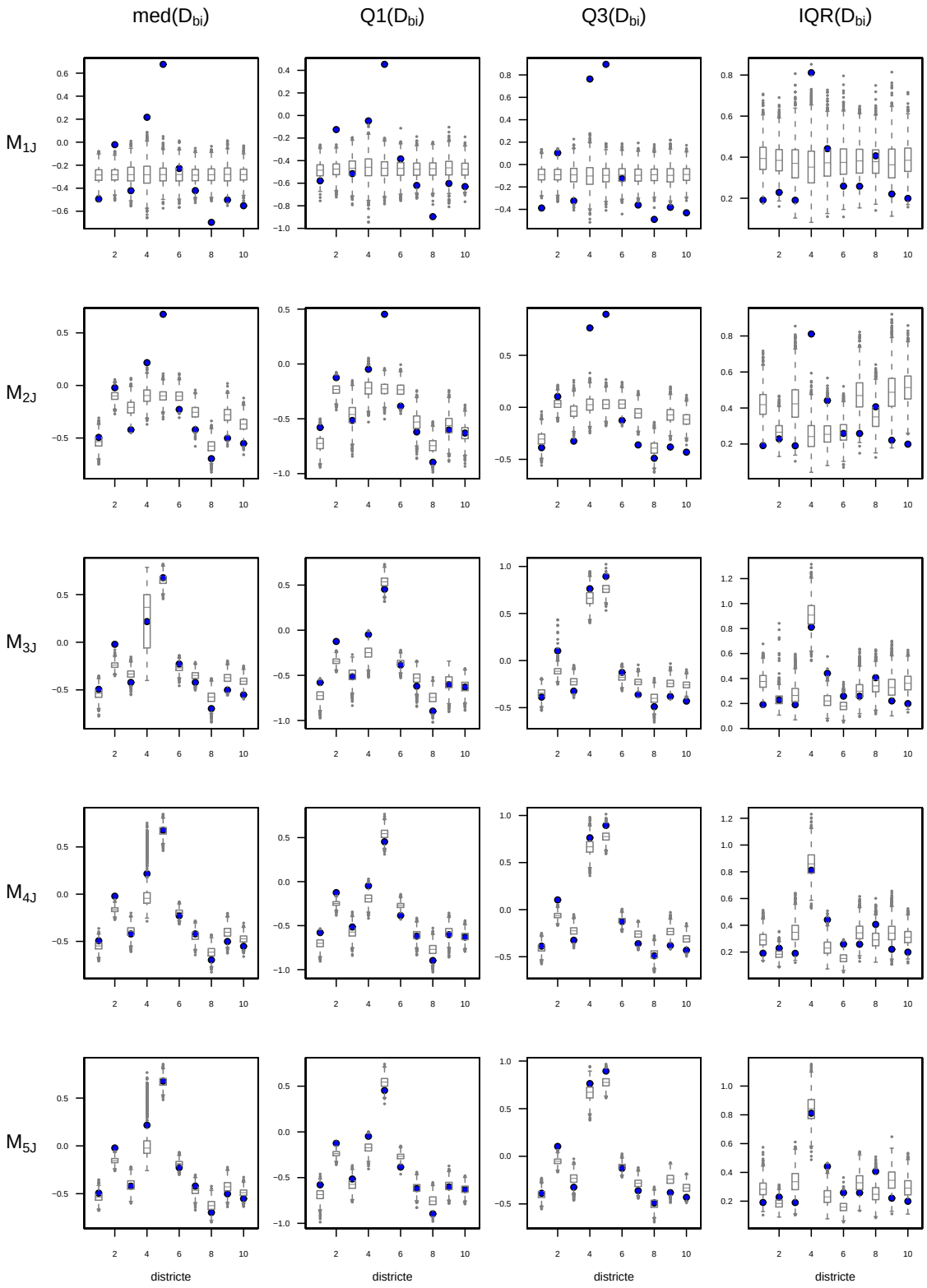

Figura 9.20: Els punts representen els valors observats per la mediana (med), el primer quartil (Q1), el tercer quartil (Q3) i el rang interquartílic (IQR) de la distribució de l'estadístic $D_{bi}(y_i) = log(y_{i,CIU+PPC}/y_{i,PSC+ERC+ICV})$ a cada un dels districte, i els diagrames de caixa representen la distribució predictiva a posteriori d'aquest estadístic per cada districte sota cada model jeràrquic considerat.

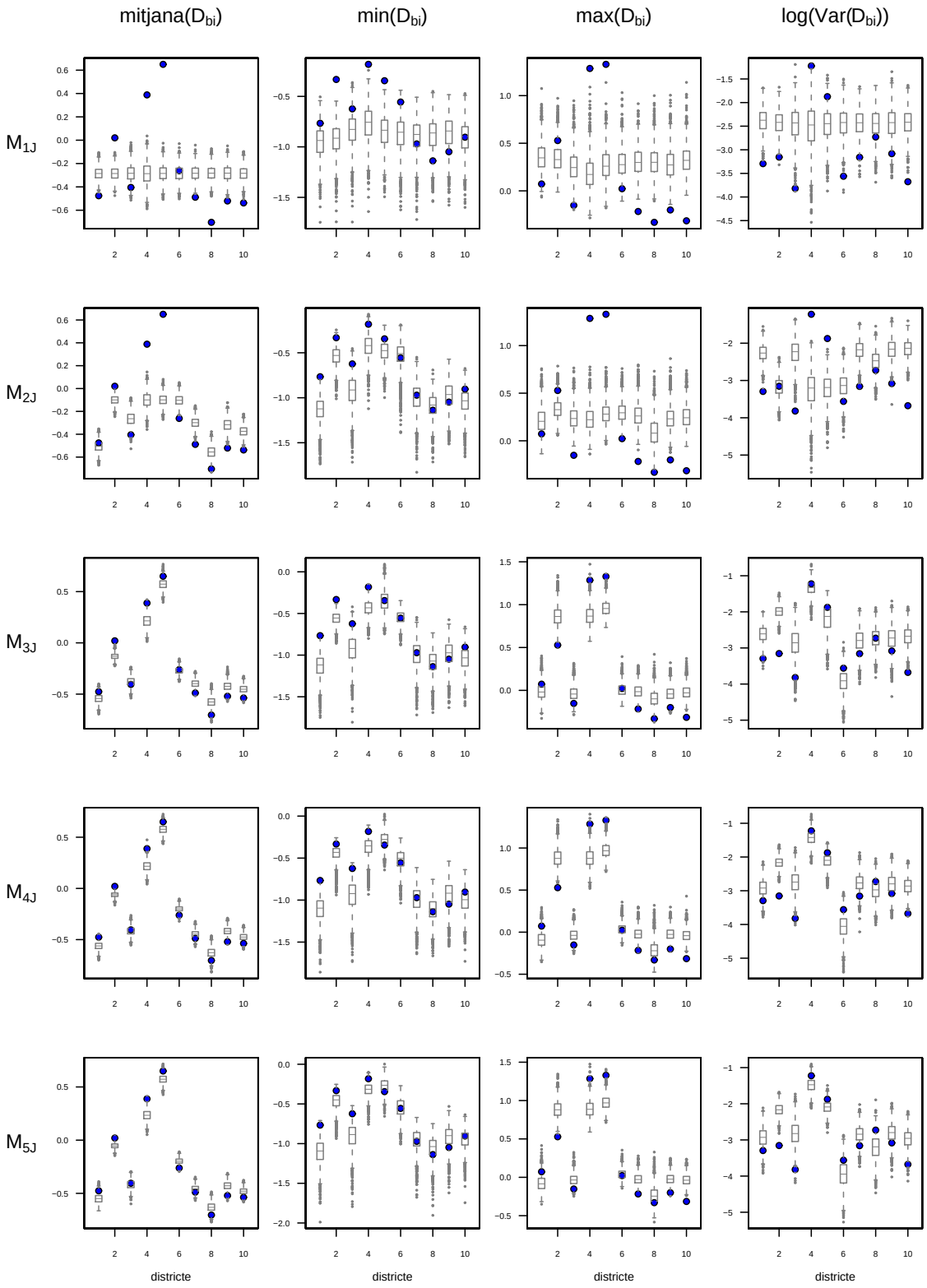

Figura 9.21: Els punts representen els valors observats per la mitjana, el mínim (min), el màxim (max) i el logaritme de la variància de la distribució de l'estadístic $D_{bi}(y_i) = log(y_{i,CIU+PPC}/y_{i,PSC+ERC+ICV})$ a cada un dels districtes, i els diagrames de caixa representen la distribució predictiva a posteriori d'aquest estadístic per cada districte sota cada model jeràrquic considerat.

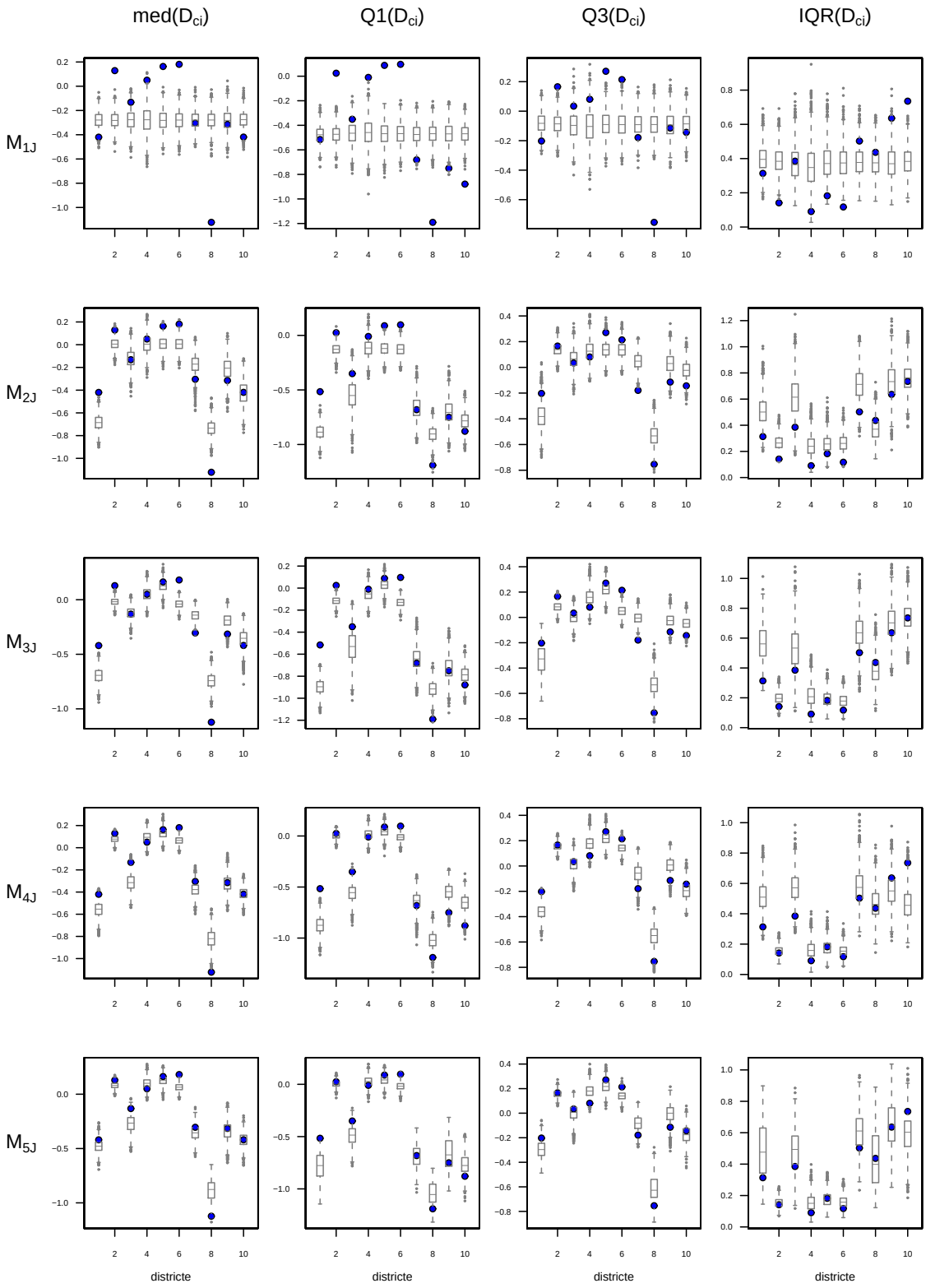

Figura 9.22: Els punts representen els valors observats per la mediana (med), el primer quartil (Q1), el tercer quartil (Q3) i el rang interquartílic (IQR) de la distribució de l'estadístic $D_{ci}(y_i) = log(y_{i,CIU+ERC}/y_{i,PSC+PPC+ICV})$ a cada un dels districtes, i els diagrames de caixa representen la distribució predictiva a posteriori d'aquest estadístic per cada districte sota cada model jeràrquic considerat.

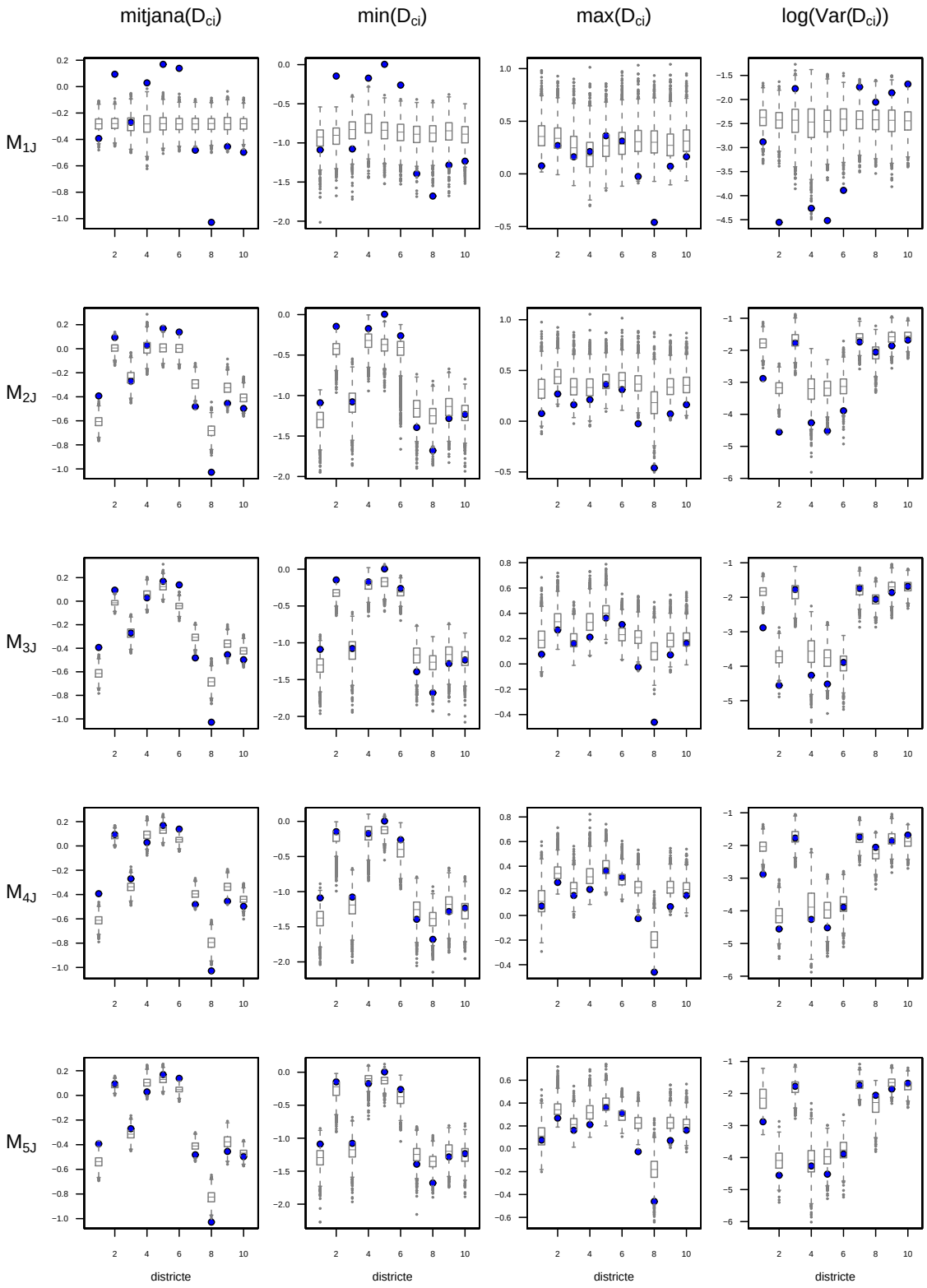

Figura 9.23: Els punts representen els valors observats per la mitjana, el mínim (min), el màxim (max) i el logaritme de la variància de la distribució de l'estadístic $D_{ci}(y_i) = log(y_{i,CIU+ERC}/y_{i,PSC+PPC+ICV})$ a cada un dels districtes, i els diagrames de caixa representen la distribució predictiva a posteriori d'aquest estadístic per cada districte sota cada model jeràrquic considerat.

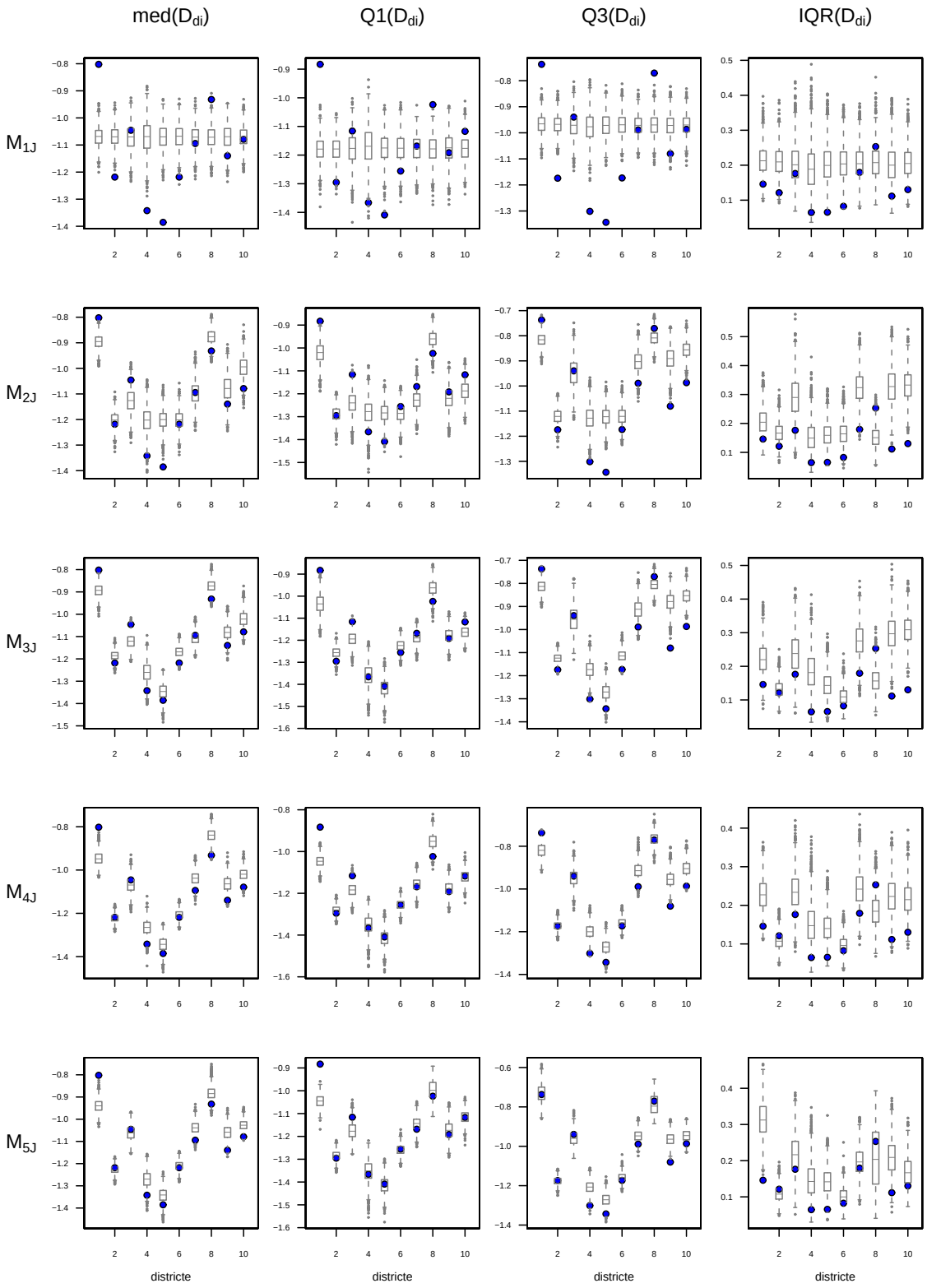

Figura 9.24: Els punts representen els valors observats per la mediana (med), el primer quartil (Q1), el tercer quartil (Q3) i el rang interquartílic (IQR) de la distribució de l'estadístic $D_{di}(y_i) = log(y_{i,abs}/N_i)$ a cada un dels districtes, i els diagrames de caixa representen la distribució predictiva a posteriori d'aquest estadístic per cada districte sota cada model jeràrquic considerat.

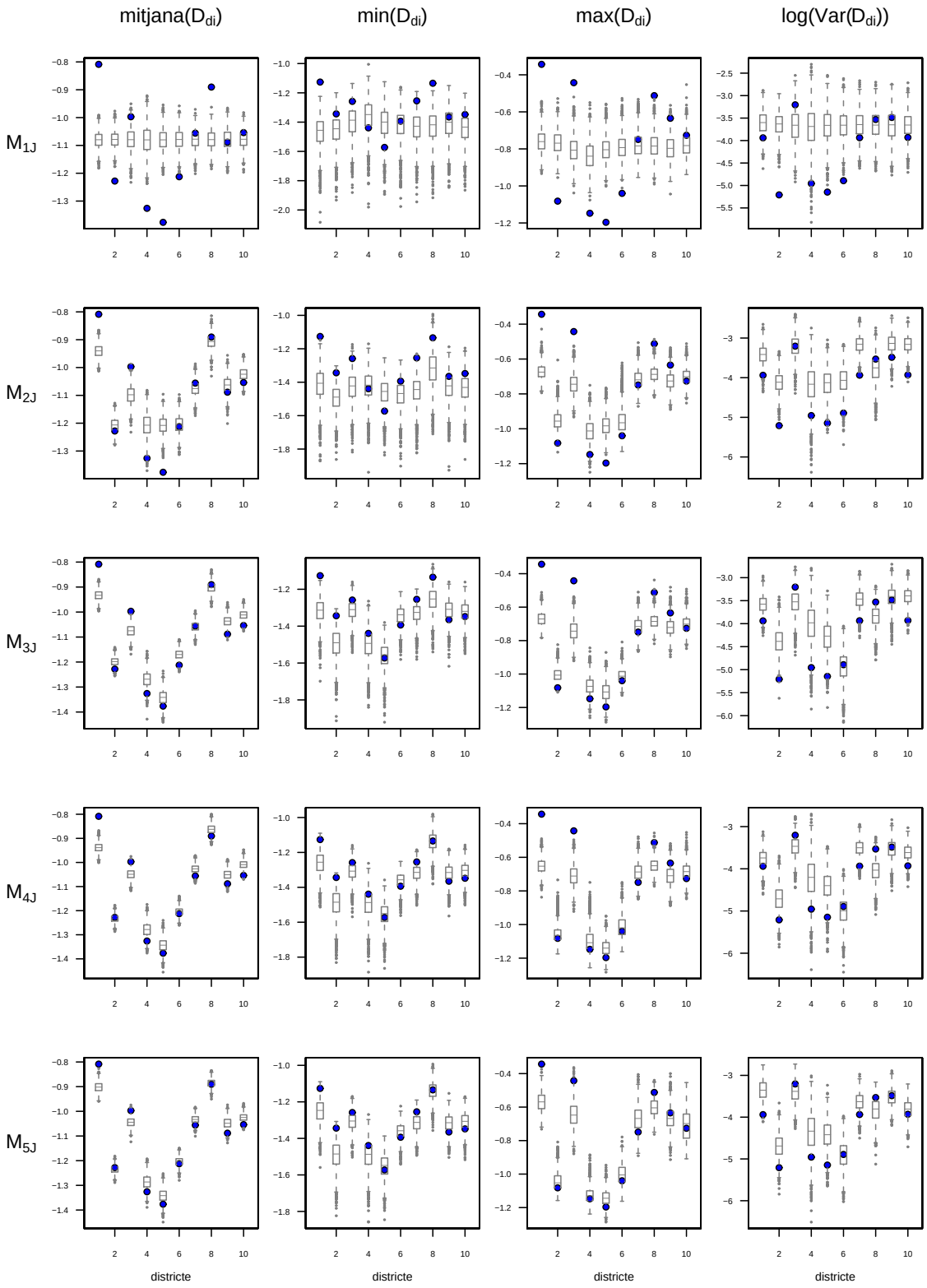

Figura 9.25: Els punts representen els valors observats per la mitjana, el mínim (min), el màxim (max) i el logaritme de la variància de la distribució de l'estadístic $D_{di}(y_i) = log(y_{i,abs}/N_i)$ a cada un dels districtes, i els diagrames de caixa representen la distribució predictiva a posteriori d'aquest estadístic per cada districte sota cada model jeràrquic considerat.

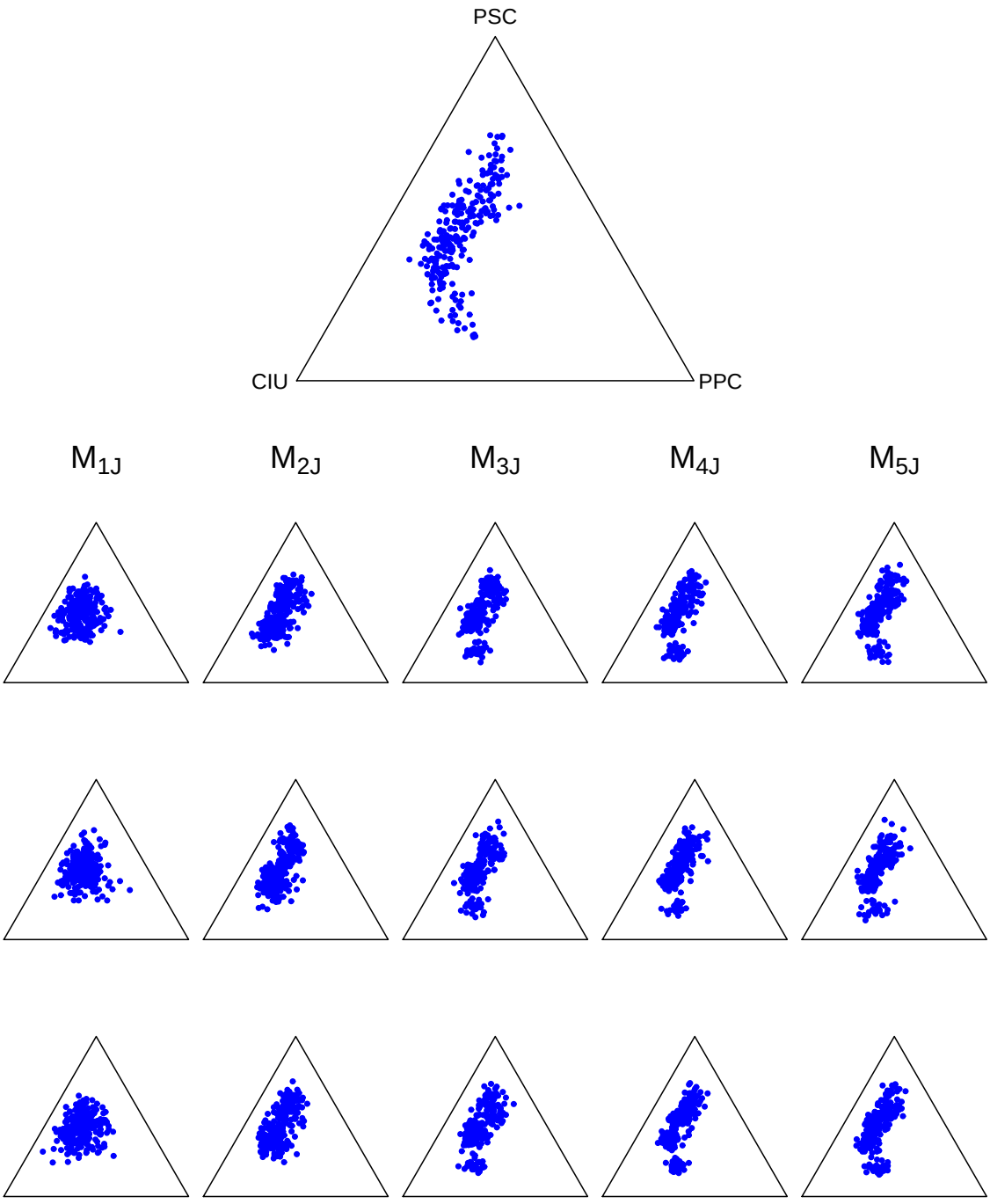

Figura 9.26: Validació gràfica dels diferents models a partir de representacions ternàries dels perfils de tres columnes de la Taula 6.2. El primer gràfic correspon als perfils observats a les eleccions al Parlament de Catalunya del 2003 a Barcelona ciutat i les altres files de corresponen a rèpliques de les dades obtingudes simulant de la predictiva a posteriori de cada un dels cinc models jeràrquics considerats.

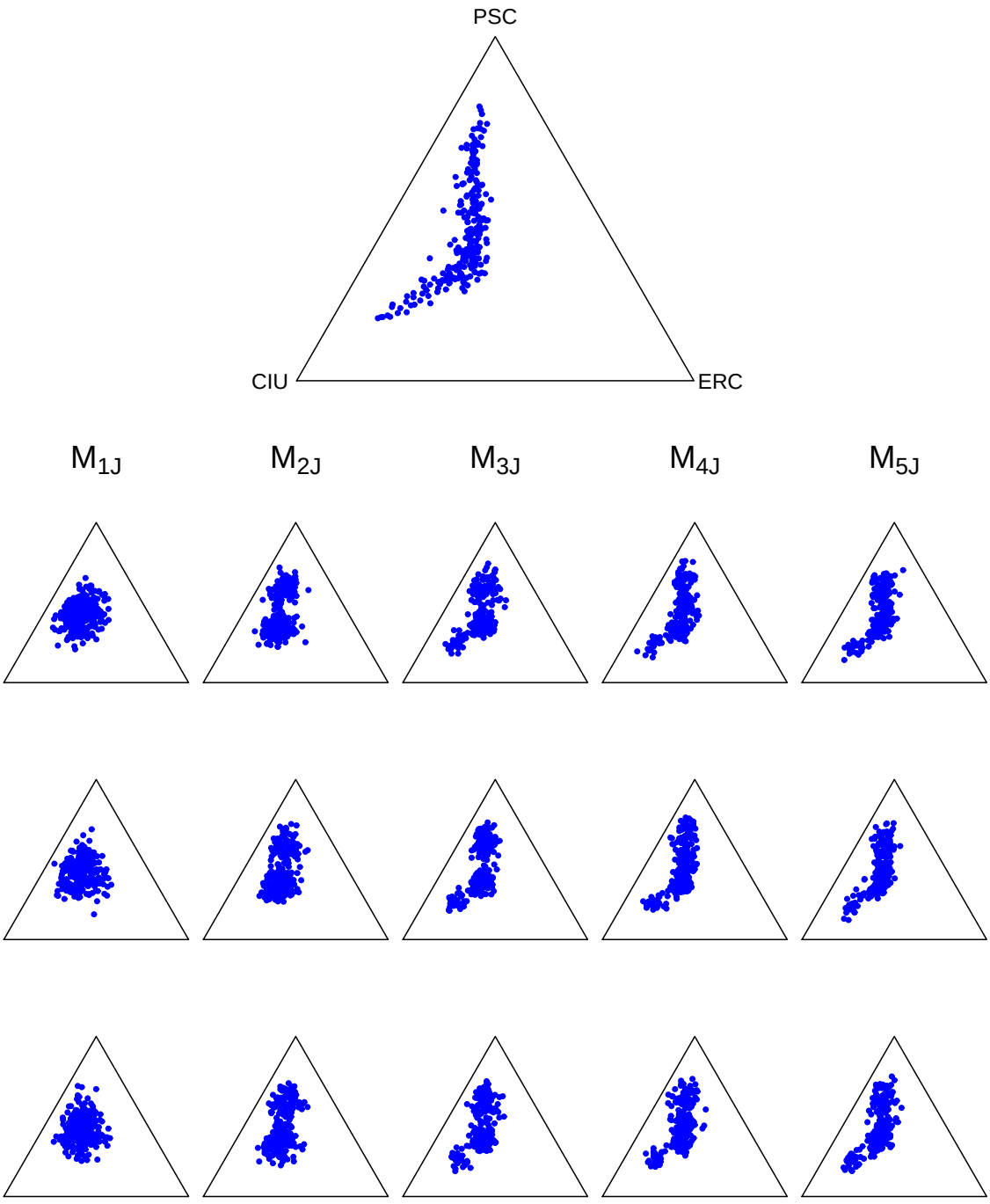

Figura 9.27: Validació gràfica dels diferents models a partir de representacions ternàries dels perfils de tres columnes de la Taula 6.2. El primer gràfic correspon als perfils observats a les eleccions al Parlament de Catalunya del 2003 a Barcelona ciutat i les altres files de corresponen a rèpliques de les dades obtingudes simulant de la predictiva a posteriori de cada un dels cinc models jeràrquics considerats.

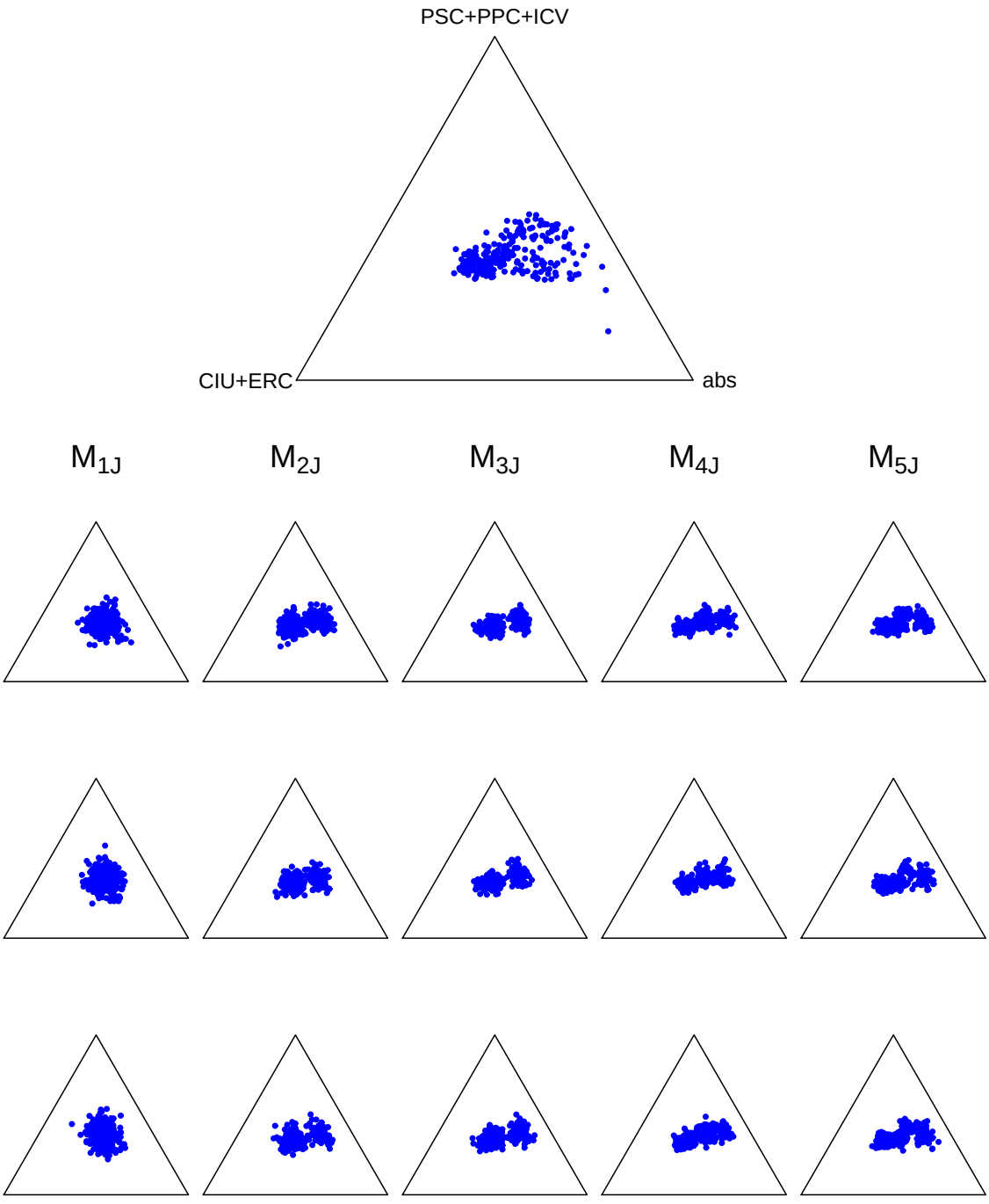

Figura 9.28: Validació gràfica dels diferents models a partir de representacions ternàries dels perfils resultants d'agregar columnes de la Taula 6.2. El primer gràfic correspon als perfils observats a les eleccions al Parlament de Catalunya del 2003 a Barcelona ciutat i les altres files corresponen a rèpliques de les dades obtingudes simulant de la predictiva a posteriori de cada un dels cinc models jeràrquics considerats.

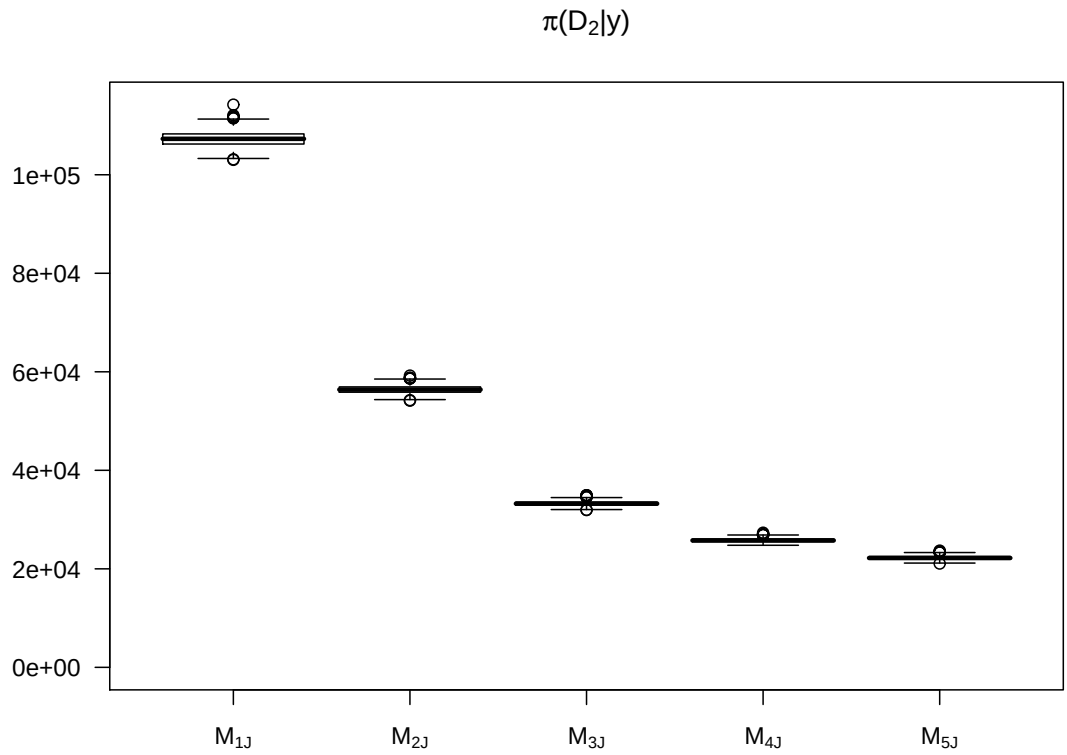

Figura 9.29: Distribucions a posteriori de la mesura de discrepància D_2 per als cinc models jeràrquics considerats.

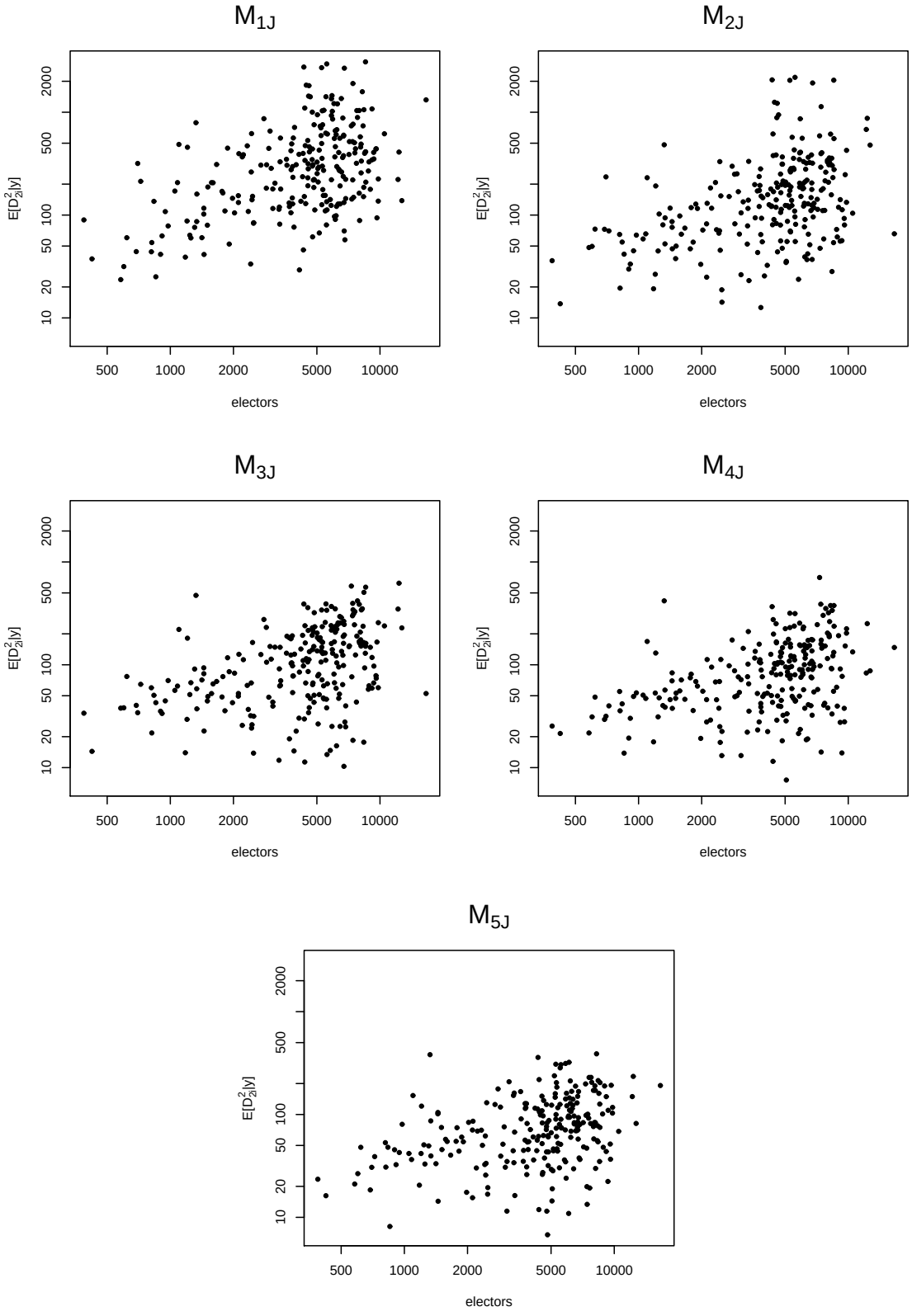

Figura 9.30: Esperança a posteriori de D_{2i}^2, $E[D_{2i}^2|y]$, per a cada zrp en funció del seu nombre d'electors per cada model jeràrquic considerat. Els eixos estan en escala logarítmica.

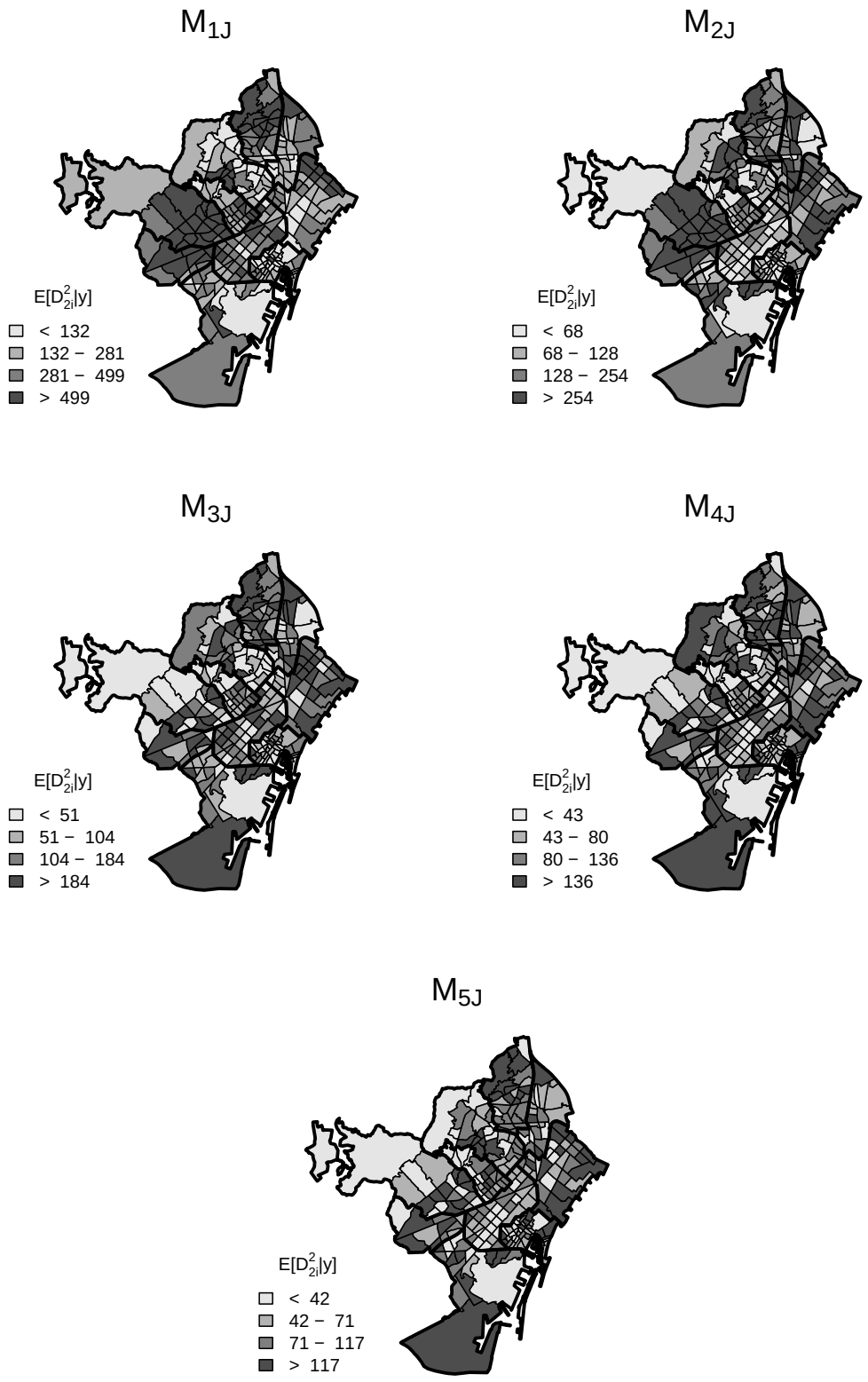

Figura 9.31: Distribució espacial de l'esperança a posteriori de D_{2i}^2 per a cada un dels cinc models jeràrquics considerats per els resultats a les eleccions al Parlament de Catalunya del 2003 a Barcelona ciutat.

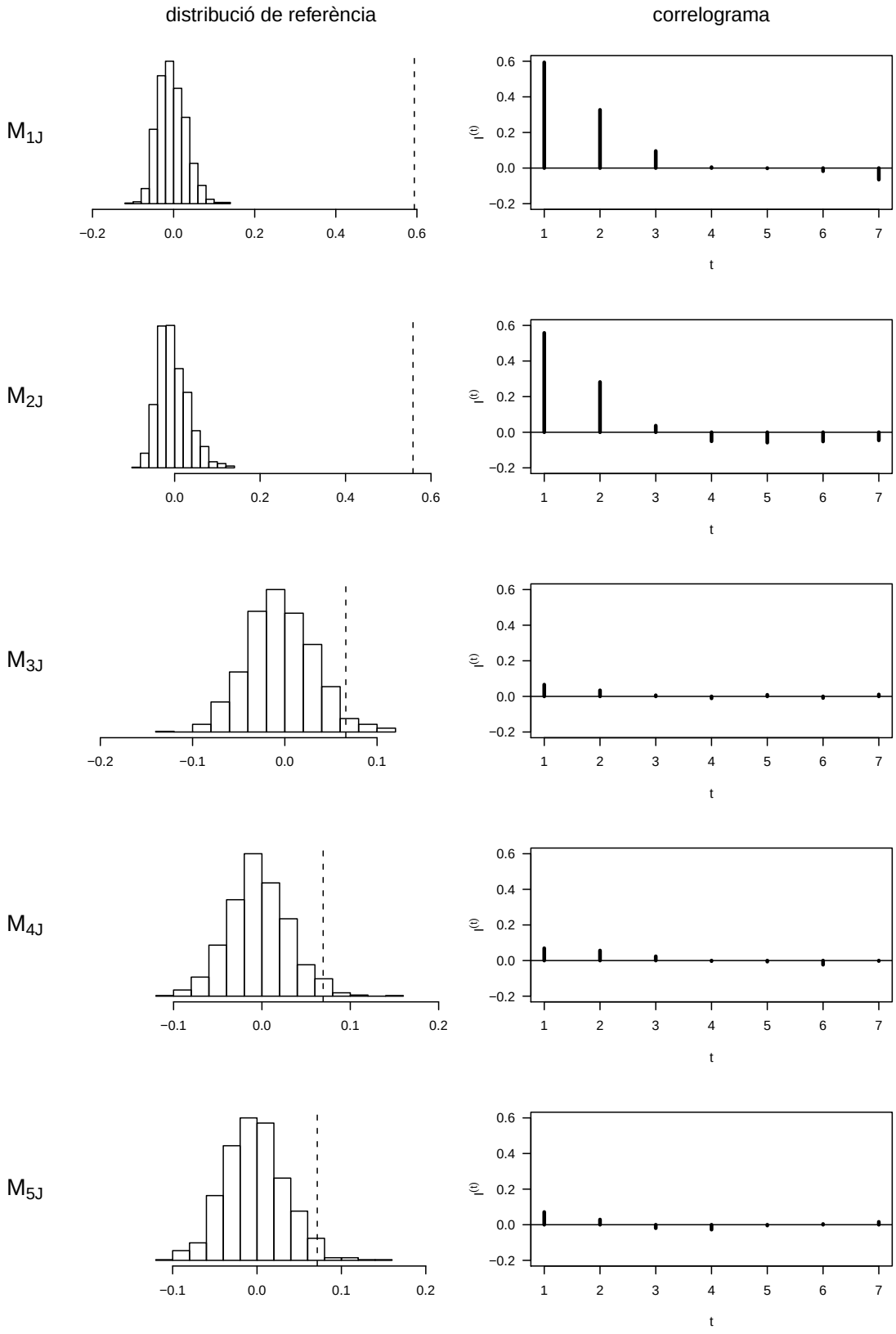

Figura 9.32: Índex de Moran, calculat a partir de $E[D_{2i}^2|y]$, la seva distribució de referència sota el test de permutacions, i els correlogrames per als cinc models jeràrquics considerats.

Capítol 10

Interpretació dels models per al 2003 a Barcelona

En aquest capítol mostrarem els resultats de tots els models presentats al Capítol 9, comparant els resultats de la inferència del cas jeràrquic i no jeràrquic, i comentarem amb més detall els resultats pels model de tres i quatre clusters.

Els paràmetres d'interès tant pel model cluster no jeràrquic com pel jeràrquic, que són els Model 5 i 6 del Captíol 7, seran:

a) la probabilitat de que una observació triada a l'atzar pertanyi al cluster r, ω_r per $r = 1, \ldots, s$,

b) el vector de variables latents ζ que recull a quin cluster pertany cada zrp, i

c) el perfil de probabilitat de cada cluster que per al cas no jeràrquic correspon a θ_r i per al cas jeràrquic a μ_r, per $r = 1, \ldots, s$,

a més a més per al cas jeràrquic també interessa:

d) el paràmetre que regula el grau d'heterogeneïtat dels clusters, τ_r per $r = 1, \ldots, s$. Com més petit sigui τ_r més heterogeneïtat hi haurà entre les zrp del cluster r-èssim i com més gran sigui més homogeneïtat hi haurà.

Com ja s'ha dit al Capítol 7, en el cas jeràrquic no analitzarem els paràmetres θ del primer nivell perquè en el nostre cas no són d'interès.

A la Taula 10.1 es presenten els valors esperats a posteriori de tots els paràmetres d'interès excepte de ζ per a cadascun dels models jeràrquics i no jeràrquics. Si comparem el valor esperat a posteriori de cada component dels perfils de probabilitat per un cluster r, que pel cas no jeràrquic correspon a $E[\theta_{rj}|y]$ i per al cas jeràrquic a $E[\mu_{rj}|y]$, per $j = 1, \ldots, 8$ i $r = 1, \ldots, s$, aquests valors són molt similars, de manera que si resumíssim els perfils del cluster amb el seu valor esperat les conclusions serien similars sota els casos no jeràrquic i jeràrquic. No obstant els models jeràrquics, que hem vist que reprodueixen molt més bé les característiques de les dades, tenen distribucions a posteriori i predictives a posteriori amb una major dispersió fruit de les característiques del model, tal i com es posarà de manifest en les Figures d'aquest capítol.

La distribució a posteriori de la probabilitat de pertànyer a cada un dels clusters, $\pi(\omega_r|y)$ per $r = 1, \ldots, s$, les trobem representades mitjançant diagrames de caixa a la Figura 10.1 pel cas no jeràrquic i a la Figura 10.2 pel cas jeràrquic. I a la Taula 10.1 hi trobem els valors esperats a posteriori. Tant els valors esperats com la variabilitat de les respectives distribucions a posteriori sòn molt similars pel cas jeràrquic i no jeràrquic.

La Figura 10.3 presenta la distribució a posteriori de τ_r, $\pi(\tau_r|y)$ per $r = 1, \ldots, s$, per a cadascun dels 5 models jeràrquics. A l'augmentar el nombre de clusters els τ_r tendeixen a tenir valors més alts. Aquest fenomen era previsible ja que τ_r modela l'heterogeneïtat dels perfils de les zrp que pertanyen al cluster r, de forma que a valors més petits més heterogeneïtat. Així a mesura que augmentem el nombre de clusters fa que aquests siguin més homogenis i per tant que els valors esperats per τ_r siguin més alts.

La classificació de cada zrp a cada cluster s'ha fet mitjançant la moda a posteriori de les variables latents ζ_i, $Mo(\zeta_i|y)$. Els resultats es mostren a la Figura 10.4 per a cada model de dos i més clusters, jeràrquics i no jeràrquics. Els resultats són molt semblants per al cas jeràrquic i no jeràrquic. Així per al model de tres clusters la majoria de zrp dels districtes de Sarrià-Sant Gervasi i Les Corts han estat assignats al cluster 1, on destaca CIU, PPC i la baixa abstenció, les zrp que voregen aquest cluster 1, bàsicament zrp dels districtes de Sants-Montjuïc, l'Eixample i Gràcia, i també moltes zrp de Ciutat Vella i Sant Martí pertanyen al cluster 2, caracteritzat per un empat de CIU i PSC i on hi destaca també ERC, i finalment les zrp més perifèriques i en especial Nou Barris pertanyen al cluster 3 on hi destaca el PSC i l'alta abstenció. De la representació dels mapes també destaca la notable agregació espacial que presenten els clusters.

Del model de tres clusters per al cas no jeràrquic, M_3, la majoria de zrp concentren la massa de $\pi(\zeta_i|y)$ en un sol cluster, en el sentit que $\pi(\zeta_i = r|y) = 1$ per un valor de r, i només trobem 5 zrp, del total de 248, que reparteixen la massa entre el segon i tercer clusters. Pel cas de tres clusters jeràrquic, M_{3J}, hi trobem 18 zrp que concentren la

massa en dos clusters, però ho fan majoritàriament entre primer i segon cluster.

A la Figura 10.5 presentem la distribució a posteriori de θ_{rj} del model no jeràrquic M_3 per $j = 1, \ldots, 8$, mitjançant diagrames de caixes, i a la Figura 10.6 hi presentem la distribució a posteriori de $\log(\theta_{rj}/\theta_{r'j})$ per $r > r'$. Aquesta darrera Figura ens permet esbrinar quines components del perfil diferencien un cluster d'un altre. Així el cluster 2 es diferencia del 1 per tenir més presència el PSC, ICV i ERC en detriment de CIU i del PPC, mentre que el cluster 3 es diferencia del 1 per tenir més presència el PSC i l'abstenció, en detriment de CIU i del PPC, i el cluster 3 es diferencia del 2 per tenir menys presència CIU i ERC.

Les Figures 10.7 i 10.8 presenten les distribucions a posteriori d'aquests perfils per el model jeràrquic M_{3J}, és a dir la distribució a posteriori de μ_{rj} per $j = 1, \ldots, 8$ i la distribució a posteriori de $\log(\mu_{rj}/\mu_{r'j})$ per $r > r'$. Els gràfics són molt similars als observats per al cas no jeràrquic; la diferència principal rau en que les caixes de les distribucions a posteriori són més amples ja que tenen en compte l'heterogeneïtat.

Els mateixos resultats per als dos models de 4 clusters els podem troabar a les Figures 10.9 i 10.10 per al cas no jeràrquic i a les Figures 10.11 i 10.12 per al cas jeràrquic. Es torna a observar les semblances entre el cas jeràrquic i no jeràrquic. Observem com el primer cluster del model M_{4J} és pràcticament idèntic al primer cluster del model M_{3J} on destaquen CIU i PPC, i que els altres dos clusters del model M_{3J} es reconverteixen en 3 clusters del model M_{4J}. Al passar del segon al tercer cluster i del tercer al quart augmenten progressivament els perfils de probabilitat del PSC i de l'abstenció, mentre que els perfils de probabilitat de CIU i ERC decreixen, i el del PPC es manté constant.

Els valors esperats a posteriori dels perfils dels clusters, presentats a la Taula 10.1, i la distribució espacial dels clusters, presentada a la Figura 10.4, són similars sota els models no jeràrquics i jeràrquics. No obstant els models jeràrquics tenen la peculiaritat d'una major dispersió de les distribucions a posteriori dels perfils fruit de les característiques del model i les dades. Si be hom podria pensar que les conclusions podrien ser similars si s'utilitza el model no jeràrquic o el jeràrquic cal tenir present que de la validació efectuada en el capítol 8 es desprén que no s'hagués escollit el mateix model, ja que partint de la validació dels models no jeràrquics, feta a la secció 9.1 s'arriba a models amb molts més clusters que a través dels models jeràrquics.

Model	cluster	CIU	PSC	PPC	ICV	ERC	altres	b+n	abs	$E[\omega\|y]$	$E[\tau\|y]$
M_1	1	0.186	0.204	0.101	0.057	0.104	0.008	0.008	0.333	-	-
M_{1J}	1	0.173	0.198	0.098	0.061	0.098	0.014	0.013	0.344	-	73.99
M_2	1	0.225	0.175	0.106	0.058	0.123	0.007	0.008	0.298	0.592	-
	2	0.116	0.257	0.092	0.054	0.069	0.010	0.007	0.396	0.408	-
M_{2J}	1	0.219	0.174	0.103	0.061	0.121	0.010	0.011	0.302	0.624	146.09
	2	0.111	0.242	0.083	0.054	0.064	0.012	0.009	0.424	0.376	152.76
M_3	1	0.302	0.122	0.176	0.038	0.091	0.007	0.009	0.254	0.123	-
	2	0.201	0.192	0.090	0.063	0.127	0.008	0.008	0.312	0.532	-
	3	0.108	0.263	0.091	0.051	0.060	0.010	0.007	0.410	0.345	-
M_{3J}	1	0.301	0.12	0.179	0.037	0.087	0.008	0.011	0.257	0.118	286.55
	2	0.202	0.187	0.089	0.066	0.127	0.009	0.009	0.311	0.522	324.22
	3	0.110	0.243	0.083	0.053	0.063	0.013	0.009	0.427	0.360	153.48
M_4	1	0.306	0.118	0.187	0.035	0.086	0.006	0.01	0.253	0.111	-
	2	0.223	0.177	0.093	0.062	0.133	0.007	0.008	0.297	0.36	-
	3	0.154	0.224	0.090	0.063	0.104	0.009	0.008	0.348	0.287	-
	4	0.099	0.272	0.089	0.048	0.052	0.010	0.007	0.423	0.241	-
M_{4J}	1	0.302	0.119	0.182	0.037	0.086	0.008	0.011	0.257	0.112	302.20
	2	0.221	0.176	0.090	0.064	0.134	0.008	0.009	0.298	0.364	503.68
	3	0.151	0.221	0.086	0.066	0.099	0.010	0.009	0.358	0.291	302.61
	4	0.098	0.25	0.081	0.047	0.052	0.012	0.008	0.451	0.233	171.40
M_5	1	0.305	0.118	0.187	0.035	0.086	0.006	0.010	0.253	0.112	-
	2	0.232	0.172	0.098	0.061	0.131	0.007	0.008	0.29	0.267	-
	3	0.172	0.204	0.078	0.065	0.122	0.008	0.007	0.343	0.289	-
	4	0.133	0.257	0.111	0.058	0.078	0.009	0.008	0.346	0.130	-
	5	0.093	0.272	0.082	0.046	0.049	0.010	0.006	0.442	0.202	-
M_{5J}	1	0.303	0.118	0.183	0.036	0.085	0.007	0.011	0.257	0.111	340.87
	2	0.232	0.171	0.097	0.062	0.132	0.008	0.009	0.289	0.265	686.88
	3	0.171	0.200	0.076	0.069	0.119	0.009	0.008	0.347	0.287	378.22
	4	0.131	0.260	0.110	0.059	0.078	0.010	0.008	0.344	0.119	677.23
	5	0.097	0.243	0.077	0.047	0.052	0.012	0.008	0.463	0.218	182.47

Taula 10.1: Esperança a posteriori dels perfils de probabilitat per als clusters de cada model, que en el cas dels models no jeràrquics correspon a $E[\theta_j|y]$ mentre que en el cas del model jeràrquic correspon a $E[\mu_j|y]$ per $j = 1, \ldots, k$. Per a cada model també es dona $E[\omega|y]$, que indica el tamany relatiu de cada cluster, i pel cas jeràrquic també es dona $E[\tau|y]$ que indica el grau d'heterogeneitat del cluster.

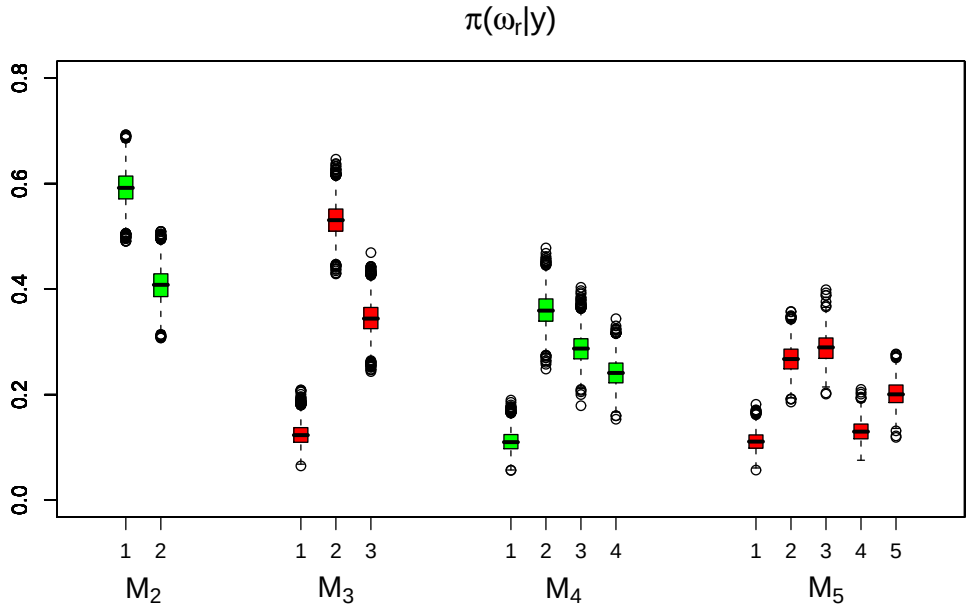

Figura 10.1: Distribució a posteriori de la probabilitat de pertànyer a cada un dels clusters, $\pi(\omega_r|y)$ per $r = 1, \ldots, s$, per als models cluster no jeràrquics considerats.

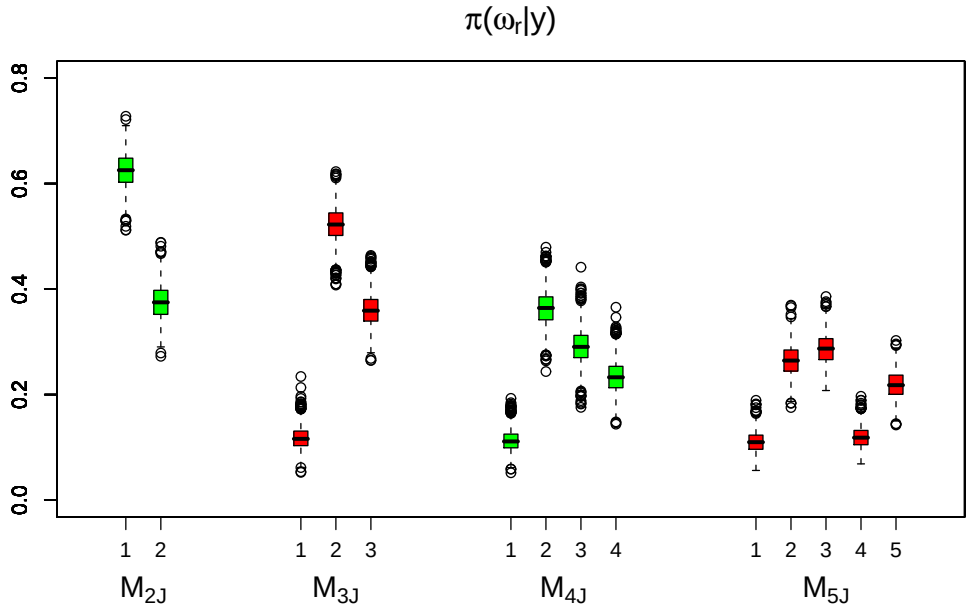

Figura 10.2: Distribució a posteriori de la probabilitat de pertànyer a cada un dels clusters, $\pi(\omega_r|y)$ per $r = 1, \ldots, s$, per als models cluster jeràrquics considerats.

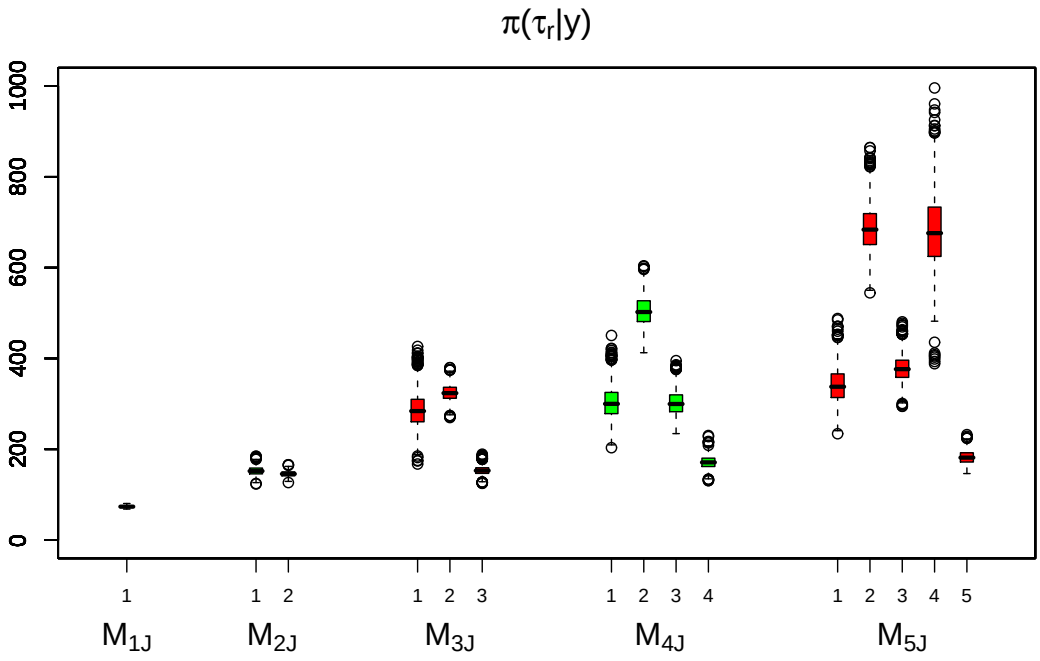

Figura 10.3: Distribució a posteriori de τ, $\pi(\tau_r|y)$ per $r = 1, \ldots, s$, per als cinc models jeràrquics considerats.

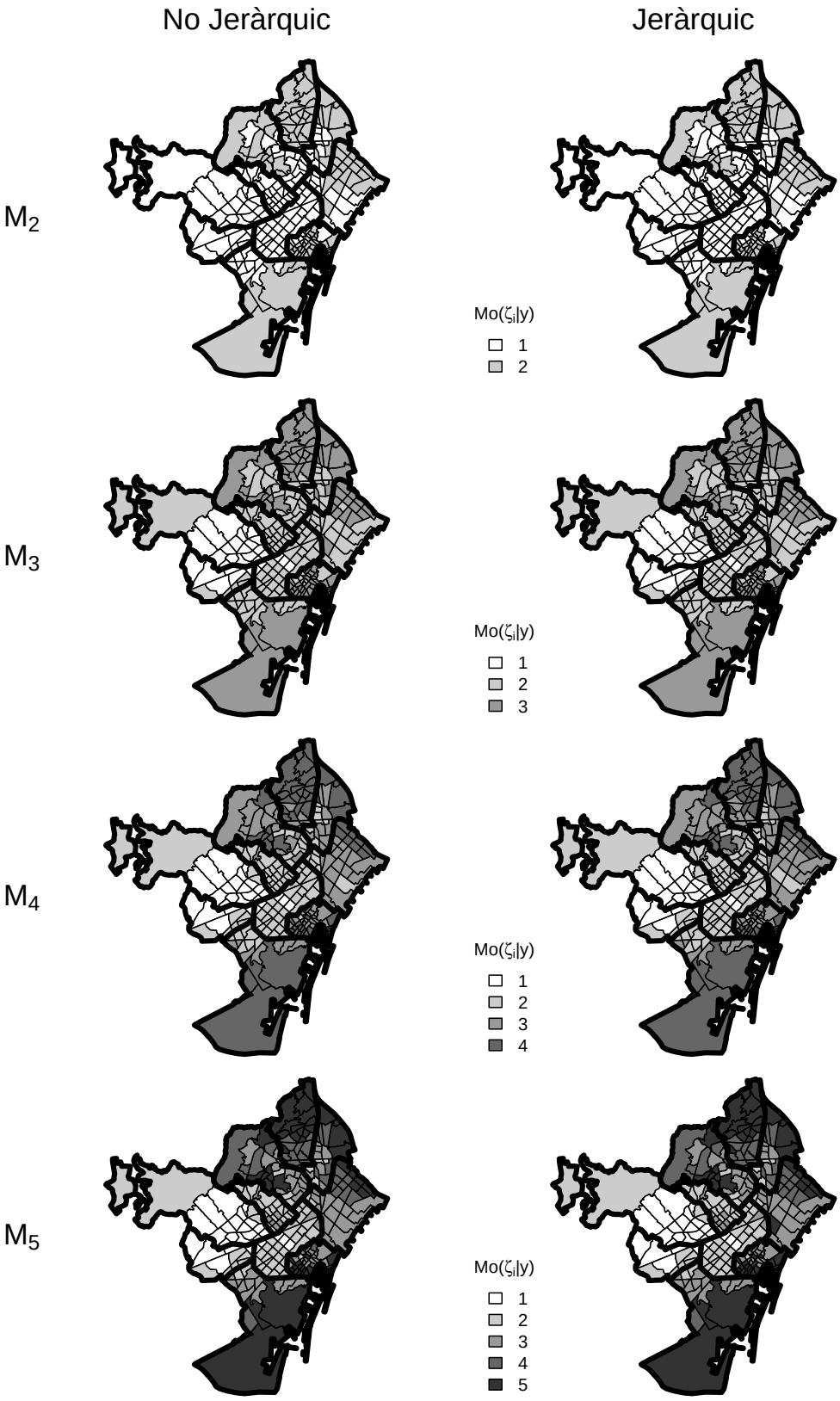

Figura 10.4: Classificació de les zrp en cada un dels clusters utilitzant com a criteri la moda a posteriori de les variables latents ζ_i, $Mo(\zeta_i|y)$, sota tots els models cluster considerats.

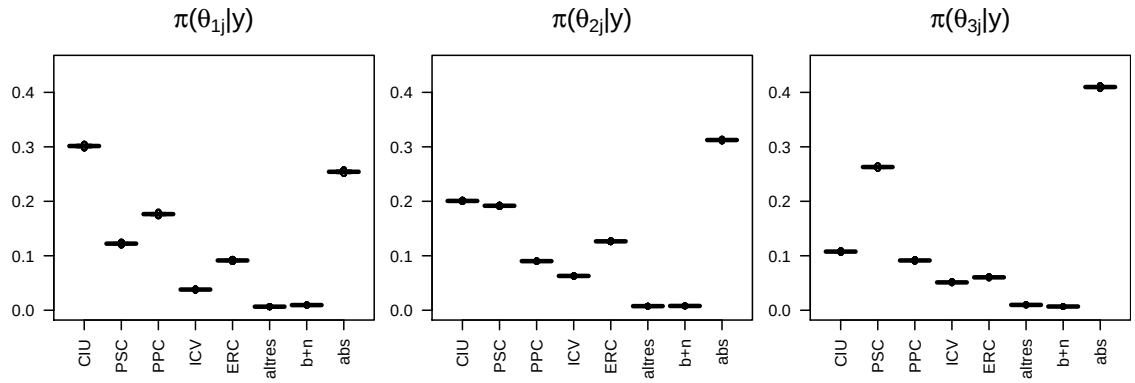

Figura 10.5: Distribució a posteriori de les components del perfil de cada cluster per al model M_3, $\pi(\theta_{rj}|y)$ per $j = 1, \ldots, 8$. Els tres clusters estan ordenats per ordre decreixent de percentatge de vot a CIU i creixent al PSC.

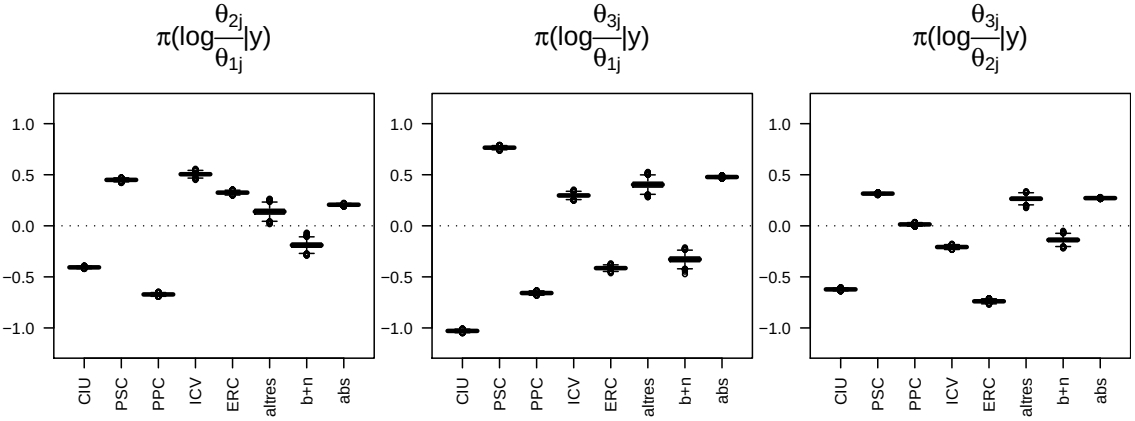

Figura 10.6: Distribució a posteriori de $\log(\theta_{rj}/\theta_{r'j})$ per al model M_3 per $r > r'$ i $j = 1, \ldots, 8$.

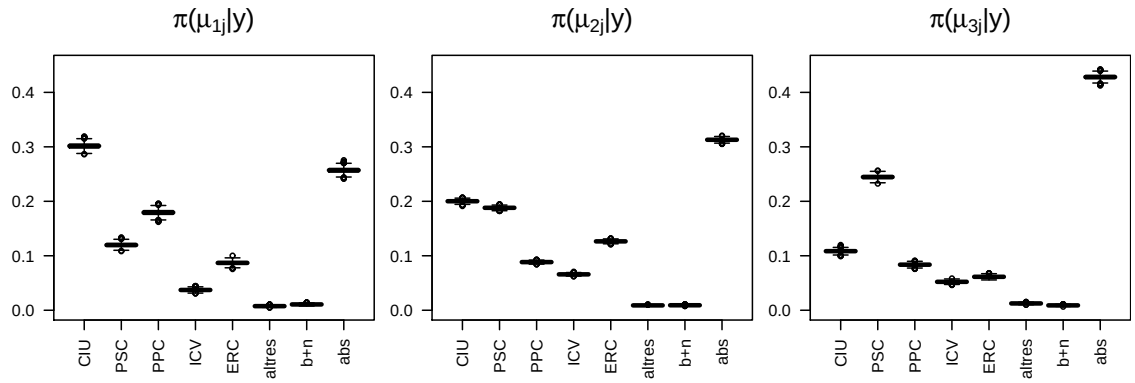

Figura 10.7: Distribució a posteriori de les components del perfil de cada cluster per al model M_{3J}, $\pi(\mu_{rj}|y)$ per $j = 1, \ldots, 8$. Els tres clusters estan ordenats per ordre decreixent de percentatge de vot a CIU i creixent al PSC.

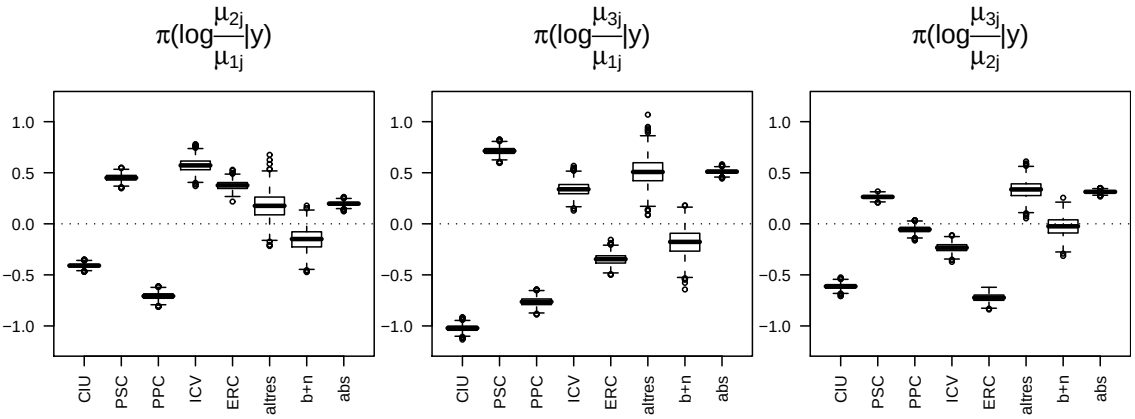

Figura 10.8: Distribució a posteriori de $\log(\mu_{rj}/\mu_{r'j})$ per al model M_{3J} per $r > r'$ i $j = 1, \ldots, 8$.

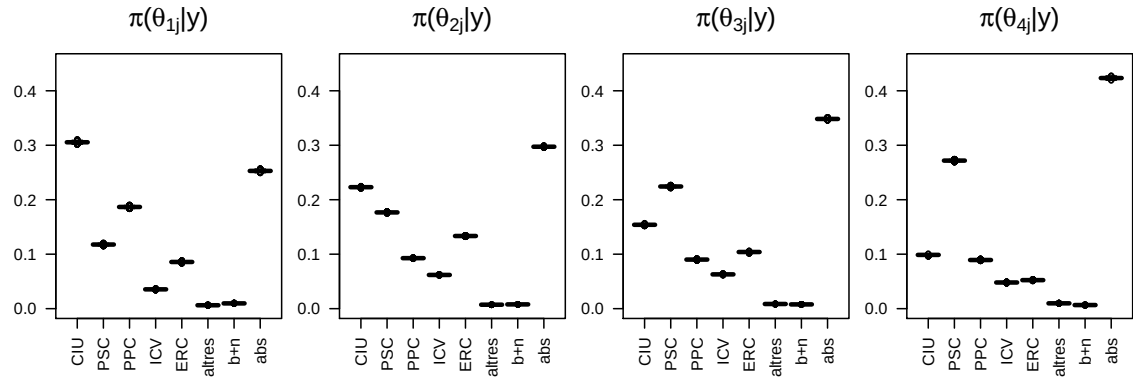

Figura 10.9: Distribució a posteriori de les components del perfil de cada cluster per al model M_4, $\pi(\theta_{rj}|y)$ per $j = 1, \ldots, 8$. Els quatre clusters estan ordenats per ordre decreixent de percentatge de vot a CIU i creixent al PSC.

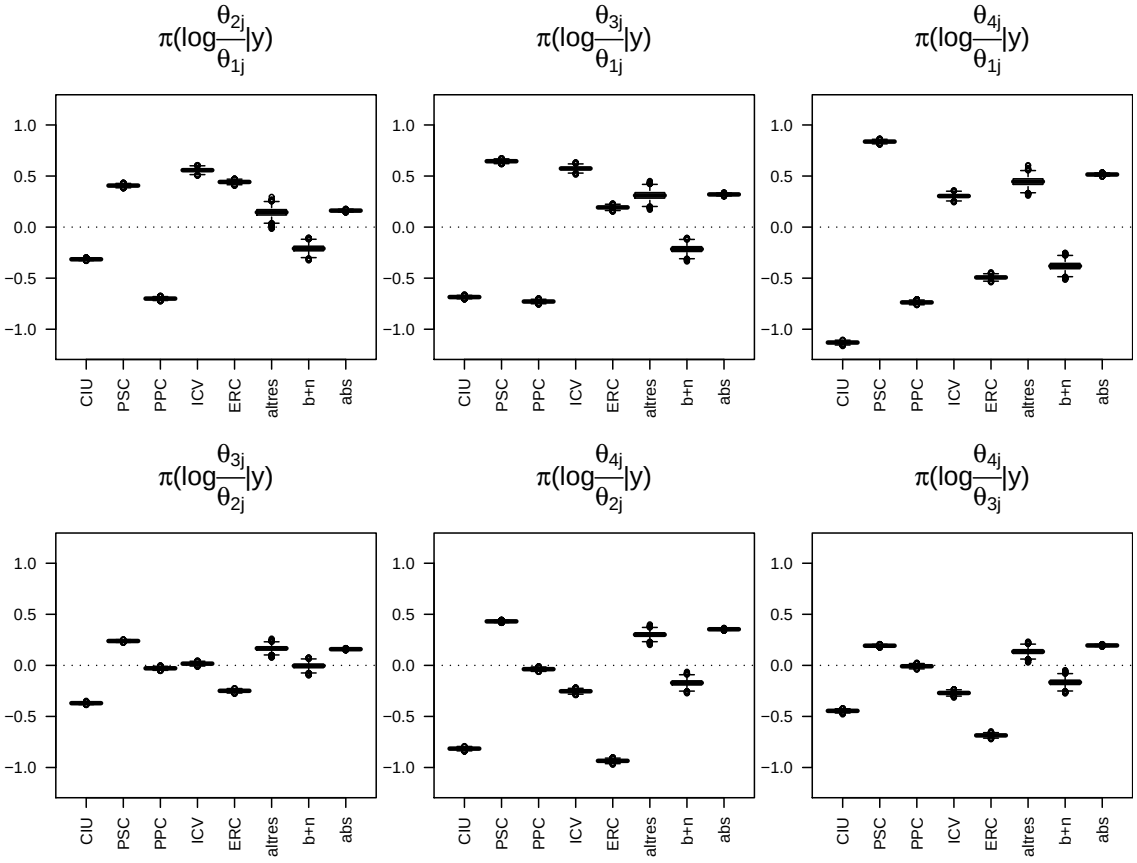

Figura 10.10: Distribució a posteriori de $\log(\theta_{rj}/\theta_{r'j})$ per al model M_4 per $r > r'$ i $j = 1, \ldots, 8$.

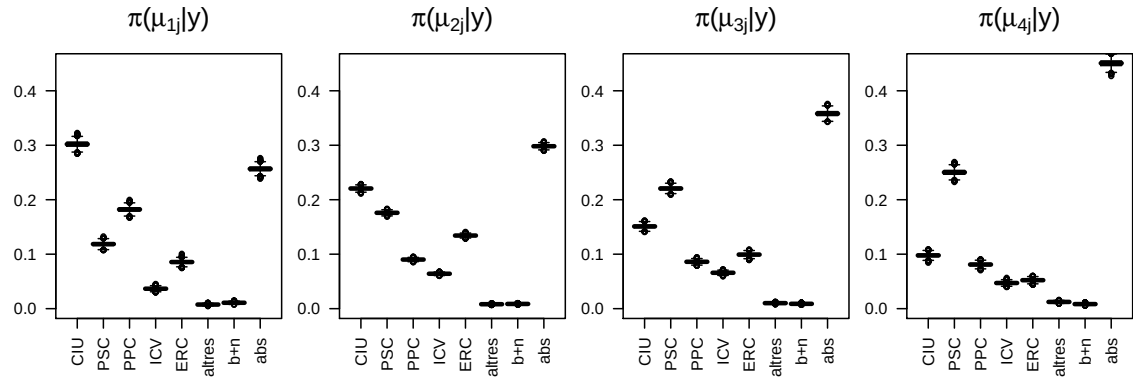

Figura 10.11: Distribució a posteriori de les components del perfil de cada cluster per al model M_{4j}, $\pi(\mu_{rj}|y)$ per $j = 1, \ldots, 8$. Els quatre clusters estan ordenats per ordre decreixent de percentatge de vot a CIU i creixent al PSC.

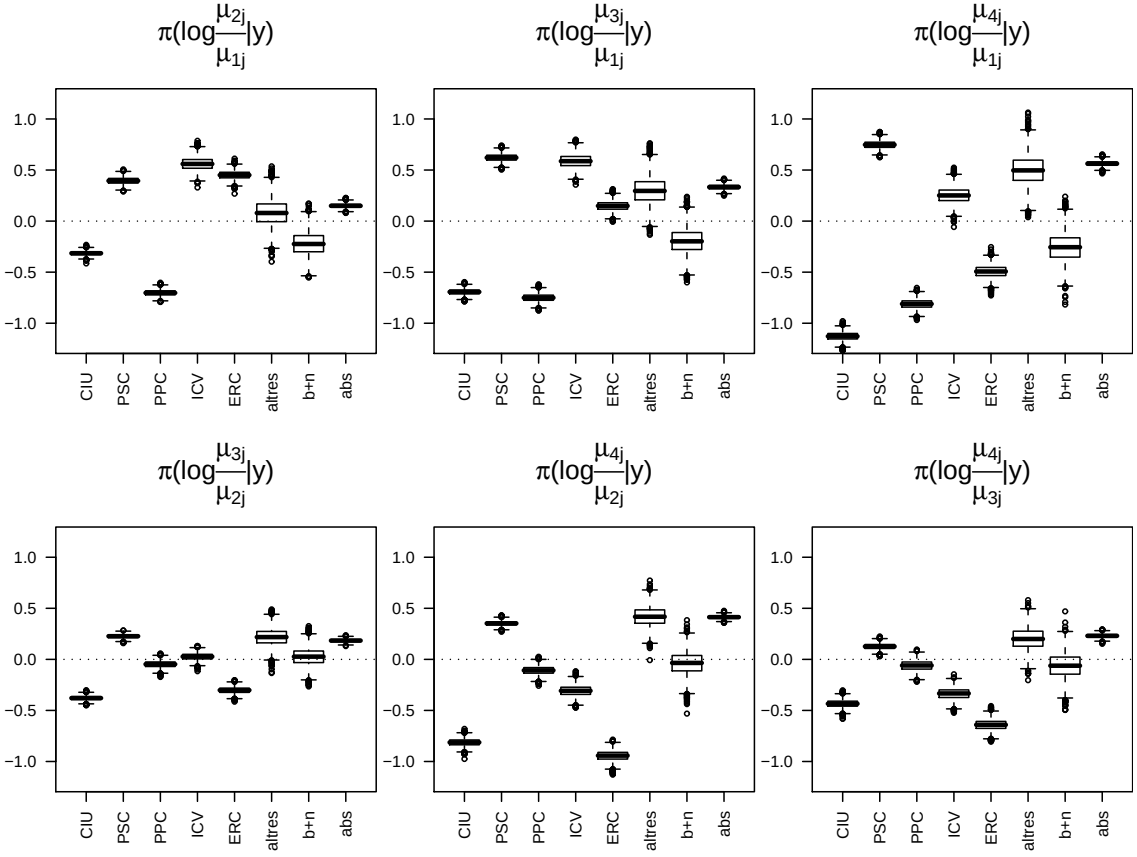

Figura 10.12: Distribució a posteriori de $\log(\mu_{rj}/\mu_{r'j})$ per al model M_{4j} per $r > r'$ i $j = 1, \ldots, 8$.

Capítol 11

Comparació i validació dels models jeràrquics per al 1992-2006 a Barcelona

Als capítols 9 i 10 hem descartat els models no jeràrquics perquè no capturen bé el grau de sobredispersió de les dades. En aquest capítol comparem els models cluster jeràrquics per a les cinc darreres eleccions al Parlament de Catalunya a Barcelona, a través d'una selecció dels gràfics més rellevants, amb l'objectiu d'escollir els models que al Capítol 12 utilitzarem per descriure l'evolució dels patrons al llarg dels diferents comicis.

A la Figura 11.1 s'analitza l'estadístic

$$D_{ai}(y_i) = log(\tfrac{y_{i,CIU}}{y_{i,PSC}}),$$

per $i = 1, \ldots, 248$ en funció del districte, la distribució de les zrp dintre d'un districte és presenta mitjançant un diagrama de caixa. Per a cadascun dels models M_{2J}, M_{3J} i M_{4J} i per a cadascuna de les cinc eleccions s'ha construït un gràfic semblant al de les dades a partir d'una rèplica de la distribució predictiva a posteriori. La primera columna d'aquest gràfic ens ofereix una anàlisi descriptiva, que ens indica que a les eleccions dels anys 1992 i 1995 CIU obtenia més vots que el PSC a tots els districtes excepte al districte de Nou Barris, mentre que als tres darrers comicis CIU perd l'hegemonia als districtes de Ciutat Vella, Sants-Montjuïc, Horta-Guinardó, Sant Andreu i Sant Martí. En aquest gràfic s'observa que el model M_{2J} està lluny de capturar el comportament de les dades mentre que el model de tres clusters, M_{3J}, ja captura bastant bé els matisos a nivell de

districte, fins al punt que si ordenem els districtes en funció de les medianes de l'estadístic calculat a partir de la rèplica, llavors aquest ordre coincideix pràcticament per a tots els anys amb l'ordre resultant de fer el mateix per a les dades observades.

Les Figures 11.2, 11.3 i 11.4 examinen la qualitat del model a través de

$$D_{bi}(y_i) = log(\frac{y_{CIU+PPC}}{y_{PSC+ERC+ICV}}),$$

$$D_{ci}(y_i) = log(\frac{y_{i,CIU+ERC}}{y_{i,PSC+PPC+ICV}}),$$

i

$$D_{di}(y_i) = log(\frac{y_{i,abs}}{N_i}).$$

Per a cadascun dels estadístics i per a tots els anys els gràfics de validació presenten conclusions similars en el sentit que el model de dos clusters no captura adequadament el comportament de les dades mentre que el model de tres clusters ja captura bastant bé l'estructura de les dades, tot i que no tant bé com ho fa el de quatre clusters. Cal decidir si considerem que el model de tres clusters ja captura les principals característiques de les dades o bé és necessari considerar el model de quatre clusters, o en altres paraules cal decidir si el procés generador de la rèplica del model M_{3J} és versemblant suposar que és el mateix procés generador que ha generat les dades.

La primera columna d'aquestes figures també ens ofereixen una anàlisi descriptiva. La Figura 11.2 s'observa que a les dues primeres eleccions la centre-dreta dominava l'esquerra en tots els districtes excepte Nou Barris, i que actualment només ho fa de forma clara als districtes de Les Corts i de Sarrià-Sant Gervasi. La Figura 11.3 mostra com els dos partits nacionalistes catalans a les darreres eleccions de l'any 2006 dominaven en els districtes de l'Eixample, Les Corts, Sarrià-SantGervasi i Gràcia, mentre que als primers anys també ho feien a Sants-Montjüic. I la Figura 11.4 no presenta grans canvis en el patró de l'abstenció al llarg dels diferents comicis.

Aprofundint més en el comportament dels estadístics $D_{ai}(y_i)$, $D_{bi}(y_i)$, $D_{ci}(y_i)$ i $D_{di}(y_i)$ hem estudiat la seva distribució per districte tot graficant mesures de tendència central com la seva mediana i mitjana, mesures de la variabilitat com el seu rang interquartílic i la seva variància i mesures de la posició com el primer i tercer quartil, així com el seu mínim i el seu màxim. Hem comparat la distribució predictiva a posteriori d'aquestes mesures per a cada districte i per cada model jeràrquic considerat, amb el respectiu valor observat.

A la Figura 11.5 hem representat, per als models M_{2J}, M_{3J} i M_{4J}, la distribució predictiva a posteriori de la mediana de $D_{ai}(y_i)$ per cada districte i cada any mitjançant un diagrama de caixa i el valor observat mitjançant un punt. L'objectiu és estudiar la compatibilitat dels valors observats amb el model, quan no siguin compatibles, estudiem en que es desvia el valor observat respecte a les prediccions del model. En el model de dos clusters el valor observat dels districtes de Ciutat Vella, Sarrià-Sant Gervasi i Nou Barris s'allunya sistemàticament en tots els anys dels valors simulats a partir del model. En el model de tres clusters trobem molta coincidència entre els valors observats i els predits pel model tot i que en la majoria de casos trobem els valors observats a les cues de manera que si calculéssim els *p valors* Bayesians associats obtindríem valors propers a 0 i a 1. En el model de quatre clusters ja no trobem valors observats marcadament allunyats dels valors simulats a partir del model.

La columna de l'esquerra de les Figures 11.9 i 11.10 representen simultàniament tres columnes de la taula original mitjançant diagrames ternaris per a cada una de les cinc eleccions, mentre que a la resta de columnes s'hi presenten els mateixos gràfics ternaris construïts a partir d'una rèplica de la distribució predictiva a posteriori per a cada model considerat. Aquestes Figures tornen a posar de manifest la capacitat dels models de tres i quatre clusters de reproduir el patró general de les dades. Pel que fa a la interpretació descriptiva de l'evolució, la Figura 11.9 mostra com en termes relatius el PPC ha guanyat pes respecte CIU i PSC i de forma més important en les zrp a on CIU tenia més pes a les primeres eleccions. I de la Figura 11.10 se'n desprén que en termes relatius ERC ha guanyat pes respecte CIU i PSC i ho ha fet de forma més important en les zrp on CIU tenia una lleugera avantatge respecte el PSC.

Per avaluar l'existència d'una possible correlació espacial de D_{2i}^2 amb la finalitat de comporvar si les desviacions del model estan estructurades espacialment o no, hem calculat l'índex de Moran, I, a partir de $E[D_{2i}^2|y]$, i l'hem contrastat amb la distribució de referència obtinguda a partir de 5000 permutacions. Això ens permet observar si la dependència espacial de D_{2i}^2 varia al canviar d'un model a un altre i si ho fa de la mateixa manera per les cinc eleccions estudiades.

A la Figura 11.11 s'hi representa la distribució de referència i el corresponent valor observat per I per als models M_{2J}, M_{3J} i M_{4J} per a cadascuna de les cinc eleccions. En tots els anys els valors observats per I sota el model M_{2J} es troben extremadament allunyats de la distribució de referència. En les tres primeres eleccions els valors observats per I sota el model M_{3J} també es troben allunyats de la distribució de referència i per aquests mateixos anys el valor observat per I sota el model M_{4J} és compatible amb la distribució de referència tot i que sempre cau a la cua de la dreta. Per a les dues eleccions més recents la dependència espacial disminueix a partir del model M_{3J}. A la Taula 11.1

Any	M_{2J}	M_{3J}	M_{4J}
1992	0.000	0.000	0.004
1995	0.000	0.003	0.034
1999	0.000	0.000	0.006
2003	0.000	0.031	0.024
2006	0.000	0.082	0.107

Taula 11.1: *p valor* associats a l'índex de Moran, calculats a partir de $E[D^2_{2i}|y]$, per als models M_{2J}, M_{3J} i M_{4J} per a cadascuna de les cinc eleccions al Parlament.

es presenta com a mesura de conflicte el *p valor* per a tots els anys i models considerats.

Analitzant els correlogrames de la Figura 11.12 s'observa que per a tots els anys el model M_{2J} presenta la típica forma indicativa d'existència d'estructura espacial, en els que les correlacions de primer ordre presenten valors elevats per després anar disminuint fins a valors al voltant de zero a partir d'un cert retard. I sota el model M_{3J} les correlacions són moderades en les tres primeres eleccions i febles en les dues darreres.

El model de tres clusters captura bé la correlació espacial en les dues darreres eleccions mentre que no ha fa bé per a les tres primeres eleccions estudiades. El fet d'observar *p valors* petits suggereixen encara una certa existència de dependència espacial fins i tot pel model de quatre clusters. No obstant, abans de decidir abordar aquest problema caldria primer plantejar-se la possibilitat de fer servir models que contemplin la correlació espacial, ja que aquests models a la pràctica són eines de suavitzat que podrien entrar en conflicte amb l'estructura de clusters utilitzada.

En aquest punt tenim la possibilitat o bé d'escollir el que creiem com a millor model per a cada any o be escollir un mateix model de compromís per a tots els anys de forma que permeti descriure millor l'evolució dels clusters i perfils d'aquests al llarg de les diferents eleccions. Al Capítol 12 analitzarem els resultats per a tots els anys basant-nos tant en el model M_{3J} com en el model M_{4J}. Això ens permetrà comparar les conclusions fruit d'escollir el model de tres clusters amb les conclusions obtingudes a partir del de quatre clusters.

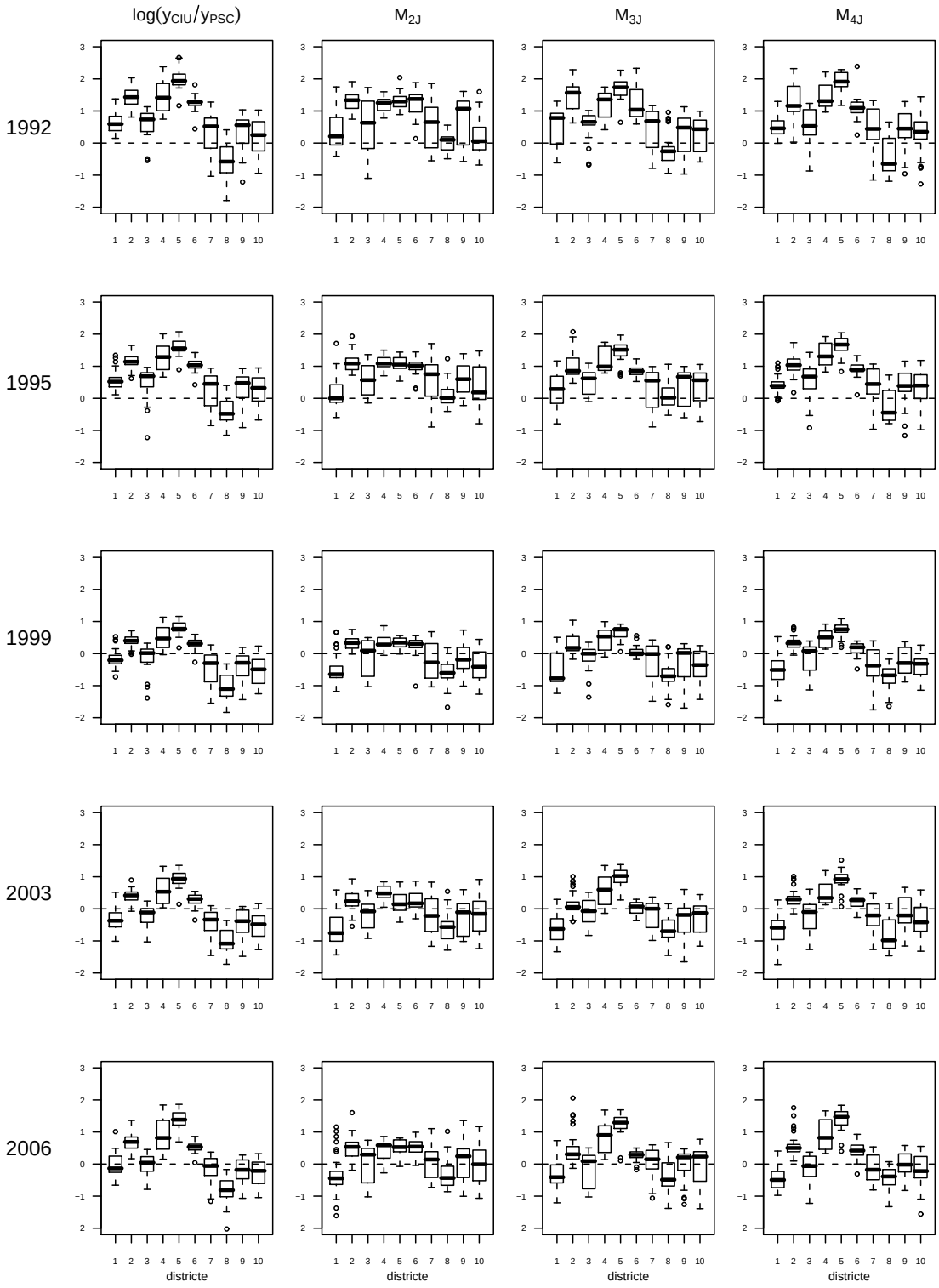

Figura 11.1: La columna esquerra presenta els valors observats per $D_{ai}(y_i)$ a les eleccions al Parlament de Catalunya del 1992 al 2006 a nivell de zrp als 10 districtes de Barcelona, i les altres columnes de gràfic presenten una rèplica de les dades obtingudes a partir de la predictiva a posteriori dels models M_{2J}, M_{3J} i M_{4J}.

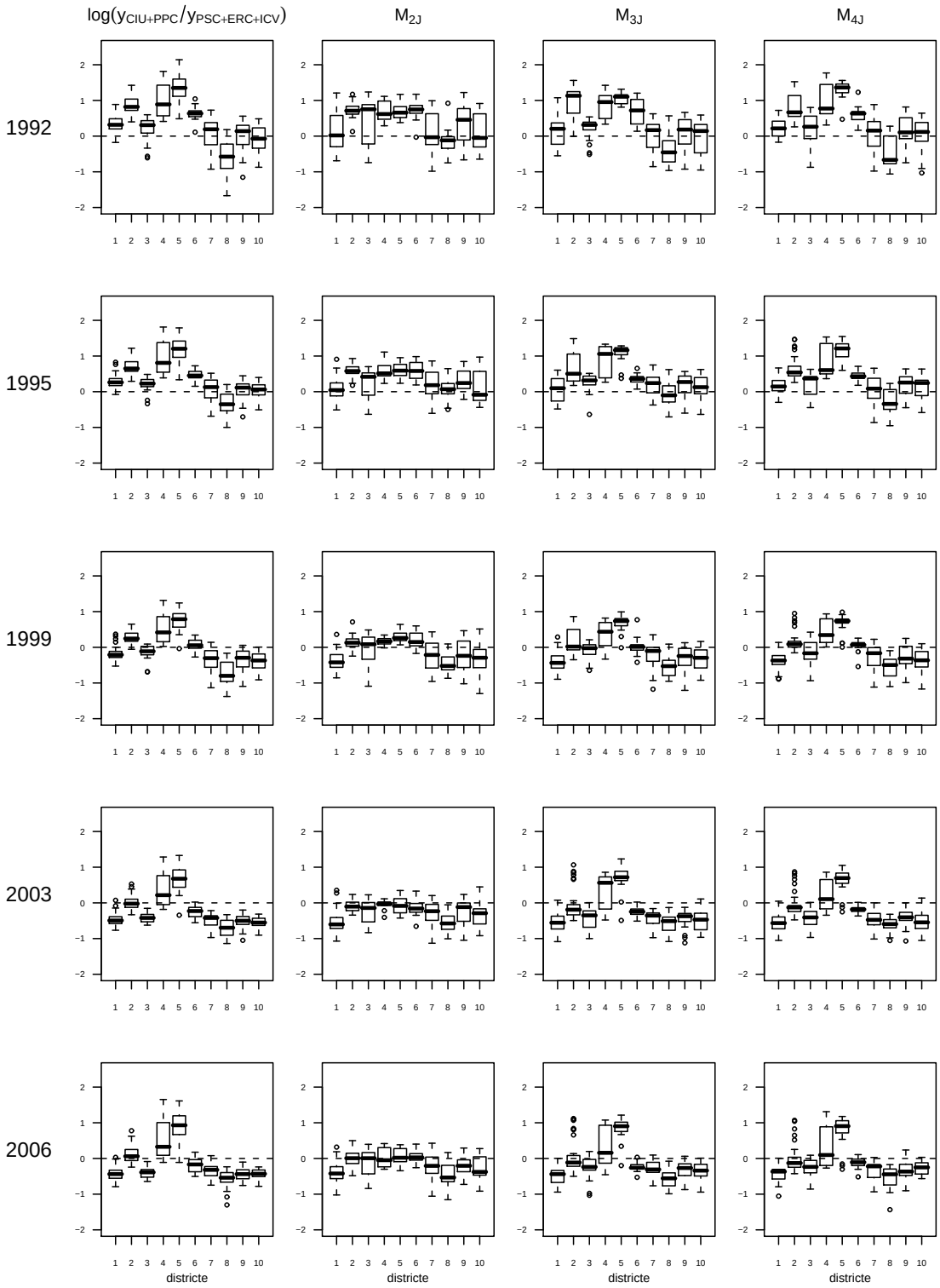

Figura 11.2: La columna esquerra presenta els valors observats per $D_{bi}(y_i)$ a les eleccions al Parlament de Catalunya del 1992 al 2006 a nivell de zrp als 10 districtes de Barcelona, i les altres columnes de gràfic presenten una rèplica de les dades obtingudes a partir de la predictiva a posteriori dels models M_{2J}, M_{3J} i M_{4J}.

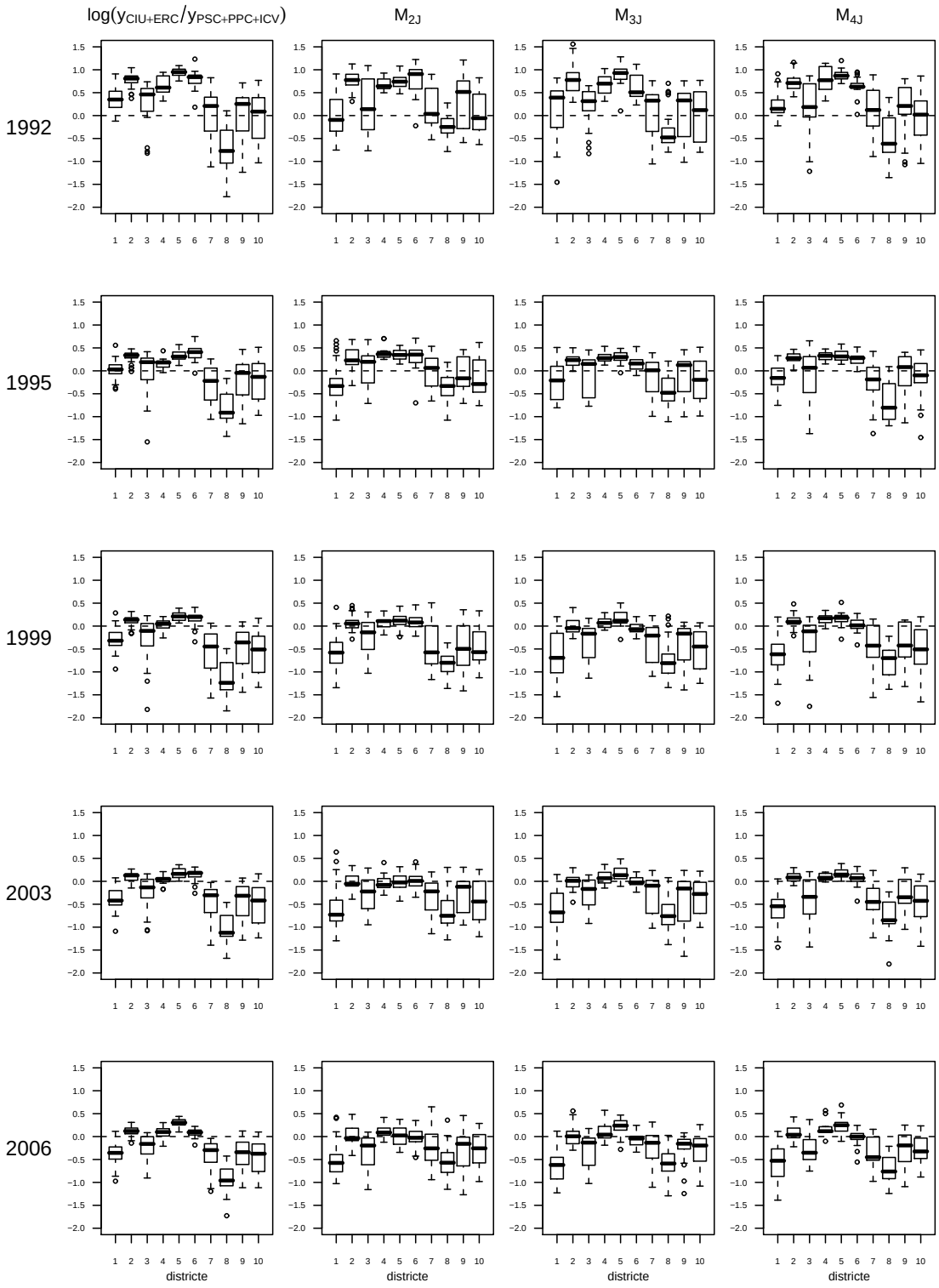

Figura 11.3: La columna esquerra presenta els valors observats per $D_{ci}(y_i)$ a les eleccions al Parlament de Catalunya del 1992 al 2006 a nivell de zrp als 10 districtes de Barcelona, i les altres columnes de gràfic presenten una rèplica de les dades obtingudes a partir de la predictiva a posteriori dels models M_{2J}, M_{3J} i M_{4J}.

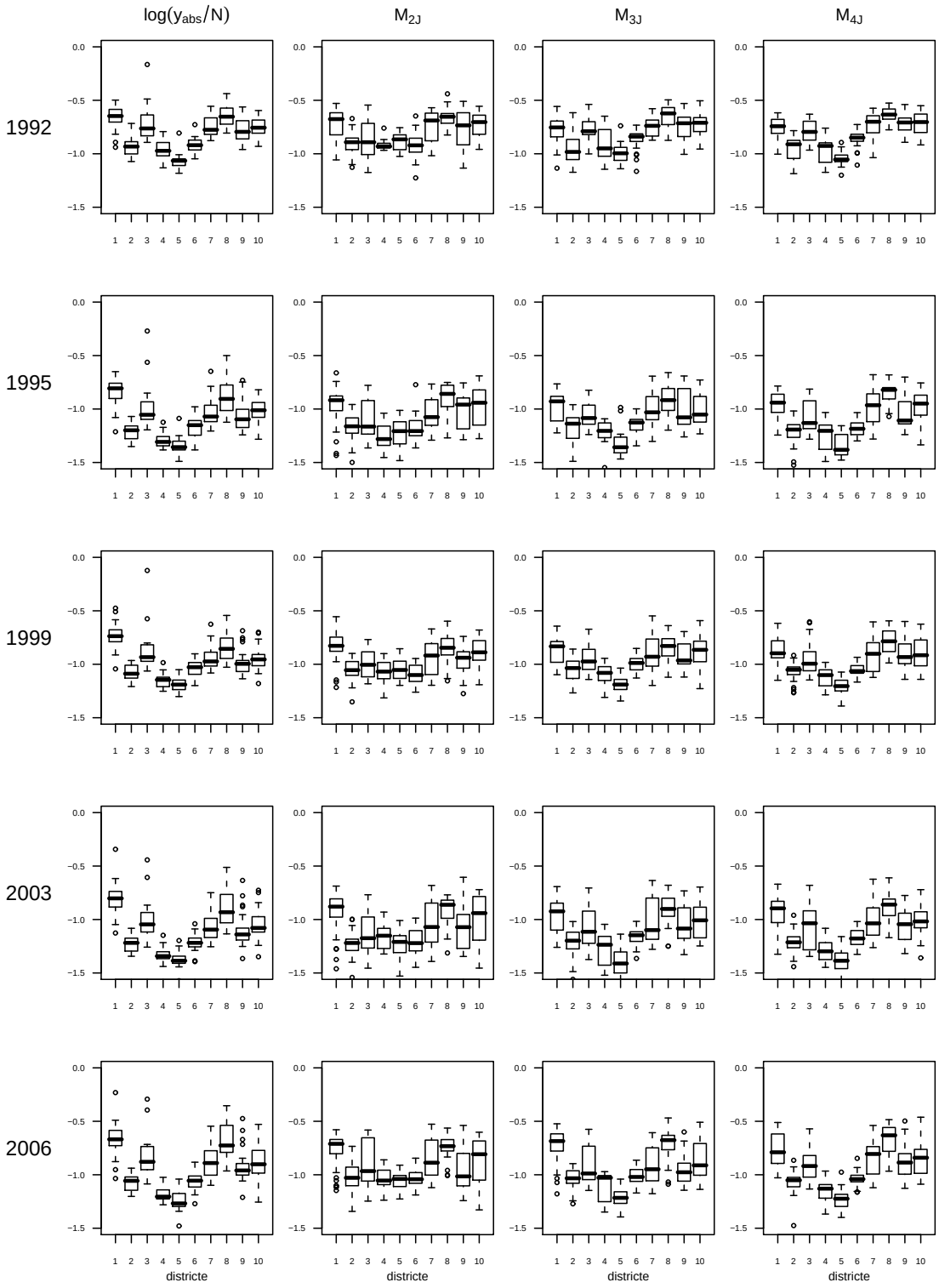

Figura 11.4: La columna esquerra presenta els valors observats per $D_{di}(y_i)$ a les eleccions al Parlament de Catalunya del 1992 al 2006 a nivell de zrp als 10 districtes de Barcelona, i les altres columnes de gràfic presenten una rèplica de les dades obtingudes a partir de la predictiva a posteriori dels models M_{2J}, M_{3J} i M_{4J}.

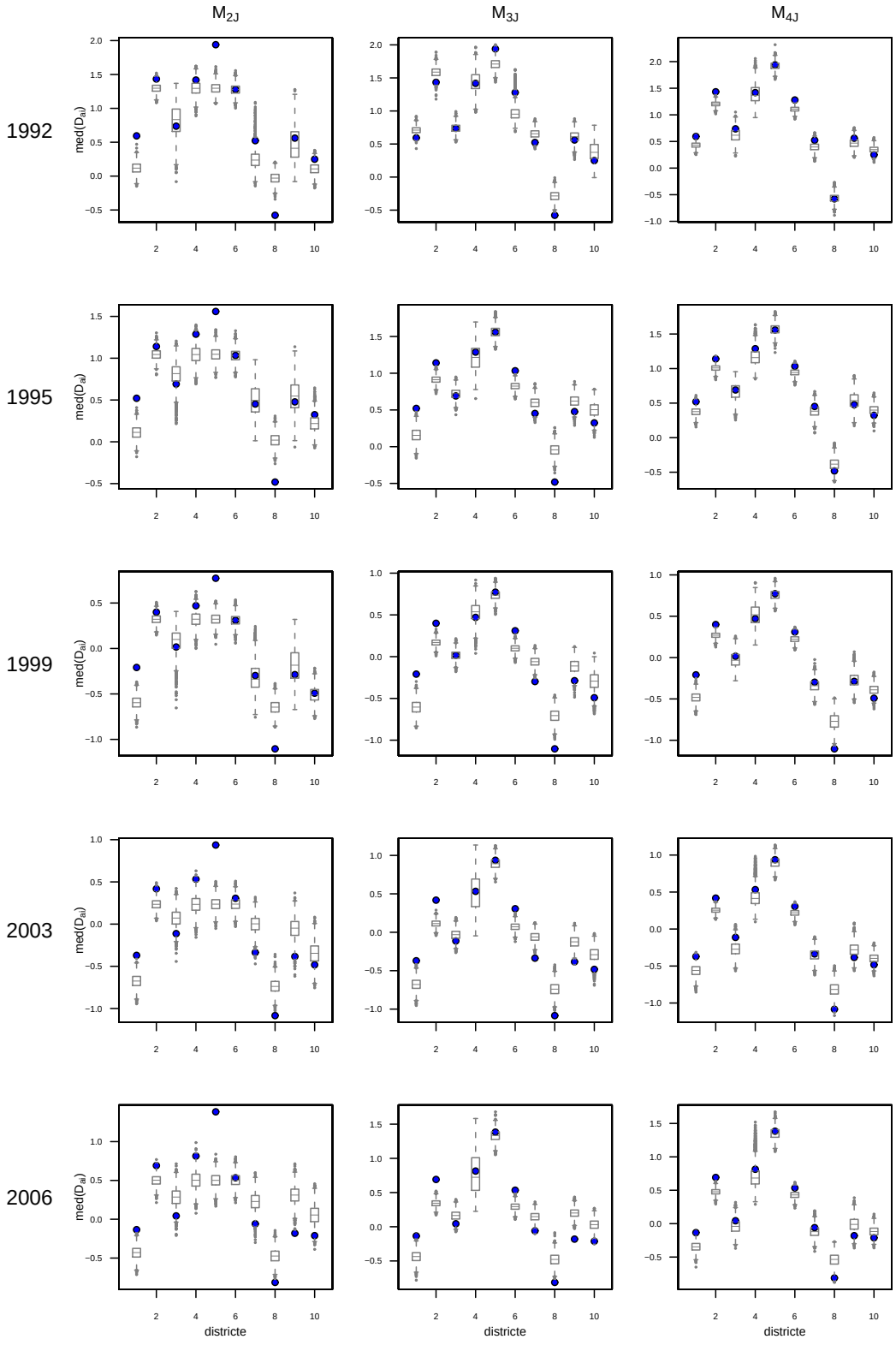

Figura 11.5: Els punts representen el valor observat de la mediana de l'estadístic $D_{ai}(y_i) = log(y_{i,CIU}/y_{i,PSC})$ a les eleccions al Parlament de Catalunya del 1992 al 2006 a nivell de zrp als 10 districtes de Barcelona, i els diagrames de caixes representen les respectives distribucions predictiva a posteriori per cada model jeràrquic considerat.

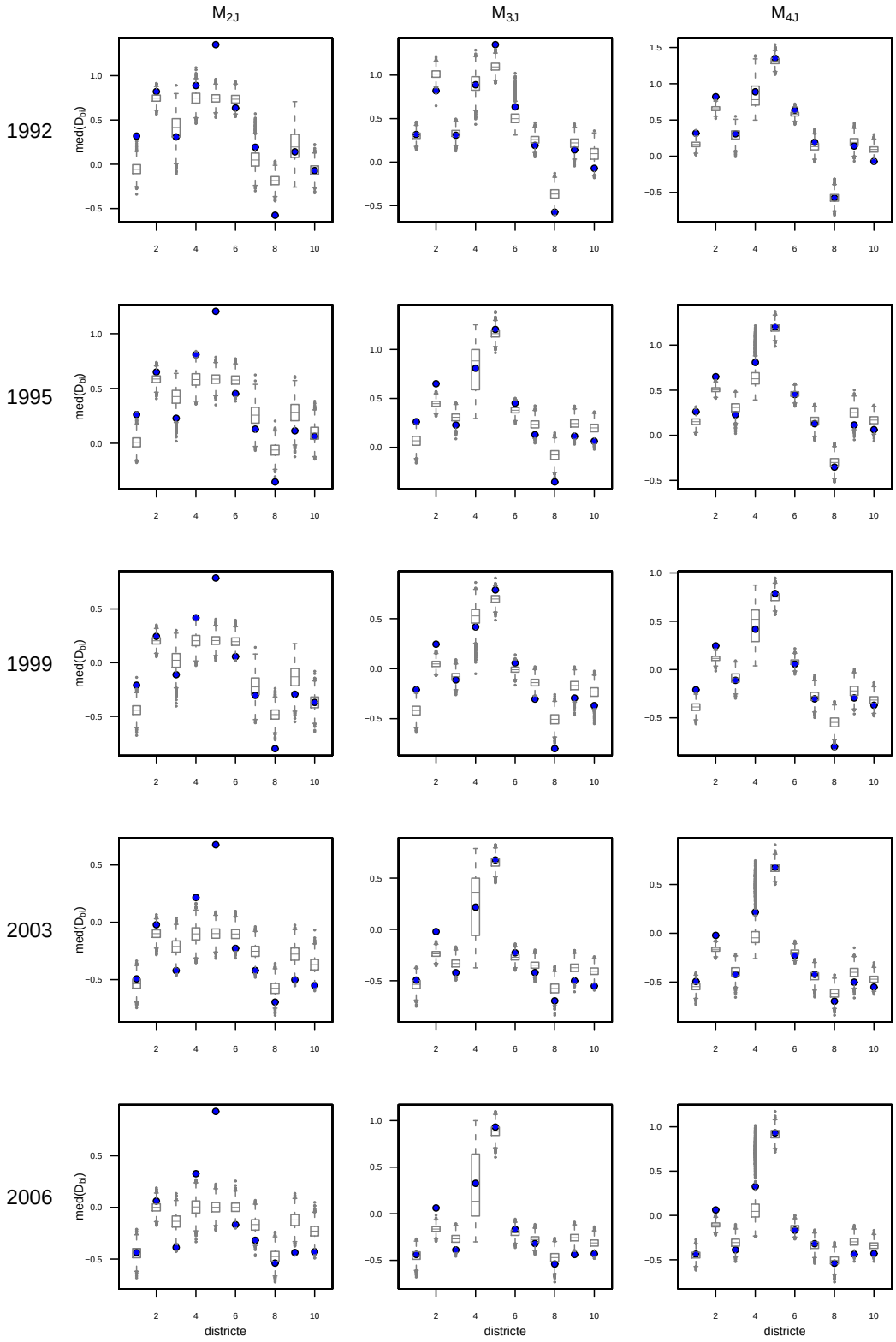

Figura 11.6: Els punts representen el valor observat de la mediana de l'estadístic $D_{bi}(y_i) = log(y_{i,CIU+PPC}/y_{i,PSC+ERC+ICV})$ a les eleccions al Parlament de Catalunya del 1992 al 2006 a nivell de zrp als 10 districtes de Barcelona, i els diagrames de caixes representen les respectives distribucions predictiva a posteriori per cada model jeràrquic considerat.

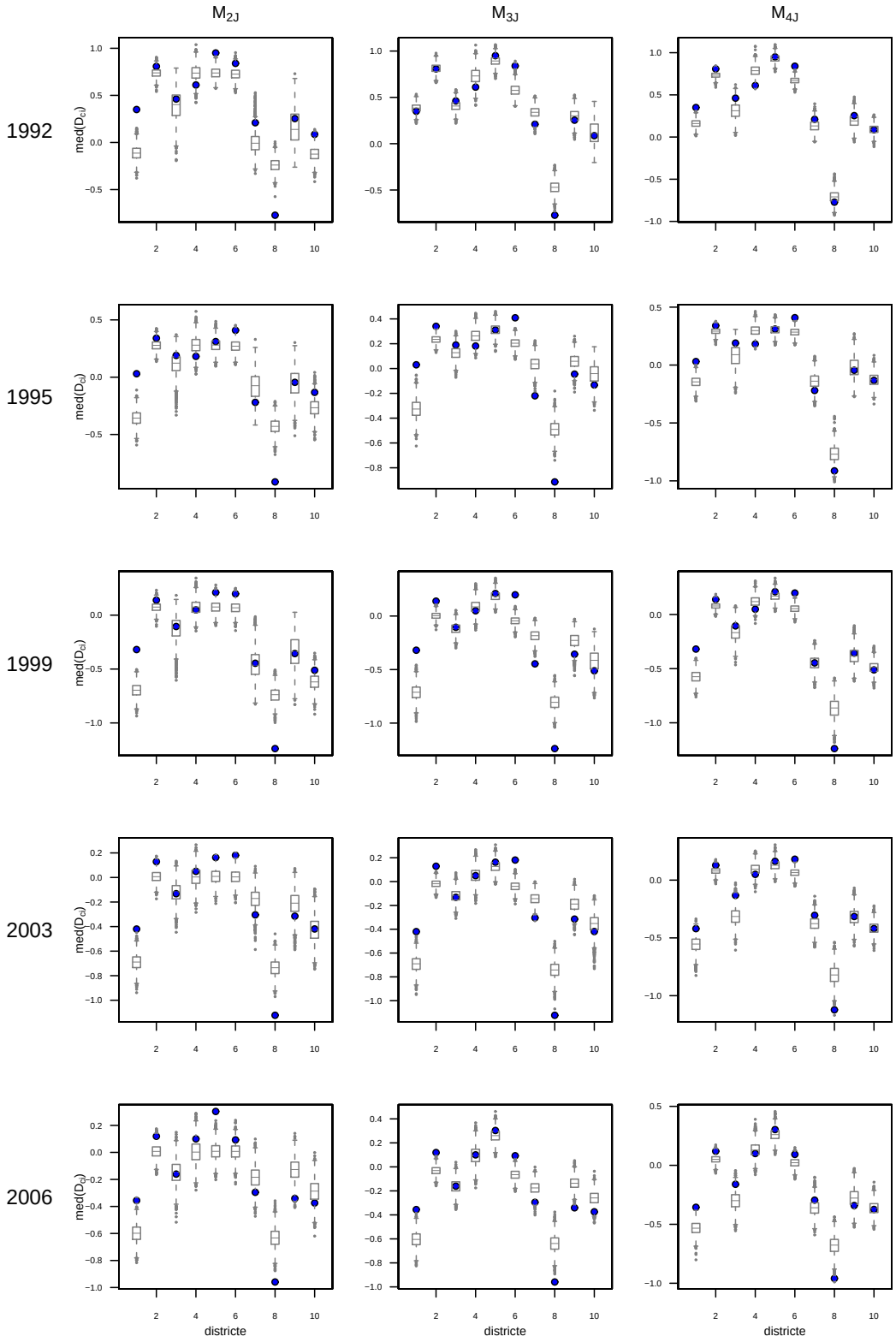

Figura 11.7: Els punts representen el valor observat de la mediana de l'estadístic $D_{ci}(y_i) = log(y_{i,CIU+ERC}/y_{i,PSC+PPC+ICV})$ a les eleccions al Parlament de Catalunya del 1992 al 2006 a nivell de zrp als 10 districtes de Barcelona, i els diagrames de caixes representen les respectives distribucions predictiva a posteriori per cada model jeràrquic considerat.

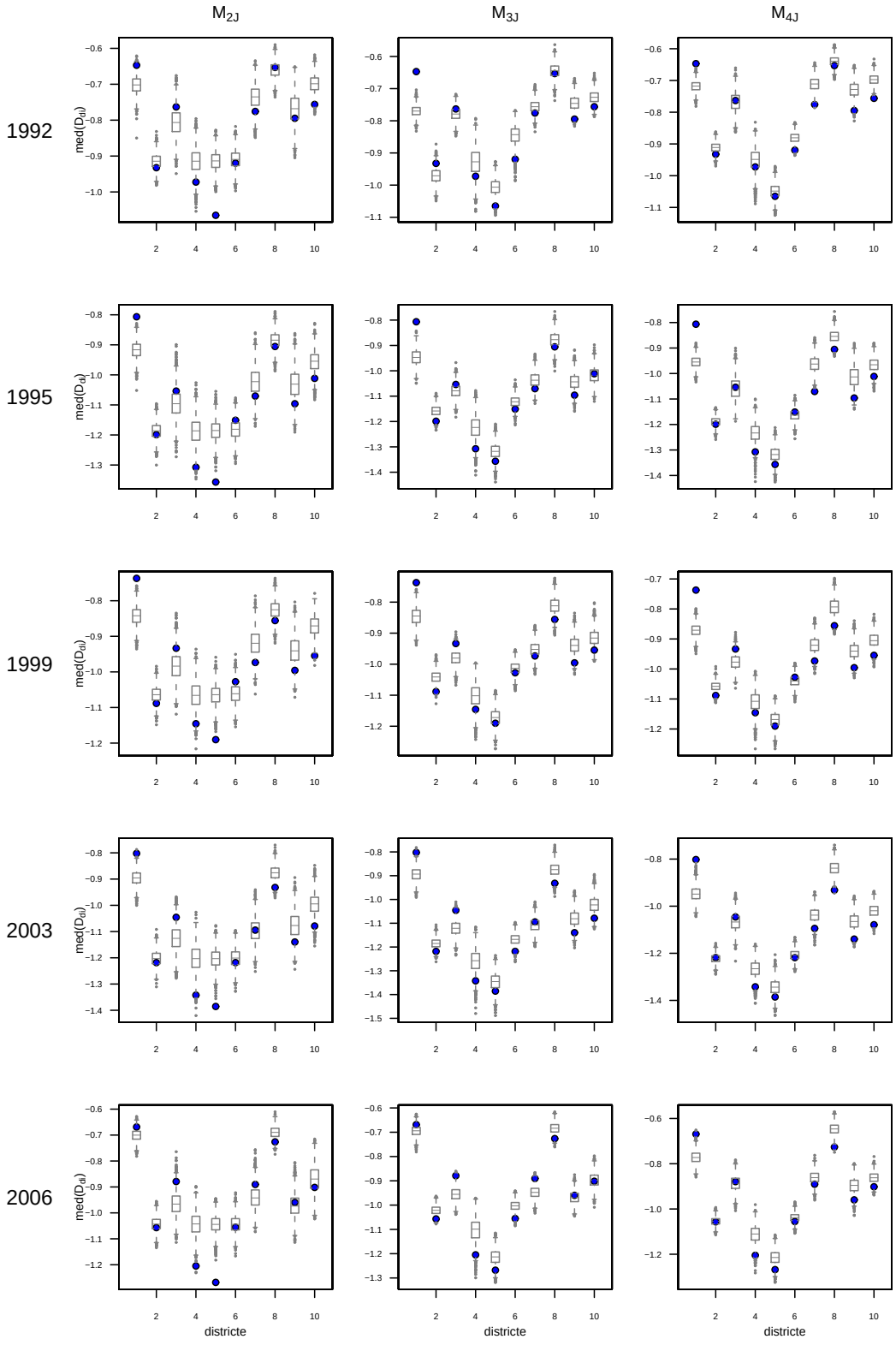

Figura 11.8: Els punts representen el valor observat de la mediana de l'estadístic $D_{di}(y_i) = log(y_{i,abs}/N_i)$ a les eleccions al Parlament de Catalunya de 1992 a 2006 a nivell de zrp als 10 districtes de Barcelona, i els diagrames de caixes representen les respectives distribucions predictiva a posteriori per cada model jeràrquic considerat.

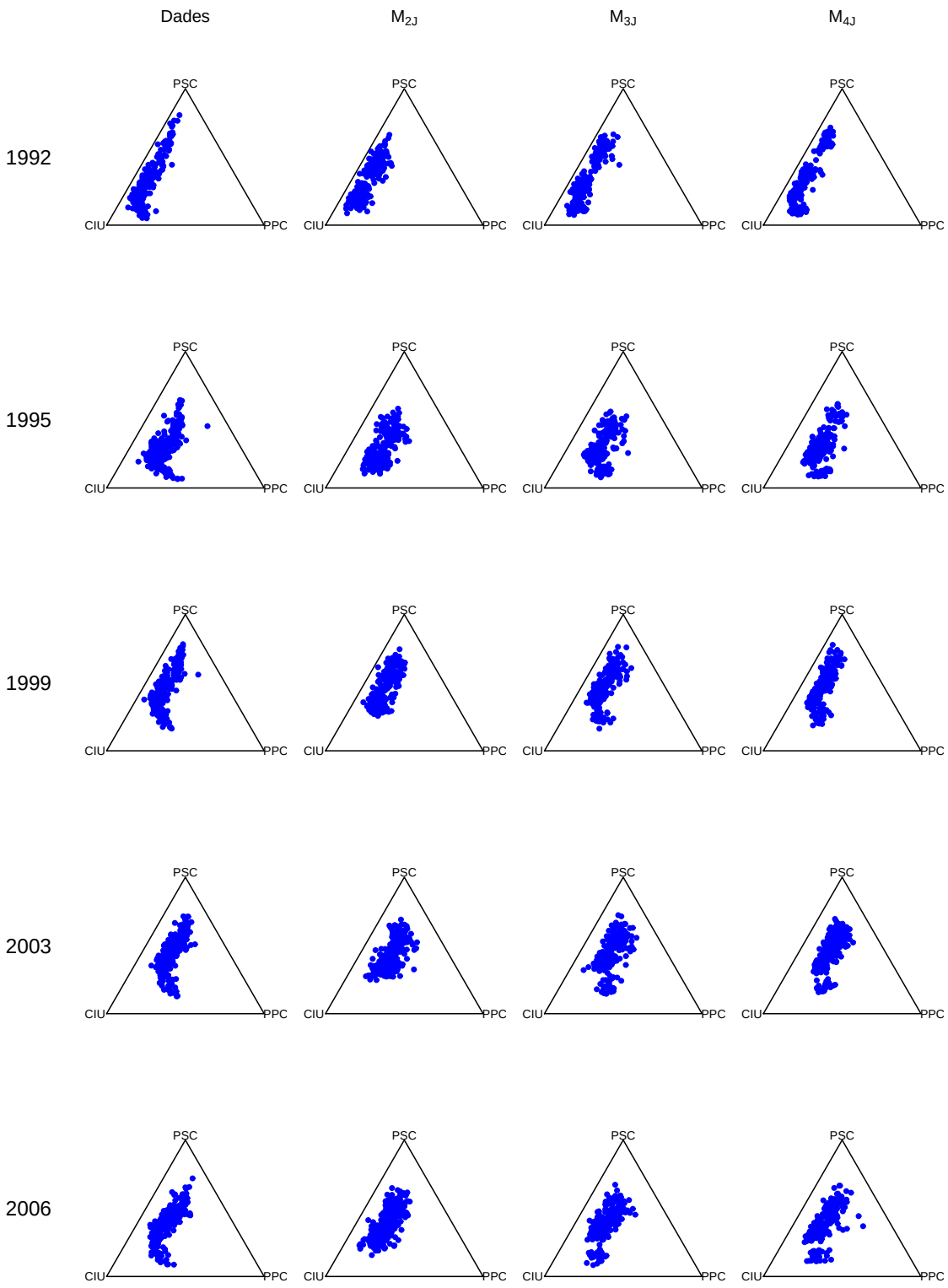

Figura 11.9: La columna esquerra presenta la representació ternària dels valors observats dels perfils de tres columnes de la taula original a les eleccions al Parlament de Catalunya del 1992 al 2006 a nivell de zrp de Barcelona, i les altres columnes de gràfic presenten una rèplica de les dades obtingudes a partir de la predictiva a posteriori dels models M_{2J}, M_{3J} i M_{4J}.

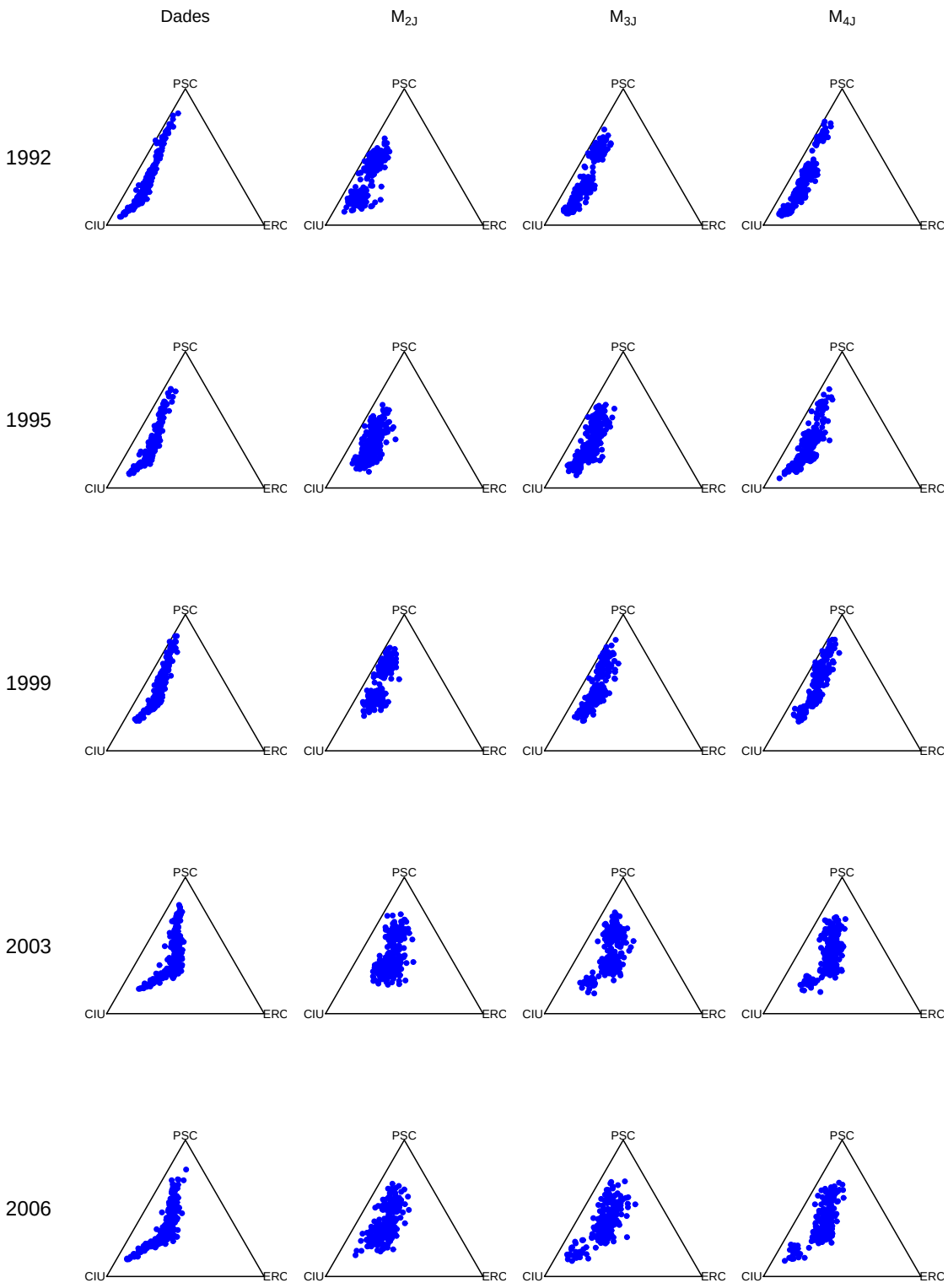

Figura 11.10: La columna esquerra presenta la representació ternària dels valors observats dels perfils de tres columnes de la taula original a les eleccions al Parlament de Catalunya del 1992 al 2006 a nivell de zrp de Barcelona, i les altres columnes de gràfic presenten una rèplica de les dades obtingudes a partir de la predictiva a posteriori dels models M_{2J}, M_{3J} i M_{4J}.

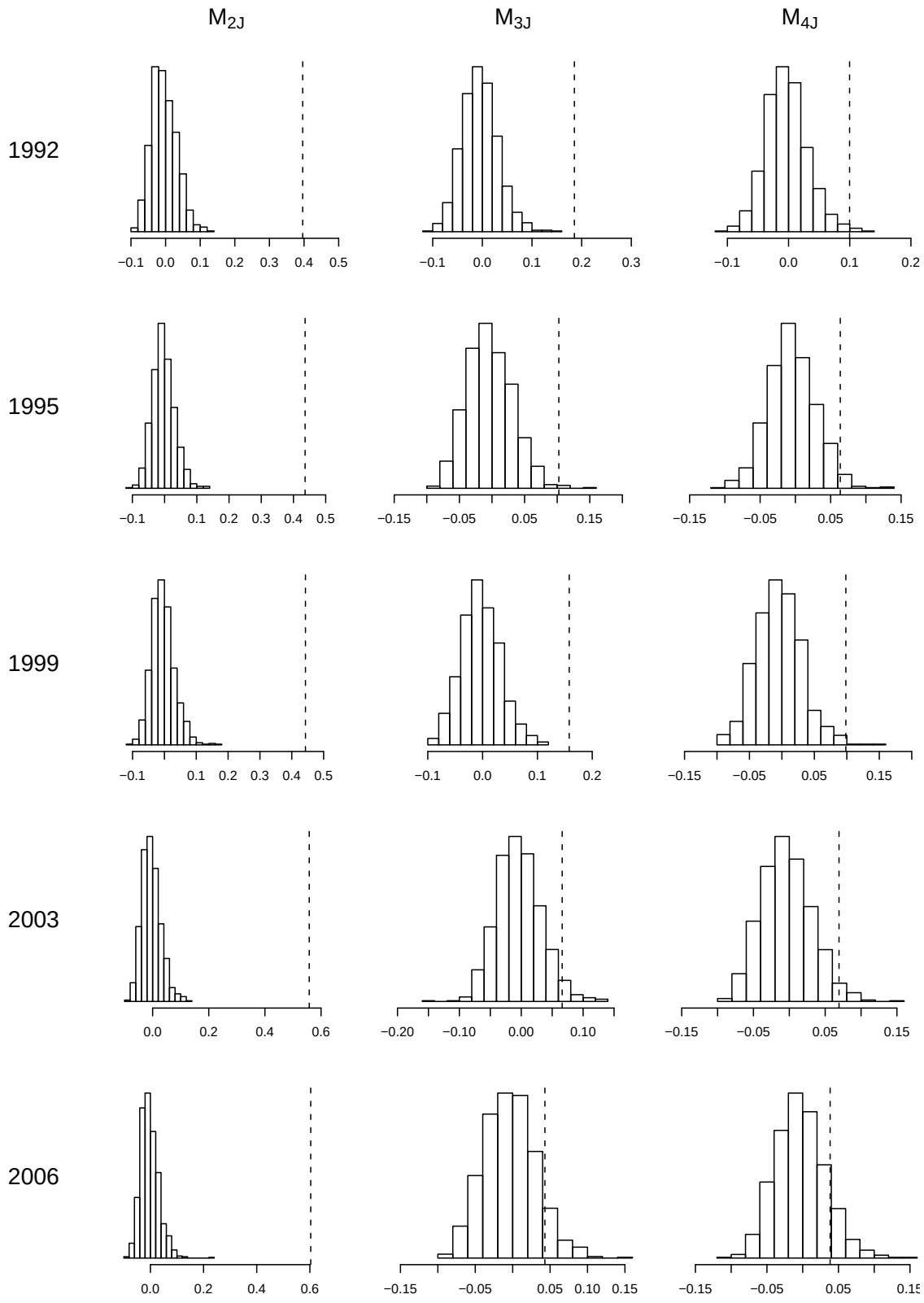

Figura 11.11: Índex de Moran, calculat a partir de $E[D_{2i}^2|y]$ i la seva distribució de referència sota el test de permutacions per als models jeràrquics M_{2J}, M_{3J} i M_{4J}.

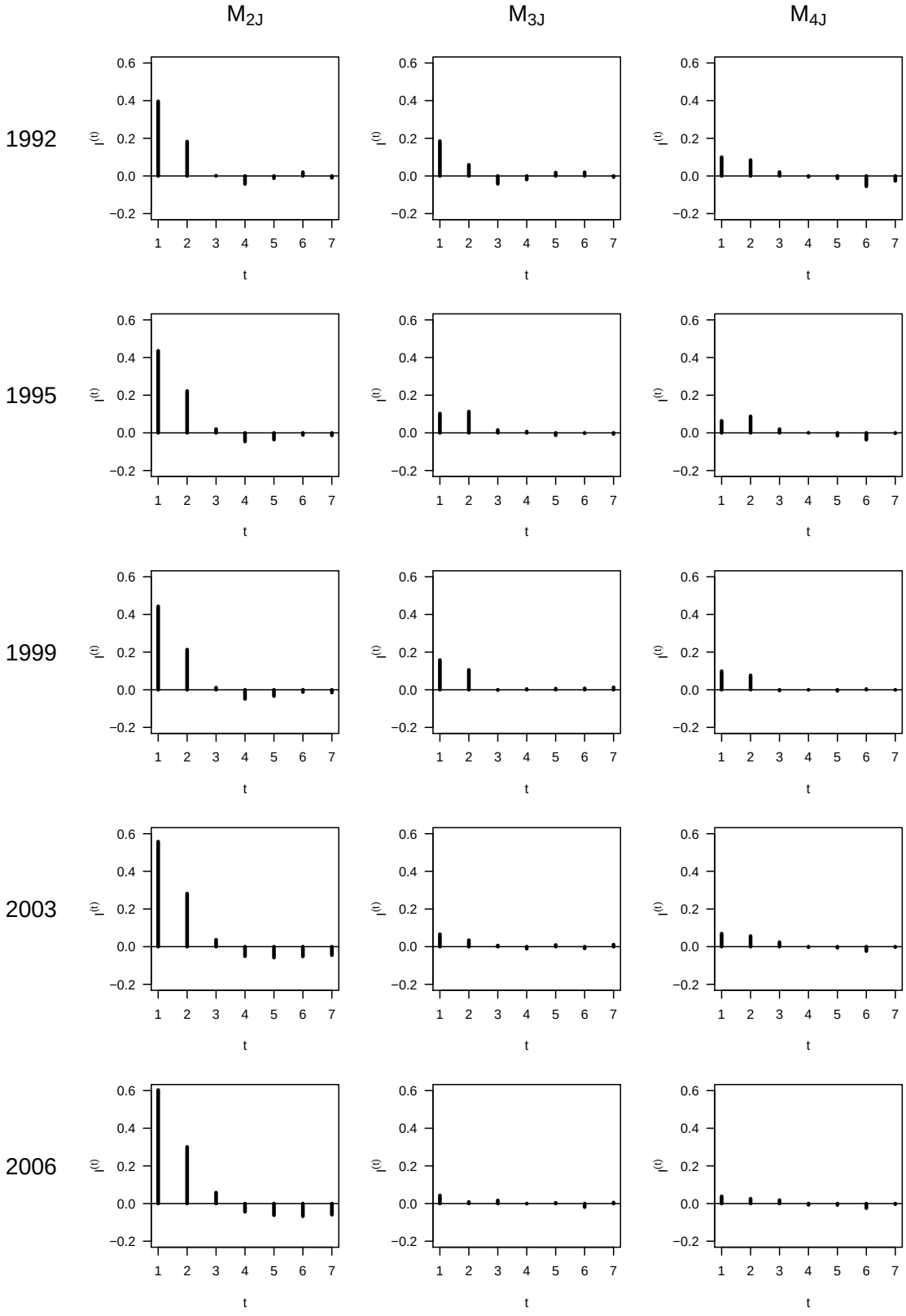

Figura 11.12: Correlogrames fins al setè ordre, calculats utilitzant l'Índex de Moran a partir de $E[D^2_{2i}|y]$, per als models jeràrquics M_{2J}, M_{3J} i M_{4J}.

Capítol 12

Interpretació dels models per al 1992-2006 a Barcelona

En aquest capítol presentarem els resultats per a tots els anys dels models M_{3J} i M_{4J}. Aquests models permetran identificar l'existència de patrons de vot, veure quines zrp pertanyen a cada patró de vot, estudiar com aquests patrons han anat variant al llarg de les diferents eleccions i explorar quines zrp estan canviant de patró i en quina direcció. Conèixer si les evolucions han estat diferents per a cada cluster serà el punt de partida per a una de les extensions apuntades al Capítol 13 en la qual es vol inferir sobre el transvasament de vots d'un partit a un altre entre dues eleccions consecutives.

Primer presentarem els resultats per al model jeràrquic de tres clusters, M_{3J}, i després per al model jeràrquic de quatre clusters, M_{4J}, presentats al Capítol 7 sota l'etiqueta de Model 6.1. També analitzarem les diferències que comporta escollir aquests dos models a nivell d'interpretació.

Els paràmetres d'interès seran:

a) la probabilitat de que una observació triada a l'atzar pertanyi al cluster r, ω_r per $r = 1, 2, 3$ en el cas de tres clusters i per $r = 1, 2, 3, 4$ pel de quatre clusters,

b) el vector de variables latents, $\zeta = (\zeta_1, \ldots, \zeta_{248})$, que recull a quin cluster pertany cada zrp,

c) el perfil de probabilitat esperada per cada cluster $\mu_r = (\mu_{r1}, \ldots, \mu_{r8})$ per $r = 1, 2, 3$ en el cas de tres clusters i per $r = 1, 2, 3, 4$ pel de quatre clusters, i

d) el paràmetre que regula el grau d'heterogeneïtat dels clusters, τ_r per $r = 1, 2, 3$ en el cas de tres clusters i per $r = 1, 2, 3, 4$ pel de quatre clusters.

12.1 Resultats del model de tres clusters: M_{3J}

A la Taula 12.1 es presenten per a cada elecció els valors esperats a posteriori de tots els paràmetres d'interès, excepte de ζ, sota el model M_{3J}. Per facilitar la interpretació dels clusters i la comparació entre anys, s'han ordenat els clusters per ordre decreixent del perfil de vot de CIU, de manera que on obté més vots CIU és al cluster 1 i on n'obté menys és al darrer cluster. Sota aquest criteri d'ordenació automàticament el PSC, *altres* i l'abstenció queden ordenats en ordre creixent i el PPC i blancs i nuls, queden ordenats en ordre decreixent. En aquesta mateixa taula també es presenten el percentatge de vots observats de cada categoria per a cada any. Si ens fixem en el valor esperat a posteriori del perfil d'ERC observem que l'any 1992 obté el valor més alt en el primer cluster i la resta d'anys en el segon cluster, mentre que ICV l'any 1992 i 1995 obté els valors més alts en el tercer cluster i en els següents anys en el segon cluster.

La distribució a posteriori de la probabilitat de pertànyer a cada un dels clusters, $\pi(\omega_r|y)$ per $r = 1, 2, 3$, les trobem representades a la Figura 12.1 mitjançant diagrames de caixa, i a la Figura 12.2 es presenta la distribució a posteriori de τ_r, per $r = 1, 2, 3$, per a cadascuna de les cinc eleccions. Del model M_{3J} destaca que el cluster 2 és el més majoritari i a l'hora el més homogeni, que el cluster 1 és el més minoritari i que el cluster 3 és el més heterogeni. Si ens concentrem en les quatre darreres eleccions s'observa com la grandària dels clusters és relativament estable, però en canvi tendeix a disminuir la homogeneïtat del cluster 1. A les eleccions del 1992 el cluster 1 i 3 tenien grandàries semblants amb una valor esperat del percentatge a posteriori del 24% i 28% respectivament, mentre que al 2003 i al 2006 el cluster 3 triplica en grandària al cluster 1.

La classificació de cada zrp, i, a cada cluster, $r = 1, 2, 3$, s'ha fet mitjançant la moda a posteriori de les variables latents ζ_i, $Mo(\zeta_i|y)$ per $i = 1, \ldots, 248$. Els resultats per al model de tres clusters es mostren a la Figura 12.3. Si ens fixem en les darreres quatre eleccions observem com la majoria de zrp dels districtes de Sarrià-Sant Gervasi i aproximadament al meitat de Les Corts han estat assignades al cluster 1. Les zrp que voregen el cluster 1, que són bàsicament zrp dels districtes de Sants-Montjuïc, l'Eixample i Gràcia pertanyen al cluster 2, i finalment les zrp de Ciutat Vella i Nou Barris pertanyen majoritàriament al cluster 3.

Com a canvi més rellevant entre anys notem que a l'any 1992 la majoria de zrp de

l'Eixample estaven classificades al cluster 1 mentre que per la resta d'eleccions en aquest districte hi predominen les zrp pertanyents al cluster 2. De la representació dels mapes també destaca la notable agregació espacial que presenten els clusters, en el sentit de que el canvi d'un cluster a un altre es fa gradualment a mesura que et desplaces per Barcelona, en funció de la distància al centre del districte de Sarrià-Sant Gervasi.

La Taula 12.2 presenta el nombre de zrp que han canviat de cluster en cadascuna de les 6 possibles direccions. Recordem que els clusters estan etiquetats seguint un gradient descendent per CIU i ascendent per PSC. En aquesta Taula s'observa que en les darreres cinc eleccions cap zrp ha canviat dos clusters a l'hora passant del cluster 1 al 3 o viceversa entre eleccions consecutives. És més només una zrp ha canviat del cluster 1 al 3 al llarg de totes les eleccions estudiades; es tracta d'una zrp de Cuitat Vella que el 1992 pertanyia al cluster 1, el 1995 al cluster 2 i el 1999 al cluster 3. Cap zrp ha passat del cluster 3 al 1 en aquests 14 anys.

A la Figura 12.4 s'hi presenten com s'han distribuït aquests canvis en l'espai i en el temps. Així de les eleccions del 1992 a les del 1995 hi ha zrp de Gràcia i de l'Eixample que passen del cluster 1 al 2 i zrp de Citat Vella que passen del 2 al 3. De les eleccions del 1995 a les de 1999 el districte de Sant Martí és el que presenta els principals canvis ja que tres de les seves zrp canvien del cluster 2 al 3, i de les eleccions del 1999 a les del 2003 són zrp de l'Eixample les que passen del cluster 1 al 2. I entre les eleccions del 1999 a les del 2006 s'observa com alguna zrp dels districtes de Horta Guinardó, Sant Andreu i Sant Martí canvien del cluster 3 al 2. A nivell global l'evolució entre el 1992 i el 2006 es pot resumir dient que part de les zrp de l'Eixample han passat de pertànyer al cluster 1 a pertànyer al cluster 2, i una cosa semblant passa a Ciutat Vella on moltes de les seves zrp han passat de pertànyer al cluster 2 a pertànyer al cluster 3, que Nou Barris ha estat l'únic districte en que cap zrp ha experimentat cap canvi de cluster i que en general han estat més nombrosos els canvis a pujar de cluster que no pas a baixar.

Si bé els mapes de classificació de la Figures 12.3 no mostren grans canvis pel que fa a la seva distribució geogràfica en les quatre darreres eleccions, la distribució dels perfils si que han variat més al llarg dels diferents comicis.

A la Figura 12.5 presentem la distribució a posteriori dels perfils de probabilitat esperats per a cada cluster, $\mu_r = (\mu_{r1}, \ldots, \mu_{r8})$ per $r = 1, 2, 3$ sota el model jeràrquic M_{3J} i a la Figura 12.6 la distribució a posteriori de $log(\mu_{rj}/\mu_{r'j})$ per $r > r'$. Aquestes figures ens permeten esbrinar quines components del perfil canvien d'unes eleccions a les altres.

En el cluster 1 hi destaca sobretot el domini de CIU i és l'únic cluster en que un partit està per damunt de l'abstenció, la segona força política, tot i que a una gran distància, ha

estat per a tots els anys el PPC amb l'excepció de l'any 1999 que ho va ser el PSC, en la resta d'eleccions el PSC ha estat la tercera força política, mentre que ERC sempre havia estat la quarta força política per damunt de ICV i l'any 2006 aquestes dues formacions s'igualen.

En el cluster 2 el partit més representat els anys 1992 i 1995 amb diferència era CIU mentre que els anys 1999 i 2003 el PSC va guanyar terreny fet que coincideix amb la presència d'en Pasqual Maragall. El 2006 aquestes dues formacions es tornen a distanciar però no ho fan en les magnituds dels dos primers comicis analitzats. També en aquest cluster 2 s'observa com en el 1995 i el 1999 el PPC era la tercera força política, en el 2003 passa a ser la quarta darrera de CIU, PSC i ERC i l'any 2006 la cinquena força política per darrera també de ICV, i de la mateixa manera que havíem observat al cluster 1 trobem que ERC estava per sobre de ICV en les quatre primeres eleccions analitzades i que l'any 2006 aquestes dues formacions s'igualen.

En el cluster 3 hi destaca sobretot l'alta abstenció essent aquesta superior al 50% els anys 1992 i 2006, també en aquest cluster 3 s'observa com en les eleccions del 1995 CIU i PSC lluitaven pel lideratge i en les posteriors eleccions ha estat clar el domini del PSC especialment al 1999 i 2003.

Pel que fa a la comparació de perfils dels clusters, segons la Figura 12.6, tenim que el cluster 2 es diferencia del 1 per tenir-hi més presència el PSC, ICV i ERC en detriment de CIU i del PPC, mentre que el cluster 3 es diferencia del 1 per tenir-hi més presència el PSC i l'abstenció, en detriment de CIU i del PPC, i el cluster 3 es diferencia del 2 per tenir-hi menys presència CIU i ERC i en el 2006 també menys presència ICV.

A la Figura 12.7 representem subconjunts de les dades observades en diagrames ternaris colorejant les zrp segons el cluster al que han estat classificats segons el model M_{3J}. I a la Figura 12.8 s'hi representen per a cada cluster les ternes del valor esperat a posteriori dels perfils de probabilitat esperats resultats d'agregar categories, així per al cluster r-èssim la terna representada a la primera columna de gràfics per a cadascun dels anys s'ha calculat com:

$$E[\frac{\mu_{r,CIU}}{\mu_{r,CIU}+\mu_{r,PSC}+\mu_{r,ERC}}, \frac{\mu_{r,PSC}}{\mu_{r,CIU}+\mu_{r,PSC}+\mu_{r,ERC}}, \frac{\mu_{r,ERC}}{\mu_{r,CIU}+\mu_{r,PSC}+\mu_{r,ERC}}|y] \tag{12.1}$$

$$E[\frac{\mu_{r,CIU}+\mu_{r,PPC}}{\mu_{r,CIU}+\mu_{r,PSC}+\mu_{r,PPC}+\mu_{r,ICV}+\mu_{r,ERC}+\mu_{r,abs}}, \frac{\mu_{r,PSC}+\mu_{r,ERC}+\mu_{r,ICV}}{\mu_{r,CIU}+\mu_{r,PSC}+\mu_{r,PPC}+\mu_{r,ICV}+\mu_{r,ERC}+\mu_{r,abs}}, \frac{\mu_{r,abs}}{\mu_{r,CIU}+\mu_{r,PSC}+\mu_{r,PPC}+\mu_{r,ICV}+\mu_{r,ERC}+\mu_{r,abs}}|y] \tag{12.2}$$

$$E[\frac{\mu_{r,CIU}+\mu_{r,ERC}}{\mu_{r,CIU}+\mu_{r,PSC}+\mu_{r,PPC}+\mu_{r,ICV}+\mu_{r,ERC}+}, \frac{\mu_{r,PSC}+\mu_{r,PPC}}{\mu_{r,CIU}+\mu_{r,PSC}+\mu_{r,PPC}+\mu_{r,ICV}+\mu_{r,ERC}+}, \frac{\mu_{r,ICV}}{\mu_{r,CIU}+\mu_{r,PSC}+\mu_{r,PPC}+\mu_{r,ICV}+\mu_{r,ERC}+}|y] \tag{12.3}$$

per $r = 1, 2, 3$.

La primera columna de les Figures 12.7 i 12.8 que presenta el pes relatiu dels partits de CIU, PSC i ERC mostra el progressiu augment del pes relatiu d'ERC en el cluster 2, tot i que també però de forma més moderada en zrp del cluster 1 i 3.

La segona columna de les Figures 12.7 i 12.8 presenta l'evolució del pes relatiu entre la centre-dreta, l'esquerra i l'abstenció i mostra com el cluster 1 s'ha mantingut fidel a la centre-dreta i el 3 a l'abstenció, mentre que l'esquerra ha vist augmentar de manera progressiva des del 1992 al 2003 el seu pes al cluster 2.

Finalment a la tercera columna de les Figures 12.7 i 12.8 s'hi agrupa per una banda els partits nacionalistes catalans CIU i ERC, per una altra els partits d'obediència espanyola PSC i PPC, i per una altra ICV. En aquesta columna s'observa com el cluster 1 sempre s'ha mantingut allunyat d'ICV i que així com al 1992 el cluster 1 era fortament catalanista en les posteriors eleccions s'ha vist lleugerament traslladat cap a la vertent més espanyolista, principalment per l'augment del PPC en aquest cluster. El cluster 2 si bé els anys 1992 i 1995 hi pesava més l'eix catalanista els darrers anys es troben equidistanciats però amb una aproximació progressiva cap a ICV. Finalment el cluster 3, amb caràcter més espanyolista, també s'ha vist atret encara que de forma més moderada pel pol d'ICV.

Una darrera anàlisi dels resultats passa per explorar l'evolució de les variables latents ζ_i per a cada zrp en cadascuna de les eleccions i identificar les zrp que han estat sempre fidels al seu cluster, les zrp que oscil·len entre dos clusters o les zrp que estan evolucionant d'un cluster cap a un altre cluster canviant el seu comportament de vot. A l'Apèndix A s'hi presenten les distribucions a posteriori de totes les ζ_i per $i = 1, \ldots, 248$ per a cada elecció, i a la secció 12.3 es presenten i s'interpreten l'evolució de les distribucions a posteriori concretes per algunes zrp sota els models M_{3J} i M_{4J}.

Model	cluster	CIU	PSC	PPC	ICV	ERC	Cs	altres	b+n	abs	$E[\omega\|y]$	$E[\tau\|y]$
1992	1	0.378	0.068	0.074	0.032	0.050	-	0.022	0.013	0.364	0.241	302.95
	2	0.268	0.120	0.038	0.041	0.048	-	0.025	0.010	0.450	0.476	358.85
	3	0.136	0.193	0.034	0.041	0.024	-	0.029	0.009	0.533	0.283	182.17
	dades	0.263	0.130	0.048	0.039	0.042	-	0.024	0.009	0.445		
1995	1	0.356	0.074	0.189	0.038	0.057	-	0.006	0.015	0.266	0.138	354.87
	2	0.288	0.127	0.102	0.066	0.074	-	0.007	0.010	0.326	0.518	342.27
	3	0.164	0.184	0.095	0.072	0.040	-	0.010	0.010	0.426	0.345	141.79
	dades	0.261	0.141	0.115	0.065	0.061	-	0.006	0.009	0.341		
1999	1	0.314	0.149	0.135	0.019	0.054	-	0.008	0.013	0.308	0.145	343.0
	2	0.237	0.216	0.069	0.031	0.064	-	0.009	0.010	0.364	0.494	381.0
	3	0.130	0.276	0.063	0.030	0.029	-	0.010	0.009	0.453	0.361	144.8
	dades	0.216	0.234	0.079	0.028	0.051	-	0.008	0.008	0.376		
2003	1	0.301	0.120	0.179	0.037	0.087	-	0.008	0.011	0.257	0.118	286.55
	2	0.202	0.187	0.089	0.066	0.127	-	0.009	0.009	0.311	0.522	324.22
	3	0.110	0.243	0.083	0.053	0.063	-	0.013	0.009	0.427	0.360	153.48
	dades	0.186	0.204	0.101	0.057	0.104	-	0.008	0.008	0.333		
2006	1	0.309	0.080	0.141	0.051	0.051	0.043	0.012	0.019	0.294	0.123	280.13
	2	0.188	0.140	0.070	0.086	0.089	0.027	0.015	0.017	0.367	0.529	338.63
	3	0.096	0.166	0.066	0.058	0.049	0.021	0.016	0.014	0.514	0.348	171.35
	dades	0.178	0.147	0.080	0.072	0.073	0.027	0.013	0.015	0.394		

Taula 12.1: Esperança a posteriori els perfils de la probabilitat esperada per a cada cluster sota el model M_{3J}, $E[\mu_{rj}|y]$ per $r = 1, 2, 3$ i per $j = 1, \ldots, 8$ o 9, del tamany relatiu de cada cluster, $E[\omega_r|y]$, i del grau d'heterogeneitat del cluster, $E[\tau_r|y]$, i el percentatge de vots de cada opció per a cada una de les cinc eleccions considerades.

Canvi	1992-1995	1995-1999	1999-2003	2003-2006
$1 \to 3$	0	0	0	0
$2 \to 3$	11	14	3	2
$1 \to 2$	24	1	8	1
$3 \to 2$	3	1	6	5
$2 \to 1$	0	2	0	2
$3 \to 1$	0	0	0	0

Taula 12.2: Canvis en la classificació de les zrp a cada cluster entre dues eleccions consecutives. Per exemple $1 \to 2$ indica el nombre de zrp que canvien del cluster 1 al cluster 2 d'una elecció a la següent.

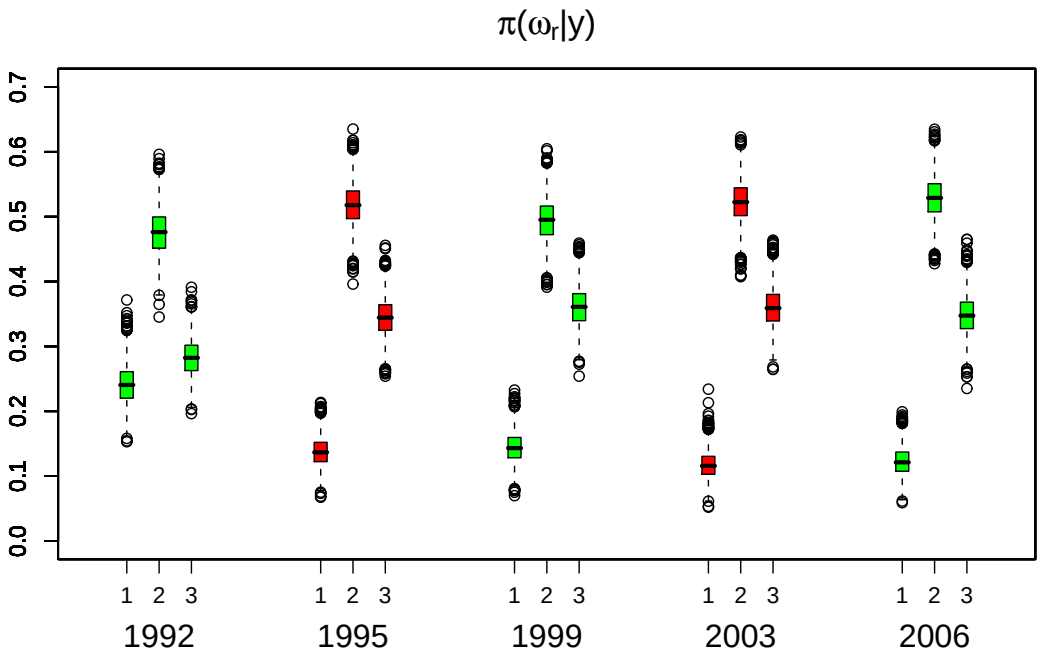

Figura 12.1: Distribució a posteriori de la probabilitat de pertànyer a cada un dels clusters, $\pi(\omega_r|y)$, per $r = 1, 2, 3$, sota el model M_{3J} pels resultats de cada una de les cinc darreres eleccions.

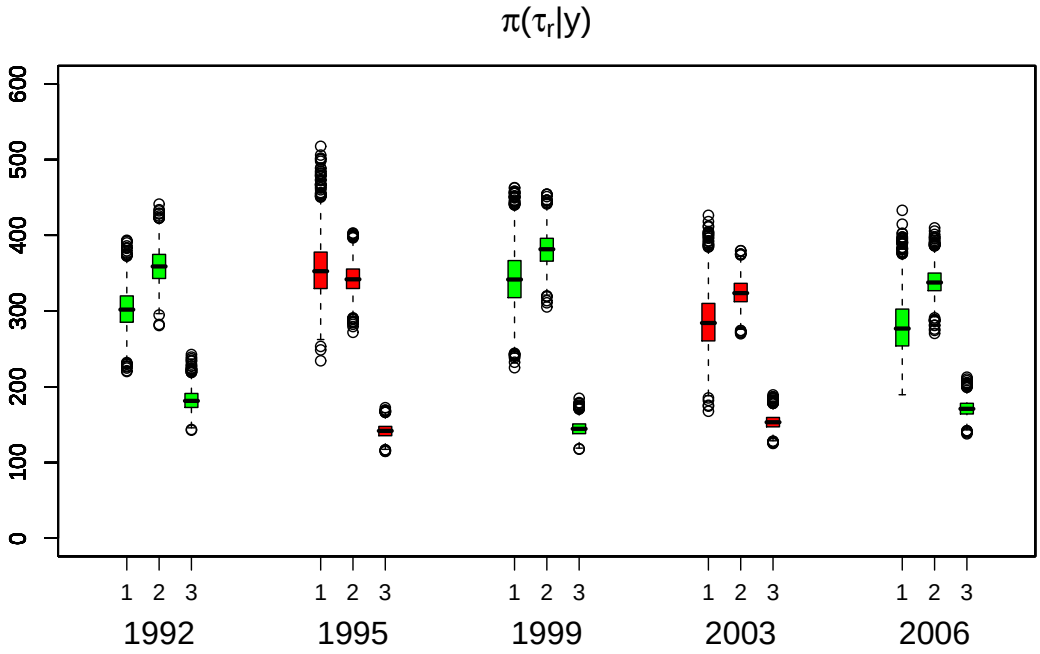

Figura 12.2: Distribució a posteriori del grau d'heterogeneïtat, $\pi(\tau_r|y)$, per $r = 1, 2, 3$, sota el model M_{3J} pels resultats de cada una de les cinc darreres eleccions.

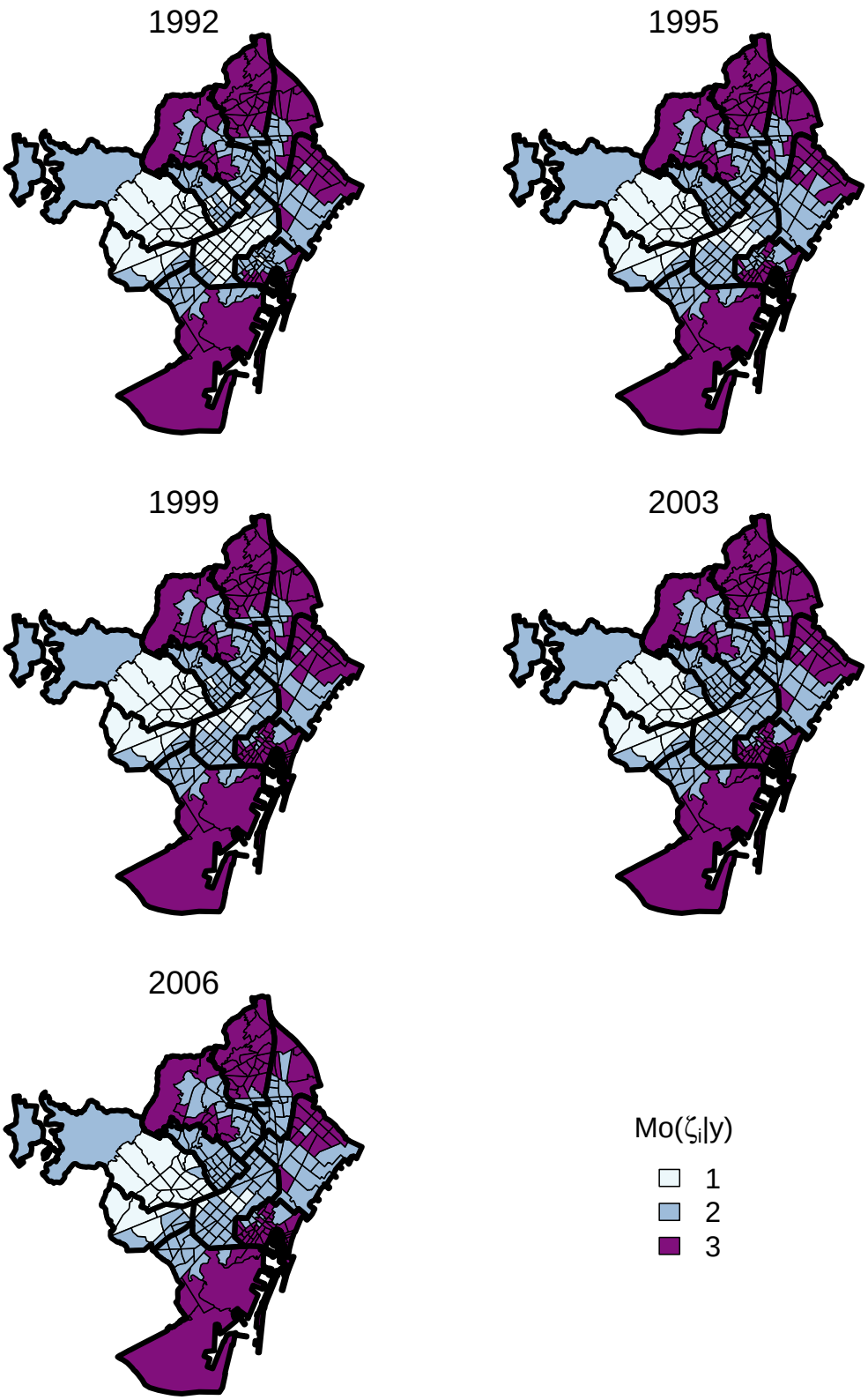

Figura 12.3: Classificació de les zrp utilitzant com a criteri la moda a posteriori de les variables latents ζ_i, $Mo(\zeta_i|y)$, sota el model M_{3J} pels resultats de cada una de les cinc darreres eleccions.

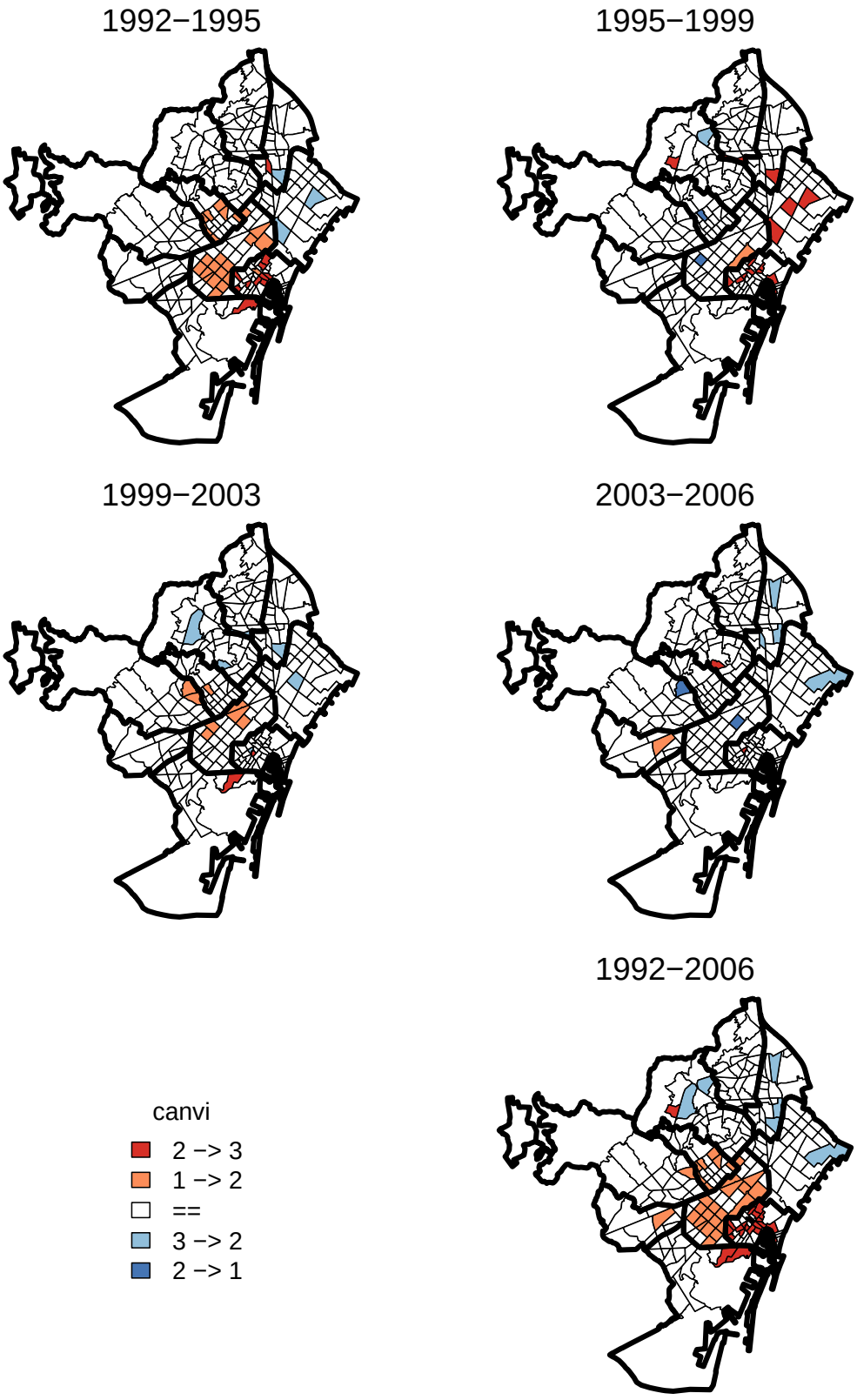

Figura 12.4: Canvis en la classificació de les zrp als tres clusters entre dues eleccions consecutives, en funció de la direcció del canvi.

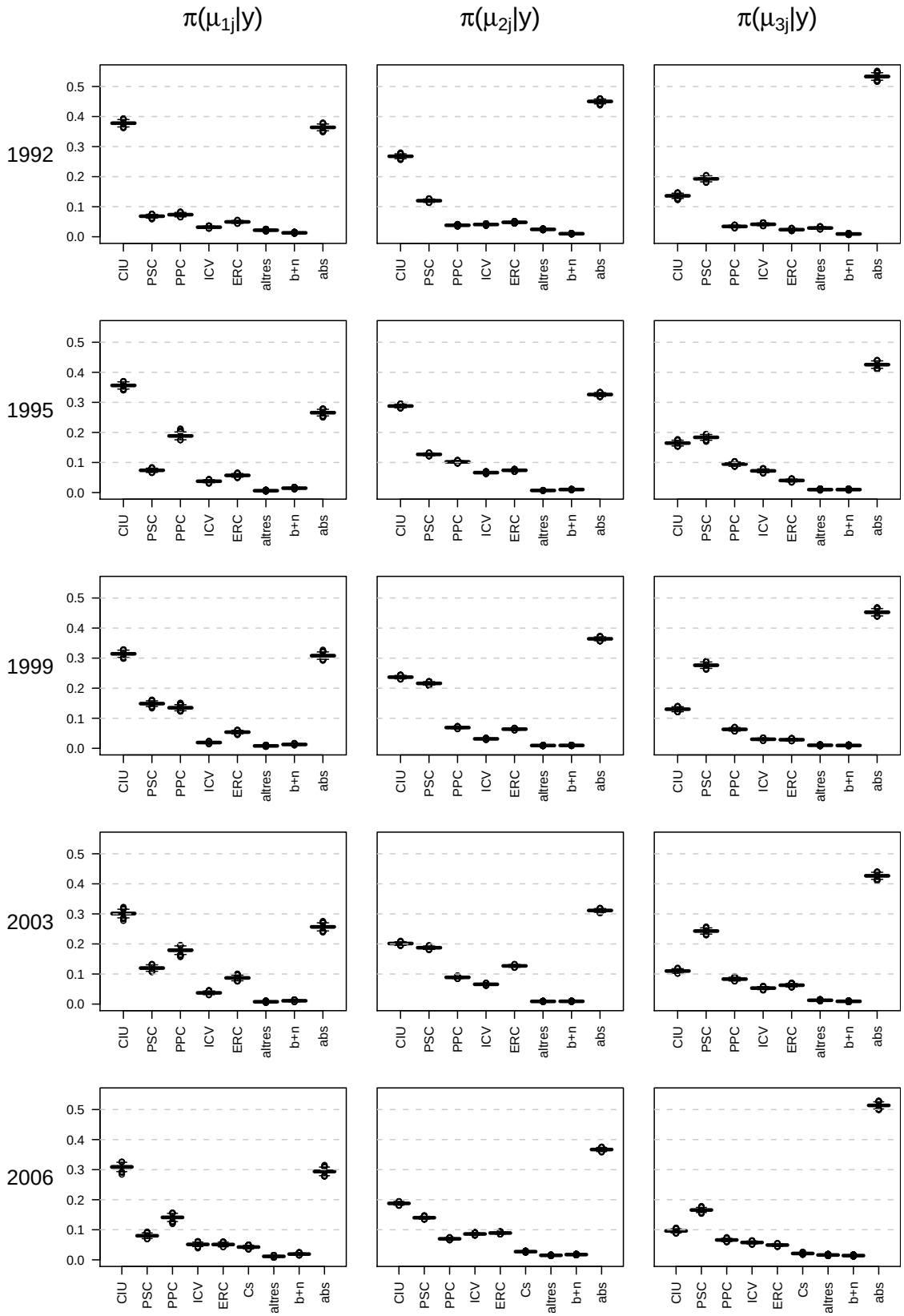

Figura 12.5: Distribució a posteriori dels components del perfil de probabilitat esperada per cada cluster sotal model M_{3J}, $\pi(\mu_{rj}|y)$ per $r = 1, 2, 3$ i per $j = 1, \ldots, 8$ o 9, pels resultats a cada una de les cinc darreres eleccions.

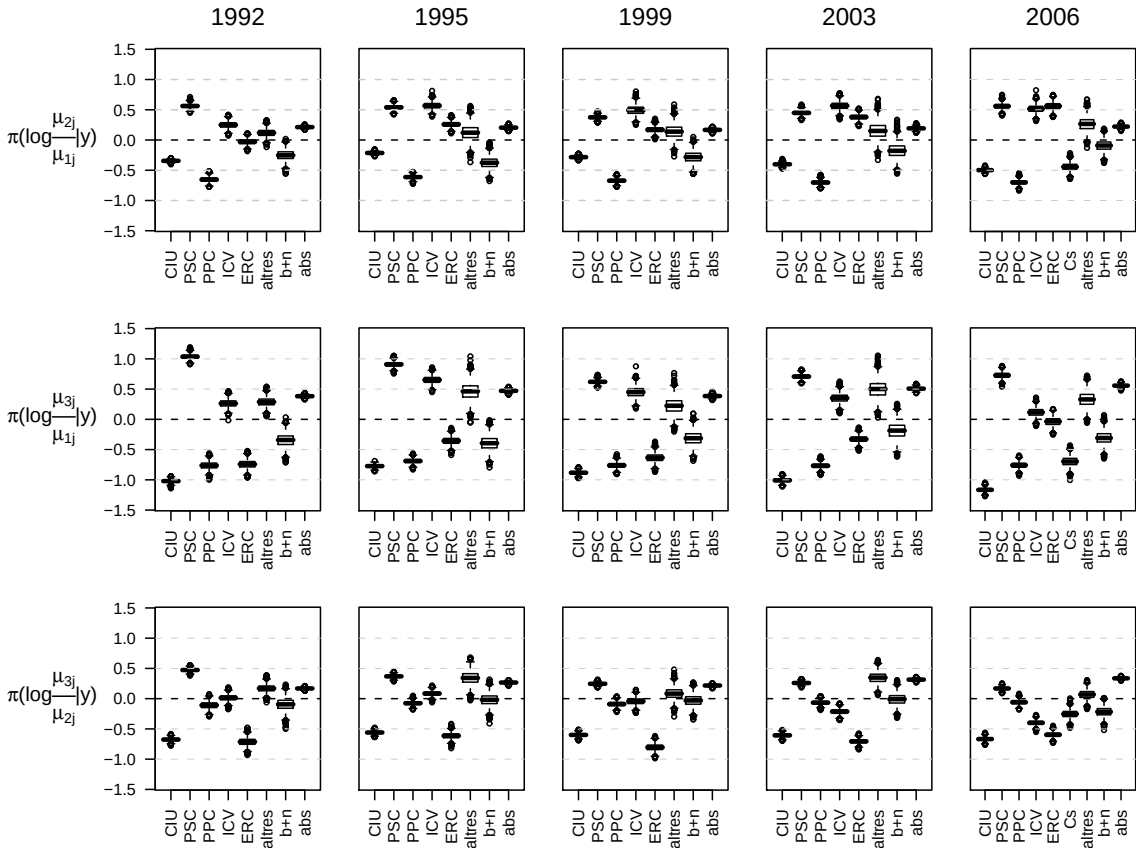

Figura 12.6: Distribució a posteriori de $\log(\mu_{rj}/\mu_{r'j})$ per $r > r'$ sota el model M_{3J}, pels resultats a cada una de les cinc darreres eleccions.

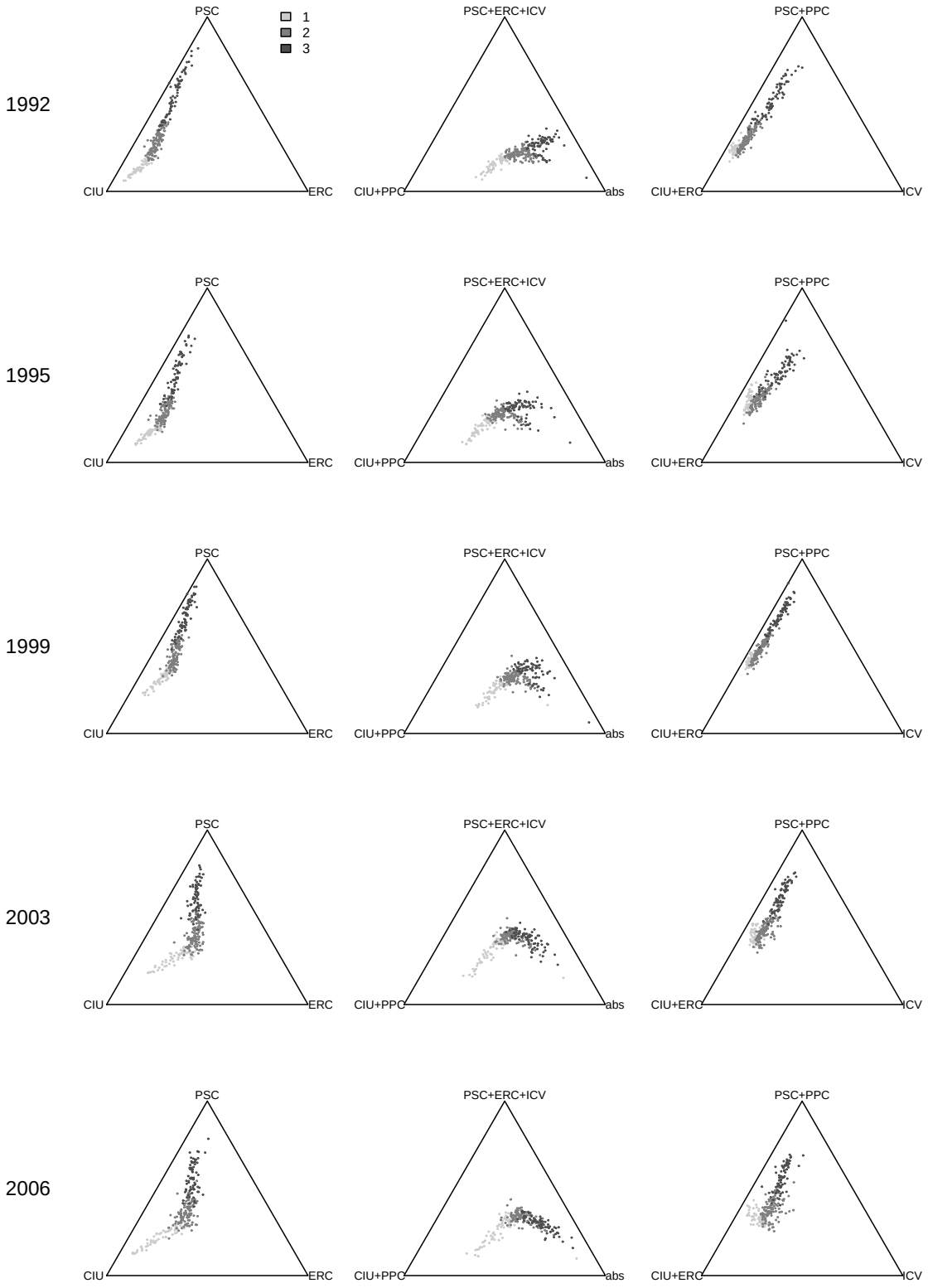

Figura 12.7: Representació ternària dels perfils resultants d'agregar columnes de la taula de dades observades. Cada zrp s'ha colorejat en funció al cluster al que han estat classificades seguint el criteri de la moda a posteriori sota el model M_{3J}.

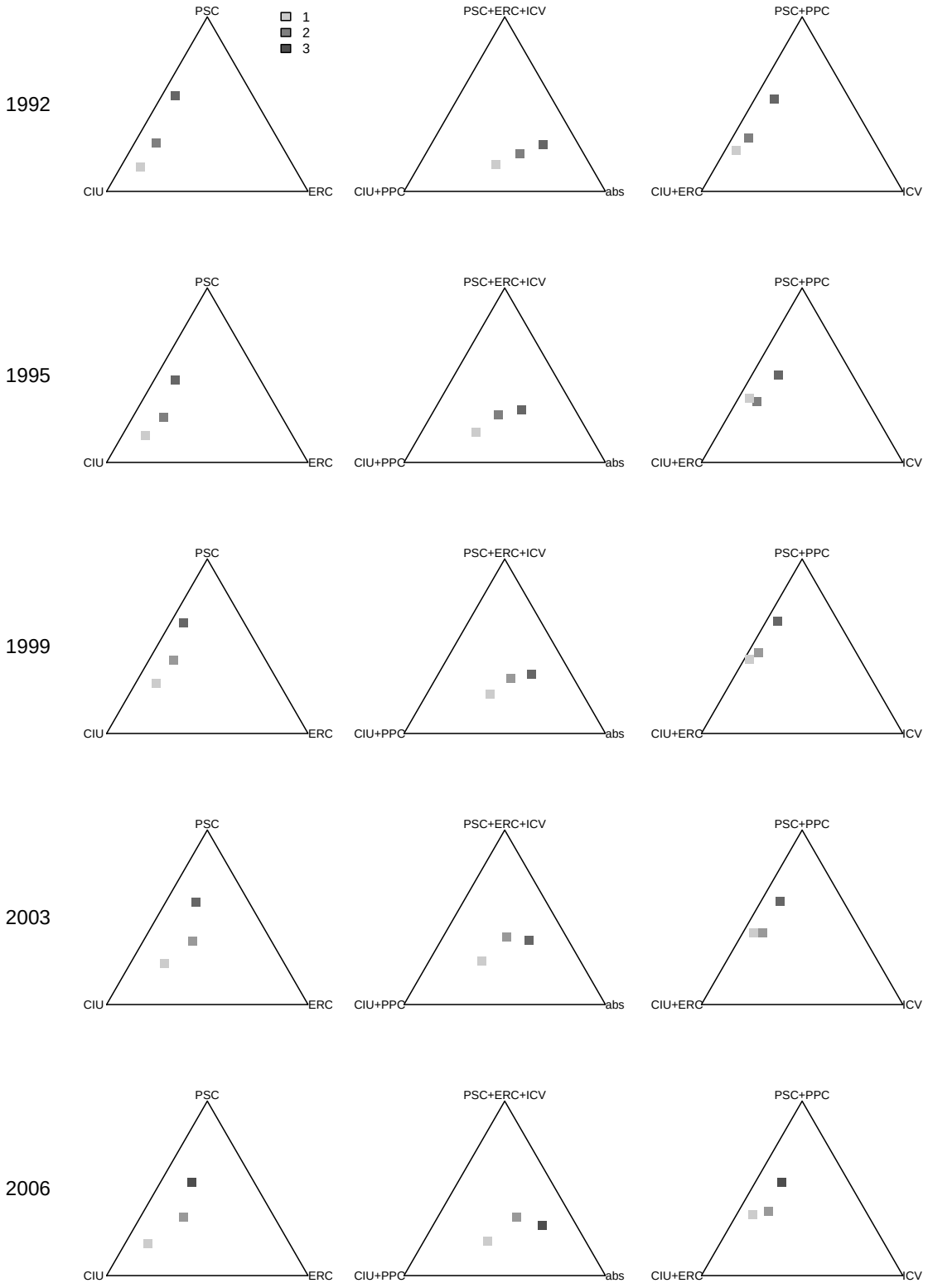

Figura 12.8: Representacions ternàries de (12.1), (12.2) i (12.3) sota el model M_{3J}. Cada punt es pot considerar com el valor més representatiu de cadascun dels tres clusters de la Figura 12.7.

12.2 Resultats del model de quatre clusters: M_{4J}

En aquesta secció presentarem els resultats per al model M_{4J} de la mateixa menera que a la secció anterior ho hem fet per al model M_{3J}.

A la Taula 12.3 es presenten per a cadascun dels anys els valors esperats a posteriori de tots els paràmetres, d'interès excepte de ζ, sota el model M_{4J}. Per facilitar la interpretació dels clusters i la comparació entre anys, s'han ordenat els clusters per ordre decreixent del perfil de vot de CIU. Sota aquest criteri d'ordenació automàticament el PSC, *altres* i l'abstenció queden ordenats en ordre creixent i el PPC i blancs i nuls queden ordenats en ordre decreixent excepte el PPC l'any 2006. En aquesta mateixa taula també es presenten per a cada any el percentatge de vots observats de cada categoria. Si ens fixem en el valor esperat a posteriori del perfil d'ERC observem que sempre obté el valor més alt en el segon cluster, i pel que fa a ICV a les eleccions del 1992 i 1995 obté els valors més alts en el quart cluster, els anys 1999 i 2003 en el tercer i l'any 2006 en el segon.

La distribució a posteriori de la probabilitat de pertànyer a cada un dels clusters, $\pi(\omega_r|y)$ per $r = 1, 2, 3, 4$, les trobem representades mitjançant diagrames de caixa a la Figura 12.9. El cluster 2 és el més majoritari seguit dels clusters 3, 4 i 1. Malgrat mantenir-se aquest ordre a les cinc eleccions, hi ha hagut canvis en les magnituds, així al 1992 el cluster 3 presentava uns grandària molt similar al cluster 2, mentre que en les posteriors eleccions va disminuint progressivament de grandària. De forma complementària el cluster 4 a les eleccions del 1992 presentava una grandària similar al cluster 1 i en les posteriors eleccions aquest darrer cluster ha vist augmentat el seu pes relatiu. La Figura 12.10 presenta la distribució a posteriori de τ_r, $\pi(\tau_r|y)$ per $r = 1, 2, 3, 4$, per al model M_{4J} per a cadascuna de les cinc eleccions. El cluster més homogeni és el cluster 2 i el més heterogeni especialment en les tres darreres eleccions ha estat el cluster 4. Globalment tots els cluster són més heterogenis en les dues darreres eleccions respecte de les dues primeres.

La classificació de cada zrp a cada cluster s'ha fet mitjançant la moda a posteriori de les variables latents ζ_i. Els resultats per al model M_{4J} es mostren a la Figura 12.11. S'observa un gradient espacial, en el sentit de que el canvi d'un cluster a un altre es fa gradualment a mesura que et desplaces per Barcelona, de manera que a Sarrià-Sant Gervasi hi trobem les zrp amb els percentatges més alts de CIU i a mida que te n'allunyes, aquests percentatges disminueixen de forma gradual.

La Taula 12.4 presenta el nombre de zrp que han canviat de cluster en cadascuna de

les 12 possibles direccions. Entre dues eleccions consecutives no s'observa cap zrp que hagi canviat tres clusters de cop passant del cluster 1 al 4 o viceversa i només una zrp, concretament de Ciutat Vella, ha canviat 2 clusters de cop passant de pertànyer al cluster 2 l'any 1995 a pertyàner al cluster 4 l'any 1999.

A la Figura 12.12 es presenten com s'han distribuïts aquests canvis en l'espai i en el temps. El districte de l'Eixample ha vist com algunes de les zrp passaven de pertànyer al cluster 1 a pertànyer al cluster 2. Moltes zrp de Ciutat Vella han passat de pertànyer al cluster 2 a pertànyer al cluster 3. Sant Andreu i Sant Martí han tingut una evolució semblant de manera que entre les eleccions del 1992 i 1995 algunes zrp d'aquests districtes passaven del cluster 4 al 3 i d'altres del cluster 3 al 2, entre les eleccions del 1995 i 1999 algunes zrp passaven del cluster 3 al 4 i del 2 al 3 i que entre el 1999 i 2006 zrp també dels districtes de Sant Andreu i Sant Martí han passat del cluster 4 al 3 i d'altres del cluster 3 al 2. En general han estat més nombrosos els canvis per pujar de cluster que no pas per baixar.

A la Figura 12.13 presentem la distribució a posteriori dels perfils de probabilitat esperats per a cada cluster, $\mu_r = (\mu_{r1}, \ldots, \mu_{r8})$ per $r = 1, 2, 3, 4$ sota el model jeràrquic M_{4J} i a la Figura 12.14 la distribució a posteriori de $log(\mu_{rj}/\mu_{r'j})$ per $r > r'$. Aquestes figures ens permeten esbrinar quines components del perfil canvien d'unes eleccions a les altres.

En el cluster 1 domina CIU inclús per sobre de l'abstenció, aquest és l'únic cluster en que un partit està per damunt de l'abstenció, la segona força política, tot i que a una gran distància ha estat per a tots els anys el PPC.

En el cluster 2 el partit més representat els anys 1992 i 1995 amb diferència era CIU mentre que els anys 1999 i 2003 el PSC va guanyar terreny fet que coincideix amb la presència d'en Pasqual Maragall. També en aquest cluster 2 s'observa com en el 1995 i 1999 PPC era la tercera força política, en el 2003 passa a ser la quarta darrera de CIU, PSC i ERC i l'any 2006 la cinquena força política per darrera també de ICV.

En el cluster 3 CIU durant els anys 1992 i 1995 era la força política més representada i en les posteriors eleccions deixa que ho sigui el PSC. En el 1992 PPC, ICV i ERC estàn igual de representats, i després de lleugeres fluctuacions en els anys intermitjos al 2006 es repeteix l'equitat entre aquests tres partits.

En el cluster 4 hi trobem un domini absolut de l'abstenció i el PSC com a força política més representada.

En la comparació dels perfils de la Figura 12.14 s'observa com el cluster 1 destaca sobre

els altres per tenir sobrerepresetnació de CIU i PPC, el cluster 2 per tenir major ERC, el cluster 4 per tenir major PSC i abstenció, i el cluster 3 està a cavall del cluster 2 i 4, en el sentit que té més representació de CIU que el cluster 4 i menys que el cluster 2, i té més representació de PSC i abstenció que el cluster 4 i menys que el cluster 2.

A la Figura 12.15 representem subconjunts de les dades observades en diagrames ternaris colorejant les zrp segons el cluster al que han estat classificats segons el model M_{4J}. I a la Figura 12.16 s'hi representen per a cada cluster les ternes del valor esperat a posteriori dels perfils de probabilitat esperats resultats d'agregar categories.

A a la primera columna de les Figures 12.15 i 12.16 presenta el pes relatiu dels partits de CIU, PSC i ERC i mostra el progressiu i marcat augment del pes relatiu d'ERC en els clusters 2 i 3, tot i que també però de forma més moderada en zrp dels clusters 1 i 4.

La segona columna de les Figures 12.15 i 12.16 presenta l'evolució del pes relatiu entre la centre-dreta, l'esquerra i l'abstenció i mostra com el cluster 1 s'ha mantingut fidel a la centre-dreta. El cluster 4 tendeix a tenir les zrp amb més alta abstenció, mentre que l'esquerra ha vist augmentar de manera progressiva des del 1992 al 2003 el seu pes al cluster 2 i 3.

Finalment a la tercera columna de les Figures 12.15 i 12.16 s'hi agrupa per una banda els partits nacionalistes catalans CIU i ERC, per una altra els partits d'obediència espanyola PSC i PPC i per una altra ICV. En aquesta columna s'observa com el cluster 1 sempre ha estat el que s'ha mantingut més allunyat d'ICV mentre que la resta de clusters i especialment en les dues darreres eleccions han experimentat un apropament cap a ICV. Si ens fiexem en el vertex catala-espanyol observem com el cluster 4 sempre ha estat marcadament espanyolista mentre els cluster 1 i 2 ho han estat catalanista, i el cluster 3 ha evolucionat de catalanista a espanyolista especialment entre el 1992 i el 1999.

Una darrera anàlisi dels resultats passa per explorar l'evolució de les variables latents ζ_i per a cada zrp en cadascuna de les eleccions i identificar les zrp que han estat sempre fidels al seu cluster, les zrp que oscil·len entre dos clusters o les zrp que estan evolucionant d'un cluster cap a un altre cluster canviant el seu comportament de vot. A l'Apèndix A s'hi presenten les distribucions a posteriori de totes les ζ_i per $i = 1, \ldots, 248$ per a cada elecció, i a la secció 12.3 es presenten i s'interpreten l'evolució de les distribucions a posteriori concretes per algunes zrp sota els models M_{3J} i M_{4J}.

Model	cluster	CIU	PSC	PPC	ICV	ERC	Cs	altres	b+n	abs	$E[\omega\|y]$	$E[\tau\|y]$
1992	1	0.398	0.058	0.088	0.027	0.045	-	0.020	0.014	0.349	0.148	434.77
	2	0.311	0.103	0.043	0.041	0.054	-	0.023	0.010	0.414	0.356	551.31
	3	0.209	0.145	0.037	0.039	0.037	-	0.027	0.009	0.497	0.345	316.28
	4	0.108	0.224	0.029	0.042	0.017	-	0.029	0.008	0.543	0.151	352.24
	dades	0.263	0.130	0.048	0.039	0.042	-	0.024	0.009	0.445		
1995	1	0.356	0.072	0.195	0.036	0.056	-	0.006	0.015	0.264	0.124	392.73
	2	0.307	0.119	0.106	0.063	0.078	-	0.006	0.010	0.312	0.385	487.85
	3	0.216	0.154	0.098	0.069	0.054	-	0.008	0.009	0.390	0.346	233.18
	4	0.126	0.221	0.085	0.079	0.028	-	0.009	0.009	0.443	0.145	255.25
	dades	0.261	0.141	0.115	0.065	0.061	-	0.006	0.009	0.341		
1999	1	0.317	0.144	0.144	0.017	0.051	-	0.008	0.013	0.306	0.121	394.87
	2	0.256	0.203	0.072	0.030	0.068	-	0.009	0.010	0.353	0.385	543.55
	3	0.175	0.259	0.066	0.032	0.044	-	0.009	0.009	0.407	0.290	300.98
	4	0.11	0.283	0.060	0.028	0.022	-	0.010	0.009	0.478	0.204	139.7
	dades	0.216	0.234	0.079	0.028	0.051	-	0.008	0.008	0.376		
2003	1	0.302	0.119	0.182	0.037	0.086	-	0.008	0.011	0.257	0.112	302.20
	2	0.221	0.176	0.090	0.064	0.134	-	0.008	0.009	0.298	0.364	503.68
	3	0.151	0.221	0.086	0.066	0.099	-	0.010	0.009	0.358	0.291	302.61
	4	0.098	0.250	0.081	0.047	0.052	-	0.012	0.008	0.451	0.233	171.40
	dades	0.186	0.204	0.101	0.057	0.104	-	0.008	0.008	0.333		
2006	1	0.311	0.079	0.144	0.050	0.050	0.043	0.011	0.019	0.293	0.116	305.59
	2	0.205	0.133	0.069	0.088	0.093	0.027	0.015	0.018	0.353	0.392	431.99
	3	0.132	0.161	0.072	0.075	0.070	0.026	0.015	0.016	0.432	0.284	348.55
	4	0.084	0.169	0.062	0.049	0.041	0.019	0.016	0.012	0.548	0.209	187.05
	dades	0.178	0.147	0.080	0.072	0.073	0.027	0.013	0.015	0.394		

Taula 12.3: Esperança a posteriori els perfils de la probabilitat esperada per a cada cluster sota el model M_{4J}, $E[\mu_{rj}|y]$ per $r = 1, 2, 3, 4$ i per $j = 1, \ldots, 8$ o 9, del tamany relatiu de cada cluster, $E[\omega_r|y]$, i del grau d'heterogeneitat del cluster, $E[\tau_r|y]$, i el percentatge de vots de cada opció per a cada una de les cinc eleccions considerades.

Canvi	1992-1995	1995-1999	1999-2003	2003-2006
$1 \rightarrow 4$	0	0	0	0
$2 \rightarrow 4$	0	1	0	0
$3 \rightarrow 4$	0	19	0	0
$1 \rightarrow 3$	0	0	0	0
$2 \rightarrow 3$	3	5	8	0
$1 \rightarrow 2$	6	2	2	0
$4 \rightarrow 3$	3	2	3	9
$3 \rightarrow 2$	9	1	1	7
$4 \rightarrow 2$	0	0	0	0
$2 \rightarrow 1$	0	2	0	0
$3 \rightarrow 1$	0	0	0	0
$4 \rightarrow 1$	0	0	0	0

Taula 12.4: Canvis en la classificació de les zrp a cada cluster entre dues eleccions consecutives. Per exemple $1 \rightarrow 2$ indica el nombre de zrp que canvien del cluster 1 al cluster 2 d'una elecció a la següent.

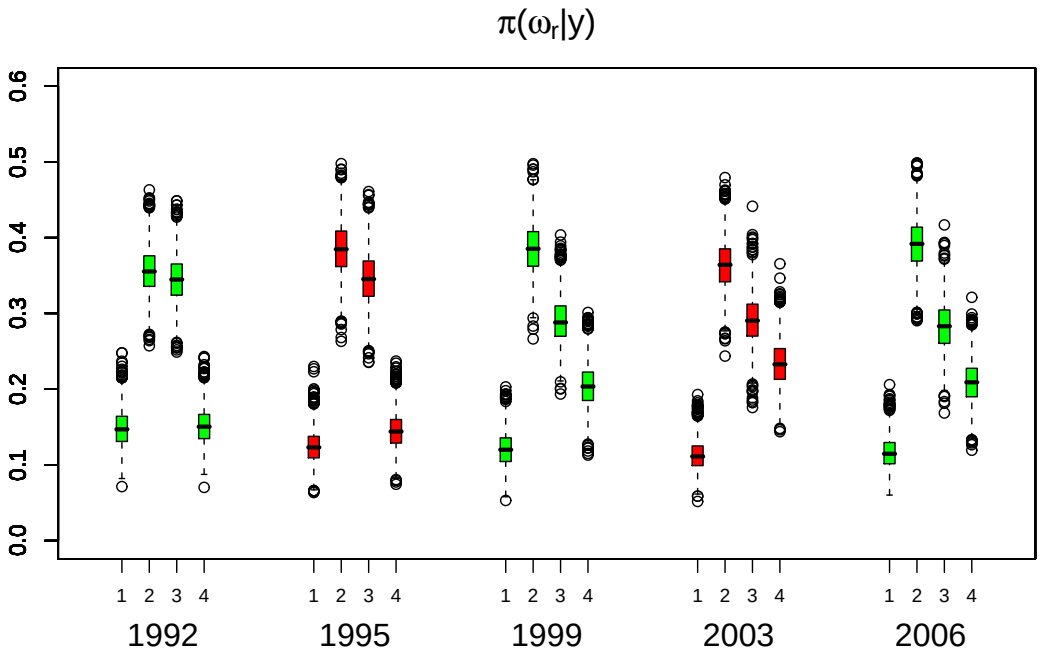

Figura 12.9: Distribució a posteriori de la probabilitat de pertànyer a cada un dels clusters, $\pi(\omega_r|y)$, per $r = 1, 2, 3, 4$, sota el model M_{4J} pels resultats de cada una de les cinc darreres eleccions.

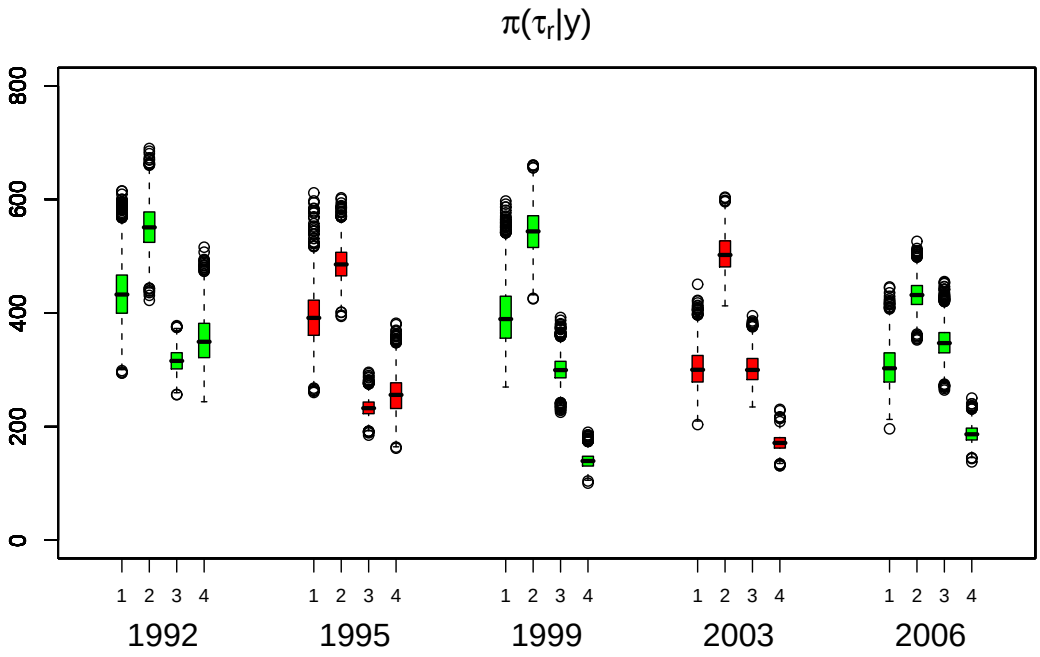

Figura 12.10: Distribució a posteriori del grau d'heterogeneïtat, $\pi(\tau_r|y)$, per $r = 1, 2, 3, 4$, sota el model M_{4J} pels resultats de cada una de les cinc darreres eleccions.

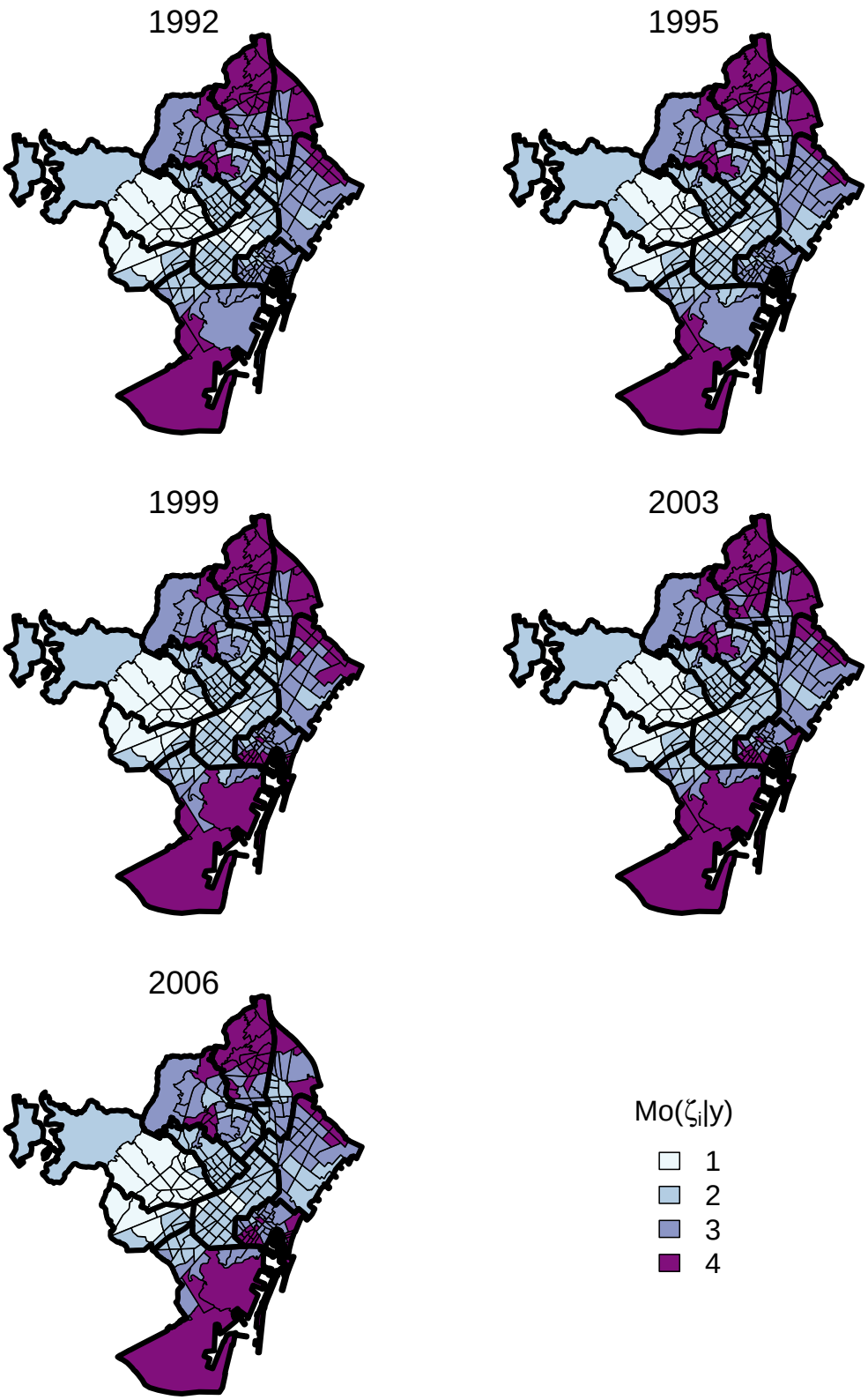

Figura 12.11: Classificació de les zrp utilitzant com a criteri la moda a posteriori de les variables latents ζ_i, $Mo(\zeta_i|y)$, sota el model M_{4J} pels resultats de cada una de les cinc darreres eleccions.

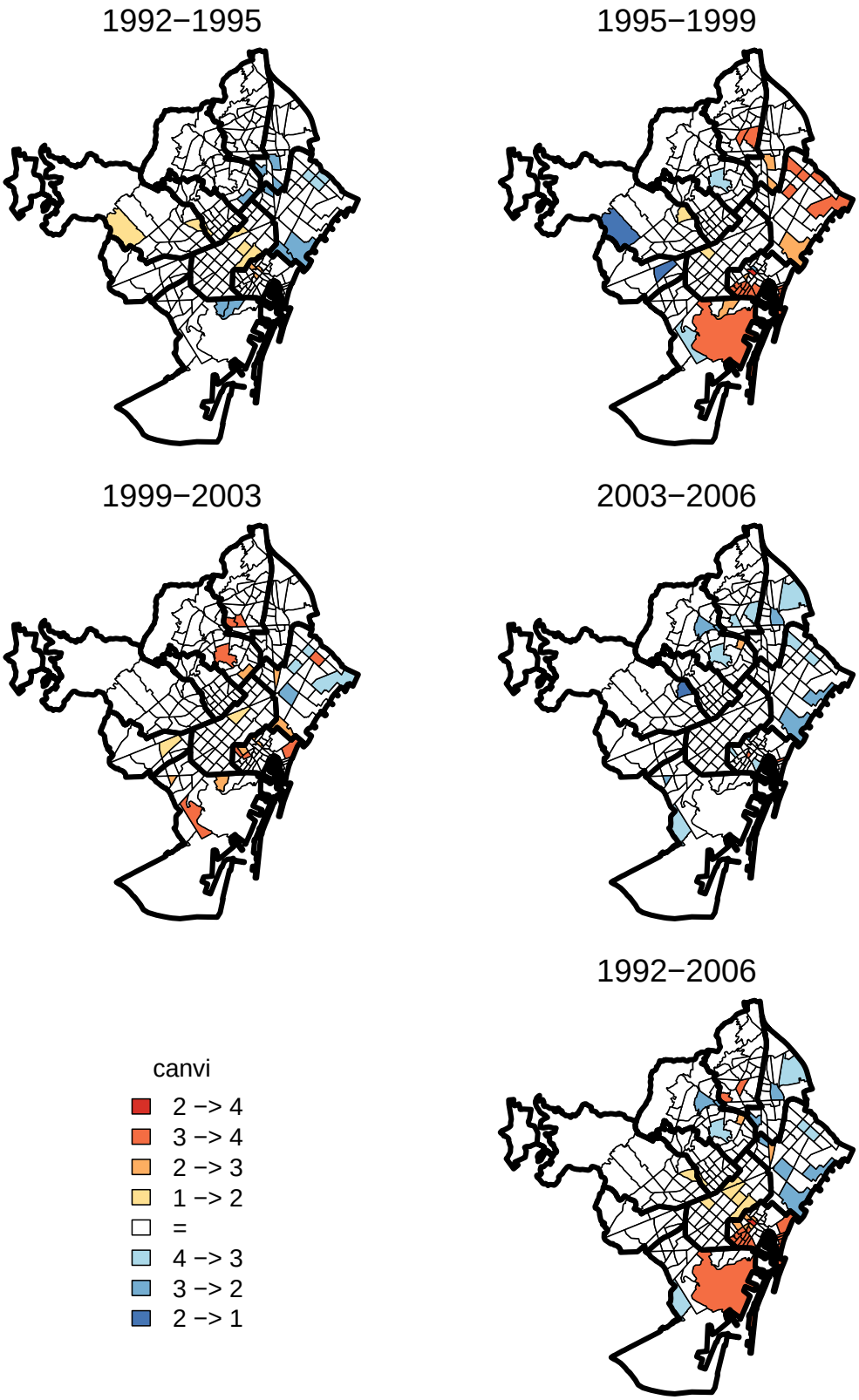

Figura 12.12: Canvis en la classificació de les zrp als quatre clusters entre dues eleccions consecutives, en funció de la direcció del canvi.

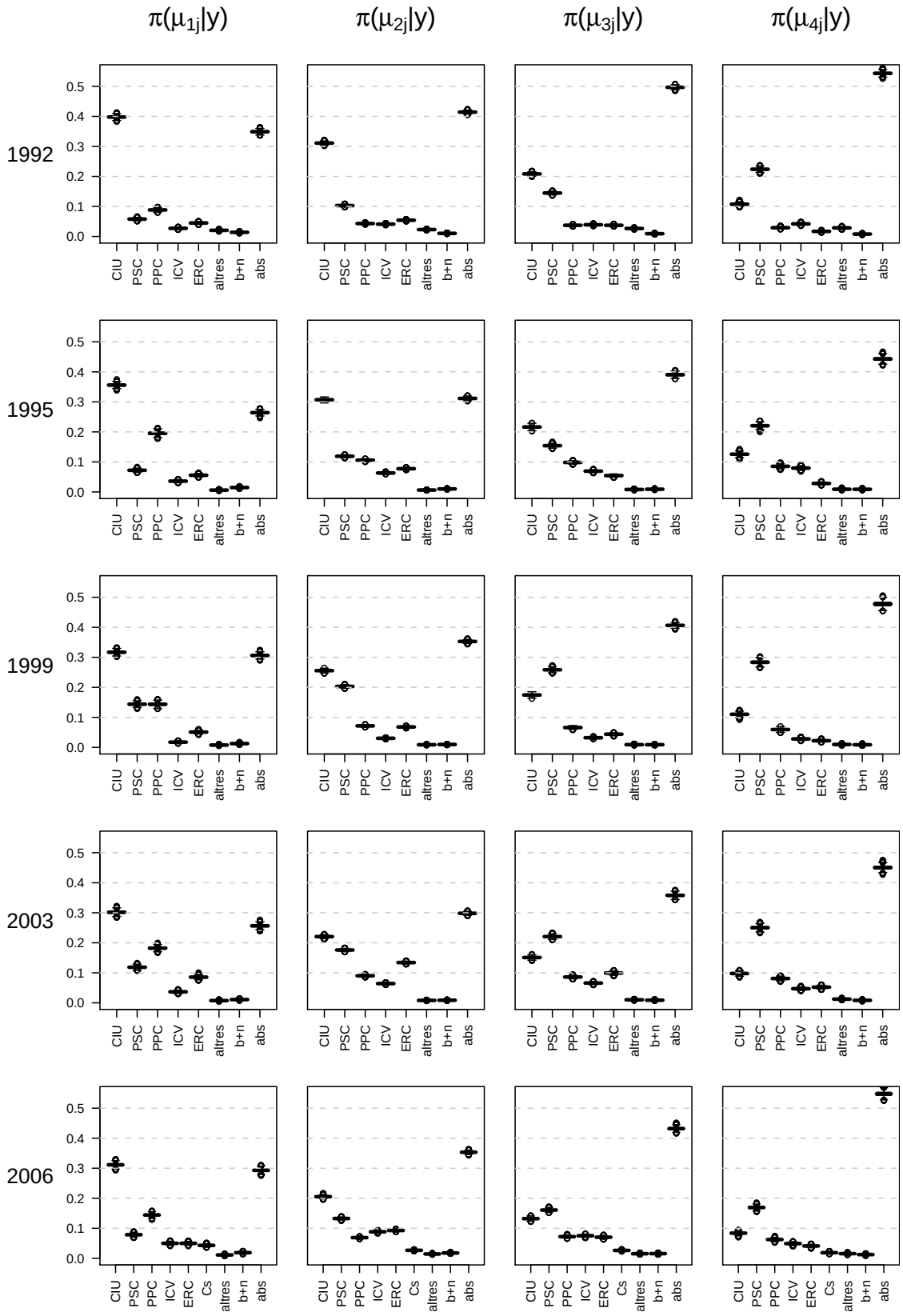

Figura 12.13: Distribució a posteriori dels components del perfil de probabilitat esperada per cada cluster sotal model M_{4J}, $\pi(\mu_{rj}|y)$ per $r = 1, 2, 3, 4$ i per $j = 1, \ldots, 8$ o 9, pels resultats a cada una de les cinc darreres eleccions.

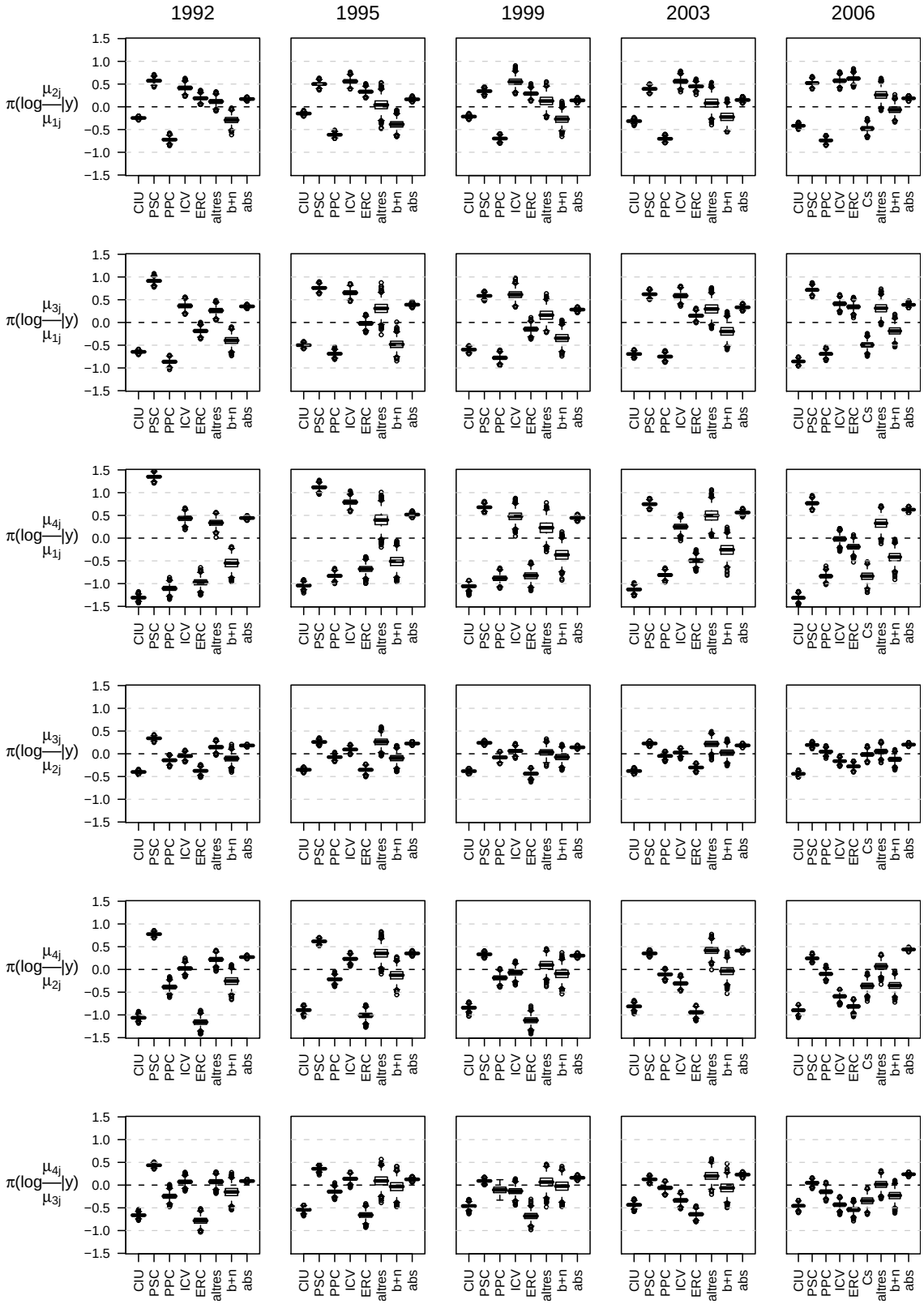

Figura 12.14: Distribució a posteriori de $\log(\mu_{rj}/\mu_{r'j})$ per $r > r'$ sota el model M_{4J}, pels resultats a cada una de les cinc darreres eleccions.

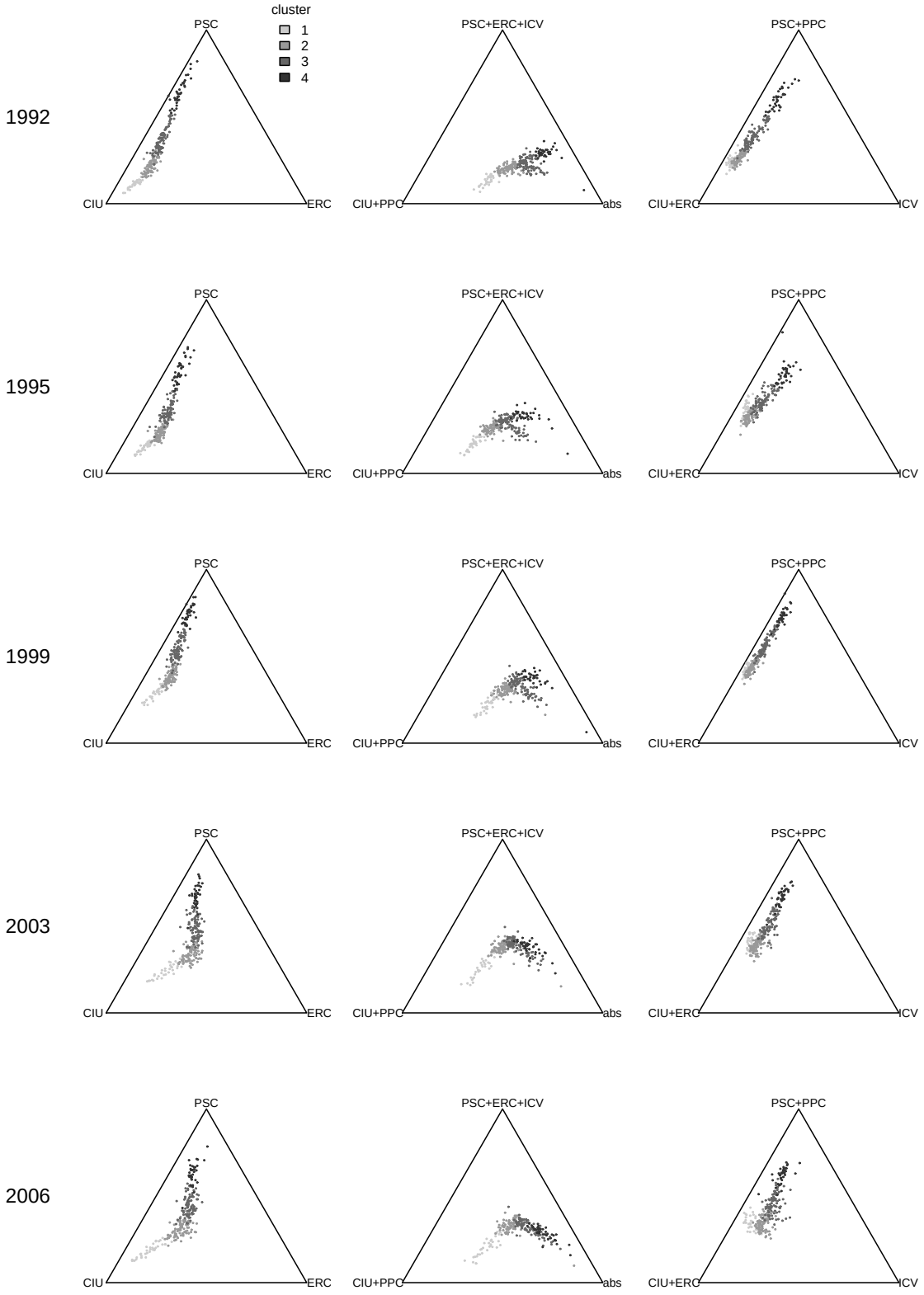

Figura 12.15: Representació ternària dels perfils resultants d'agregar columnes de la taula de dades observades. Cada zrp s'ha colorejat en funció al cluster al que han estat classificades seguint el criteri de la moda a posteriori sota el model M_{4J}.

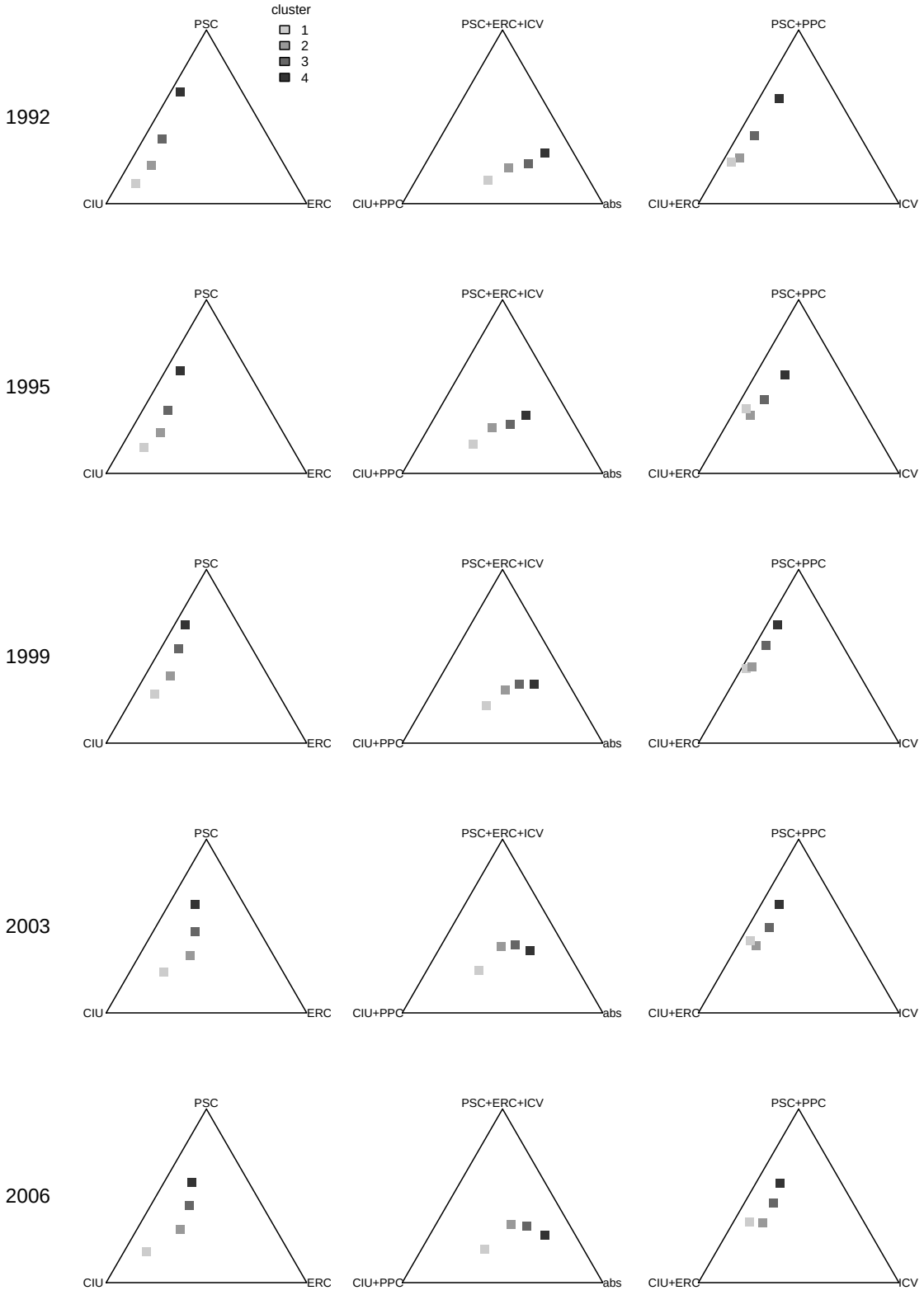

Figura 12.16: Representacions ternàries de (12.1), (12.2) i (12.3) sota el model M_{4J}. Cada punt es pot considerar com el valor més representatiu de cadascun dels tres clusters de la Figura 12.15.

12.3 Comparació dels resultats dels models M_{3J} i M_{4J}

En aquesta secció comparem els resultats obtinguts utilitzant el model M_{3J} amb els resultats obtinguts utilitzant el model M_{4J}, i discutim fins a quin punt el model escollit afecta les conclusions.

S'observa com els valors esperats a posteriori del grau d'heterogeneïtat, $E[\tau_r|y]$, per a cada cluster prenen valors més alts en el model de quatre clusters que en el model de tres clusters. Aquest fenomen era previsible ja que τ_r modela l'heterogeneïtat dels perfils probabilitat de les zrp que pertanyen al cluster r, de forma que a valors més petits més heterogeneïtat, i per tant augmentar el nombre de cluster porta a que les zrp de cada cluster siguin més homogènies.

Analitzant les distribucions a posteriori dels perfils de probabilitat esperada per cada cluster a les Figures 12.5 i 12.13, observem que els perfils del cluster 1 d'ambdós models, M_{3J} i M_{4J}, cada vegada s'assemblen més. De fet a les quatre darreres eleccions són tan iguals que els clusters 2 i 3 del model M_{3J} són els que es trenquen en els clusters 2, 3 i 4 del model M_{4J}.

Tal i com hem etiquetat els clusters el fet de passar del cluster r al $r+1$ està associat a un decreixement dels vots per CIU i PPC i a un creixement de vots per el PSC i l'abstenció, i tant ERC com ICV tenen més percentatge de vot en els clusters intermitjos, és a dir el cluster 2 pel model M_{3J} i els cluster 2 i 3 pel model M_{4J}, que en els clusters extrems.

Sota ambdós models la distribució espacial dels clusters presenta un clar patró en el que el cluster 1 està localitzat al districte de Sarrià-Sant Gervasi i a mesura que et distancies del centre d'aquest districte va augmentant el n^o del cluster. En altres paraules al allunyar-te d'aquest epicentre, disminueix el nombre de votants a CIU i augmenta el nombre de votants al PSC i l'abstenció.

A la pràctica per transmetre els resultats dels perfils de probabilitat esperada que caracteritzen cada cluster, així com les diferències entre perfils de diferents clusters, resulta més còmode i entenedor el model M_{3J}, i és en aquest sentit que creiem que la utilització del model M_{3J} malgrat ser un pel més simplista queda justificada.

No obstant si a part de caracteritzar els clusters ens interessés sobretot el comportament de cada zrp i la identificació de les zrp que han estat sempre fidels al seu cluster, les zrp que oscil·len entre dos clusters o les zrp que estan evolucionant d'un cluster cap a un altre cluster canviant el seu comportament de vot, llavors en aquest cas el model de

quatre clusters permet fer-ne una anàlisi més precisa.

En aquest sentit i a mode il·lustratiu les figures 12.17 i 12.18 mostren l'evolució de les variables latents ζ_i per a una zrp concreta de cada districte sota els models M_{3J} i M_{4J}, aquestes figures ens revelen la distribució de probabilitat a posteriori de cada zrp de pertànyer a cadascun dels clusters. Si ens fixem per exemple en la zrp nº 1, que pertany al districte de Ciutat Vella, llavors a partir del model M_{3J} no identificaríem cap evolució del seu patró de vot, en canvi a partir del model M_{4J} s'evidencia una evolució gradual del seu patró de vot passant de pertànyer al cluster 3 durant les primeres eleccions a pertànyer al cluster 4 en les darreres eleccions.

De la mateixa manera, si ens fixem en la zrp nº 51 del districte de l'Eixample, que inclou l'Hospital Clínic, observem que a partir del model M_{3J} s'intueix un canvi de comportament a l'any 1999, ja que fins a aleshores tota la massa de probabilitat es concentrava en el cluster 1 i en aquest any tot i dominar el cluster 1 comparteix la massa de probabilitat amb el cluster 2, per finalment concentrar tota la massa de probabilitat en el cluster 2 a les eleccions del 2003. Analitzant aquesta zrp a través dels resultats del model M_{4J} ja es detecta aquest inici de canvi de comportament a partir de les eleccions de 1995.

A l'apèndix B s'hi presenta l'evolució de les distribucions a posteriori de ζ_i per totes les 248 zrp sota els models M_{3J} i M_{4J}.

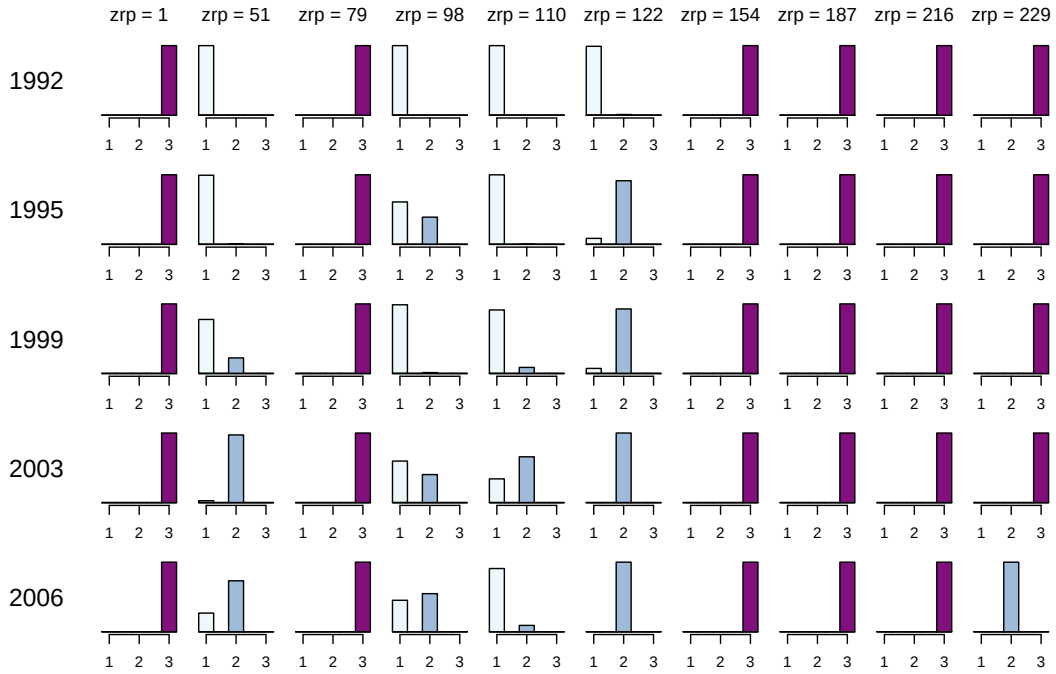

Figura 12.17: Distribució a posteriori de ζ_i, $\pi(\zeta_i|y)$, sota el model M_{3J} per una serie de 10 zrp escollides de forma que n'hi ha una per cada districte.

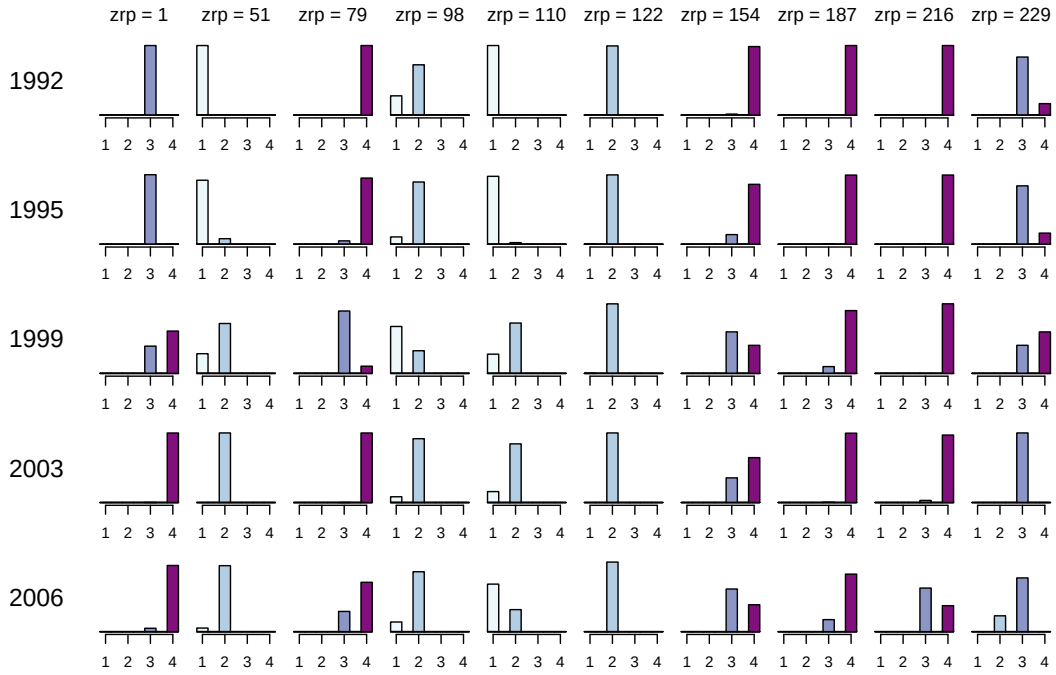

Figura 12.18: Distribució a posteriori de ζ_i, $\pi(\zeta_i|y)$, sota el model M_{4J} per una serie de 10 zrp escollides de forma que n'hi ha una per cada districte.