Essays on Political Science Applications of the Mixture Index of Fit

by

Juraj Medzihorsky

Submitted to

Central European University

Department of Department of Political Science

In partial fulfillment of the requirements for the degree of Doctor of Philosophy

Supervisor: Professor Tamás Rudas

Budapest, Hungary

2015

I, the undersigned [Juraj Medzihorsky], candidate for the degree of Doctor of Philosophy at the Central European University Department of Political Science, declare herewith that the present thesis is exclusively my own work, based on my research and only such external information as properly credited in notes and bibliography. I declare that no unidentified and illegitimate use was made of work of others, and no part the thesis infringes on any person's or institution's copyright. I also declare that no part the thesis has been submitted in this form to any other institution of higher education for an academic degree.

Budapest, 31 August 2015

_____

Signature

Acknowledgments

First and foremost, I would like to thank my advisor Tamás Rudas for his patient guidance. For their help and encouragement, I would like to thank my committee members Gábor Tóka and Jason Wittenberg. For insightful comments on the drafts I am grateful also to [alphabetically] Constantin Manuel Bosancianu, Mihail Chiru, Zsolt Enyedi, Zoltán Fazekas, Pavol Hardos, Martin Mölder, Roland Schmidt, Daniela Širinić, and István Gergő Székely, as well as to the editors and two anonymous reviewers at *Political Analysis*. For helpful discussions I would like to also thank [alphabetically] Andrew Bennett, Colin Elman, Adam N. Glynn, Ákos Horváth, Luis Schiumerini, Jason Seawright, Gábor Tusnády, and the members of the Political Behavior Research Group at CEU. The list would not be complete without acknowledging the invaluable multifaceted help provided by András Bozóki, Levente Littvay, and Carsten Q. Schneider.

Abstract

This thesis proposes new applications of the Rudas-Clogg–Lindsay mixture index of fit and log-linear models that improve inferences in several areas of substantive research in political science. These include problems from electoral research–detection of electoral fraud from digit distributions, allocations of seats according to votes, territorial variability of electoral support and competition, and analysis of voter transitions with aggregate data–as well as analysis of roll call data in the study of legislative politics, and statistical analysis of political text. The improvements are due to the fact that the methods allow to abandon conventional assumptions known to be difficult or false. Most importantly, the mixture index abandons the assumption that the whole population is described by the model. Furthermore, the index also allows to abandon the assumption that the data was stochastically sampled. Log-linear models allow to represent associations in multivariate categorical data without assuming continuity or requiring transformations that produce it. The thesis is accompanied by an R package named `pistar` that implements procedures for the application of the mixture index of fit in a variety of settings.

Table of Contents

List of Tables

List of Figures

Introduction

This thesis proposes new applications of the Rudas–Clogg–Lindsay mixture index of fit and log-linear models that improve inferences in several substantively motivated research problems by handling three methodological issues common in political science–unobserved heterogeneity, non-stochastic samples, and categorical data. Statistical methods conventionally used in political science assume that the whole population is described by the same model. In many settings this homogeneity assumption is at best difficult. Furthermore, the conventional methods assume that the data is stochastically sampled. Yet, in many political science problems the data is better understood as a population. Consequently, substantive inferences can be adversely affected by the use of the conventional methods. The mixture index offers a general framework that can be applied to a variety of models in order to abandon the assumption of homogeneity, and where desired, also the assumption of stochastic sampling. Furthermore, political scientists conventionally apply techniques for continuous data to discrete data, sometimes in settings where more appealing alternatives are available. Log-linear models offer one such alternative.

The main contribution of this thesis is the introduction to political science of the mixture index of fit, accompanied by a package for the R language for its application named `pistar` and made freely available online. The index is applied to five different problems from the domain of political science–electoral fraud detection, seat allocation in elections,

Table 1: The main methodological themes of the thesis across its chapters.

Chapter	Application	The Mixture Index	Log-Linear Models
2	Electoral fraud detection with digit distributions	✓	✓
3	Seat allocation in elections	✓	×
4	Territorial distribution of electoral support	×	✓
5	Voter transitions in aggregate data	✓	✓
6	Roll call analysis	✓	✓
7	Text analysis	✓	✓

analysis of voter transitions from aggregate data, roll call analysis, and analysis of political text. Chapter 4 differs, as it instead applies the dissimilarity index, which in Chapter 2 is introduced within a single general framework with the mixture index of fit. The secondary contribution is the application of log-linear models to represent multivariate associations in categorical data in five of the six research problems. Table 1 shows how the two themes appear through the chapters.

The applications cover different research areas, and some readers might not be equally interested in all of them. For that reason, the material is composed into chapters based on substantive topics as opposed to methodological ones. Each chapter is designed as self-contained, and can be read without having read any of the other chapters. This also facilitates the publication of the chapters as journal articles, as is already the case of Chapter 2 (Medzihorsky, 2015a). To further streamline the text, references to other chapters as well as footnotes are avoided where possible, and non-essential material is removed to appendices. Thus, a reader interested only in one of the investigated substantive problems might read the relevant chapter, and then consult for technical details Chapter 1, which introduces the mixture index of fit and the procedures for its application as well as their implementation in the `pistar` package.

An alternative organizing perspective is provided by the cycle of representative democratic politics. Chapters 2–5 focus on problems related to elections–electoral fraud in Chapter 2, translation of votes into seats in Chapter 3, territorial variability of electoral support and competition in Chapter 4, and voter transitions and loyalty in Chapter 5. The prospective and incumbent elected representatives are the focus of Chapters 6 and 7, the first of which is on the analysis of legislative behavior from roll call data, and the second on statistical analysis of political text.

Chapter 2 improves digit-based electoral fraud detection with the mixture index, the dissimilarity index, and log-linear models. Existing digit-based fraud detection methods assume that the distributions of digits in fraud-free returns are known ex ante, and compare them to the observed distributions. The comparisons are within the null hypothesis significance testing framework, which has several features that render its use in this context problematic. Most importantly, the tests can be sensitive to sample size, they are designed for stochastic samples, which the election results are not, and they require to set the significance level, a task at best very difficult in this context. The mixture index and the dissimilarity index offer more appealing ways to compare the distributions. The indexes rest on assumptions known to be true, are easy to interpret and compute, and allow to ask the quantitative question of how much fraud was there, instead of the qualitative question whether there was fraud. Furthermore, the strong assumption about fraud-free distributions can be relaxed in settings where multiple sets of digits are available for inspection using log-linear models.

In Chapter 3 the D'Hondt method, a widely-used procedure for apportioning seats to parties according to the votes they received is given an interpretation from the perspective of the mixture index of fit. The key to this interpretation is the fact that the quantity minimized by the D'Hondt method, the D'Hondt index, is a simple function of the mixture index. From this perspective, the method splits the votes into two classes–proportionally

represented ones and discarded ones, while maximizing the fraction of the represented votes. This perspective has several advantages. First, it enables a new kind of residual analysis that can inform substantive inferences. The analysis can be applied to seat allocations generated by any seat apportionment procedure, and rests on inspecting the discarded votes. Its use is illustrated on a set of 16 British general elections from 1950–2010. Second, it allows to generalize the index to settings with partially observed vote. Such analysis is illustrated with the example of the 1982 Brazilian federal lower house elections, in which a double-digit percent of votes was declared invalid, although there are arguments that many of the invalidated votes were cast in good faith. Finally, the mixture interpretation of the D'Hondt method corrects several arguments about the method in the literature on electoral formulas and proportionality.

Chapter 4 focuses on the problem known as the measurement of 'party nationalization.' In a body of research, a party is considered as nationalized if it enjoys the same level of electoral support across all the territories of the same level in a country, and a party system is considered as nationalized if this holds for all the (relevant) parties. A variety of measures of party nationalization has been proposed, both on the party and the system levels. Yet, none is consensually adopted. The chapter argues that this is because the distribution of votes over territories considered as corresponding to a nationalized party system is one under which party and territory are independent. However, since there is an infinite number of ways how a bivariate distribution can diverge from independence, a universal measure is not possible. The chapter proposes to use more specific measures, and offers one–the index of residential segregation of voters. The segregation index is a well-known special case of the dissimilarity index. The underlying idea is intuitively appealing–the index captures the smallest fraction of the population that would need to relocate in order for the reality to conform perfectly to an ideal-typical model such as complete integration. This general approach–a model that represents an ideal type, and an easy-to-interpret measure

of distance from the reality to the model–can be applied to other problems from electoral research. The dissimilarity index can be combined with log-linear models to inspect not only the spatial, but also the temporal variability of electoral support, which the chapter demonstrates on Canadian general elections from 2006–2011. The dissimilarity index can also be combined with mixture models to inspect the territorial variability of electoral competition, which the chapter demonstrates on the 2015 British general elections and on Belgian general elections from 1946–1995.

Chapter 5 develops new methods based on the mixture index for the analysis of voter transitions from aggregate data. The first is a measure of voter transitions that rests on splitting the votes into those by voters who always chose the same party in the inspected set of elections and those cast by other voters. Log-linear models allow to take into account any other categorical attributes of elections. This is shown on the example of simultaneous elections for multiple offices, a scenario under which voters can be loyal to a party only within a single batch of elections, or only for a specific office. The chapter demonstrates this approach on the 2004 and 2008 presidential and gubernatorial elections in Montana. Furthermore, the mixture approach can be extended to two popular restrictive models of voter transitions, the 'uniform' and 'proportional' swing, as demonstrated on the 1966 and 1970 British general elections.

Chapter 6 applies the mixture index to a variety of problems in the study of legislative politics that rely on roll call data. The index is used to define an easy-to-interpret measure of partisan voting–the largest possible fraction of votes cast independently of party. The measure applies to any categorical characteristic of the legislators, such as the territory they represent. In case of multiple relevant legislator characteristics, log-linear models can be used to describe the votes, and the mixture index to pick the most fitting description. Furthermore, roll call votes are frequently scaled with a variety of methods to generate legislator positions in a low-dimensional space typically interpreted in policy or ideological

terms. In such analysis, the mixture index can be applied in tasks such as model selection or the detection of differential item functioning. The methods are demonstrated using congressional roll calls related to the Civil Rights Act of 1964.

Chapter 7 introduces the mixture index to text analysis. Statistical text analysis is used in a variety of domains to diverse ends. Its use in political science has been growing considerably over the last decade, following the increase in the amount of political text in machine-readable formats, and the development of new methods. A considerable portion of the methodological innovations originates in computer science. Notably, it includes model evaluation techniques. Computer science applications of statistical text analysis usually prioritize predictive performance, a fact reflected in the model evaluation techniques. However, in political science as well as in some other fields, description and exploration often have priority over prediction. Yet, no corresponding model fit metrics are currently in wide use. Instead, text models are evaluated in political science by the substantive interpretability of the computed quantities. However, relying on models that ex post appear substantively interpretable, but fit the data poorly, might contaminate substantive inferences. In this context, the mixture index of fit is a particularly appealing goodness-of-fit measure, as it captures the model's descriptive performance in an easy-to-interpret quantity, can be applied to a variety of models, and rests on assumptions known to be true. In the chapter, the index is applied to two families of statistical models of text popular in political science, topic models and scaling models, with demonstrations on a corpus of ten U.S. party platforms from 1996 and 2000. Furthermore, a new text-analytic approach is introduced that allows to investigate the associations between document content and other document attributes. The approach rests on log-linear models combined with the mixture index.

The thesis concludes by summarizing the contributions and findings and discussing the limitations of the present work, as well as avenues for future research.

Chapter 1

The Mixture Index of Fit and the `pistar` Package

This chapter discusses the Rudas–Clogg–Lindsay π^* mixture index of fit. The existing procedures for the computation of the index in a variety of contexts are presented. The presentation is accompanied by the introduction of `pistar`, an R package that implements the index for a variety of statistical models.

1.1 Introduction

In model criticism goodness of fit plays a crucial role. Model fit is conventionally assessed by comparing the observed distribution with one expected under the model of interest. Such comparisons rest on measures of distance from the observed to the expected distribution, and take the form of measures of fit or tests of fit. This chapter discusses the Rudas–Clogg–Lindsay π^* mixture index of fit, a latent class measure which quantifies model fit with the smallest fraction of the population outside of the model. As such, the index is related to

other latent class and mixture approaches to model fit, but differs in the fact that it rests on assumptions known to be always true.

The chapter proceeds as follows. Section 1.2 outlines the main issues in the goodness-of-fit evaluation, and discusses latent class and mixture approaches to model fit. The mixture index of fit is discussed in detail in Section 1.3. Section 1.4 discusses the procedures for the computation of the index in the existing literature and their implementation in the `pistar` package. The source of the version of the package used in this thesis is available in an online replication archive at the Harvard Dataverse as Medzihorsky (2015c), and the latest version on GitHub as Medzihorsky (2015b). Techniques for interval estimates of the index, and the generalization of the index to settings with missing data are discussed in Section 1.5. The chapter concludes by a short summary of some existing statistical and substantive applications of the index not discussed in this chapter.

1.2 Model Fit Evaluation

Conventional techniques for the goodness-of-fit evaluation typically rely on comparing the observed distribution to the one expected under the model, and take the form of either a measure or of a test (Rudas, 2002). The former quantify the fit, and the latter give a categorical answer, usually from the $\{\text{reject}, \text{retain}\}$ set. In either case, the procedure rests on a fit statistic, which has the general form of

$$d(\mathrm{o}, \mathcal{M}),$$

where d is a measure of distance from the observed distribution o to the model $\mathcal{M}$. In case of the quantitative assessment the distance is used to evaluate the model.

If, on the other hand, d is used as a test statistic, then its value for the data is

compared to its distribution under a restrictive model. This distribution is either known, or approximated analytically or with simulations. Perhaps the best known such tests are those in the framework of Null Hypothesis Significance Testing (NHST). Under NHST, the statistical significance of the deviation from the null hypothesis is evaluated by comparing the observed value of the statistic to its distribution under the 'null hypothesis' (null model). Typically, the comparison rests on computing the fraction of samples with an equal or larger value of the test statistic if the model is true, also known as the p-value. This value is then compared to a threshold, conventionally one of $\{0.05, 0.01, 0.001\}$, to provide the investigator with a decision to either reject or retain the null model. In this way, a quantitative summary of model fit is used to arrive at a qualitative judgement on the model.

Tests of the kind outlined above can be sensitive to sample size in the sense that they can give different qualitative answers for the same model for two samples with identical probability distributions but different sizes. For that reason, under large samples simpler models can be rejected in favor of more complex ones even if in practice the simpler models meet the practical demands. Just the same, under small samples simple models can be retained even if their fit is far from satisfactory in practice. Crucially, the testing framework rests on the assumption that the data at hand is a stochastic sample, of which at least in principle many can be obtained by the researcher, and it is embedded in the test that it will give the wrong answer in a known fraction of the samples if the null hypothesis is true. A more detailed discussion of this aspect of NHST is given in Chapter 2 in the context of electoral fraud detection from digit distributions.

In both the measure- and test-type procedures, a crucial role is played by the selection of the fit statistic. While there is only one way how two distributions can be identical, there are infinitely many ways of measuring the distance from one of the distributions to the other. Consequently, if a quantitative assessment of model fit is desired as opposed to a test of the type described above, the investigator can choose a measure based on substantive

and operational concerns.

Finally, regardless of whether they take the form of measures or of tests, the conventional goodness-of-fit evaluation procedures rest on the assumption that the model describes the whole population (see e.g. Rudas et al., 1994; Rudas, 2002; Imai and Tingley, 2012). This assumption, known hereafter as *the homogeneity assumption*, is known to be at best difficult in many applied contexts. In the past two decades a body of research has emerged that aims to relax the assumption in the context of model comparison and hypothesis testing using finite mixture (or latent class) models. In this research, the homogeneity assumption is relaxed by representing the model by some, but not all components in a finite mixture model (or classes in a latent class model). The mixture weights of the components of interest are taken for the goodness-of-fit statistics. Some of this research (Imai and Tingley, 2012; Kamary et al., 2014) rests on comparing two models, $\mathcal{M}_1$ and $\mathcal{M}_2$. Formally,

$$\mathrm{o} = \alpha \mathrm{m}_1 + (1-\alpha)\mathrm{m}_2,\ \mathrm{m}_1 \in \mathcal{M}_1, \mathrm{m}_2 \in \mathcal{M}_2,$$

where o is the observed distribution, m_1 and m_2 distributions that belong to the models $\mathcal{M}_1$ and $\mathcal{M}_2$, and α and $1-\alpha$ their mixture weights and serve as the goodness-of-fit measures. Unlike these methods, the mixture index of fit (Rudas et al., 1994; Clogg et al., 1995; Rudas, 1998a, 2002) does not offer only a relative assessment of model fit, but also an absolute one–it decomposes the observations into those described by the model and an unspecified residual component. The next section introduces the index in detail.

1.3 The Mixture Index of Fit

The mixture index of fit is a goodness-of-fit measure which is highly general in the sense that it provides a framework under which any population can be inspected (Rudas, 2002).

It is defined in a latent class context. Specifically, the population is considered as composed of two latent classes, one of which belongs to the model of interest and the other of which is not specified. Formally,

$$\mathrm{o} = (1-\pi)\mathrm{m} + \pi\mathrm{r},\ \mathrm{m} \in \mathcal{M}, \mathrm{r}\ \text{unrestricted}, \pi \in [0,1]$$

where o is the observed distribution, m one that belongs to the model $\mathcal{M}$, r the unrestricted one, and π its size. For any population-model pair there is a size of the out-of-the-model class π for which the description holds true. However, given that the second class is not restricted, the description will always fit perfectly for any π on $[\pi^*, 1]$, where π^* is the smallest such size for which the representation fits perfectly and the value of the mixture index of fit. The index is formally defined as

$$\pi^* = \pi^*(\mathrm{o}, \mathcal{M}) = \inf\left\{\pi : (1-\pi)\mathrm{m} + \pi\mathrm{r},\ \mathrm{m} \in \mathcal{M}, \mathrm{r}\ \text{unrestricted}, \pi \in [0,1]\right\}. \tag{1.1}$$

The index has a straightforward intuitive interpretation as the smallest fraction of the population that cannot be described by the model.

The mixture index has a host of widely-appealing features. Chief among them is that it rests on assumptions known to be always true. Not only does the index abandon the assumption of homogeneity, it does not require to assume that the data is a stochastic sample. Furthermore, the index is not sensitive to sample size in the conventional sense. Consequently, it can be used to compare fit not only across models applied to the same sample, but also across samples of different sizes. Given that the framework is always valid, it always provides validly-defined residuals, and allows for a new kind of residual analysis which can inform substantive inferences.

Table 1.1: A two-by-two table.

	Y=0	Y=1
X=0	a	b
X=1	c	d

1.4 Procedures for the Mixture Index

1.4.1 Two-by-Two Tables

The simplest procedure for the computation of π^* in multivariate context is to models defined by odds ratios in two-by-two tables. The odds ratio is a widely used measure of association in contingency tables. In the context of a two-by-two contingency table shown in Table 1.1 the odds ratio (also known as 'cross product ratio') is calculated as

$$OR = \frac{ad}{bc},$$

(see e.g. Rudas, 1998b). In this context, independence is defined as an odds ratio of one.

Clogg et al. (1995) define a simple algorithm for the calculation of π^* for any model defined as an odds ratio in a two-by-two table. If the observed odds ratio is larger than the desired one, the residual cases are either in cell a or d. Just the same, if the observed ratio is smaller than the desired one, the residuals are in cell b or c. Since both the numerator and the denominator are products of two cell values, in both cases the greatest effect can be achieved by removing cases from the smaller cell value in the product. This procedure is implemented in the `pistar` package in the `pistar.2by2()` function. The user can supply the desired odds ratio as the `alpha` argument of the function.

1.4.2 Models for Contingency Tables

Rudas et al. (1994) define an algorithm for the computation of π^* for any model under which the observed and expected values can be represented as contingency tables of sample fractions. The algorithm combines an Expectation-Maximization (EM) (Dempster et al., 1977) algorithm routinely used to fit latent class models to contingency tables with a binary search (i.e., line splitting) algorithm. The EM algorithm is used to fit the mixture of the model and residual components for any given mixing weights. The binary search algorithm is used to search across the mixing weights until the highest such weight of the model component is found for which the fit of the mixture is still perfect in terms of the likelihood ratio statistic.

An alternative to the use of the binary search algorithm in this context has been proposed by Grego (2010). It rests on subtracting a small positive constant, which can in a given application be considered as practically zero, from the likelihood ratio statistic for the mixture for given mixing proportions, and use a general-purpose one dimensional root-finding algorithm (Brent, 1973) to find the model mixture weight for which the likelihood ratio statistic equals the small positive constant.

Both the Rudas et al. (1994) and Grego's (2010) algorithms are available in the `pistar` package through the `pistar.ct()` function. The algorithm is selected by setting the `method` argument of the function to `"split"` for that of Rudas et al. (1994) and to `"uniroot"` for Grego's (2010). The `pistar.ct()` function can be used with any user-supplied model. The model is supplied to the `fn` argument as a function which inputs only the contingency table with the data, and outputs a list with two elements, `fit` which contains the contingency table with the expected values and `param` which contains other quantities of interest such as model parameters. A computationally more efficient implementation of the Rudas et al. (1994) algorithm specifically for log-linear models is available as the `pistar.ll()` function. The EM and binary search algorithm of Rudas et al. (1994) are also available as self-standing

functions `rcl.em()` and `rcl.s()`, respectively.

1.4.3 Models of Univariate Distributions

A variety of procedures can be used to apply the mixture index to models of univariate distributions. In the discrete case, the univariate distribution can be represented in a contingency table of counts of possible values. Consequently, the algorithm for contingency tables described in the previous subsection can be applied. In the continuous case, the contingency table algorithm can be applied if the observed (and expected) continuous distributions can be discretized with an acceptable loss of information.

The `pistar` package implements in the `pistar.uv()` function different algorithms for the application of the mixture index in the univariate case. These rest on the finding of Rudas (1999) that

$$\pi(\mathrm{o}, \mathcal{M}) = 1 - \frac{1}{\max \frac{\mathrm{m}}{\mathrm{o}}},\ \mathrm{m} \in \mathcal{M},$$

where m is the density under the model and o the observed one. The `pistar.uv()` function minimizes this function using a general-purpose optimization algorithm. The default is the Nelder and Mead (1965) algorithm, but the user can choose other algorithms accessible with the `optim()` function of R. In the continuous case, the densities are discretized into $1 \times K$ contingency tables with kernel density estimation (Scott, 1992), using the Gaussian kernel and a K of 1000 as the defaults.

1.4.4 Independence under Multivariate Normality

As shown by Rudas et al. (1994), in bivariate Normal distribution the π^* mixture index is related to the Pearson correlation coefficient ρ. Specifically,

$$\pi^* = \sqrt{\frac{1 - |\rho|}{1 + |\rho|}}, \tag{1.2}$$

where π^* is the largest fraction of population that can be described by independence under bivariate normality. This finding has been extended by Knott (2005) to the multivariate Normal distribution, using an iterative algorithm for the estimation. The `pistar` package implements both, the analytical solution of Rudas et al. (1994) in the `pistar.bvn()` function and the iterative algorithm of Knott (2005) in the `pistar.mvn()` function.

1.4.5 Linear Regression with Uniform and Normal Error

Rudas (1999) develops the application of the mixture index in linear regression with Uniform or Normal error structure. In this case, the restrictive model of interest is

$$y = X\beta + \epsilon,$$

where y is the vector of observed values, X the design matrix, β the parameter vector, and ϵ the error, which is either distributed normally,

$$\epsilon \sim Normal(0,\ \sigma),$$

or uniformly,

$$\epsilon \sim Uniform(-\delta, +\delta),\ \delta \geq 0.$$

The observed distribution is approximated with the model

$$y = \mu_x + \epsilon,$$

where μ_x is the expected value of y for the given value of x, and just as in the case of the restricted model ϵ is distributed either normally or uniformly.

In both cases, the values of the β under the mixture index can be computed using the

Chebyshev or minimax regression, as shown by Rudas (1999). Under the Normal error, the mixture index is

$$\pi^* = 1 - \frac{\sigma}{s} \exp\left(\frac{-\min_\beta \max_x (X\beta - \hat{\mu}_x)}{2(s^2 - \sigma^2)}\right),$$

where s and $\hat{\mu}_x$ are the estimates of σ and μ_x under the observed distribution. Under the Uniform error, the mixture index is

$$\pi^* = \frac{1}{\tilde{\delta}} \min_\beta \max_{x_i} |\beta x_i - \hat{\mu}_{x_i}|,$$

where $\tilde{\delta}$ is the estimated half-width of the Uniform error distribution in the observed distribution, and x_i the i^{th} row of the design matrix X. The mixture index can be used for model selection under the Uniform error even if there is no estimate $\tilde{\delta}$. Specifically, the fit of two models can be compared with

$$\frac{\delta_1 - \delta_2}{\delta_1},\ \delta_1 \geq \delta_2,$$

where δ_1 and δ_2 are the half-widths of the error distributions. These procedures are not implemented in the current version of the `pistar` package.

1.4.6 Logistic Regression

Verdes and Rudas (2003) apply the mixture index to logistic regression with continuous predictors. The key step in the application is the transformation of the joint densities of the predictors conditional on the observed outcome from continuous into a discrete ones, which allows to apply algorithms for the computation of π^* in contingency table models. To this aim, they apply the Averaged Shifted Histograms (ASH) (Scott, 1992). They find that the procedure is computationally demanding due the use of ASH, and sensitive to the numbers of bins used for the joint histogram. The procedure is not implemented in the

current version of the `pistar` package.

1.4.7 Some Alternative Estimation Procedures

In the context of contingency tables and generalized linear models two alternatives were proposed to the algorithm of Rudas et al. (1994) which aim to be computationally less expensive in application (Verdes, 2002). The first was proposed by Xi and Lindsay (1996) for contingency tables and relies on formulating the estimation as a constrained maximization problem and applying a Sequential Quadratic Programming (SQP) algorithm to solve it. Specifically, for a log-linear model

$$\log m = X\beta, \tag{1.3}$$

where m is the vector of expected values, X the design matrix, and β a parameter vector π^* can be formulated as the following constrained maximization problem:

$$\text{maximize} \sum_r e^{X_{r,\cdot}\beta}, \text{ given that } X\beta \leq \log o,$$

where r indexes the rows of the design matrix X and o is the vector of the observed values. The downside of this approach is the potential dependence on starting values, which requires the investigator to try several different ones (Xi, 1996; Verdes, 2002).

The second is more general, and based on the finding of Rudas (1999) that for any generalized linear model $\mathcal{M}$ and an observed density (or its smoothed version) o the mixture index of fit is

$$\pi^* = 1 - \sup_{\mathrm{m}\in\mathcal{M}} \inf_{\mathrm{supp\ m}} \frac{\mathrm{o}}{\mathrm{m}}.$$

Consequently, for the log-linear model (1.3) it can be considered as the following minimax

problem

$$\frac{1}{1-\pi^*} = \min_r \max_\beta \left\{ \frac{e^{X_{r,.}\beta}}{o_r} \right\},$$

which can be solved by general algorithms for solving minimax problems (Verdes, 2002).

1.5 Additional Topics

1.5.1 Interval Estimation

In some contexts, the investigator might be interested in obtaining interval estimates for π^* to reflect the uncertainty associated with stochastic sampling. In the context of contingency tables, Rudas et al. (1994) propose a procedure to calculate the lower end of a 95% confidence interval by using their algorithm to find the value of π for which the likelihood ratio statistic equals 2.7. This procedure can be applied with functions `pistar.ct()` and `pistar.ll()` by setting their `lr_eps` argument to 2.7 or any other desired value. In the context of independence under bivariate normality confidence intervals can be obtained plugging any of the conventional confidence intervals for Pearson ρ into (1.2).

A more general proposal has been made by Dayton (2003, 2008), who suggested to use the jackknife resampling (Efron, 1982) to obtain interval estimates for π^*. In the `pistar` package, uncertainty estimates for the index as well as for model parameters are available with jackknife in all functions for the mixture index by setting their `jack` argument to `TRUE`.

1.5.2 Generalization to Missing Data

Rudas (2005) generalized the mixture index to settings with missing data in the context of survey data recorded in contingency tables. The generalization applies to unit non-response,

no contact, and coverage error. It rests on the mixture representation

$$(1-\zeta)\mathrm{o}+\zeta\mathrm{u}=(1-\pi)\mathrm{m}+\pi\mathrm{r}, \tag{1.4}$$

where o is the distribution that describes the observed and u the unobserved data, ζ the mixing weight of the latter, m the distribution under the model, r an unspecified residual distribution, and π its mixing weight. The mixture representation (1.4) is not restrictive in the sense that for any problem there are values of ζ and π for which it holds true. Consequently, rather than a model, it is a framework that can be applied to any sample.

For any fixed value of ζ, the 'best case' value of π^* can be computed. If the rate of missingness is not known, the investigator can profile the function

$$\pi\left(\zeta\right)=\min\left\{\pi:(1-\zeta)\mathrm{o}+\zeta\mathrm{u}=(1-\pi)\mathrm{m}+\pi\mathrm{r},\mathrm{m}\in\mathcal{M},\ \mathrm{u},\mathrm{r}\ \text{unrestricted}\right\}$$

by inspecting the values of π^* for a set of ζ values. The $\pi\left(\zeta\right)$ can be also used to find the lowest such ζ for which π^* is zero.

Rudas and Verdes (2015) extended the approach to item non-response, by adding further mixture components to (1.4). For example, if two items A and B are considered, the extended representation is

$$\zeta_{AB}\mathrm{o}_{AB}+\zeta_{A}\mathrm{o}_{A}+\zeta_{B}\mathrm{o}_{B}+\zeta\mathrm{u}=(1-\pi)\mathrm{m}+\pi\mathrm{r}, \tag{1.5}$$

where o_{AB}, o_{A}, and o_{B} are the distributions for those units for which both or only of the items were observed, and ζ_{AB}, ζ_{A}, and ζ_{B} are the sizes of these components such that

$$\zeta_{AB}+\zeta_{A}+\zeta_{B}=1-\zeta.$$

In case the number of items is large, this approach can create computational difficulties, as for Q items there are 2^Q possible observed data patterns. Rudas and Verdes (2015) suggest in such cases to assume identical distributions for some of the components. The computational approach applied by Rudas and Verdes (2015) is to formulate the problem as constrained optimization and apply a general algorithm for solving this type of problems.

1.6 Conclusion

The π^* mixture index of fit is a latent class measure of model fit which abandons the assumption that the whole population is described by the model, and considers the population as composed of two classes, one of which is perfectly described by the model and the other is not. The index rests on assumptions known to be always true, and provides a framework under which any population can be inspected. In this framework, fit is measured by the size of the smallest fraction of the population which cannot be described by the model.

Procedures for the computation of the index in a variety of settings have been proposed in the literature, several of which are implemented in the `pistar` package for the R language. These procedures include, but are not limited to, the algorithm for models fit to contingency tables of sample fractions (Rudas et al., 1994), odds ratios in two-by-two tables (Clogg et al., 1995), independence under multivariate normality (Rudas et al., 1994; Knott, 2005), and models of univariate distributions. Table 1.2 shows the procedures for the application of the mixture index discussed in Section 1.4 and whether they are implemented in the version of the `pistar` package used in this thesis.

In addition to the statistical problems discussed in this chapter, the index has been applied to a variety of statistical and substantive problems not discussed here, including item response models (Rudas and Zwick, 1997; Hernández et al., 2006; Formann, 2006; Revuelta, 2008), Guttman scaling (Tractenberg et al., 2012), latent class analysis (Formann,

Table 1.2: Procedures for the π^* mixture index implemented in the current version of the `pistar` package (0.5.2.2).

Application	In `pistar`
Two-by-two tables	✓
Contingency tables	✓
Discrete univariate distributions	✓
Continuous univariate distributions	✓
Independence under bivariate normality	✓
Independence under multivariate normality	✓
Regression with Uniform and Normal error	×
Logistic regression	×

2000, 2003a,b), and robust statistics (Ispány and Verdes, 2014).

Chapter 2

Latent Class and Log-Linear Election Forensics

Digit-based election forensics typically relies on null hypothesis significance testing, with undesirable effects on substantive conclusions. This chapter proposes an alternative free of this problem. It rests on decomposing the observed numeral distribution into the 'no fraud' and 'fraud' latent classes, by finding the smallest fraction of numerals that either needs to be removed or reallocated to achieve a perfect fit of the 'no fraud' model. The size of this fraction can be interpreted as a measure of fraudulence. Both alternatives are special cases of measures of model fit–the mixture index of fit and the dissimilarity index, respectively. Furthermore, independently of the latent class framework, the distributional assumptions of digit-based election forensics can be relaxed in some contexts. Independently or jointly, the framework and the relaxed assumptions allow to dissect the observed distributions using models more flexible than those of existing digit-based election forensics. Reanalysis of Beber and Scacco's (2012) data shows that the approach can lead to new substantive conclusions.

Note: An earlier version of this chapter has been accepted to *Political Analysis* as Medzihorsky, J. (2015) 'Election Fraud: A Latent Class Framework for Digit-Based Tests.'

2.1 Introduction

Methods for electoral integrity evaluation include statistical ones, some of which focus on digit distributions in electoral returns and are known as 'election forensics.'[1] Digit-based election forensics (DBEF) promises to detect some kinds of electoral fraud by an inexpensive inspection of the electoral returns. Two features typify it. The first is the assumption that the distribution of numerals in fraud-free returns is known and different from that in fraudulent ones, hereafter referred to as the *strong distributional assumption*. The second is the evaluation of elections by testing the statistical significance of the deviation from the null hypothesis that the observed results are fraud-free.

The validity of DBEF has been questioned, mainly due to its distributional assumption. Independently of this, there are issues with the use of null hypothesis significance testing (NHST) detrimental to the usefulness of DBEF. The main contribution of the present chapter is an alternative statistical approach based on latent classes free of these issues. The secondary contribution is an independent method to relax the strong distributional assumption in some contexts. The main finding is that adopting one or both can lead to new substantive conclusions, and provide a new perspective on the enterprise of DBEF.

The core assumptions of DBEF and the criticism leveled at them are summarized in Section 2.2. Section 2.3 critiques the use of NHST in DBEF, and Section 7.6 introduces

[1]The term 'election forensics' has been coined by Mebane (2006a). Other examples of this approach are found e.g. in Mebane (2006b, 2007, 2008, 2010a,b); Mebane and Kalinin (2009); Buttorf (2008); Breunig and Goerres (2011); Pericchi and Torres (2011); Beber and Scacco (2012), and as a part of a more general approach in Leemann and Bochsler (2014). For an overview of electoral integrity evaluation methods and the place of election forensics in this context see Alvarez et al. (2009) and Alvarez et al. (2012).

a latent class based alternative. An independent approach to relax the distributional assumption of DBEF implemented with log-linear models is presented in Section 7.8. Section 2.6 shows, reanalyzing Beber and Scacco's (2012) data, that the proposed methods can lead to new substantive conclusions as well as a new perspective on the enterprise DBEF.

2.2 Distributional Assumptions in DBEF

Existing DBEF methods are based on the assumption that the distribution of numerals in fraud-free vote counts is known and different from that in fraudulent ones. Typically, this distribution is derived from Benford's law (BL), an observation that for certain kinds numbers the frequencies of digits at each position resemble a logarithmic distribution (Newcomb, 1881; Benford, 1938). Under BL the probability of leading digit $d \in \{1, \ldots, 9\}$ is

$$P(D_1 = d) = \log_{10}\left(1 + \frac{1}{d}\right),$$

and of digit $d \in \{0, \ldots, 9\}$ in position $j \in \{2, \ldots, J\}$

$$P(D_j = d) = \sum_{k=10^{j-2}}^{10^{j-1}-1} \log_{10}\left(1 + \frac{1}{10k + d}\right).$$

Not all DBEF uses BL–Beber and Scacco (2012) expect under no fraud uniformly distributed numerals in the last digits of three or more digit numbers. In practice, this is not radically different, since the Benford distribution with increasing digit order approaches uniformity, getting close to it already at the third digit (see Table A.1 in the Appendix).

The strong distributional assumption of DBEF has been subject to several critiques. Since the present chapter proposes a way to bypass these issues in some contexts, they are only briefly summarized here. The relevance of BL for fraud-free vote counts is

argued for using formal and empirical evidence. First, it has been proved that digits in numbers randomly drawn from a random mixture of distributions converge to a logarithmic distribution (Hill, 1995), and observed that many numbers that represent naturally occurring phenomena follow it. Second, in applications such as forensic accounting and scientific fraud detection, BL-based methods have been considered successful.[2] It is contentious that elections, free of fraud or not, share the relevant features with such processes (Deckert et al., 2011). Third, studies that attempt to simulate fraud-free elections show that numerals in the simulated vote counts follow BL (e.g. Mebane, 2006a) or uniformity (Beber and Scacco, 2012). However, there is little evidence that the simulated processes bear sufficient resemblance to their empirical counterparts, and given the lack of well-formed models the value of these simulations is not known (see Deckert et al. 2011 and also Mebane 2011). Fourth, some elections for which there is considerable other evidence of fraud or its absence are evaluated congruently by DBEF (Beber and Scacco, 2012). The value of this validation depends on whether the sample represents typical fraudulent and fraud-free elections, which at best has a large uncertainty attached.

The expectation that BL (or uniformity) will not hold under fraud is argued for using evidence from experiments and simulations. First, a considerable amount of experimental evidence shows that when asked to generate random numbers, humans produce numbers with numeral distributions different from Benford's or uniform (see e.g. Nickerson, 2002; Beber and Scacco, 2012). Since fabrication of electoral returns is a similar process, it is argued, it should show in the distribution of the digits. The experiments have used a range of subjects of different nationalities, ages, and education. Yet, it is not known how well do their findings extend to those with very different backgrounds, or to fraudsters who operate

[2]A brief discussion of digit-based forensics outside of elections is provided by Beber and Scacco (2012). Interested readers can find a comprehensive inter-disciplinary bibliography on Benford's law and its application at the Benford Online Bibliography `http://benfordonline.net/list/alphabetical`.

under different priorities and constraints. Also, there is little to prevent the fraudsters from using simple tools such as dice or pseudo-random number generators. Second, simulated fraudulent elections produce vote counts with numerals that do not follow BL (e.g. Mebane, 2006a). It merits skepticism whether the simulations adequately represent their empirical counterparts.

These issues are crucial for the validity of the existing DBEF methods, and Section 7.8 shows how they can be in some contexts bypassed. Yet, there are also issues related to the statistical techniques used in DBEF that are of equal importance. These are introduced in the following section, and resolved in Section 7.6.

2.3 Statistical Issues in DBEF

Existing DBEF studies differ in the order of the inspected digits, their expected distributions, and the statistics used. Yet, all compare the observed digit distribution to the one expected under no fraud, typically by testing the statistical significance of the deviation from the null hypothesis that the observations are drawn from the expected distribution. The null hypothesis significance testing framework is a venerable statistical workhorse, fruitfully applied to a wide range of problems. However, a large and diverse body of literature finds issues with its features, often repeatedly and independently (see e.g. Ziliak and McCloskey, 2008). Several of these features have within the context of DBEF unappealing effects on substantive inferences.

First, NHST assumes that the data were generated by stochastic sampling. However, elections are unique events that cannot even in principle be repeated, and their returns are better understood as population data. Relatedly, it is embedded in NHST that with a known frequency the test will reject the null hypothesis when it is true (i.e., commit a Type I error). In the context of DBEF it means that some non-fraudulent elections will be

labeled as fraudulent, and regardless of whether fraud is present, the more the investigator tries to uncover it, the more evidence of it she will find. Since allegations of fraud often arise in new democracies and developing countries (see e.g. Norris et al., 2014), and can reflect the biases of actors that make them, this can lead to the confirmation of these biases. Lowering the Type I error rate comes at the price of the power of the test, and its choice should be based on operational concerns. Since fair elections are fundamental to democratic legitimacy, and raising and amplifying unsubstantiated suspicions of electoral fraud can undermine it, setting the rate is at best an extremely sensitive task. In short, NHST is designed for settings where stochastic samples are repeatedly taken from a population, and the Type I error rate is chosen based on operational concerns, neither of which applies well in DBEF.

Third, inferences are limited if the test rejects the model. The test is not informative if the model is not true, because it rests on the comparison of the value of the test statistic with its distribution if the model is true. Furthermore, if all restrictive models are rejected, there are no validly defined residuals. In existing DBEF the restrictive models are of fraud-free elections, and no alternative ones are considered. Thus, if the model of no fraud is rejected, there is no alternative model, and no residuals.

Finally, test statistics have features that can be unappealing in some contexts. Perhaps the best known test statistic, Pearson χ^2, is sensitive to sample size in the sense that it might lead to different conclusions for two samples with identical densities, but different sample sizes. In the context of DBEF this might lead to rejections of models of no fraud with large samples. This has been recognized in the digit based forensic literature and different solutions have been devised.

One group of solutions rests on offering different test statistics (Leemis et al., 2000; Giles, 2007; Tam Cho and Gaines, 2007; Judge and Schechter, 2009). As illustrated by Judge and Schechter's (2009) analysis, these statistics can with the same data on the same level of

statistical significance lead to different conclusions. Simply, alternative test statistics offer different trade-offs in terms of their sensitivity, but do not resolve the above mentioned issues embedded in the NHST framework.

Another group of solutions relies on Bayesian inference. Pericchi and Torres (2011) and Jiménez and Hidalgo (2014) build on the χ^2 test, but use adjusted p-values and compute Bayes factors, obtaining an appealing quantity–the probability of the hypothesis given the data. This approach builds on the χ^2 test and inherits some of its assumptions, and furthermore requires the use of priors. Cantú and Saiegh (2011) use a naive Bayes classifier, trained on synthetic data using significance testing, and generate for each inspected set of digits a posterior probability of being fraudulent. In contrast, the latent class approach introduced in the present chapter does not require the use of priors, does not rely on NHST in any way, and rests on assumptions known to be true.

The above outlined issues stem from the fundamental setup of the null hypothesis significance testing framework, and cannot be resolved by choices within it, such as of a different test statistic or significance level. The next section outlines an alternative framework based on latent class analysis that can replace NHST in DBEF.

2.4 Latent Class DBEF

In a set of electoral returns fraud can affect between 0% and 100% of the reported numbers. Accordingly, the observed distribution of numerals O is composed of a distribution N that belongs to the non-fraudulent class, and a distribution F that belongs to the fraudulent one, the size of which $\zeta \in [0, 1]$,

$$O = (1 - \zeta)N + \zeta F. \tag{2.1}$$

Existing DBEF methods restrict the digit distribution only for the non-fraudulent results, leaving that of the fraudulent ones unspecified. Because of this flexibility, the model in (2.1) describes the data perfectly under values of ζ on $[\zeta_L, 1]$, where ζ_L is the lowest such size of the unrestricted component that will still result in perfect fit. If a non-fraudulent distribution describes the data perfectly, then ζ_L is zero. The value of ζ_L depends on whether to achieve a perfect fit of the no-fraud model the unrestricted component is to be removed or reallocated. In the former case ζ_L is a special case of the π^* mixture index of fit (Rudas et al., 1994) and in the latter of the Δ dissimilarity index (Gini, 1914). Standard measures of model fit assume that a single model describes the whole population. The indexes abandon this idea, and assume that the population is composed of two classes, units for which the model holds true and those for which it does not.

2.4.1 The π^* Mixture Index of Fit

If the perfect fit is to be achieved by removing the unrestricted component, then under the scaled distribution of the non-fraudulent class $(1 - \zeta)N$ the probability of any of the digits is not higher than its observed one. After substituting ζ with π, for any set of observed digits the fit of this model is perfect for all π on $[\pi^*, 1]$, where π^* is the lowest such size that will still result in perfect fit. This quantity can be interpreted as the smallest fraction of the inspected digits that cannot be described as free of fraud. As such, it is a special case of the π^* mixture index of fit (Rudas et al., 1994; Rudas, 1998a, 1999, 2002), a latent class based measure of model fit with wide applicability.

The quantity of interest is π^*, the smallest share of cases for which the model does not hold. It can be understood as a measure of distance from the observations O to the model $\mathcal{M}$, the smallest such π that decomposes the observed density perfectly into an element M

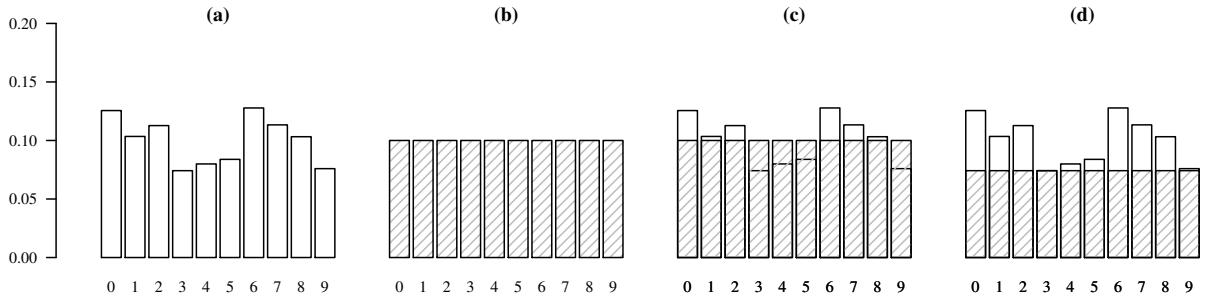


Figure 2.1: Illustration of the mixture index of fit decomposition of an observed set of digits (a) using a discrete uniform model (b). The highest ratio of the model density over the observed density is for digit '3' (c). The resulting latent class model (d) has a π^* value of 0.26. Simulated data.

from the model and an unspecified component U,

$$\pi^*(O, \mathcal{M}) = \inf\{\pi \colon O = (1-\pi)M + \pi U,\ M \in \mathcal{M},\ U \text{ unspecified}\}.$$

Conventional models of fraud-free digit distributions lack free parameters, which eases the application of the index. To obtain the value of the index, the scaled model density $(1-\pi)M$ needs to be 'below' the observed density O while 'shrinking' as little as possible, that is only as much as it needs to fit in the cell where it fits the worst. This is the cell with the highest ratio of the model density over the observed density. The solution is to multiply the model density by the inverse of this ratio. The value of π^* is

$$\pi^* = 1 - \frac{1}{\max\limits_{i=1,\ldots,N} \left\{ \frac{M_i}{O_i} \right\}},$$

as shown more generally by Rudas (1999). The procedure is illustrated in Figure 2.1 with simulated data.

2.4.2 The Δ Dissimilarity Index

If the perfect fit of the model is to be achieved by reallocating some of the observations, then ζ_L in (2.1) is the sum of the absolute values of the residuals divided by twice the sample size. This is because the residuals are symmetric in the sense that the sum of the absolute values of the positive residuals is equal to that of the negative residuals. In this setting ζ_L is a special case of the Δ dissimilarity index (Gini, 1914), defined for a contingency table with cell counts O_i, predicted values M_i, and sample size N as

$$\Delta = \frac{\sum_{i=1}^{N} |O_i - M_i|}{2N}.$$

While the Δ index is not usually presented in latent class terms, in that case the interpretation of the in-model component is different than under the π^* index. Whereas under π^* the unscaled 'in-model' distribution N is perfectly described by the model, it is not under Δ. Instead, the scaled 'in-model' distribution $(1 - \zeta)N$ is interpreted as the largest such fraction that does not need to be reallocated to achieve a perfect fit of the model if the rest (ζF) is reallocated to this aim.

The dissimilarity index can be given a straightforward interpretation in the context of electoral fraud. In an election, usually the returns for each of the options from each territory have to be reported. Thus, also the total number of their last digits is fixed. If the fraudsters substitute a true number by a fraudulent one, its last digit can be changed and thus reallocated. Then, Δ can be interpreted as the smallest fraction of digits that would need to be changed to their presumed original values in order to observe the distribution thought to characterize the absence of fraud.

2.4.3 The Appeal of π^* and Δ in Election Forensics

In the context of digit-based election forensics the π^* and Δ indexes are appealing with a host of features.[3] First, they do not assume homogeneity or stochastic sampling, and consequently their underlying assumptions are always true. Electoral returns are better understood as a population, and this allows to treat them as such. Second, they do not rest on rejecting or retaining a null hypothesis, and do not run the risks of Type I and II errors. Consequently, regardless of how many elections will be inspected, none will be falsely labeled as fraudulent or fraud-free. The amount of fraud will simply be over- or under-estimated for some. In short, the indexes allow to ask the quantitative question of how much fraud there was, instead of just the qualitative one of whether there was fraud.

Third, few model fit statistics have an equally straightforward interpretation and are easy to reason about as do π^* and Δ. Fourth, both indexes are independent of sample size in the sense that for any of them the value is the same for any two datasets with the same observed probability distribution regardless of their size. This allows to compare the fit of the model to datasets of different sizes. Fifth, the units in the unrestricted component can be interpreted as residuals. Thus, unlike in the conventional approach, the residuals are defined in a way that is always valid, and are available for interpretation regardless of how badly does the model fit (see esp. Clogg et al., 1995).

Additional inferential leverage is available under stochastic sampling. The uncertainty attached to sample size can be represented with confidence intervals for the fit statistic, e.g. via jackknife, as shown for π^* by Dayton (2003). From a strict frequentist perspective, if the data are not a stochastic sample, such interval estimates are not meaningful. From a pragmatically Bayesian perspective, given their low cost they can be treated as approx-

[3]These features are discussed in detail in the context of π^* by Rudas et al. (1994); Clogg et al. (1995), and Rudas (1998a, 2002), with the exception of the use of jackknife and the use as a test statistic in NHST.

imations of credible intervals under weakly informative priors. In addition, the indexes can serve as test statistics in NHST using simulated reference distributions. Given the discussed features of NHST this might not be a procedure of first choice.

Finally, the present chapter applies the indexes by decomposing the observations into a component that belongs to a restrictive model and an unrestricted component. The decrease in power due to the use of the unrestricted component is offset by considerably lighter assumptions. Should restrictive models of fraud processes be available, the indexes can be applied either to them, or to mixtures of models of fraudulent and fraud-free processes. In such settings the indexes could no longer be automatically interpreted as degrees of fraudulence, but the other advantages they have over NHST would remain.

2.5 Relaxing the Distributional Assumptions of DBEF

Most existing criticism of DBEF is directed at its strong distributional assumption. This assumption can be relaxed in a way independent of the proposed latent class approach. Specifically, where multiple sets of numerals are available, the investigator can ask whether they can be described by the same probability distribution under the *relaxed distributional assumption* that numerals in fraudulent and fraud-free results are distributed differently. Under this assumption, if a single probability distribution describes all the relevant subsets, either all or none of them are deemed fraudulent. Other evidence can decide which of these interpretations is more appropriate. The ability of such procedures to detect fraud depends on more evidence than the existing DBEF, but with much weaker and less controversial assumptions.

Statistically, such inspections can be done in multiple ways, including log-linear models (see e.g. Agresti, 2002, 314-56), which are appealing due to their flexibility. An intuition can be gained from the following example. In an election suspect of fraud observers were

Table 2.1: Classification of last digits from ward returns by the numeral and whether the ward has been inspected by election observers.

	Observers	
Numeral	Yes	No
0	$d_{1,1}$	$d_{1,2}$
1	$d_{2,1}$	$d_{2,2}$
2	$d_{3,1}$	$d_{3,2}$
3	$d_{4,1}$	$d_{4,2}$
4	$d_{5,1}$	$d_{5,2}$
5	$d_{6,1}$	$d_{6,2}$
6	$d_{7,1}$	$d_{7,2}$
7	$d_{8,1}$	$d_{8,2}$
8	$d_{9,1}$	$d_{9,2}$
9	$d_{10,1}$	$d_{10,2}$

deployed, and reported little evidence of fraud in the visited wards. This data can be represented as a contingency table shown in Table 5.1, where d_{ij} is the count if i^{th} numeral in the j^{th} category. The simplest log-linear model which can applied is

$$\log d_{ij} = \lambda,$$

under which all numerals are equally likely regardless of whether they are from a ward with observers. If the proportion of wards with observers is not 50%, it is useful to introduce the 'observer' parameters λ_j,

$$\log d_{ij} = \lambda + \lambda_j,$$

which is equivalent to expecting the digits to be uniformly distributed within each group of wards, but not necessarily overall. The model of independence for this data is

$$\log d_{ij} = \lambda + \lambda_i + \lambda_j,$$

which allows also the frequency of each numeral to differ by including the numeral parameters

λ_i. Under this model a single probability distribution estimated from the data describes the numerals, regardless of whether they came from the wards with observers. Assuming that fraud manifests itself in digit distributions, and that the observer reports are highly credible, this would temper suspicions of fraud. If the model does not fit well, under the same assumptions the allegations of fraud would appear more credible.

Typically, the situation tends to be more complex than in the above example. Most electoral returns contain at least the following information on the digits of interest–numeral, party (candidate), and who carried a given territory. This allows to use the independence model

$$\log d_{ijk} = \lambda + \lambda_i^D + \lambda_j^P + \lambda_k^R,$$

where λ_i^D are the numeral (digit), λ_j^P the party, and λ_k^R the result parameters. The usefulness of this model is limited as it restricts the numbers of won/lost territories to be the same for all parties. Substantively interesting models will lie between the independence model and the saturated model

$$\log d_{ijk} = \lambda + \lambda_i^D + \lambda_j^P + \lambda_k^R + \lambda_{ij}^{DP} + \lambda_{ik}^{DR} + \lambda_{jk}^{PR} + \lambda_{ijk}^{DPR},$$

which fits perfectly by definition. Substantive inferences can be drawn by comparing the fit of the models that include one or more of the two-factor interaction (association) terms $\{\lambda_{ij}^{DP}, \lambda_{ik}^{DR}, \lambda_{jk}^{PR}\}$. If a model without association terms involving numerals describes the data well, then a single probability distribution fits the numerals under all party-result combinations. The use of this approach is shown in the following section on empirical examples of elections believed to be fraudulent as well as fraud-free.

The above examples deal in accord with the focus of this chapter with strictly descriptive questions of differences between distributions, and by extension of presence of fraud. However, in some settings the log-linear analysis can be extended to handle causal questions.

For instance, considering the first example, conditional on the observer-allocation procedure and other available data, it might be possible to estimate the effect of observer presence on electoral integrity. Log-linear models can be used to test hypothesized causes of fraud, provided the variables of interest are either discrete, or can be discretized with acceptable loss of information.

2.6 Empirical Demonstration

Beber and Scacco (2012) validate their diagnostic procedures using both elections strongly suspect of fraud and elections widely considered fraud-free.[4] This section reanalyzes their data with the proposed approach as well as with a conventional approach, represented by Pearson's χ^2 test, the most common NHST method in the digit-based forensic literature. I opt for the no-fraud model of uniformity, since as Beber and Scacco (2012) demonstrate, it is better theoretically founded than the Benford-based alternatives. Each set of electoral returns is first analyzed under the strong distributional assumption of DBEF and then relaxing this assumption.

2.6.1 Sweden 2002

The Swedish parliamentary elections of 2002 were selected by Beber and Scacco (2012) as an example of elections believed to be fraud-free, and identified as such by them. In the reanalysis, I classify the last digits by numeral, party, and ward result. Detailed results are reported in the Appendix, Section A.3.1. Under the strong distributional assumption both latent class indexes show that uniformity describes the inspected sets of numerals well and

[4]Beber and Scacco's (2012) data are available online as Beber and Scacco (2011). The data used in the analysis, as well as the replication code are available online Replication materials are available online as Medzihorsky (2015c).

thus no suspicions of fraud are raised. NHST leads to the same conclusion, but at the price of more difficult assumptions. Under the relaxed distributional assumption the extent of contamination by fraud seems similar in the inspected subsets. Given other evidence that indicates the absence of fraud, we can conclude that this extent is practically zero, and the detected small departures from the model are due to other causes.

2.6.2 Nigeria 2003

The Nigerian 2003 presidential elections were selected by Beber and Scacco (2012) as strongly alleged of fraud, and the inspected returns (from the Plateau state) were flagged accordingly by their diagnostics. I reanalyze them classified by numeral, party, and polling station result. Section A.3.2 in the Appendix reports the findings in detail. Under the strong distributional assumption the latent class diagnostics lead to a similar substantive assessment as NHST. Under the relaxed distributional assumption it appears that fraud contaminated all inspected party-result subsets of digits roughly equally, since all are relatively close to a common distribution. Considering other evidence on the election presented by Beber and Scacco (2012) this suggests that all inspected returns were affected by fraud and/or clerical errors to a similar extent.

2.6.3 Senegal 2000 and 2007

The Senegalese presidential elections of 2000 and 2007 differ in that the former are reported to be largely free of fraud, and the latter marred with it (Beber and Scacco, 2012). This makes them an especially interesting test case for inspection under the relaxed distributional assumption–if this assumption holds and the reports are accurate, then digit distributions should vary across elections.

Table 2.2: Last digits in election results in Senegalese presidential elections of 2000 and 2007 for the winner (Wade in both cases) and the second-placed candidate (Diouf in 2000 and Seck in 2007). Only numbers with three or more digits included. N=15,172. Source: author's calculation. Data source: Beber and Scacco (2012).

	2000		2007	
	1st	2nd	1st	2nd
0	405	265	822	212
1	386	244	764	198
2	373	222	705	189
3	395	261	761	174
4	409	232	757	195
5	368	233	701	159
6	367	213	714	170
7	395	206	705	156
8	353	209	665	201
9	361	191	689	147

Beber and Scacco (2012) pool the numerals into those from winner's (A. Wade) returns and those from selected other columns, and retain the hypothesis of no fraud for 2000 and reject it for 2007. I inspect also the returns of the runners up (A. Diouf in 2000 and I. Seck in 2007). Table 2.2 reports the digits by numeral, candidate, and year (information on ward result is not provided by Beber and Scacco (2012)).

The fit of uniformity to the inspected subsets of returns is reported in Table 2.3. The χ^2 test rejects at the 5% level the null for winner's last digits in 2000, but not in 2007. However, the sample is much larger in 2007, and the test is sensitive to sample size–if the winner's digits would in 2000 have the 2007 sample size the null would be rejected at the 5% level, and would be retained if in 2007 they would have the 2000 sample size.[5] For the returns of the runners up–not inspected by Beber and Scacco (2012)–uniformity is rejected by the χ^2 test at the 1% level in both elections. A somewhat different picture

[5]For the 2000 density with the 2007 sample size χ^2 is 17 and p-value 0.049 and for the 2007 density with the 2000 sample size χ^2 is 14 and p-value 0.121 (one million simulations from the reference distribution).

Table 2.3: Fit of the uniform distribution to ten subsets of last digits in the Senegalese electoral returns. Reference distributions of the test statistics obtained with one million simulations. Fraction sizes in %. Jackknifed confidence intervals for π^* and Δ. 'Other col.' refers to returns pooled by Beber and Scacco (2012).

		N	χ^2	p	π^*	95% ci	Δ	95% ci
2000	1st (Wade)	3,812	8.89	0.45	7	(0, 17)	2	(1, 4)
	2nd (Diouf)	2,276	22.97	0.01	16	(5, 27)	4	(2, 6)
	First Two	6,088	23.02	0.01	9	(2, 17)	3	(1, 4)
	Registered	8,058	7.18	0.62	5	(0, 11)	1	(0, 2)
	Other Col.	10,334	5.34	0.80	3	(0, 9)	1	(0, 2)
2007	1st (Wade)	7,283	26.82	<0.01	9	(2, 15)	3	(1, 4)
	2nd (Seck)	1,801	24.08	<0.01	18	(6, 31)	5	(3, 8)
	First Two	9,084	37.09	<0.01	8	(2, 14)	3	(2, 4)
	3rd (Dieng)	1,091	15.15	0.09	16	(0, 32)	5	(2, 8)
	Other Col.	2,892	26.54	<0.01	15	(5, 25)	4	(2, 6)

is provided by the π^* and Δ indexes–the distances from uniformity were similar for the winner in both elections (π^* of 7% vs. 9% and Δ of 2% vs. 3%) and for the runner up of 2000 and the second and third candidate of 2007 (π^* of 16%, 18%, and 16%, and Δ of 4%, 5%, and 5%), respectively. These findings can be interpreted in at least two ways. First, assuming uniformity characterizes fraud-free results, fraud appears similarly prevalent in both elections in the returns for the top two candidates. Alternatively, one can abandon this assumption for the case of Senegal in the observed period.

Under the relaxed distributional assumption, if fraud was relatively low in 2000, but much more present in 2007, then the digit distributions should be different for the two elections, even if the uniform distribution does not characterize the fraud-free last digits. This can be tested with a series of log-linear models reported in Table 2.4, starting with the independence model. The second model allows the candidates (defined by their placement) to carry different number of territories across elections, imposing the same probability distribution of numerals on all year-candidate combinations. The following three models allow additional interactions.

Table 2.4: Model fit for five log-linear models fit to the Senegalese data. N=15,172. Fraction sizes in %. Jackknifed confidence intervals.

	χ^2	df	p	π^*	95% ci		Δ	95% ci	
Independence	596.73	28	<0.01	12	(10,	14)	8	(8,	9)
Candidate-Year, Numeral	26.53	27	0.49	4	(0,	8)	1	(1,	2)
Candidate-Year, Numeral-Candidate	13.25	18	0.78	2	(0,	4)	1	(0,	2)
Candidate-Year, Numeral-Year	23.13	18	0.19	3	(1,	4)	1	(1,	2)
Candidate-Year, Numeral-Year, Numeral-Candidate	8.39	9	0.50	1	(0,	3)	1	(0,	1)

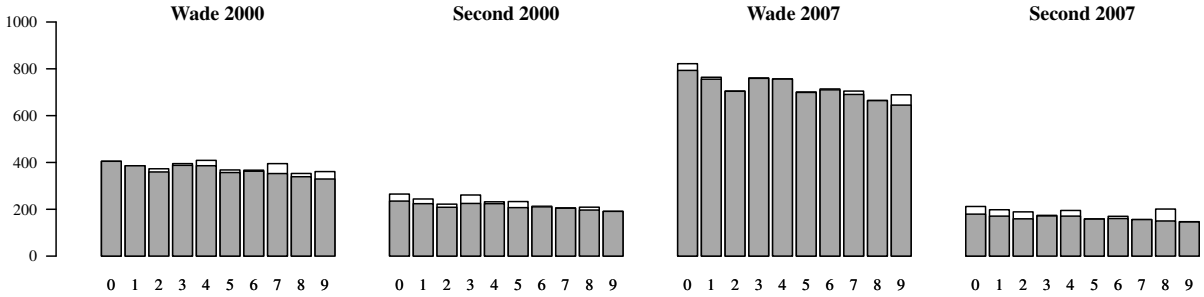


Figure 2.2: Fit under π^* of the second log-linear model (year-candidate, numeral) to the Senegalese data. N=15,172. Model distribution in grey, residuals in white.

The second model is the simplest with near perfect fit–only only 1% of the observations need to be reallocated or 4% removed for perfect it, as shown in Figure 2.2. Under the relaxed distributional assumption both elections seem comparably fraudulent. How to negotiate this inference with the reports of high prevalence of fraud in the second election, but of low one in the first? This might be a fruitful venue for the reanalysis of the reports–it is possible that they underestimate the extent of fraud in 2000 and/or overestimate it in 2007. However, a simpler and more troubling answer is readily available. Namely that–at least for the case of Senegal in the observed period–fraudulent and fraud-free electoral returns are characterized by practically identical probability distributions of last digits. In other words, that the distribution of last digits is not informative with regards to the presence of fraud. This of course means calling into question the enterprise of digit-based election forensics.

2.7 Limitations

The value of the empirical demonstration of the proposed solutions to the problems of digit-based election forensics depends on the degree to which the inspected electoral returns represent typical fraudulent and fraud-free elections. This degree has at best a very large uncertainty attached. For that reason, it is common in the literature to validate the methods also using simulations (see Mebane, 2006a; Deckert et al., 2011; Beber and Scacco, 2012). As discussed in Section 2.2, the value of this is not known. I report the performance of π^* and Δ in such simulations in the Appendix, and do not claim them to be a validation of the methods. Moreover, even if a DBEF method would be convincingly shown to be valid for past elections, it can be invalidated by deliberate behavior on the part of the fraudsters. In case they would deem it worth the costs, the fraudsters can adopt a variety of simple tools that will allow them to fabricate numbers with any digit distribution they desire.

2.8 Conclusion

Among the methods for electoral integrity evaluation digit-based election forensics differentiates itself with its promise to inexpensively evaluate elections requiring only the relevant vote counts. This promise might seem too good to be true, and indeed faces serious challenges. The present chapter focuses on whether it can deliver on its promise, and prioritizes solutions that can resolve some of the challenges over the question how well does DBEF perform compared to its alternatives.

Digit-based election forensics relies on the strong assumption that a known probability distribution describes digits in the absence of fraud, but not in its presence. Typically, it evaluates elections with a test of statistical significance of null hypothesis of no fraud based on this assumption. Two independent sets of issues related to the strong distributional assumption and the use of null hypothesis significance testing decrease the usefulness of

DBEF. The present chapter proposes an alternative to NHST free of its issues, and an independent approach that allows to relax the distributional assumption in some contexts. The former is based on decomposing digit distributions into fraudulent and fraud-free components, and the later on comparing multiple sets of digits using log-linear models.

Reanalysis of Beber and Scacco's (2012) data finds that in two instances the proposed methods lead to similar substantive conclusions, based however on less restrictive assumptions. The finding from the final and arguably the most interesting test are different, and can be interpreted as going against other evidence on the case, or as evidence that even the relaxed distributional assumption is inadequate. The second interpretation suggests that the enterprise of digit-based election forensics is not feasible, since digits can easily be distributed the same way in fraudulent and fraud-free results even if the fraudsters do not deliberately attempt this. Choosing between these two interpretations requires additional information. Given that fair elections are crucial for democratic legitimacy, this can be read as a note of caution for the use of digit-based election forensics.

Chapter 3

The Generalized D'Hondt Index

The D'Hondt method is a widely used seat apportionment procedure. The method minimizes the maximum ratio of seats over votes, also known as the D'Hondt index, and sometimes used to evaluate seat distributions produced by other methods. Despite the method's widespread use, some of its properties are not well understood. This chapter shows that the method divides the votes into two classes, one of which is represented proportionally, and the other not at all, while minimizing the size of the unrepresented class, and can therefore be understood in the context of the mixture index of fit. This allows to generalize the D'Hondt index and method to situations with partially observed vote distributions. Moreover, a new kind of residual analysis becomes available that rests on inspecting the unrepresented votes. The residual analysis is illustrated with 16 British general elections from 1950 to 2010, and the generalization to partially observed votes with the Brazilian federal lower house election of 1982.

3.1 Introduction

The D'Hondt method is a widely used procedure for translating votes into seats where proportionality is desired. In this context, proportionality is understood as the equivalence of the probability distributions of seats and votes over some categories, such as parties or territories. The method minimizes the largest ratio of seat share over vote share among the categories. The ratio is also known as the D'Hondt index (Taagepera and Shugart, 1989), and is sometimes used as a measure of deviation from proportionality also for seat distributions generated by other procedures. The main argument of this chapter is that the D'Hondt index is a simple function of the mixture index of fit (Rudas et al., 1994). The main finding is that the D'Hondt method implies a latent class model that splits the votes into two classes, one of which is translated into seats proportionally and the other not at all, while minimizing the size of the latter class. The secondary finding is that the mixture formulation of the index has certain features lacking in the existing formulation and seen as desirable in the literature, as well as some additional appealing features, chief among which is a new kind of residual analysis. The tertiary finding is that the mixture formulation allows to generalize the index so that it applies also in settings where the vote distribution is observed only partially.

The chapter proceeds as follows. Section 3.2 introduces the problem of disproportionality in seat allocations, and the D'Hondt method. Section 3.3 shows the formal relationship between the D'Hondt index and the mixture index of fit and gives the D'Hondt method a mixture formulation. A new kind of residual analysis enabled by the mixture formulation is introduced in Section 3.4, using an example of 16 British general elections from 1950–2010. The generalization of the index to settings with partially observed vote distributions is defined in Section 3.5, demonstrating it on the Brazilian 1982 federal lower house elections. Additional implications of the findings for the research on seat allocation procedures and

the measurement of disproportionality are discussed in Section 3.6. The chapter concludes by summarizing and contextualizing the findings. For simplicity, the chapter refers to the set of relevant categories as 'parties,' but all the arguments apply for any set of categories deemed relevant.

3.2 Disproportionality and the D'Hondt Method

An observed allocation of seats according to votes can be considered as

$$s = f(v, l),$$

where s and v are probability distributions of seats and of votes over a set of parties, l the size of the legislature, and f a seat apportionment function. When choosing f, it is often seen as desirable that the resulting allocation is not far from proportionality. There is only one way how an allocation of seats can be proportional–if for each party its seat share equals its vote share. A necessary condition for a seat apportionment function to deliver proportionality under all vote distributions is that the number of seats equals the number of votes multiplied by a positive integer. Since electorates usually greatly exceed legislatures in size, in most cases proportionality is unobtainable. A variety of methods to minimize and measure disproportionality has been proposed (see e.g. Balinski and Young, 1982; Taagepera and Shugart, 1989; Gallagher, 1991; Monroe, 1994; Taagepera and Grofman, 2003). This is because there is only one way for a seat allocation to be proportional, but an infinite number of ways in which it can diverge from proportionality (Gallagher, 1991).

Any seat allocation procedure that minimizes some measure of distance from proportionality can be considered as a model-fitting procedure

$$f\colon o \xmapsto{\min d(o,m)} m,$$

where d is the loss function, o the observed density (i.e., the vote distribution), and m the seat allocation (i.e., a density which belongs to the model).

Let n^p be the number of parties, n_i^s be the number of seats and n_i^v the number of votes of party $i \in \{1, \ldots, n^p\}$, $l = \sum_i n_i^s$ to be the size of the legislature and $e = \sum_i n_i^v$ to be the size of the electorate. The D'Hondt method (hereafter f_{DH}) is widely used where minimizing disproportionality is desirable, and consists of the following algorithm:

1. For each party $i \in \{1, \ldots, n^p\}$ divide its vote count by $\forall x \in \{1, \ldots, l+1\}$.

2. Pool all the $n^p \times (l+1)$ values from the first step and find the l largest ones.

3. For each party give it as many seats as there are values obtained by dividing its vote in the first step among the l largest values identified in the second step.

The quantity minimized by the method is $\max(s_i/v_i)$. It is sometimes represented as $\max(A_i)$, where A_i is the 'advantage ratio,' which takes the value of 1 if a party is neither 'advantaged' ($A_i > 1$) nor 'disadvantaged' ($A_i < 1$) (Taagepera and Shugart, 1989; Gallagher, 1991). Plainly, the method minimizes the largest ratio of seats to votes among the parties.

Other algorithms for minimizing $\max(s_i/v_i)$ have been proposed and used (Gallagher, 1991). Perhaps the best know is the Jefferson method, which in fact predates the D'Hondt method (see e.g. Balinski and Young, 1982). Since all of these methods minimize the same index, the arguments presented in the rest of this chapter hold for them as well, but for reasons of brevity only the D'Hondt method is discussed, as it is the most widely used one.

3.3 Relationship between π^* and D'Hondt Index

The quantity minimized by the D'Hondt method is

$$\delta = \max(s_i/v_i),$$

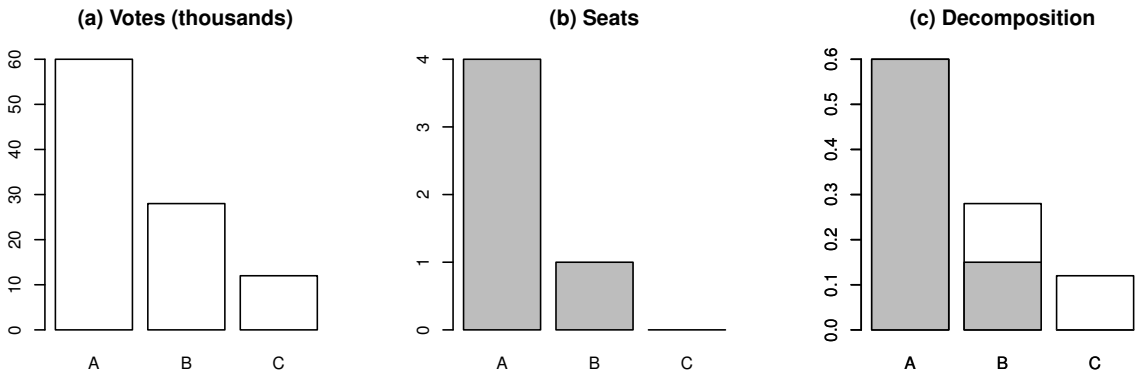


Figure 3.1: The D'Hondt method as 'shrinking' the probability distribution of seats 'below' the probability distribution of votes. Data from Gallagher's (1991, 35) example of five seats distributed to three parties with 60, 28, and 12 thousand votes.

and the D'Hondt method is a procedure for finding such s that minimizes δ. Since both s and v are discrete probability distributions, each sums to one, and the minimum value of $\max(A_i)$ is 1, for $s = v$. Consequently

$$\forall i\colon s_i/\delta \leq v_i,$$

and

$$\sum_{i=1}^{n^p}(v_i - s_i/\delta) = 1 - 1/\delta. \tag{3.1}$$

Perhaps a more intuitive presentation of this relationship is a graphical one. As shown in Figure 3.1, the method 'shrinks' s so that it fits 'below' v, but only as much as necessary to achieve this. From this perspective, the D'Hondt method splits the votes into two classes, those represented proportionally and those not at all, while minimizing the size (3.1) of the latter class. Thus, the D'Hondt method is related the mixture index of fit, defined as

$$\pi^*(o, \mathcal{M}) = \mathsf{inf}\{\pi\colon o = (1-\pi)m + \pi r,\ \pi \in [0,1],\ m \in \mathcal{M},\ r\ \mathsf{unspecified}\},$$

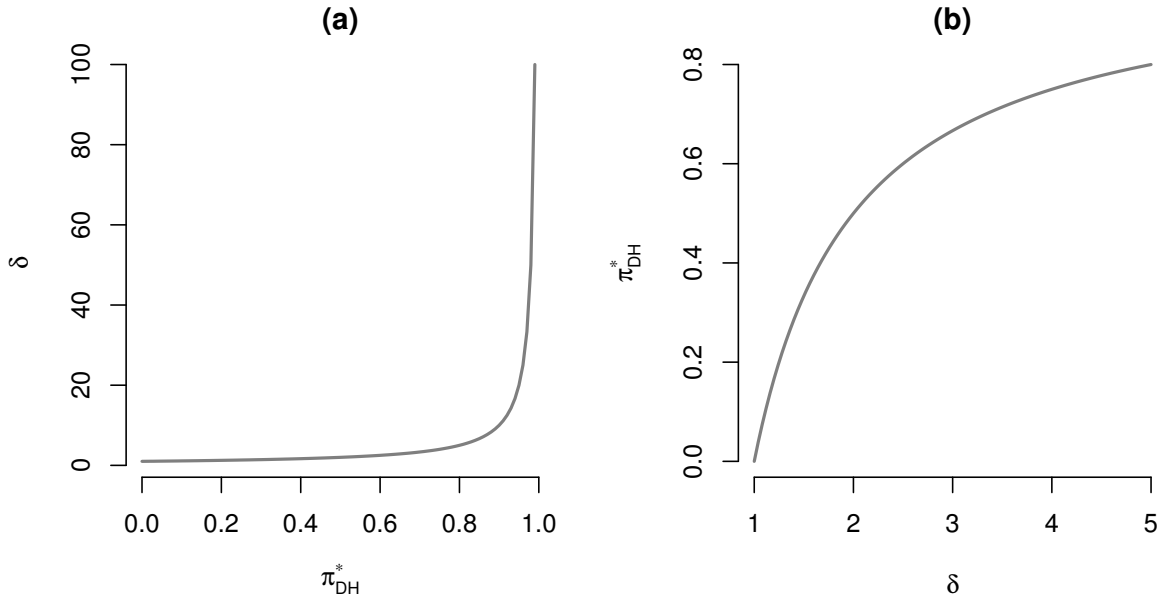


Figure 3.2: The relationship between the D'Hondt δ and π^*_{DH}, δ as a function of π^*_{DH} (a) and π^*_{DH} as a function of δ (b).

where o is the observed distribution, m an element from a model $\mathcal{M}$, and r a residual component (Rudas et al., 1994; Clogg et al., 1995; Rudas, 1999, 2002). The π^* mixture index measures the fit of a model with the smallest fraction of the population not described by the model. In the case of the D'Hondt method, the model is proportionality. A general formulation of the relationship between the ratio of model density over observed density has been given by Rudas (1999).

To detail the relationship between the D'Hondt index and the mixture index of fit, consider that for a discrete distribution the mixture index of fit is

$$\pi^* = 1 - 1/\max(m_i/o_i),$$

where m is the probability under the model, and o the observed one. Substituting s for m

and v for o gives $\max(s_i/v_i)$ which is δ and the D'Hondt π^* can be defined as

$$\pi^*_{DH}\colon \pi^*(v,s) = 1 - 1/\delta, \tag{3.2}$$

of which the D'Hondt index is a simple function

$$\delta = \frac{1}{1-\pi^*_{DH}},$$

shown in Figure 3.2. The following two sections use these insights to present a new kind of residual analysis made possible by the mixture formulation of D'Hondt, and a generalization of the D'Hondt π^* for partially observed vote.

3.4 Residual Analysis: Whose Votes were Discarded?

One of the appealing features of the π^* mixture index of fit is the residual analysis it enables (Rudas et al., 1994; Clogg et al., 1995, 1997). In the context of D'Hondt, π^* residuals are defined as

$$r_i = v_i - \pi^*_{DH} s_i,$$

and interpreted as votes for i^{th} party which were not translated into seats under proportionality as a fraction of total votes. These residuals can be transformed into fractions of their party's votes

$$w_i = r_i/v_i \tag{3.3}$$

and interpreted as party's share of 'discarded votes,' that is votes which were not proportionally translated into seats. Since by definition at least one party is represented perfectly proportionally under the D'Hondt method, at least one of the party-weighted residuals w_i will be zero.

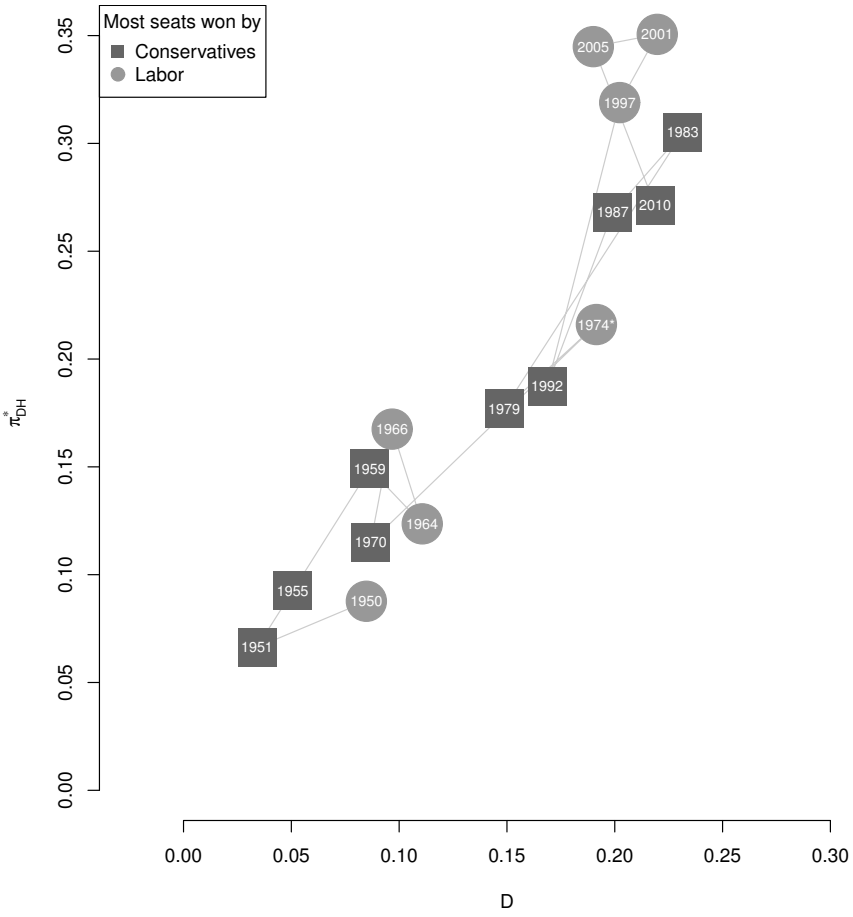


Figure 3.3: Disproportionality of seat allocations in 16 British general elections from 1950–2010 (from 1974 only the October election is included) as measured by the dissimilarity index (x-axis) and π^*_{DH} (y-axis). Source: author's calculation. Data source: compiled by Calvo and Rodden (2015) from Caramani (2000) and *The Guardian*.

As an illustration, an analysis of this kind is performed on 16 British general elections from 1950 to 2010. Figure 3.3 shows the values of π^*_{DH} for these elections against the values of a special case of the dissimilarity index (Gini, 1914) known as the D index of distortion (Loosemore and Hanby, 1971),

$$D = 1/2 \sum_{i=1}^{n^p} |v_i - s_i|.$$

A straightforward interpretation of D is the smallest fraction of seats that would need

to be redistributed given the same distribution of seats over parties in a legislature equal in size to the electorate in order to achieve proportionality. Taagepera and Grofman (2003) identify D (together with the Gallagher's (1991) index) as possessing the largest number of features considered as desirable by them. Some of the features, such as lying on the $[0, 1]$ interval, are shared by π^*_{DH}, which makes it an appealing comparison. However, these measures capture different concepts of disproportionality.

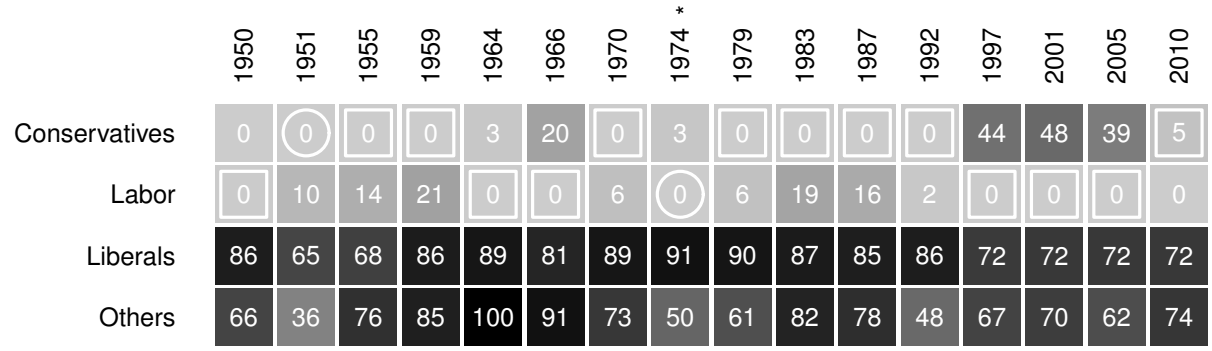


Figure 3.4: Discarded vote percentages (π^* residuals) in 16 British general elections (from 1974 only the October election is included). Rectangular frames indicate parties with the largest shares of both seats and votes, circular ones parties with the largest seat shares but without the largest vote shares. Source: author's calculation. Data source: compiled by Calvo and Rodden (2015) from Caramani (2000) and *The Guardian*.

Figure 3.4 presents the party-weighted residuals (3.3) for the inspected elections. While by definition at least one party will not be underrepresented at all, the 1950 elections illustrate that this can be the case for multiple parties. Also, while for most of the inspected elections the party with the largest seat share was not underrepresented, the 2010 elections illustrate that this is not necessarily the case.

3.5 Generalization to Partially Unobserved Vote

The mixture index of fit has been generalized to missing data, both for known and unknown rates of unobserved units (Rudas, 2005; Rudas and Verdes, 2015). The generalization allows to define a more general index of disproportionality, π^*_{GDH}, of which π^*_{DH} is a special case. Under the generalized index the distribution of votes over parties can take a different meaning, and denote not the observed vote, but a hypothetical vote under different conditions. Examples of such conditions are lack of procedural errors by the voters, wider suffrage, mandatory voting, or perfect access to polling stations. The definition of the unobserved vote might differ in various research contexts–it might consist for instance of invalid ballots, abstentions, or legally disenfranchised populations. In all such cases, the precision with which the rate of missing observations might be known varies depending on the available information. Thus, it might be of interest to inspect the value of the index not only for a fixed rate of missing observations, but also for a wider set of rates.

The generalization is based on the fact that for any electorate it is true that

$$h = (1 - \rho)o + \rho u,$$

where h is the distribution of votes over parties in case all voters of interest–a category which can be given different definitions under different contexts–would vote, o the observed one, u the unobserved one, and ρ the fraction of unobserved units. In the case of π^*_{DH}, ρ is assumed to be 0, and h equal to v. Lifting this assumption allows to define the generalized index

$$\pi^*_{GDH}(o, s, \rho) = \mathsf{inf}\{\pi \colon (1 - \pi)s + \pi r = (1 - \rho)o + \rho u,\ \pi \in [0, 1],\ \rho \in [0, 1]\}. \tag{3.4}$$

The generalized D'Hondt π^* is a 'best case' value of π^*_{DH} for a given unobserved vote rate ρ,

and can be computed using the following algorithm. Starting from the observed distribution of votes, in as many steps as there are missing votes, update the vote distribution by adding a vote to a party so that it produces the lowest $max(s_i/v_i)$ of all possible options.

A related quantity, one that corresponds to 'worst case' scenario, can be obtained by adding the rate of missingness to the 'discarded' component, resulting in the 'worst case' value

$$\pi^*_{DH}(1-\rho)+\rho.$$

Furthermore, it is possible to simulate various scenarios that lie in between the above two, and compare the observed π^*_{DH} statistic and/or residuals to their distribution under these scenarios.

In case the rate of missing data is not fixed, it might be of interest to inspect the values of π^*_{DH} for a range of missing data rates. Taking advantage of the fact that with increasing rate of missing votes π^*_{DH} can only decrease, this can be done by exploring the values of π^*_{DH} over a grid of ρ values (Rudas, 2005).

The remainder of this section illustrates the use of the proposed procedures on the example of Brazilian 1982 elections to the lower house of the parliament. The election results are shown in Table 3.1. Brazil has regularly experienced double-digit percents of invalid votes in federal parliamentary elections from 1962 to 1998 (see e.g. Nohlen, 2005). A large fraction of these invalid votes is a result of incorrect invalidations as well as voting procedures challenging for low information voters (Hidalgo, 2010). In both cases, the voters tried to cast valid votes in good faith, but failed. Thus, the invalid votes can be considered as unobserved.

Table 3.1: Brazilian lower house elections of 1982–vote and seat distributions, and a hypothetical seat distribution under pure D'Hondt. Percents rounded. Source of votes and seats is Nohlen (2005), distribution under pure D'Hondt calculated by the author.

	Votes		Seats		D'Hondt	
	Count	%	Count	%	Count	%
PDS	17,775,738	43	235	49	207	43
PMDB	17,666,773	43	200	42	206	43
PDT	2,394,723	6	23	5	28	6
PTB	1,829,055	4	13	3	21	4
PT	1,458,719	4	8	2	17	4
Invalid	7,330,871	15				
Valid	41,125,008	85				
Reg	58,871,378	–				
Cast	48,455,879	82				

The D'Hondt π^* for the observed seat allocation is 0.12. In other words, 88% of the votes was translated proportionally, and the remaining 12% discarded. Table 3.2 reports the π^*_{DH} residuals for the observed seat allocation. Only the most successful party is represented proportionally, and the smaller the party, the larger share of its votes is discarded.

Table 3.2: Brazilian parliamentary election of 1982, discarded votes (π^*_{DH} residuals) by party as shares of total and party vote. Values rounded. Source of votes and seats: Nohlen (2005).

	Count	Total vote %	Party vote %
PDS	0	0	0
PMDB	2,538,485	6	14
PDT	654,970	2	27
PTB	845,716	2	46
PT	853,588	2	59

Under the 'best case' scenario–i.e., π^*_{GDH}–only 3% of the total vote would be discarded.

In Figure 3.5, panel (c) shows the decomposition of the votes ignoring the invalid votes and panel (d) including them under the 'best case' scenario. As shown in Figure 3.6, π^*_{GDH} becomes practically zero with the missing rate of 50%.

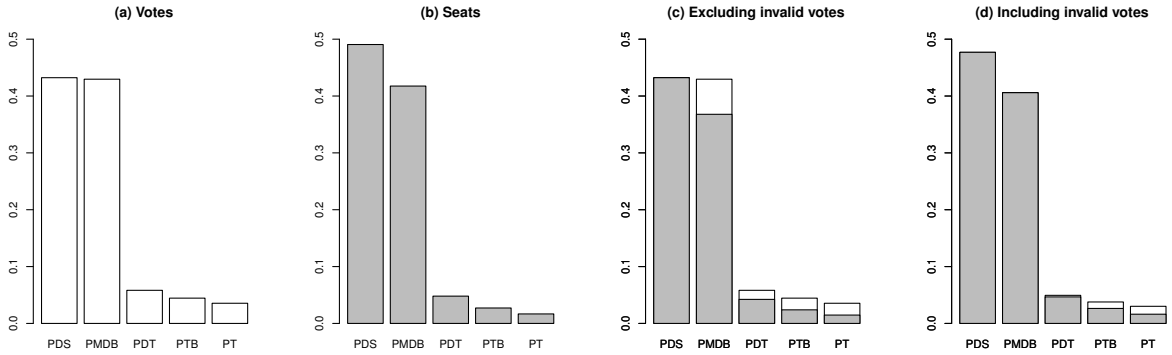


Figure 3.5: Probability distributions of (a) votes, (b) seats in the 1982 Brazilian lower house elections. Panel (c) shows the π^*_{DH} decomposition under the observed vote and panel (d) under the π^*_{GDH} decomposition ('best case' scenario). Source of votes and seats: Nohlen (2005).

On the other hand, under the 'worst case scenario' the value of π^*_{DH} increases about twofold, to 0.25. It might be of interest to compare the observed value of the index and/or the associated residuals to their distributions under various hypothetical scenarios. Consider the scenario where all possible distributions of the missing votes over the 5 parties are equally likely, which can be approximated by sampling the densities from a Dirichlet distribution with all concentration parameters set to 1 and multiplying them by the number of unobserved votes. In this case, about 91% of the simulated index values are larger than 0.12 (π^*_{DH}), and about 50% of the party weighted discarded votes of the second party (PMDB) are larger than the observed value of 6%.

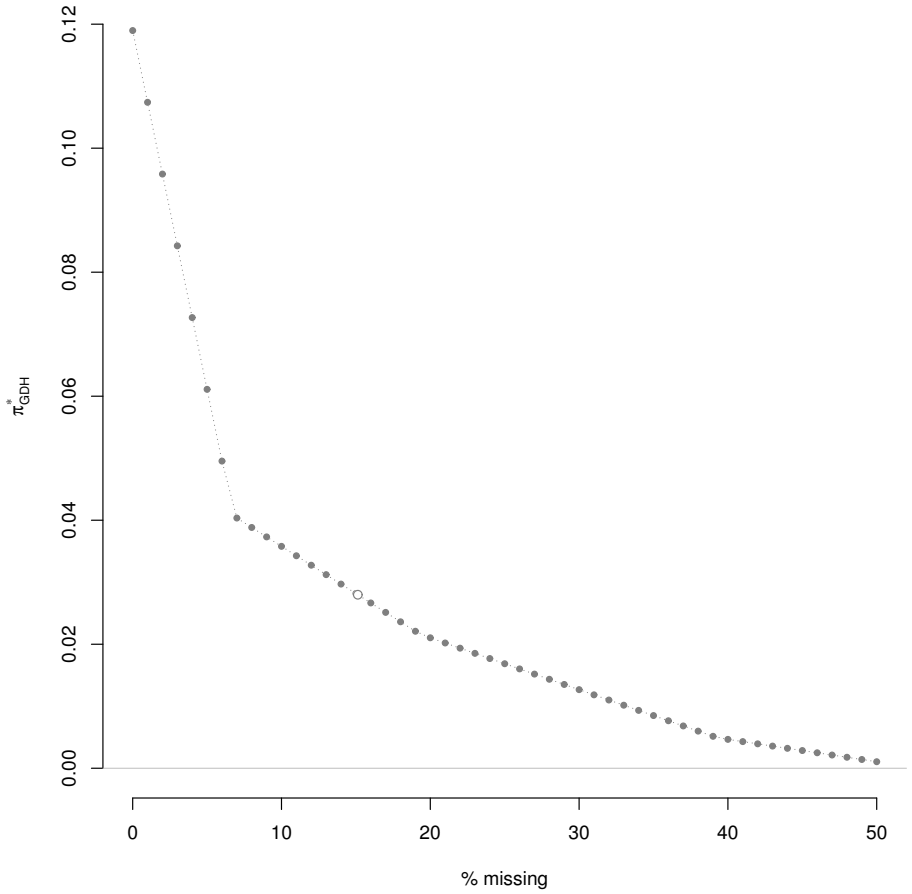


Figure 3.6: Values of the generalized D'Hondt π^* for the 1982 Brazilian lower house elections over a regular grid of missing data rates. The value for the observed rate (15.1%) indicated by a superimposed hollow point.

3.6 Some Additional Implications

It has been recognized in the literature that the D'Hondt method minimizes the maximum overrepresentation understood as the ratio of seat shares over vote shares, known also as the D'Hondt index. Section 3.3 shows that the D'Hondt method has a non-intuitive property of minimizing the share of non-represented votes, quantity designated here as π^*_{DH}, of which the D'Hondt index is a simple function. This finding leads to the reconsideration of several

arguments made about the D'Hondt method and index in the literature.

Gallagher (1991, 42) argues that "[t]he [D]'Hondt method does not work by trying to minimize some overall quantity." In the π^* formulation of the method this does not hold, as it minimizes the fraction of non-represented votes by maximizing the share of proportionally represented votes. Also, Gallagher (1991) considers as problematic cases when the largest ratio of seat over vote shares belongs to a small party. In the provided example of the 1983 Italian general election, it is 2.085 and belongs to a party with 7.6% of the votes. However, the D'Hondt index captures the overall underrepresentation as captured by the value of π^*_{DH}, which is 0.52.

Monroe (1994, 141) argues that under the D'Hondt index most observed seat allocations seem very close to proportionality. This is perhaps due to the fact that largest ratio of seat over vote shares is somewhat less straightforward to interpret than some of the alternative measures. However, as shown in (3.2) and illustrated in Figure 3.2, the relationship between δ and π^*_{DH} is curvilinear, and π^*_{DH} reaches the mid-point already with a δ of 2.

Taagepera and Grofman (2003) fail the D'Hondt index on 8.5 of their 12 criteria. The π^* formulation of the index does meet four of these 8.5 criteria. Specifically, it takes into account all the seat and vote shares (the criterion of 'informational completeness') meeting also the criterion of 'uniformity,' lies on $[0, 1]$ (fourth criterion), has a value of zero if $s = v$ (fifth criterion), and has a value of one if all parties which receive seats did not receive any votes and all the parties which obtained votes did not receive seats (sixth criterion). Thus, the π^* formulation of the index meets 8 of their 12 criteria, which is considerably more than what Taagepera and Grofman (2003) indicate for the D'Hondt index.

3.7 Conclusion

The D'Hondt method has been widely used to allocate seats where proportionality is desired. The method minimizes a quantity known as the D'Hondt index, which is the largest ratio of seat shares over vote shares for a given set of categories. The present chapter shows that this quantity is a simple function of the mixture index of fit, and the D'Hondt method can be interpreted as a procedure that splits the votes into two classes, those represented proportionally and those not at all, while minimizing the size of the second class. This perspective reveals certain non-intuitive properties, and allows to generalize to settings with partially observed votes.

Although the mixture index of fit formulation of the proportional seat allocation problem is given here the 'D'Hondt' designation, it holds for all methods (e.g. Jefferson's) which minimize the same objective function. Just the same, the argument is presented in terms of seats allocated to parties, but holds for similar problems, such seats allocated to states in federal bodies, or the measurement of the representativeness of collective bodies with regards to some member characteristics, such as gender or ethnicity.

Extending similar, but more limited arguments raised earlier (e.g. by Monroe, 1994; Pennisi, 1998; Taagepera and Grofman, 2003), this chapter argues that the measurement of disproportionality rests on the comparison of two distributions–one of which belongs to a model and the other describes the inspected data. In other words, the measurement of disproportionality can be thought of as the evaluation of model fit, and its measures as badness-of-fit statistics. The data in this context are the votes, the model is the legislature, and the goal is to distribute the seats to parties in a way that minimizes some measure of distance from the vote distribution to the seat distribution.

Finally, it has been recognized in the literature (Gallagher, 1991; Pennisi, 1998) that there is only one way how a seat allocation can be proportional, but many ways how it

can diverge from it. Paraphrasing the classical insight of Goodman and Kruskal (1954) on independence and association, there is only one concept of proportionality, but infinitely many possible concepts and measures of disproportionality, and their choice should be guided by substantive and operational concerns. This chapter redefined one concept of disproportionality and its measure, formulating the generalized D'Hondt π^*, which might appeal to some of these concerns.

Chapter 4

Analysis of Electoral Support with the Dissimilarity Index

This chapter presents a general framework for the analysis of electoral support based on the dissimilarity index. It rests on inspecting the fraction of votes that would need to be cast differently for the reality to conform to ideal-typical models, such as stable support or territorially homogeneous competition. Measures are formulated within the framework, which relate to those of electoral volatility and party nationalization and regionalization, are easy to interpret and use, allow comparisons across observations and theories, and account for party and territorial electorate sizes. The measures include the index of residential segregation of votes, which is compared with other measures from the literature on a set of 1495 elections in 119 countries from 1789 to 2013. Two families of models are used with the framework–log-linear models to capture territorial and spatial variability of electoral support and latent class models to inspect territorial heterogeneity of electoral competition. Their use is demonstrated on general elections in Canada (2006–2011), UK (2015), and Belgium (1946–1995).

4.1 Introduction

In party research among the most widely used concepts are electoral volatility and party nationalization, yet research on their measurement is in stark contrast. A single measure of volatility–the Pedersen index (Pedersen, 1979)–has been prevailing for three decades (see also Powell and Tucker, 2014). No such consensus exists on how to measure party nationalization (see e.g. Bochsler, 2010; Golosov, 2014; Lago and Montero, 2014). The main argument of this chapter is that the dissimilarity index (Gini, 1914) can be used to measure other concepts while retaining the appealing features of the Pedersen index, which is its special case. The main contribution is a general framework in which concepts are measured with distances of observations from ideal-typical models. Where votes are the observations, the dissimilarity index allows to measure the distance as their fraction that would need to be cast differently for the reality to conform to the model. The resulting measures are easy and intuitive to interpret and use, and allow for comparisons across observations and theories. Among the quantities defined in the framework is the index of residential segregation of voters, which offers an alternative to the existing measures of party nationalization and regionalization in some contexts. In the chapter, two families of techniques are used to model the ideal types statistically. Log-linear models are used to represent types of territorial and spatial variability of electoral support, and latent class analysis patterns of electoral competition.

Section 4.2 shows how the dissimilarity index generalizes the Pedersen index to any number of elections, and how it can be used to measure residential segregation of voters. The residential segregation index is discussed in the context of research on party nationalization in Section 4.3, showing that the conventional definition of party nationalization corresponds to independence between party and territory, and thus its measurement can be considered as one of association. The residential segregation index is compared with the existing

measures of nationalization on a set of 1495 elections from 119 countries from 1789 to 2013. Section 7.8 extends the use of the dissimilarity index to associations between territory, party, and election with log-linear models, illustrating it on the example of Canadian General Elections of 2006, 2008, and 2011. Section 7.6 formulates a new approach to the analysis of local competition patterns based on latent class analysis, and illustrates its use on the UK general election of 2015 and 17 post-1945 general elections in Belgium.

4.2 The Dissimilarity Index

The success of the Pedersen index rests in its simplicity and clarity of its interpretation. The index captures the lowest possible fraction of voters that changed parties in a pair of elections, provided the same voters took part in both. More generally, it is the smallest fraction of votes that would need to be cast differently in order for the two elections to show the same division of the vote among the options. Formally,

$$\text{PI} = \frac{1}{2}\sum_{j} |s_{j,k=1} - s_{j,k=2}|,$$

where $s_{j,k}$ is the vote share of j^{th} party in the k^{th} election. This quantity always lies on $[0, 1]$, which eases comparisons.

The Pedersen index is a special case of the Gini (1914) index of dissimilarity (Johnston, 1980),

$$D = \frac{\sum_c |o_c - m_c|}{2\sum_c o_c},$$

where the data is a contingency table with N cells, $\mathbf{o} = \{o_1, \ldots, o_N\}$ are the observed values and $\mathbf{m} = \{m_1, \ldots, m_N\}$ those expected under the model. The value of the Pedersen index is equal to that of D if either election from the pair is considered as the model for the other, or if the averages are considered as a model for the pair combined.

The Pedersen index can be generalized to more than two elections, applying

$$D = \frac{\sum_j \sum_k |v_{j,k} - m_{j,k}|}{2 \sum_j \sum_k v_{j,k}},$$

where $v_{j,k}$ is the observed number of votes for j^{th} party in k^{th} election and $m_{j,k}$ the expected one if the division of the vote between the parties is the same in all the elections. in all the inspected elections. The index can be calculated for subsets of the data–its value of for y^{th} party is

$$D = \frac{\sum_i |v_{i,j=y} - m_{i,j=y}|}{2 \sum_i v_{i,j=y}},$$

and for z^{th} territory

$$D = \frac{\sum_j |v_{i=z,j} - m_{i=z,j}|}{2 \sum_j v_{i=z,j}}.$$

The expected values $m_{j,k}$ can be computed using the log-linear model

$$\ln \hat{v}_{j,k} = \lambda^0 + \lambda^P_j + \lambda^E_k,$$

where λ^0 is the grand mean, λ^P party coefficients, and λ^E election coefficients. This model is one of independence, i.e., of no association between party and election. Maximum likelihood estimation of log-linear models (see e.g. Agresti, 2002, 314-56) minimizes the sum of the absolute values of the residuals under the condition that these are symmetrical in the sense that they sum up to zero–i.e., it minimizes the dissimilarity index.

Since this lifts the restriction on the number of categories, any set of interest can be used in place of elections. For territories, the corresponding log-linear model is

$$\ln v_{i,j} = \lambda^0 + \lambda^T_i + \lambda^P_j,$$

where λ^T are territory coefficients. Again, this is a model of independence between the

two set of categories (territories and parties). In this context the index can be interpreted as the smallest fraction of voters that would need to vote differently and/or elsewhere for all parties to record the same shares locally as they did nationally. In other words, it is a distance measure from the observed results to those expected under party system nationalization. The dissimilarity index has a long tradition of use in research on residential segregation (Ducan and Ducan, 1955; Massey and Denton, 1988). If the value of the index is one, each territory is inhabited by members of a single group, i.e., there is perfect segregation, and if it is zero, the group composition of each territory is identical to the aggregate one. Massey and Denton (1988) identify the index as the best of the 20 surveyed measures in terms of capturing 'evenness,' one of their five dimensions of residential segregation together with 'exposure,' 'concentration,' 'centralization,' and 'clustering,' and recommended it as the measure of first choice due to its ease of interpretation and strong correlation with the other surveyed measures. The present application of the index can be interpreted as a measure of *residential segregation of voters* or *votes*, depending on the context. It captures association between party and territory, and is thus related to measures of party nationalization.

4.3 Party Nationalization

Comprehensive accounts of the history of the concept of party nationalization offer Caramani (2004), Morgenstern et al. (2009), Bochsler (2010), Golosov (2014), and Lago and Montero (2014). From the perspective of measurement, several features of the literature are salient. Party nationalization is defined as homogeneity, uniformity or equality of electoral support across territories.[1] Its observable implication is that each party receives within each territory

[1]Morgenstern et al. (2009, 2014)label party nationalization defined in this way as 'static nationalization,' to distinguish it from 'dynamic nationalization,' which implies uniformity of change in party's local vote

the same fraction of votes as it does nationally. Yet, despite this consensus, several measures are in use (Tables 4.2 and 4.3) and in some settings they give different answers. The reasons for this lie in somewhat counterintuitive statistical properties of the problem.

Table 4.1: Referential conventions for equations in Tables 4.2 and 4.3. A different set of conventions is used throughout the chapter.

T	The number of territories.
s	The national vote share of a party.
v_i	The local vote count of a party.
l_i	The local vote total.
p_i	The local vote share of a party.
$\bar{p}$	The mean local vote share of a party.
r_i	The rank of a local vote share of a party.
c_i	Party's local vote as a fraction of its national vote.
m_i	The local number of seats.
e_i	An indicator variable showing whether the party did enter the race in the territory.

shares. A somewhat different perspective on 'dynamic nationalization' is offered by Mustillo and Mustillo (2012).

Table 4.2: Measures of party system nationalization available only for the whole party system. The equations use referential conventions reported in Table 4.1, which to better follow the practices of the party nationalization literature differ from those used in the rest of this chapter.

Measure	Formula
Indicator of Party Aggregation (Chhibber and Kollman, 1998)	$\frac{1}{\sum_j s^2} - \frac{1}{T}\sum_i \left(\frac{1}{\sum_j p_{i,j}^2}\right)$
Inflation Score (Cox, 1999)	$\frac{100}{\sum_j s^2}\left(\frac{1}{\sum_j s^2} - \frac{1}{T}\sum_i \left(\frac{1}{\sum_j p_{i,j}^2}\right)\right)$
Inflation Index (Moenius and Kasuya, 2004)	$\frac{100}{(1/T)\sum_i\left(1/\sum_j p_{i,j}^2\right)}\left(\frac{1}{\sum_j s^2} - \frac{1}{T}\sum_i \left(\frac{1}{\sum_j p_{i,j}^2}\right)\right)$
Weighted Inflation Index (Moenius and Kasuya, 2004)	$100\left(\frac{\left(\sum_i l_i\right)/\left(\sum_j s^2\right)}{\sum_i\left(l_i/\sum_j p_{i,j}^2\right)} - 1\right)$
Local Entrant Measure (Lago and Montero, 2014)	$\sum_j \left(s_j \frac{\sum_i (m_i e_{i,j})}{\sum_i m_i}\right)$

Table 4.3: Measures of party nationalization applicable to individual parties. Party system values can be obtained by unweighted or weighted averages of party values. The equations use referential conventions reported in Table 4.1, which to better follow the practices of the party nationalization literature differ from those used in the rest of this chapter.

Measure	Formula
Mean Absolute Deviation (Rose and Urwin, 1975)	$\frac{1}{T}\sum_i \lvert p_i - s\rvert$
Mean Squared Deviation	$\frac{1}{T}\sum_i (p_i - s)^2$
Variance	$\frac{1}{T-1}\sum_i (p_i - s)^2$
Lee index (Lee, 1988)	$\frac{1}{2}\sum_i \lvert p_i - s\rvert$
Variability Coefficient (Ersson et al., 1985)	$\frac{1}{\bar{p}}\sqrt{\frac{1}{T-1}\sum_i (p_i - \bar{p})^2}$
Normalized Variability Coefficient (Golosov, 2014)	$\frac{1}{\bar{p}\sqrt{T}}\sqrt{\frac{1}{T-1}\sum_i (p_i - \bar{p})^2}$
Standardized and Weighted Variability Coefficient (Ersson et al., 1985)	$\frac{\sqrt{T}}{s}\sqrt{\frac{1}{T-1}\sum_i (p_i - s)^2}$
Index adjusted for Party size and number of Regions (Caramani, 2004)	$\sqrt{\frac{T\sum_i \lvert p_i - \bar{p}\rvert}{2(T-1)\sum_i p_i}}$
Cumulative Regional Inequality (Rose and Urwin, 1975)	$\frac{1}{200}\sum_i \lvert p_i - c_i\rvert$
Territorial Coverage Index (Caramani, 2004)	$\frac{1}{T}\sum_i e_i$
Index of Party Regionalization (Golosov and Ponarin, 1999)	$\sqrt{\frac{T\sum_i \lvert p_i - \bar{p}\rvert}{2(T-1)\sum_i p_i}}$
Coefficient of Party Regionalization (Golosov, 2014)	$\frac{1}{T-1}\left(T - \left((\sum_i p_i)^2 / \sum_i p_i^2\right)\right)$
Index of Party Nationalization (Golosov, 2014)	$1 - \frac{1}{T-1}\left(T - \left((\sum_i p_i)^2 / \sum_i p_i^2\right)\right)$
Normalized Gini Coefficient (Golosov, 2014)	$\frac{2\sum_i (p_i r)}{(T-1)\sum_i p_i} - \frac{T+1}{T-1}$
Party Nationalization Score (Jones and Mainwaring, 2003)	$\frac{2T+1}{T} - \frac{2\sum_i (p_i r)}{T\sum_i p_i}$
Weighted Party Nationalization Score (Bochsler, 2010) (l and p ordered by p/l)	$2\frac{\sum_i \left(l_i\left(\sum_{k=1}^{i} p_k - p_i/2\right)\right)}{\sum_i l_i \sum_i p_i}$
Scaled Party Nationalization Score (Bochsler, 2010)	$WPNS^{\left(\frac{1}{\log_{10} E}\right)};\ \ E = (\sum_i l_i)^2 / \sum_i l_i^2$

4.3.1 Measurement of Party Nationalization as a Statistical Problem

Measures of party nationalization typically use data that can be represented as a table of votes by territory and option. The set of options can consist of parties, but can also include abstaining, casting of invalid ballots, or, in the context of referendums, answers to questions. The data usually comes from official electoral reports, and less commonly from sample surveys. An example is shown in Table 5.1, which reports returns from an election in which two parties competed across four territories.

Table 4.4: Example of electoral returns from a fictitious election.

	Party		
Territory	A	B	
1	111	96	207
2	111	107	218
3	81	98	179
4	93	87	180
	396	388	784

Table 4.5: Hypothetical distribution of votes (rounded) under nationalization.

	Party		
Territory	A	B	
1	105	102	207
2	110	108	218
3	90	89	179
4	91	89	180
	396	388	784

Under the consensus on the observable implications of nationalization, for the same party and territory totals as in Table 5.1 the distribution in Table 5.2 is perfectly nationalized. In fact, it is the only such distribution. Just the same, it uniquely describes independence of territory and party for the given set of marginals–under nationalization party is not associated with territory.

Table 4.6: An example two-by-two table.

	$y = 0$	$y = 1$
$x = 0$	a	b
$x = 1$	c	d

But what distribution corresponds to the opposite of nationalization? This question might seem easy, but deceivingly so. The reason is related to the concept of association. If nationalization means no association, then its absence is association. However, association lacks a statistical definition equally clear as that of independence. This can be shown with a simple example. For a two-by-two cross-classification in Table 4.6, it might appear intuitive that association would be perfect if $x = y$ or $x = 1 - y$, i.e., if $b = c = 0$ or $a = d = 0$, respectively. However, there are many measures of association based on different concepts of it (see e.g. Goodman and Kruskal, 1954, 1959). Some distributions correspond to perfect association under some, but not all measures, but the lack of association corresponds to the same distribution under all. For instance, under a well-known measure of association in cross-classifications, the odds ratio, the value of the statistic ($or = \frac{ad}{bc}$) reaches the lowest or the highest possible value and the association is perfect if any of the four cells is zero (see e.g. Rudas, 1998b). In short, the existing measures of party nationalization differ not necessarily because some or all of them would be biased or invalid, but because they measure different concepts.

Another perspective on the issue might be provided by considering the measurement of party nationalization as the measurement of distance d from the observed distribution $\mathcal{O}$ to the hypothetical distribution under nationalization $\mathcal{N}$,

$$d(\mathcal{O}, \mathcal{N}).$$

In this formulation, $\mathcal{N}$ can be considered as a model, and d as a measure of its fit to data from $\mathcal{O}$. Thus, a more general statement is

$$d(\mathcal{O}, \mathcal{M}),$$

where $\mathcal{M}$ is a distribution that belongs to the model, and is either fixed, or selected to optimize d. The measure summarizes how much does the observed world differ from its hypothetical state in a way deemed relevant.

Some existing measures of party nationalization are formulated somewhat differently, with the goal to assign a score of one to a nationalized distribution and a score of zero to the one farthest from it (e.g. Jones and Mainwaring, 2003; Bochsler, 2010; Golosov, 2014). For

$$d(\mathcal{O}, \mathcal{N}) \in [0, 1],$$

this can be achieved with

$$f(\mathcal{O}, \mathcal{N}) = 1 - d(\mathcal{O}, \mathcal{N}),$$

where f is a measure with the desired property. This highlights the issue that desiring to measure nationalization and independence is the same–there are many ways in which a distribution can diverge from it. Thus, it is useful to define and label the measures in terms of the kind of deviation from independence they capture.

The two notions–that perfect nationalization is equivalent to independence of party and territory, and that any measure of nationalization measures the distance of the data from perfect nationalization–allow us to reconsider the measures of party nationalization.

4.3.2 Choice of Territories

Party nationalization can be understood as lack of association between two sets of categories– voting behaviors and territories, respectively. In most contexts, the first set is straightforward to define. However, the situation can be less clear regarding the second one. Furthermore, it can be seen as desirable for a measure of party nationalization to be independent of the level of aggregation (precincts, districts, regions etc.) on which the votes are inspected (Bochsler, 2010; Morgenstern et al., 2014). This has a two-fold motivation. Some measures are sensitive to the number of territorial units. Also, nationalization is seen as independent of the specific territorial divisions.

(i) The number of territories constrains the range of values of some measures. The Gini coefficient of inequality has in this context a $1-1/n^T$ limit for maximum concentration, and consequently a $1/n^T$ for its smallest possible complement. Thus, a party with a perfectly concentrated score will have a different score under systems with different numbers of territories. This is seen as a hindrance to comparisons (Bochsler, 2010), and can be handled by rescaling the score (Golosov, 2014). If we consider the statistic as a distance from the observations to the model, this ceases to be a problem. The sensitivity simply means that the measure preserves a piece of information about the system, and both the scaled and unscaled measures are comparable across cases, but contain different information. For the usefulness of the measurement, the choice should be made on substantive concerns.

(ii) Different nationalization scores are given for the same party by some indices if the results are inspected on different levels of aggregation even if these are not sensitive to the number of units (Bochsler, 2010). From the perspective of association between territory and vote, this is an appropriate feature of the measure. Two different levels of aggregation can be differently associated with unobserved variables associated with the vote. Consider the example of a party supported exclusively by a group that composes an equal share of inhabitants in each region, but lives in segregated communities that correspond to electoral

wards. The party is likely to receive a similar level of support across regions, but not across wards. While the upper level units of aggregation are not associated with support for the party, the lower level units are, because they are associated with the membership in the group.

From the substantive point of view there is another issue–not all levels of aggregation are equally interesting. Suppose that the data would be available on the level of households, a unit of aggregation common in the social sciences. Would we expect under perfect nationalization households to be split between the parties the same way as the electorate? Hardly. Substantive concerns should drive the choice of the territory type and concepts of party nationalization that see it as independent of the type of territories are of limited usefulness.

4.3.3 Concentration-Based Indices

The latest research on the measurement of party nationalization has produced several measures based on the concept of concentration, motivated by the notion that it is the opposite of nationalization (see e.g. Golosov and Ponarin, 1999). These are the Party Nationalization Score (PNS) (Jones and Mainwaring, 2003), weighted PNS (WPNS) and standardized PNS (SPNS) (Bochsler, 2010), normalized PNS (NPNS), or indices of party nationalization and party system nationalization (IPN and IPSN) (Golosov, 2014). These indices measure the distance from concentration

$$d(\mathcal{O}, \mathcal{C}),$$

where $\mathcal{C}$ is a distribution that corresponds to perfect concentration. Under $\mathcal{C}$, each party receives all its votes in a single territory. Considerable progress has been made with these measures in terms of making them sensitive to different unit sizes or scaling them to the

$[0, 1]$ interval. However, three sets of issues limit their usefulness.

First, these measures are seen as appealing indices of regionalization, since regionalization is accompanied by concentration (Golosov and Ponarin, 1999; Golosov, 2014). Yet, concentration can occur also in the absence of regionalization. Consider the example of two parties that both receive support in three units only, collecting the same share of votes in each unit in a system with ten districts of equal size and turnout. However, one of the parties receives its support in three neighboring units, the other from units scattered across the country. Their concentration-based scores are identical, but few would consider them as equally regionalized. This occurs under any measure that does not account for the location of the units. Measures that take location into account are available in other fields (see e.g. Massey and Denton, 1988).

Second, in practice perfect concentration is extremely unlikely to occur. Usually, there are many more territories than parties, and the maximum possible concentration would be far from perfect according to the standard measures. In such context the perfect concentration model is known to be false even before inspecting the data. As "[a]ll models are false, but some are useful" (Box, 1976), this of course is not a sufficient reason for discarding it, and its usefulness depends on substantive concerns.

Finally, and most importantly, from the statistical point of view concentration is not the only opposite of independence, and from the substantive point of view other concepts of distance might appeal. The concentration-based measures, as well as the criteria declared as desirable of them (Bochsler, 2010; Golosov, 2014) build on research on wealth and income inequality, assuming that inequality has the same meaning regardless of whether the context is a distribution of income across households or of party's votes across territories. However, there are many concepts of inequality, and no guarantees the same concept will be useful across contexts.

4.3.4 Party Nationalization and Residential Segregation

Just as the measures of party nationalization, the residential segregation index can be understood as a measure of distance of the observed distribution of votes across territories from the hypothetical nationalized one. As shown in Figure 4.1, for a large inspected set of constituency-level results from 1495 elections which took place in 119 countries between 1789 and 2013 (see the Appendix, Section B.1), it correlates strongly with some of the measures. The correlations with some are negative, as these measures increase with decreasing distance from the nationalized distribution. Regardless of whether parties or party systems are inspected, the correlation is especially strong with the WPNS and PNS, both of which are based on the Gini coefficient of inequality, and are considerably less straightforward to interpret. This does not mean it can supplant these measures–the consensus on the definition of party nationalization is deceptive, and various measures measure different concepts. This is illustrated with the inspected data–several measures (MSD, Lee, SWVC) correlate only weakly, and others only moderately with the remaining ones.

Rather than attempting to use the proposed index to approximately answer the broad and fuzzy question of nationalization, it is more appealing to use it to get a clear answer to the narrower one of residential segregation. Its additional advantages are the ease of calculation, clarity of interpretation, and the fact that it always lies between zero and one, which makes comparisons especially easy. As a bonus, it is closely related to other measures from party research, and a part of a more general framework.

4.4 Log-Linear Analysis of Electoral Support

As shown in Section 4.2, the dissimilarity index generalizes the Pedersen index to any number of elections, and can also be used to measure the distance of the observations from independence of two sets of categories–parties, and territories or elections. With log-linear

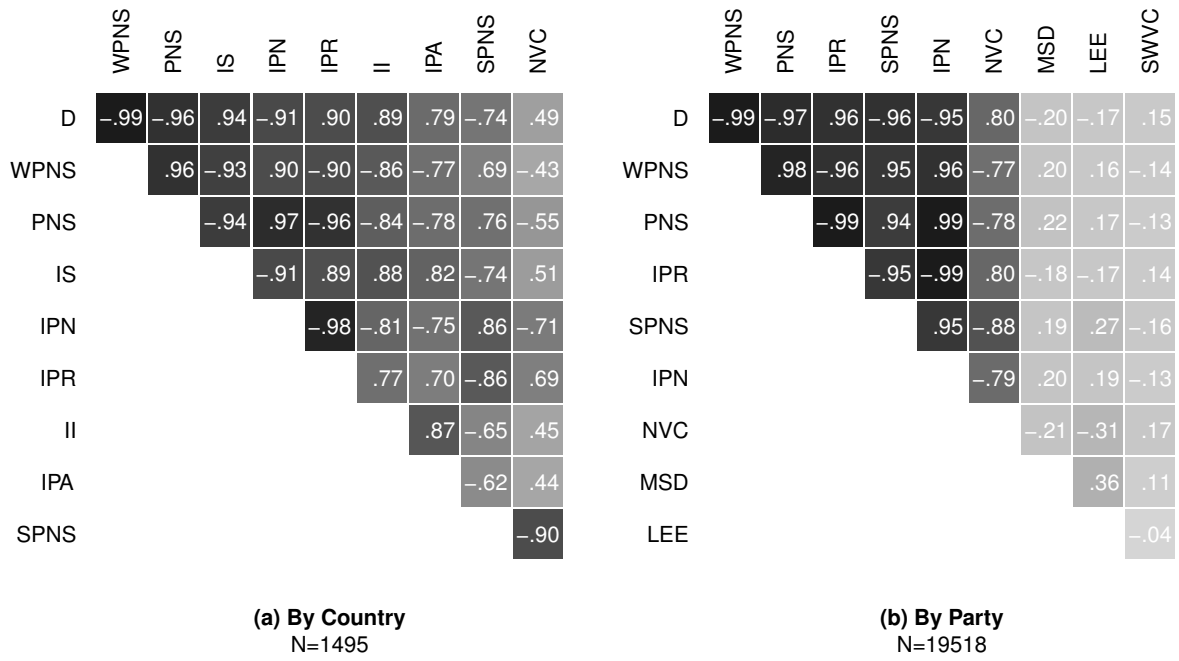


Figure 4.1: Correlations between measures used in party nationalization research. Order and color by the absolute value of Pearson's ρ with the dissimilarity index (D). The measures: Weighted Party Nationalization Score (WPNS), Party Nationalization Score (PNS), Inflation Score (IS), Index of Party Nationalization (IPN), Index adjusted for Party size and number of Regions (IPR), Inflation Index (II), Indicator of Party Aggregation (IPA), Standardized Party Nationalization Score (PNS), Normalized Variability Coefficient (NVC) Mean Standard Deviation (MSD), Lee index (LEE), Standardized and Weighted Variability Coefficient (SWVC). Data source: Kollman et al. (2014).

models, this can be extended to any number of categories. Perhaps the most appealing extension is to votes cross-classified by territory, party, and election.

Table 4.7: Hierarchically nested log-linear models for votes cross-classified by territory (T), party (P), and election (E).

			Models													
m_0	$\ln\hat{v}_{i,j,k}$	=	λ^0													
m_1			λ^0	+	λ_i^T	+	λ_j^P	+	λ_k^E							
m_2			λ^0	+	λ_i^T	+	λ_j^P	+	λ_k^E	+	$\lambda_{i,j}^{TP}$					
m_3			λ^0	+	λ_i^T	+	λ_j^P	+	λ_k^E	+	$\lambda_{i,k}^{TE}$					
m_4			λ^0	+	λ_i^T	+	λ_j^P	+	λ_k^E	+	$\lambda_{j,k}^{PE}$					
m_5			λ^0	+	λ_i^T	+	λ_j^P	+	λ_k^E	+	$\lambda_{i,k}^{TE}$	+	$\lambda_{j,k}^{PE}$			
m_6			λ^0	+	λ_i^T	+	λ_j^P	+	λ_k^E	+	$\lambda_{i,k}^{TE}$	+	$\lambda_{i,j}^{TP}$			
m_7			λ^0	+	λ_i^T	+	λ_j^P	+	λ_k^E	+	$\lambda_{j,k}^{PE}$	+	$\lambda_{i,j}^{TP}$			
m_8			λ^0	+	λ_i^T	+	λ_j^P	+	λ_k^E	+	$\lambda_{i,k}^{TE}$	+	$\lambda_{j,k}^{PE}$	+	$\lambda_{i,j}^{TP}$	
m_9			λ^0	+	λ_i^T	+	λ_j^P	+	λ_k^E	+	$\lambda_{i,k}^{TE}$	+	$\lambda_{j,k}^{PE}$	+	$\lambda_{i,j}^{TP}$	+ $\lambda_{i,j,k}^{TPE}$
			Terms													
λ^0	*Grand mean*															
λ^T	*Territory*		Electorate size can vary across territories.													
λ^P	*Party*		National vote totals can vary across parties.													
λ^E	*Election*		National electorate size can vary across elections.													
λ^{TP}	*Territory-Party*		Party vote can vary across territories–voter segregation.													
λ^{TE}	*Territory-Election*		Territorial electorate sizes can vary across elections.													
λ^{PE}	*Party-Election*		National party totals can vary across elections–'nationalized' volatility.													
λ^{TPE}	*Three-way*		Vote count can vary across territory-party-election combinations.													

Table 4.7 shows ten log-linear models for votes cross-classified by territory, party (or more generally option), and election. Each lifts a different set of restrictions on the expected count of votes $\hat{v}$ from i^{th} constituency for j^{th} party in k^{th} election. Under the null model m_0 all the counts are the same. Usually, national party totals differ, as do the total numbers of votes across elections and electorate sizes across constituencies. These three kinds of variability are captured by the independence model m_1 with territory, party, and election terms $\left\{\lambda^T, \lambda^P, \lambda^E\right\}$. Under this model, the vote share of each party stays the same across territories and elections. Substantively, it corresponds to perfect de-segregation

('nationalization') with no volatility.

Segregation is allowed by the territory-party interaction λ^{TP}. Model m_2, which includes it as the only interaction, corresponds to 'electoral continuity' understood as stable local patterns of electoral support (see Bartels, 1998; Wittenberg, 2006, 2013). Volatility is allowed by including the party-election interaction λ^{PE}. It is the only interaction in model m_4, which corresponds to nationalization with volatility. Under this model, not only electoral support, but also its change can be thought of as nationalized.

Territorial electorate sizes can change in time as well, due to reasons such as population movement or redistricting, or, if abstainers are excluded, changes in turnout. This can be captured by territory-election interaction λ^{TE}, which allows the local electorates to vary in size across elections with a different rate than the national electorate. The final, saturated, model m_9, includes all two-way interactions and the three-way interaction $\lambda^{TPE}_{i,j,k}$. It will fit perfectly by definition, and consequently its distance from the data will be zero under any metric.

The log-linear models are related to that of Morgenstern and Potthoff (2005), which extends work by Stokes (1965, 1967) and Kawato (1987). It is a linear model

$$\hat{s}_{i,j,k} = \mu_j + \beta^T_{i,j} + \beta^E_{j,k} + \beta^{TE}_{i,j,k},$$

where $s_{i,j,k}$ is the fraction of votes from i^{th} territory for j^{th} party in k^{th} election, and all the terms β are random coefficients ('effects') estimated separately for each party. Using local vote fractions assigns the same weight to all territories, which is inferentially inexpensive if local electorates are large and of similar sizes. However, if they are not, the informativeness of the data with regards to the quantities of interest might vary across territories, and weighting might improve the inferences (see Bartels, 1998; Alemán and Kellam, 2008). The log-linear models avoid this completely–by using vote counts they take into account both

party and district marginals.

In Morgenstern and Potthoff's (2005) approach, the main quantities of interest are variance parameters interpreted under the assumption that the model is true. Contrastingly, the proposed approach does not require this assumption. Instead, it treats them as ideal types, and draws inferences based on their distances from the observations. The distance measures are special cases of the dissimilarity index. Thus, unlike the variance parameters, they always range from zero to one, and have a straightforward interpretation as the lowest share of votes in a given category that would have to be cast differently in order for the model to describe the reality perfectly.

4.4.1 Variability of Electoral Support in Space and Time: Canada 2006–2011

Early 21$^{\text{th}}$ century Canada is due to temporal and spatial variability in electoral support of some of the parties an interesting example for the log-linear approach. One of the two previously largest parties–the Liberal Party–has declined and the New Democratic Party a rose to the second place (Figure 4.2), and the support of the Bloc Québécois is strongly regionalized. The data consists of returns from all 308 constituencies in the 2006, 2008, and 2011 General Elections (Kollman et al., 2014) for five parties with more than 1% of the national vote and an aggregate category for the remaining parties.

Figure 4.3 shows the fit of eight log-linear models $m_{1,\ldots,8}$ from Table 4.7 to the Canadian data in terms of the shares of votes that would need to be cast differently in order for the reality to conform to the models. The fits reveal that the data are relatively far from the models that do not allow volatility and segregation (the independence model and territory-election interaction only models). Allowing volatility (party-election interaction) does not improve the fits by much. On the other hand, allowing residential segregation

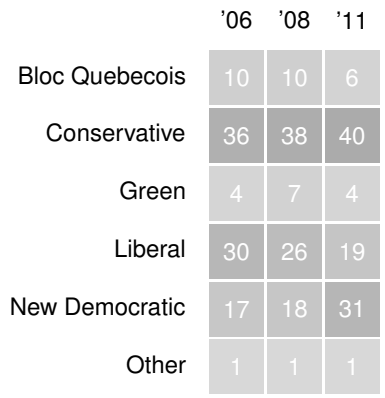


Figure 4.2: Vote fractions in rounded percents obtained in three Canadian General Elections. Data source: Kollman et al. (2014).

of voters (territory-party interaction) improves the fits markedly–even under the simplest model with this interaction the misfit is only 10%.

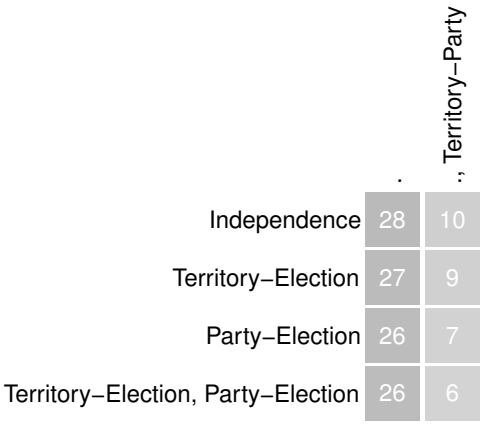


Figure 4.3: Model fit of eight log-linear models to the Canadian data in terms of the dissimilarity index (in rounded percents). Models that allow territory-party interaction (segregation) on the right. Data source: Kollman et al. (2014).

Partial fits reported in Figure 4.4 offer a more detailed picture. The independence model fits badly to the votes of Bloc Québécois and votes in the Other category. The simplest model that allows residential segregation markedly improves the fit both overall and for these two categories. Under this model, the largest remaining misfit is in 2011, for all parties

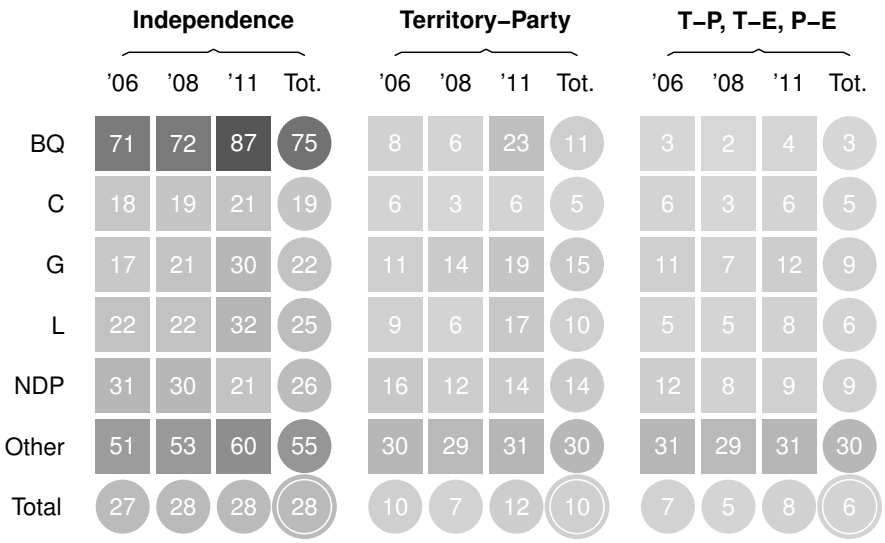


Figure 4.4: Partial and overall model fits for three selected models of the Canadian parliamentary elections of 2006, 2008, and 2011 in terms of the dissimilarity index (in rounded percents). Data source: Kollman et al. (2014).

except the Conservatives, due to the losses of the Liberals and the gains of the NDP. This is further evidenced by the fact that allowing volatility (and territory-election interaction, which has only a small effect on fit) improves the fit only slightly overall, but considerably for the Liberals and NDP. In short, the Canadian party system appears considerably marked by residential segregation of votes, chiefly due to BQ, and only moderately with volatility, which fairly 'nationalized' and affecting the Liberals and NDP the most.

4.5 Latent Class Analysis of Competition Patterns

In some elections, the local patterns of competition are as whole far from the national one, but groups of territories show strong similarities. For example, some districts might be dominated by a regional party that lacks support elsewhere, such as the Canadian Bloc Québécois discussed in the previous section. Such scenarios can be modeled by introducing more patterns of competition, and sorting the territories into groups defined by them.

Substantively, this corresponds to systems composed of homogeneous sub-systems. From this perspective, territorial heterogeneity of electoral support is assessed as the number of patterns needed to describe the election. The dissimilarity index can be applied to compare the fit of models with different numbers of patterns.

Statistically, the local distribution of votes in i^{th} territory can be modeled with a latent class model

$$\hat{v}_{i,\cdot} \sim Multinomial\left(\theta_{g[i]},\ \sum_j v_{i,j}\right),$$

where θ_g is a simplex of party fractions under g^{th} pattern. Both the patterns and the group memberships of territories are unobserved and estimated from the data. Multinomial latent class models can be estimated using the EM algorithm (Dempster et al., 1977), as implemented e.g. in the R (R Core Team, 2014) package `mixtools` (Benaglia et al., 2009) used here. The Multinomial-Poisson transformation (see e.g. Baker, 1994) means that if there is one pattern, the model distribution is identical to the one under the log-linear territory-party independence model.

The approach is illustrated here for a single election, using the 2015 UK general elections as an example, and for a series of elections from the same country, using the example of 17 Belgian general elections from 1946 to 1995.

4.5.1 Diverse Local Patterns of Competition under FPTP: UK 2015

The winner-takes-all character of the first-past-the-post system can obscure the fact that patterns of electoral competition can differ substantially even across constituencies won by the same party. How many patterns of competition describe the election can be answered by comparing the fit of latent class models with different numbers of components using the dissimilarity index. The UK general election of 2015 is an interesting example, as in

addition to the Northern Irish parties winning all 18 NI constituencies the Scottish National Party won 56 out of the 59 Scottish constituencies. The data (Healy, 2015) consists of returns from all 650 single-member constituencies for 11 parties which won at least one constituency (with the Speaker's votes classified as Conservative), shown in Table 4.8, and a 12^{th} category aggregating votes for the remaining parties.

Table 4.8: Parties that won seats in the 2015 UK general election.

Conservative Party	Con
Democratic Unionist Party	DUP
Green Party of England and Wales	GP
Labour Party	Lab
Liberal Democrats	LD
Plaid Cymru	PC
Scottish National Party	SNP
Sinn Féin	SF
Social Democractic and Labour Party	SDLP
Ulster Unionist Party	UUP
United Kingdom Independence Party	UKIP

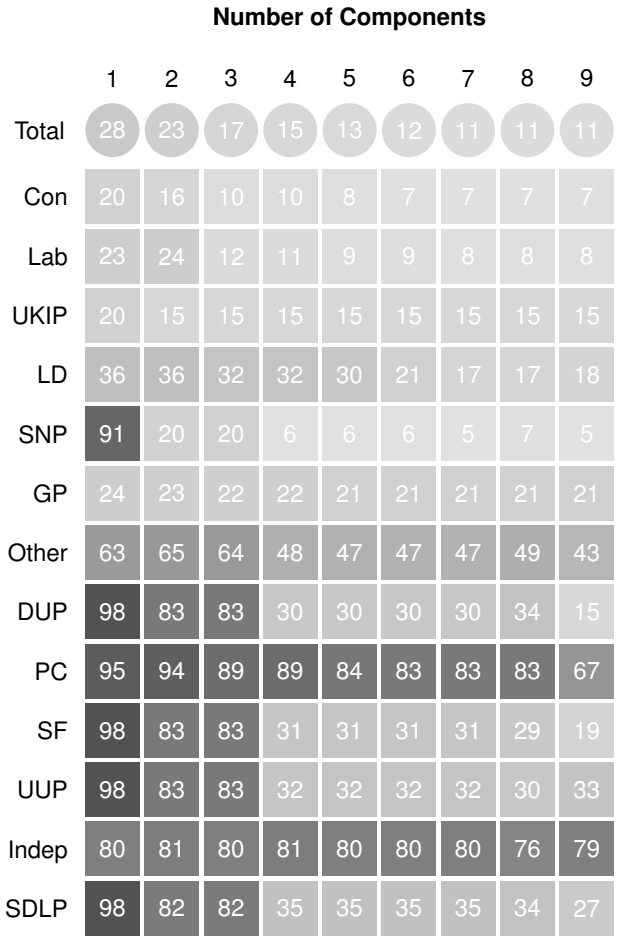


Figure 4.5: The values of D for nine mixture decompositions of the party-by-constituency vote in the British general election of 2015. Parties in descending order by their national vote totals. Values reported in rounded percents, overall values in circles and partial in squares. Data source: Healy (2015).

Figure 4.5 reports the fit of nine models with increasing number of patterns to the data, both overall, and for individual parties. Increasing the number of patterns from one to four practically halves the index, from 28% to 15%. Further patterns improve the fit only marginally. The party-level values show that introducing the second pattern captures the SNP vote, and the fourth the Northern Irish vote. Introducing the third pattern decreases misfit among voters of the major parties, which can be better understood by inspecting the

vote distributions associated with the groups.

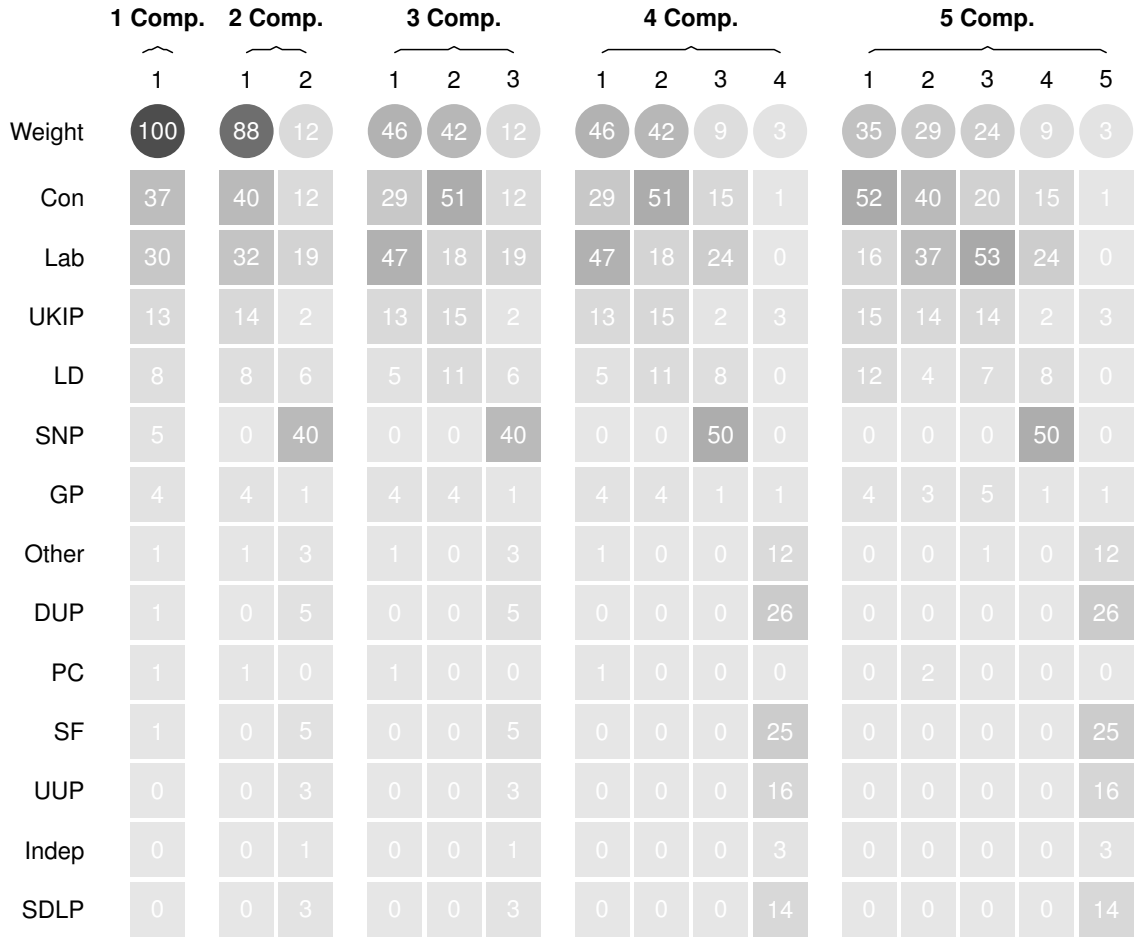


Figure 4.6: Five mixture decompositions of the party-by-constituency vote in the British general election of 2015. Proportions in rounded percents. Groups' shares of districts in circles, party shares of votes under the patterns in squares. Data source: Healy (2015).

Vote distributions under the models with one to five patterns are shown in Figure 4.6 and the geographic distribution of four of them in Figure 4.7. Under the two- and three-component models Scottish and Northern Irish constituencies belong to the smallest group, and to separate ones under four or more patterns. Under the three- and four-pattern models the largest group are English and Welsh constituencies tightly won by Labour, and the second largest those won with wide margins by Conservatives. This might appear somewhat

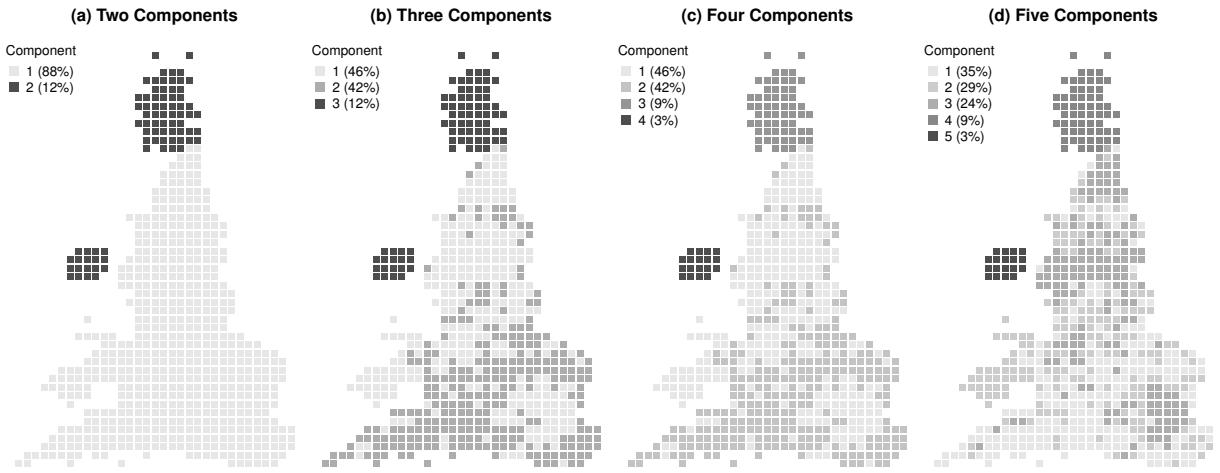


Figure 4.7: Four latent class models of the constituency vote patterns in the UK general election of 2015. Total values in circles, party values in squares. Component weights in rounded percents. Data source: Healy (2015).

counter-intuitive given that Conservatives won more districts than Labour. The five-pattern model clarifies this–the two largest groups are constituencies won by Conservatives with wide and narrow margins respectively, and the third component are constituencies won by Labour with wide margins. This captures the fact that Conservatives did better than Labour in constituencies in which one of these parties led by a narrow margin over the other.

Substantively, the UK 2015 general election does not show a single pattern of competition within constituencies, and is thus far from a 'nationalized party system.' A better description is that there are three sub-systems. The first is England and Wales, where the some constituencies are dominated by Conservatives or Labour, and they are tied for the lead in others, while UKIP and the Green Party maintain a relatively even presence. The second is Scotland with SNP enjoying a wide margin over the second Labour, and the third is Northern Ireland with its own four parties splitting the constituencies.

4.5.2 Changing Local Patterns of Competition: Belgium 1946–1995

The Belgian party system is notable for its split into Flemish and Wallonian ones from the late 1960s on. The data consists of constituency-level results from 17 post-1945 Belgian parliamentary elections (Table B.2) available from CLEA (Kollman et al., 2014). Given that voting was compulsory, an additional category of abstentions and invalid votes is included.

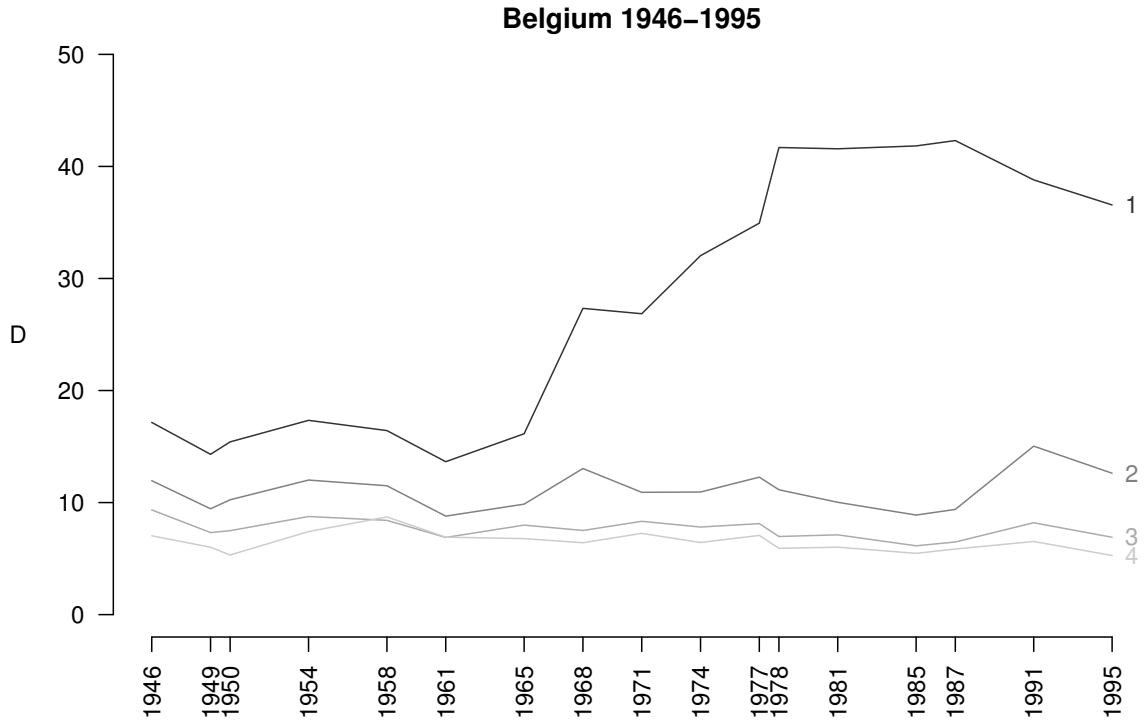
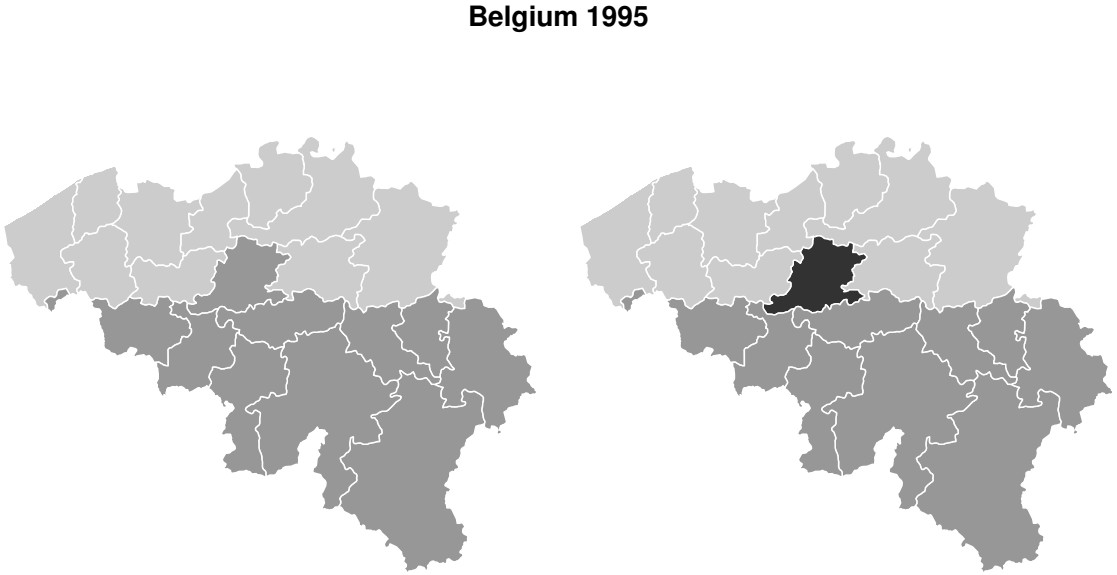


Figure 4.8: Model fit of four latent class decompositions of constituency vote patterns applied to 17 Belgian elections in terms of the D dissimilarity index reported in percents. Number of components in the model indicated on the right. Data source: Kollman et al. (2014).

As Figure 4.8 shows, the single national pattern fits much worse after 1968, with its misfit doubling from the 1950s to the 1980s. However, the two-pattern model fits well–only about 10% of votes would need to be cast differently in any of the inspected election for it to fit perfectly, and adding patterns does not improve the fit by much. This is illustrated in



(a) Two Components (b) Three Components

Figure 4.9: Belgian 1995 federal elections–grouping of districts under two- and three-pattern model. Data source: Kollman et al. (2014).

Figure 4.9 for the 1995 elections–the first two groups correspond to Wallonia and Flanders, and if a third pattern is introduced, it describes only one constituency, which includes the capital. This can be seen also by inspecting the party fractions under each model, reported in Figure 4.10–the largest two groups correspond to Flemish and Wallonian parties. In short, the Belgian party system is after late 1960s well described as composed of two sub-systems with little overlap, within both of which the degree of residential segregation of voters is relatively low.

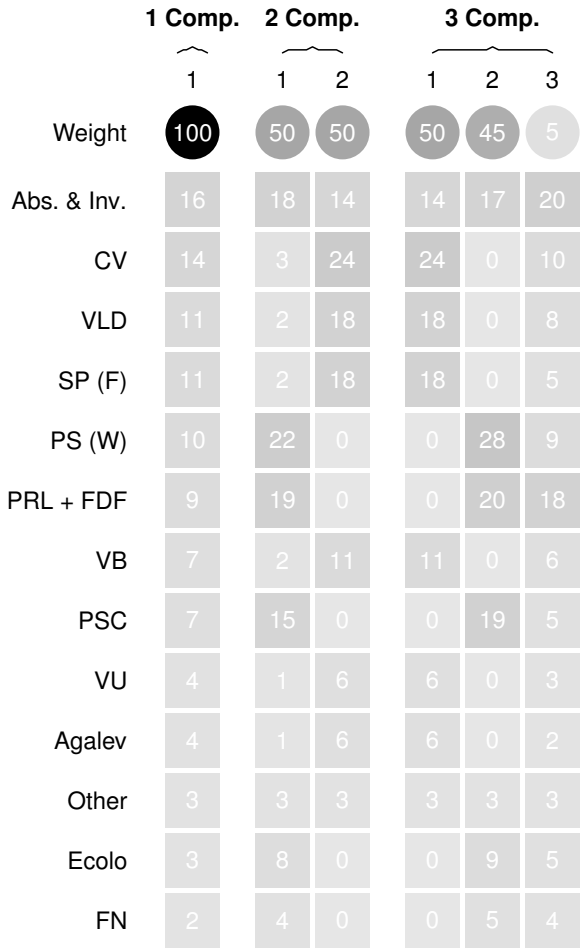


Figure 4.10: Three models of the 1995 Belgian federal elections. Component weights (in circles) and option proportions (in squares) in rounded percents. Data source: Kollman et al. (2014).

Table 4.9: Belgian parties in the 1995 general election.

Agalev		Agalev
Christian People's Party	*Christelijke Volkspartij*	CV
Ecolo		Ecolo
National Front	*Front National*	FN
Liberal Reformist Party - Democratic Front of Francophones		PRL + FDF
Socialist Party (Wallonia)	*Parti Socialiste*	PS (W)
Christian Social Party	*Parti Social Chrétien*	PSC
Socialist Party (Flanders)	*Socialistische Partij*	SP (F)
Flemish Block	*Vlaams Blok*	VB
Flemish Liberals and Democrats	*Vlaamse Liberalen en Democraten*	VLD
People's Union	*Volksunie*	VU

4.6 Conclusion

The present chapter proposes a general framework for the analysis of electoral support. In the framework, the quantities of interest are distances of the observed electoral returns from hypothetical distributions which can be understood as ideal types. The distances are measured with the dissimilarity index, a special case of which is widely used as the Pedersen index. In this application, the index is the lowest fraction of votes that would need to be cast differently for the observed results to conform to an ideal type. This quantity is easy and intuitive to interpret and use, and allows for comparisons across observations and theories.

Applied to the problem of homogeneity of electoral support, the framework yields an index of residential segregation of voters. The segregation index offers an alternative to the existing measures of party nationalization and regionalization, with several of which it strongly correlates when applied to a large and diverse set of 1495 elections in 119 countries from 1789 to 2013. It correlates especially with measures based on the Gini coefficient of inequality, however, it has a much clearer interpretation than them.

Where relationships between party support, territory, as well as changes in support over time are of interest, log-linear models can be used to model a variety of scenarios,

including volatility, voter segregation, and changes in local electorate sizes. Substantive inferences can be drawn from comparisons of the fit of the models. Where local patterns of electoral competition are of interest, latent class analysis can be applied, and a substantively informative statistical description identified using the dissimilarity index.

Chapter 5

Minimum Mixture Models of Voter Transitions in Aggregate Data

The rate of voters who switched party from one election to another is often measured from aggregate data. The existing methods involve trade-offs between how substantively informative the generated quantities and how strong and testable the underlying assumptions are. This chapter proposes a new latent class approach to the analysis of aggregate electoral data based on the Rudas–Clogg–Lindsay mixture index of fit. It is light on assumptions, but highly flexible, and provides a measure of transitions applicable to any number of elections and parties. The measure is easy to compute and interpret. The approach is extended to conditionally constant voting, and proportional and uniform swing. The use is demonstrated on data from the 2004 and 2008 elections in Montana and the 1966 and 1970 UK general elections.

5.1 Introduction

Voters can change parties from one election to another, or split their tickets in simultaneous elections. Due to ballot secrecy researchers rarely observe the transitions directly, and rely on other data. Sample surveys are not always available, but aggregate data usually is. A variety of methods has been designed to extract the information on voter transitions from it. When choosing between them, researchers face a trade-off between how substantively informative the generated quantities and how strong and testable the underlying assumptions are.

This chapter proposes a new approach to the analysis of aggregate electoral data based on the mixture index of fit (Rudas et al., 1994). The approach is assumption-light like the method of bounds (Duncan and Davis, 1953) and the Pedersen (1979) volatility index, but considerably more flexible. It rests on decomposing the votes into two classes–those that fit a restrictive model such as uniform swing, and those that do not. The primary contribution of the chapter is a measure of voter transitions from aggregate data–the smallest possible share of inconstant voters. The measure is intuitive to interpret, easy to compute, rests on light assumptions, and provides a one-number summary for any number of elections and parties. The approach can be extended to more complex models of voting behavior, such as conditionally constant voting and proportional and uniform swing. In such settings, it allows to draw new substantive findings based on the analysis of residuals.

In this chapter, *constant voting* is understood as sticking with an option in a series of elections. The options are usually defined as parties, but can include also abstentions, invalid ballots, party groups, and non-partisan candidates, or, in the context of referendums, answers to questions. For simplicity, the chapter refers to the options as parties and the units of observation as votes. Correspondingly, three types of behaviors are understood as voter transitions. Choosing different parties in non-simultaneous elections is referred to simply as *inconstant voting*. Choosing different parties in simultaneous elections and

splitting the constituency and list votes in mixed-member systems are referred to also as *ticket splitting*. If the elections are simultaneous on a regular basis, they are referred to as being in the same *batch*, and if they are for the same post, as being to the same *office*.

The chapter first discusses in Section 5.2 the problem of inferring individual behavior from aggregate data, focusing on the usefulness-assumptions trade-offs in the existing methods. The minimum mixture measure of voter transitions is introduced in Section 5.3. The application of the mixture index of fit is extended to conditionally constant voting in Section 5.4. As an example, the 2004 and 2008 presidential and gubernatorial elections in Montana are inspected. The extension to proportional and uniform swing is shown in Section 5.5, with examples from the 1966 and 1970 British general elections.

5.2 The Cross-Level Inference Problem

Inferring voter transitions from electoral returns attempts to uncover individual behavior from aggregate data, i.e., cross-level inference. Electoral returns can be represented as contingency tables with known marginals, but unknown cell values. Table 5.1 shows an example with returns from two elections in a single district. A variety of methods for cross-level inference has been proposed, all of which involve trade-offs between how substantively useful are the generated quantities and how strong and testable the underlying assumptions.

Table 5.1: Returns from two elections from a single district with ten thousand voters.

	President		
Senate	A	B	
A	a	b	9
B	c	d	1
	7	3	

Table 5.2: An alternative representation of the data from Table 5.1.

	Election	
Party	Sen.	Pres.
A	9	7
B	1	3

5.2.1 The Method of Bounds

Table 5.3: Returns from two elections from a single district with ten thousand voters. Lower and upper bounds on cell values according to the method of bounds.

	President		
Senate	A	B	
A	$[6, 7]$	$[2, 3]$	9
B	$[0, 1]$	$[0, 1]$	1
	7	3	

The method of bounds provides for each cell of the contingency table the smallest and largest possible values given the observed set of marginals (Duncan and Davis, 1953; Shively, 1991). It does not rest on untestable assumptions, however some of its extension do (Achen and Shively, 1995). Table 5.3 shows the intervals for Table 5.1. How wide, and thus substantively useful, the generated intervals are depends on the marginals of the table. In this way, the lack of problematic assumptions comes at the price of limited substantive usefulness of the generated quantities in some circumstances.

5.2.2 The Pedersen Index

The Pedersen index is not always interpreted as a measure of individual behavior from aggregate data, but in some settings it can be be considered as such. The index measures the overall amount of change in electoral support from one election to another. It is straightforward and intuitive to interpret, and easy to computate. The index is defined for a pair of elections, and equals to half the sum of absolute differences in fractions of vote obtained in two elections by the parties. Formally,

$$PI = \frac{1}{2} \times \sum_i |x_{i,t=1} - x_{i,t=2}|,$$

where $x_{i,t}$ is the fraction of votes for i^{th} party in t^{th} election. If the same voters are eligible in both elections, the index captures the lowest possible fraction of voters that changed parties. Its value for the example in Table 5.1 is 0.2, i.e., at least 20% of the voters voted differently.

The Pedersen index is a special case of the Gini (1914) index of dissimilarity (Johnston, 1980),

$$D = \frac{1}{2} \times \frac{\sum_l |o_l - m_l|}{\sum_l o_l}, \tag{5.1}$$

where $\{o_l\}_{l=1}^{L}$ are the observed values and $\{m_l\}_{l=1}^{L}$ values expected under the model in an L–cell ($l$ong) contingency table. In this context, the contingency table is not the table of party-to-party transitions such as Table 5.1, but a table of votes cross-classified by party and election, such as Table 5.2.

Although the Pedersen index is usually computed from vote fractions, using (5.1) vote counts yield the same result. For a pair of elections, the Pedersen index equals D under two models. First, if the probability distribution of votes over parties from one election is the model for the other election. Second, if the average of the elections is the model for the

pair combined.

Unlike the Pedersen index, the dissimilarity index provides one-number summaries for more than two elections. Regardless of whether the same voters were eligible in the elections, the summary can be interpreted as the lowest fraction of votes that would need to be cast differently in order for the party shares to be stable. From this perspective, the stability of electoral support is understood as lack of association between party and election. This definition of stability corresponds to the log-linear model of party-election independence

$$\ln v_{i,t} = \lambda^0 + \lambda_i^P + \lambda_t^E, \tag{5.2}$$

where $v_{i,t}$ is the count of votes of i^{th} party in t^{th} election, λ^0 the main term, and λ^P and λ^E the party and election terms, respectively.

5.2.3 Ecological Inference

A wide and diverse group is known as 'ecological inference' methods. Unlike the two methods discussed above, they aim not only for bounds, but also for probabilistic point and interval estimates of the unobserved cell values. It is beyond the scope of this chapter to provide a comprehensive review, which can be found e.g. in Cleave et al. (1995), Freedman (1999), and Glynn and Wakefield (2010). Despite their diversity, the methods involve adopting a set of strong and potentially problematic assumptions to deliver highly substantively informative quantities with relatively low levels of statistical uncertainty attached (see also Freedman et al., 1991; Gelman et al., 2001; Greiner and Quinn, 2009).

5.2.4 Entropy-Maximizing

'Entropy-maximizing' (e.g. Johnston and Hay, 1983; Johnston and Pattie, 2000, 2003) is related to the ecological inference methods, but promoted as an alternative to them. It

Table 5.4: Illustration of the entropy-maximizing method. Should integer cell values be desirable, the fractions can be rounded.

	President		
Senate	A	B	
A	6.3	2.7	9
B	0.7	0.3	1
	7	3	

consists of filling the cell values so that the entropy of the table is maximized under the constraints given by its marginals. Standardly, the marginals include in addition to the ones observed in the elections also two-way marginals from other sources, such as sample survey estimates of ticket splitting on the highest level of aggregation. The outcome of the method for Table 5.1 is shown in Table 5.4.

Since entropy is maximized under marginal independence, the method implies restrictive log-linear models (Berg, 1988). The assumption of marginal independence is not testable in this context, and carries serious inferential consequences. Consider the example in Table 5.4, where marginal independence means lack of association between party choice in the two elections, which is untestable, and lacks intuitive appeal. The method is designed for settings where two-way margin of party-to-party transitions is estimated from a sample survey. Nevertheless, even in that case, the method implies a model of independence– between the transitions and territory. Again, this assumption is not testable, and might lack appeal in many applied settings. In short, just as the ecological inference methods, entropy-maximizing offers substantively informative quantities with relatively low statistical uncertainty attached to them. Just the same, it comes at the price of strong and untestable assumptions likely to be false in some settings.

5.3 Measuring Vote Transitions with Mixtures

The previous section has outlined the trade-offs involved in the existing methods for estimation of voter transitions from aggregate data. The ecological inference methods and entropy-maximizing provide quantities which are substantively informative at the price of strong and untestable assumptions. Consequently, if the assumptions are not met, the methods can lead to erroneous substantive inferences. On the other hand, the method of bounds in its standard form rests on assumptions known to be true, but in many contexts produces intervals too wide to be substantively useful. The Pedersen index similarly rests on relatively light assumptions, but produces a single number, which can be considered as the lowest value of an unobserved quantity possible within given restrictions.

As shown in the previous section, the dissimilarity index provides an analogue of the Pedersen index for more than two elections. However, unlike the Pedersen index, it cannot always be interpreted as the lowest fraction of inconstant voters, even if the same voters took part in the elections. A measure that generalizes to any number of elections, but retains this interpretation can be defined with latent classes, based on the fact that any set of votes from multiple elections contains two classes–votes cast by constant voters and votes cast by others. From this perspective, the smallest possible size of the latter class is the lowest boundary for the fraction of inconstant voters, and in that sense is a minimum mixture measure of voter transitions.

The minimum mixture measure is a special case of the π^* mixture index of fit (Rudas et al., 1994), a statistic which measures model misfit by the smallest fraction of the population that cannot be described by the model. Formally, the π^* index is a measure of distance from the observed distribution $\mathbf{o}$ to model $\mathcal{M}$,

$$\pi^*(\mathbf{o}, \mathcal{M}) = \mathsf{inf}\{\pi \colon \mathbf{o} = (1-\pi)\mathbf{m} + \pi\mathbf{r},\ \pi \in [0,1],\ \mathbf{m} \in \mathcal{M},\ \mathbf{r}\ \mathsf{unspecified}\},$$

where **m** is an element from the model and **r** an unrestricted residual component. The mixture index of fit is highly general, and can be applied to any restrictive model. The resulting flexibility means that in electoral data analysis its use need not be restricted to constant voting.

In the context of vote transitions, assuming that the same voters were eligible in the elections the underlying model is

$$v_{i,t} = c_{i,t} + r_{i,t}, \tag{5.3}$$

where v is the observed vote count, which is composed of votes c by constant voters, and votes r by other voters. The proportion of constant voters is

$$\pi^c = \frac{\sum_i \sum_t c_{i,t}}{\sum_i \sum_t v_{i,t}},$$

and of the residual voters

$$\pi^r = 1 - \pi^c = \frac{\sum_i \sum_t r_{i,t}}{\sum_i \sum_t v_{i,t}}.$$

Constant voting is independence of party and election, and c is described by the log-linear model (5.2). Should the one of the elections include more voters than the other, the constant voting is no longer captured by the model of party–election independence (5.2) . Instead, the more restrictive model

$$\ln c_{i,t} = \lambda^0 + \lambda_i^P, \tag{5.4}$$

applies, under which the parties vote counts as opposed to shares are the same across elections. The distribution of the votes by incosistent voters is however not restricted. Consequently, the decomposition (5.3) will describe the observed votes perfectly for any $\pi^r \in [\pi^*, 1]$.

The substantive interpretation of the measure depends on whether the same voters took

Table 5.5: Returns from two elections from a single district with ten thousand voters.

	Election		
Party	Sen.	Pres.	Same
A	9	7	7
B	1	3	1

part in the elections. First, if the same voters were eligible in the elections, but different subsets have abstained or cast invalid ballots in each, the categorization by party can be generalized and these categories added to it. In such case, the minimum mixture measure is the smallest fraction of inconstant eligible voters. Second, the set of eligible voters might change from election to election, which is rather likely if the elections are not simultaneous. If the extent of it is not low enough to ignore, the model of party-election independence (5.2) can be considered more generally as one of stability of electoral support, and the minimum mixture measure as its index. In such case, the measure is not a fraction of inconstant voters, but of votes that would need to be removed for the party vote shares to be constant across the elections.

The calculation of the measure is highly intuitive, as illustrated by the example in Table 5.5. Each margin of the contingency table with unobserved values represents a distribution of votes over the same set of parties. The procedure rests on splitting each marginal distribution into two parts–shared and not shared–while maximizing the size of the former. Taking advantage of the fact that the component is shared by all margins, this is achieved by taking for each party its lowest vote count across the elections. In this case $\min\{9,7\} = 7$ and $\min\{1,3\} = 1$, giving the following distribution as the largest possible of constant voters $c = \{7,1\}$ and the corresponding distribution of switchers $s = \{2,0\}$. Table 5.6 further illustrates the decomposition–the votes are split into two classes, those who lie on the main diagonal and those who do not, while maximizing the sum of the

Table 5.6: Decomposition of the vote under the minimum mixture measure of voter transitions.

	Constant		Split		All	
	P=A	P=B	P=A	P=B	P=A	P=B
S=A	7	0	0	2	7	2
S=B	0	1	0	0	0	1

Table 5.7: Fictitious returns for three simultaneous elections from a single district with ten thousand voters.

	Election		
Party	1	2	3
A	5	4	4
B	4	3	3
C	1	2	1
abst.	0	1	2

former.

The mixture measure has in some settings the same value as the dissimilarity index and the Pedersen index. This is not the case generally, as illustrated by the example in Table 5.7. The Pedersen index, including abstentions, equals 0.2 and 0.1 for the two pairs of consecutive elections, and the overall value of the dissimilarity index is 0.1. However, the minimum mixture measure equals 0.2, as at most four thousand voters voted for A, three thousand for B, and one thousand for C in all three elections. Thus, in case the three elections used a single ticket, the minimum rate of split tickets is 20%.

This section has discussed settings where constant voting and stability of electoral support are understood as lack of association between party and election. In such applications, only two sets of categories into which the votes fall are considered–party (or, more generally, option) and election. In practice, researchers might want to consider additional

categorizations, such as the office for which the elections were held, or the constituencies in which the votes were cast. Furthermore, in some cases researchers might want to consider restrictive models of voter transitions, such as the proportional or the uniform swing. The following two sections present such extensions and demonstrate them on the example of the 2004 and 2008 Montana presidential and gubernatorial elections and the 1966 and 1970 UK general elections.

5.4 Conditionally Constant Voting

The previous sections have discussed constant voting simply as sticking with a party in a series of elections. In that respect, only two attributes of each vote were considered–party and election. However, voters' choices can be constant conditional on some other factors. An example of this are polities where elections to multiple offices regularly take place together in batches. In such settings, voters can vote for the same party across offices within batches, but switch parties across them. Vice versa, they can stick to a party's candidates for a specific office across batches, but with other parties for different offices. The former voting is hereafter called *batch-constant*, the latter *office-constant*, and unconditionally constant voting as *fully constant*. This Section presents an application of the mixture index to models of conditionally constant voting, using office- and batch-constant voting as examples.

The approach is illustrated on the Montana elections of 2004 and 2008, in which a substantial number of tickets was split–the Democratic candidate Brian Schweitzer won both gubernatorial contests, but the Republican candidates George W. Bush and John McCain carried the state's Electors. Moreover, the Democrats did better in both contests in 2008 than in 2004. Finally, disaggregated returns (Ansolabehere et al., 2014) allow to explore not only the extent of cross-office and cross-batch transitions, but also how it varies

Table 5.8: Returns for the major parties Montana elections of 2004 and 2008. The 'other' category contains the lowest number of voters who in a given election either voted for a third-party candidate, abstained, cast an invalid ballot, or weren't eligible to vote. The office-party totals differ from the overall ones due to several missing precinct-level returns. Each columns sum to 478,977. Source: Ansolabehere et al. (2014).

	2004		2008	
	President	Governor	President	Governor
Democrat	173,060	224,226	231,430	318,274
Republican	265,064	204,436	242,265	157,924
Other*	40,853	50,315	5,282	2,779

with place. The vote totals in the data are shown in Table 5.8.

To simplify the exposition, most of the following analysis assumes an ignorably low turnover in eligible voters. Furthermore, the election (office-batch) marginals are set to equal. This is achieved by including the lowest possible counts of voters who either were not eligible, abstained, cast an invalid ballot, or voted for third party candidates. Since the data does not contain the numbers of eligible voters, which would have been preferable, the size of the category is simply calculated so that within each of the 56 counties the four columns sum up to the same number. These simplifications are not required by the method, and the next section presents examples that do not use them.

The three types of voting–fully constant, batch-constant, and office-constant–imply different patterns of association between party, office, and batch. The patterns can be captured with the log-linear models shown in Table 5.9. *Fully constant* voting is sticking to a party independently of the office and batch. The corresponding log-linear model is (a), where the count of votes c for the candidate of i^{th} party for the j^{th} office in the t^{th} batch is a function of the main term λ^0 and the party term λ^P. If each election (office-batch combination) contains the same number of voters, the same values are expected under the independence model (b), where λ^O and λ^B are the office and batch terms, respectively.

Table 5.9: Log-linear models for votes cross-classified by party, office, and batch (election).

Constancy	Functional Form				
Full	$\ln c_{i,j,t}$	$= \lambda^0 + \lambda_i^P$			(a)
	$\ln c_{i,j,t}$	$= \lambda^0 + \lambda_i^P$	$+\lambda_j^O + \lambda_t^B$		(b)
Batch	$\ln c_{i,j,t}$	$= \lambda^0 + \lambda_i^P$		$+\lambda_{i,t}^{PB}$	(c)
Office	$\ln c_{i,j,t}$	$= \lambda^0 + \lambda_i^P$			$+\lambda_{i,j}^{PO}$ (d)

Terms		
λ^0	*Main*	The 'intercept.'
λ^P	*Party*	Party shares vary.
λ^O	*Office*	Electorate sizes varies with office.
λ^B	*Batch (year)*	Electorate sizes varies with batch.
λ^{PB}	*Party-batch interaction*	Party shares vary with batch.
λ^{PO}	*Party-office interaction*	Party shares vary with office.

Should the number of observations in each column differ, the interpretation of (b) would change into one of the stability of electoral support, as discussed in the previous Section. The expected values under the model can be calculated simply by taking the minima of the rows of Table 5.8.

Batch-constant voting means that the party is associated with the batch, but not with the office. The corresponding log-linear model (c) now includes also the party-batch interaction λ^{PB}. The expected values can be calculated by taking the row-wise minima separately for the first pair of columns and for the second pair of columns of Table 5.8.

Office-constant voting means choosing the party depending on the office, but not on the batch. This corresponds to the log-linear model (d), which includes the party-office interaction λ^{PO}. The expected values can be calculated by taking the row-wise minima separately for the first and third column and the second and the fourth column of Table 5.8, respectively.

The three models cannot describe 30, 15, and 16% of the aggregated data in Table 5.8, respectively. The number of units from each of the four elections is the same, and they

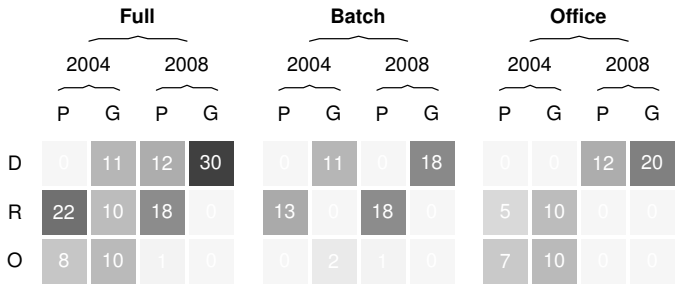


Figure 5.1: Cell residuals under the three constant voting models fit to the aggregated Montana data as percent of the column total. Data source: Ansolabehere et al. (2014).

include not only votes, but in the 'other' category also abstentions. Consequently, different interpretations of these values are possible. If the same voters were observed in the four columns, the values are fractions of the voters. If, however, only the observations from the same batch are by the same voters, they are fractions of tickets. From the substantive point of view, the rate of fully constant voting does not appear high–at least about one in three voters was not fully constant. However, the value is lower for both the batch- and office-constant voting, under which only about one in six voters was not constant.

To guide the substantive interpretation of the statistical findings, the mixture index provides also a new kind of residual analysis (Rudas et al., 1994; Clogg et al., 1995, 1997; Rudas, 2002). Figure 5.1 shows which observations in the aggregated Montana data do not fit the three models. For fully constant voting, most of them are tickets split between the Republican presidential nominees and the Democratic gubernatorial candidate. The batch-constant model highlights this–practically all its residuals are tickets split this way. The office-constant model simply shows that the Democrats did better in 2008 than in 2004.

In some settings, researchers might wish to take into account also the place of the vote. In the context of constant voting, the most direct way is by fitting the models from Table 5.9 separately to each territory. This is equivalent to adding terms to the models,

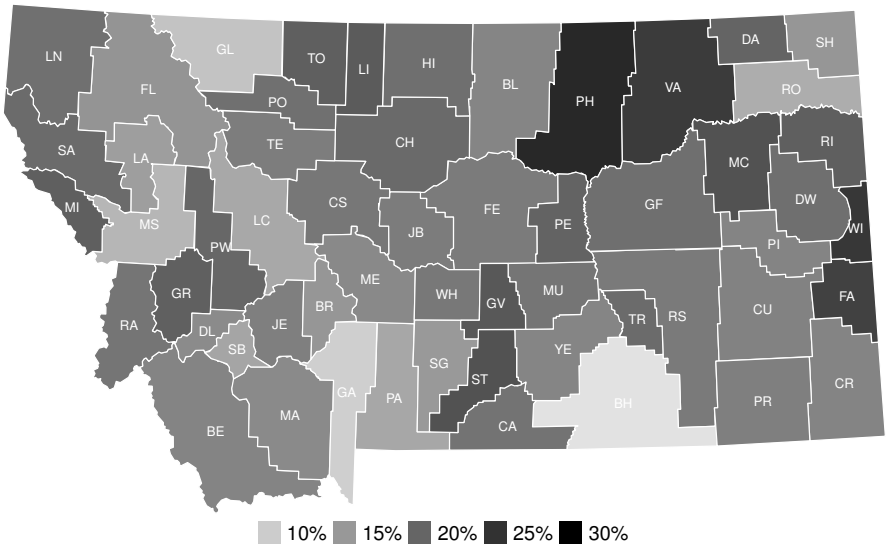


Figure 5.2: County residuals under batch-constant voting in Montana (2004–2008). Data source: Ansolabehere et al. (2014).

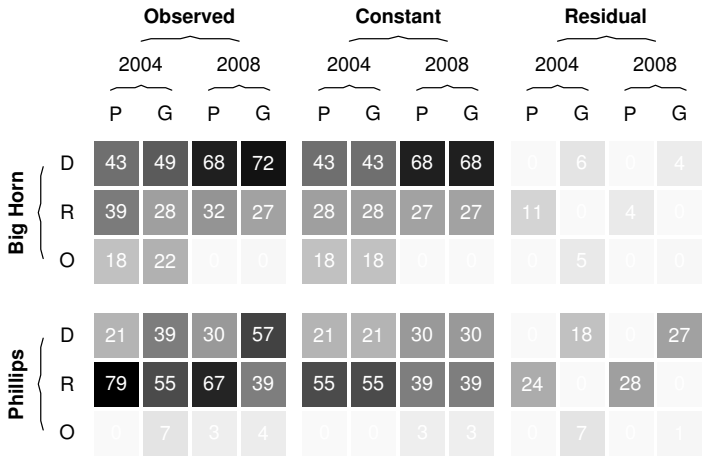


Figure 5.3: Decomposition of the vote from two selected Montana counties into a batch-constant and a residual class. The model fits the best in Big Horn county (4×5144 votes, residuals=8%) and the worst in Phillips county (4×2133, 26%). Data source: Ansolabehere et al. (2014).

Table 5.10: Descriptive statistics for county residual fractions under the three models of constant voting fit to the Montana data. Data source: Ansolabehere et al. (2014).

	Statistics						
Model	Min.	1$^{\text{th}}$ Qu.	Median	Mean	3$^{\text{rd}}$ Qu.	Max.	Var.
Fully c.	22	28	31	32	33	65	39
Batch-c.	8	16	18	18	20	26	12
Office-c.	6	11	14	15	17	52	45

and for models (a), (c), and (d) takes the form of

$$\ln c_{i,j,t,k} = \cdots + \lambda_k^T + \lambda_{ik}^{PT},$$

where k indicates the territory, $\{\dots\}$ is the sum of the terms already in the model, λ^T are territory terms which allows the numbers of observations to vary across territories, and λ^{PT} party-territory interaction which allows the party shares to vary across territories.

Increasing the number of classifications adds marginals, which can bring the value of the measure closer, but not farther, to the unobserved true rate of inconstant voting. In other words, lowering the level of aggregation will either preserve or increase the index. For the county-level data the values of the index for the three models are 31%, 15%, and 16% respectively. The distributions of the residuals are summarized in Table 5.10.

The inspection of the aggregate data has shown a pattern of splitting the tickets between the Republican presidential nominees and the Democratic gubernatorial candidate. The county-level data shows the same pattern. Notably, although the median residual fraction under batch-constant voting is larger than under office-constant one, there are fewer and smaller large values, and overall the values are closer to the mean (Table 5.10, see also Figure C.1 in the Appendix). In other words, the lowest rate of ticket splitting is fairly even across the counties. This is illustrated also by Figure 5.2, which shows the county

residual fractions under batch-constant voting. Clusters of counties with higher rates of ticket splitting are in Montana's northeast, and with lower rates in the western half of the state.

The decomposition of the votes into the batch-constant and residual components in the two counties with the lowest and highest residual fractions is shown in Figure 5.3. The index is the smallest (8%) in the Big Horn (BH) county, in which Democrats won all four inspected elections with similar shares, and the highest (26%) in Phillips (PH) county, where Republicans won with similar shares all contests except the 2008 gubernatorial election.

5.5 Constant Voting and Swing

In some contexts, the net change in electoral returns experienced by a party is known as 'swing.' In systems with two major parties alternating in power one party's net gain is often close to the negative gain of the other party. Such systems typically have districts of small magnitude, and even small changes in the overall vote shares can produce dramatic reversals in seat shares. The extent of this depends on the distribution of the change across constituencies. In this context, two kinds of swing are distinguished–'proportional' and 'uniform' (see e.g. Miller, 1972; McLean, 1973; Johnston and Hay, 1982).

Both types of swing are usually understood through changes in party's shares of local valid votes. This is a convenient simplification that facilitates analysis and its presentation. However, it can unappealingly affect inferences. If constituency turnout and/or electorate size change, the difference between the shares in two elections cannot be readily interpreted in terms of individual behavior. Relatedly, due to differences in turnout and/or electorate size, vote totals can vary across the constituencies. Yet, using the shares implicitly assigns the same weight to the constituencies. The issue is bypassed by defining the swing types in terms of vote counts as opposed to shares.

Proportional swing means that a party's local net gain of votes is proportional to its local vote in the previous election. For example, if a party received 30% in one district and 50% in another in the first election, and 36% and 60% in the following election, the swing is proportional in the sense that the new vote shares can be represented as the old shares multiplied by the same number. In terms of vote counts,

$$v_{i,t=2,k} = p_i \, v_{i,t=1,k},$$

where i indicates the party, t the election, and k the territory, and p the swing weight. The weight is always non-negative, weights below one mean loss, and above one gain of votes.

Under uniform swing, the net gain has in all constituencies the same size in terms of share of local votes cast. If a party received 5% of votes in one constituency and 70% in another in the first election, and 10% and 75% in the second, the swing is considered uniform, in the sense that the gain is in all constituencies the same fraction of the votes cast. In terms of vote counts,

$$v_{i,t=2,k} = v_{i,t=1,k} + u_i \sum_i v_{i,t=1,k},$$

where u_i is the swing weight of the i^{th} party. The party weights sum to zero, and for two parties $u_1 = -u_2$.

The labels 'proportional' and 'uniform' are not sufficiently specific–both types can be considered as uniform and proportional. Under both types, a party's swing weight (p_i or u_i) has the same value across constituencies, i.e., is uniform. Furthermore, even under uniform swing the net gain is proportional, to the number of votes cast. Nevertheless, since these labels are widely adopted, they are used throughout the chapter. More precise labels can be devised using the above definitions of the two types of swing. Consider the generalized

model

$$v_{i,t=2,k} = p_i \, v_{i,t=1,k} + u_i \sum_i v_{i,t=1,k},$$

which includes both types. The swing weights capture the association of a party's local vote with its previous local vote and the previous total vote, respectively. The uniform swing model restricts the 'proportional' weight p to one and the proportional swing model restricts the 'uniform' weight u to zero. In other words, the proportional swing is independent from local vote totals.

Proportional and uniform swing are restrictive models, and as a consequence will not always describe the reality perfectly. Thus, the observed votes can be divided into three latent classes–those cast by constant voters, those cast by swing voters, and those cast by other voters–and the size of these classes will not be known ex ante. The mixture index can be applied to obtain the lowest possible size of the residual class.

The previous discussion of swing relied on modeling the new votes as functions of previous votes. It is useful to consider the models more generally. Formally,

$$v_{i,t,k} = m_{i,t,k} + r_{i,t,k},$$

where m are votes described by the restrictive models, r the residual votes, and the mixture index minimizes the size of the residuals

$$\pi^* = \frac{\sum_i \sum_t \sum_k r_{i,t,k}}{\sum_i \sum_t \sum_k v_{i,t,k}}.$$

The votes that belong to the restrictive models are

$$m_{i,t,k} = c_{i,t,k} + s_{i,t,k},$$

where c are votes cast by constant voters, and s by those who follow the swing. The restrictions on the distributions of votes cast by the constant voters and by the swing voters are different under the two swing models.

Furthermore, it is useful to define the model of voting constant conditionally on constituency. This model reflects the fact that even if all voters are constant, their distribution over parties can differ across constituencies. The log-linear form is

$$\ln c_{i,t,k} = \lambda^0 + \lambda^P_i + \lambda^C_k + \lambda^{PC}_{i,k},$$

where λ^{PC} is the party-constituency interaction.

The proportional swing model can be considered as a generalization of this model. Under proportional swing, the support of the parties is allowed to differ not only across constituencies, but also across elections. Its log-linear formulation is

$$\ln m_{i,t,k} = \lambda^0 + \lambda^P_i + \lambda^E_t + \lambda^C_k + \lambda^{PE}_{i,t} + \lambda^{PC}_{i,k},$$

where and λ^{PE} is the party-election and terms. Although swing is typically of interest for two parties in a pair of elections, the log-linear formulation applies to any number of parties and elections.

The uniform swing cannot be defined as a log-linear model, since the association between the returns and the total number of votes is linear. Instead, it is the linear model

$$m_{i,t,k} = c_{i,k} + s_{i,t,k},$$

where c are votes cast by constant voters, and s by the swing voters. For two parties and two elections, the swing gain of one party equals the loss of the other party. This allows to

use the linear model

$$m_{i,t,k} = c_{i,k} + m_{\cdot,\cdot,k} \times u \times g_{i,t},$$

where $c_{i,k}$ are party-constituency fixed coefficients ('effects'), $m_{\cdot,\cdot,k}$ the sum of constant and swing votes in k^{th} constituency, u the rate of votes from swing voters, and $g_{i,t}$ an indicator whether the i^{th} party won the t^{th} election.

The mixture index has two-fold use in this context–it provides quantities that can be given substantive interpretation, and it aids model selection. The main quantity of interest is the lowest share of votes that cannot be attributed to either constant or to swing voters. If it can be assumed that the same set of voters took part in both elections, this quantity is also the lowest possible rate of voters who are neither constant nor follow the national swing. Furthermore, the residuals can be inspected to identify the categories of votes not described by the underlying theory.

The models of no swing, proportional swing, and uniform swing are restrictive, and will thus not describe reality perfectly. In any election some swing is practically guaranteed, and will not be perfectly proportional or uniform. The question is to what extent are the models analytically useful. Conventionally, models are selected with tests of statistical significance of the deviation from the null hypothesis that the model is true or that a parameter is zero. Whatever the many demerits of such approach, in this setting it is practically guaranteed to reject any restrictive models. The reason lies in the fact that the conventional tests such as the χ^2 or the t-test are sensitive to the sample size, and election results are typically large. The mixture index of fit is not affected by these issues.

Well-known cases of strong regularities in net vote gains across constituencies are mid-20^{th} century British general elections (Butler and Stokes, 1969; Miller, 1972; McLean, 1973). The 1966 and 1970 British general elections are especially well suited to demonstrate the method. The 1970 Conservative and Labour swings of $\pm 5\%$ brought a change of government,

Table 5.11: Conservative and Labour votes in the 1966 and 1970 British general elections in the 614 constituencies where both parties fielded candidates in both elections. Excluded voters are those who either cast their ballot for a third party or an invalid one, or abstained. Sources: Kollman et al. (2014); Kimber (2015).

	1966	1970
Conservatives	11,031,122	12,707,042
Labour	12,948,487	12,037,087
Excluded*	10,882,868	13,377,166
Eligible voters	34,862,477	38,121,295

and in both elections the two parties took more than 85% of the votes. Returns for 628 of 630 constituencies are available in the Constituency-Level Data Archive (Kollman et al., 2014). In 614 of these constituencies, Labour and Conservatives fielded candidates in both elections. The vote totals from these constituencies are shown in Table 5.11. Three erroneous entries in the CLEA data were corrected using another dataset (Kimber, 2015).

Figure 5.4 shows Conservative shares of two-party vote across the constituencies. The difference between the 1966 and 1970 shares appears to fall into a relatively narrow range, and its size independent of the 1966 share. The 1966 and 1970 shares correlate strongly (ρ=0.98) and the standard deviation of their difference is 3%. Furthermore, the difference between the 1970 and 1966 Conservative shares correlates weakly (ρ=0.07) with their 1966 shares and moderately (ρ=0.25) their 1970 shares. Practically the same holds for the ratio of 1966 Conservative votes over the 1970 votes (ρ=0.09 and 0.21, respectively). This evidence suggests that both the proportional and the uniform swing might be good models for the elections.

Before considering the fit of the proportional and uniform swing models, it is useful to consider two more restrictive models–independence and no swing–to serve as a baseline for comparisons. Under the independence model the parties local vote shares are constant across

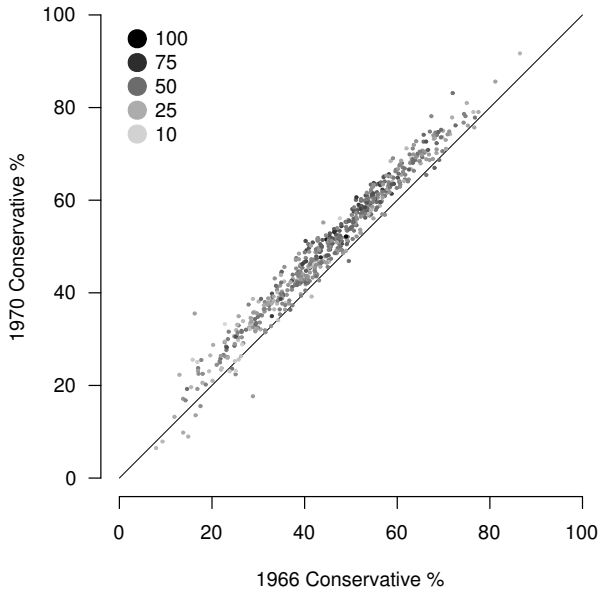


Figure 5.4: Conservative constituency-level shares of the two-party vote across 614 constituencies contested by Conservatives and Labour in 1966 and 1970. Color according to the two-party vote size in 1970 (examples for thousands). Sources: Kollman et al. (2014); Kimber (2015).

Table 5.12: Comparison of three models of swing fit to the UK data. Component weights in rounded percents.

Swing	No. Param.	Constant	Swing	Residual
None	1,228 (1,845)	93	–	7
Proportional	1,232 (1,851)	90	5	5
Uniform	1,229	92	3	5

territories and elections. Substantively, it corresponds to no volatility and no residential segregation of voters. It fits badly, and describes at best 73% of the votes. It appears that party choice is associated with territory and/or election. The second model–constant voting without swing–includes the party-constituency association, and fits the data markedly better, describing at best 93% of the votes.

The proportional swing model lifts a further restriction–not only can party support vary

Table 5.13: Descriptive statistics for constituency residual fractions under the three models fit to the UK data. Data sources: Kollman et al. (2014); Kimber (2015).

	Statistics						
Swing	Min.	1th Qu.	Median	Mean	3rd Qu.	Max.	Var.
None	2	5	7	7	9	36	9
Proportional	1	3	4	5	6	35	10
Uniform	2	4	5	6	7	35	10

across constituencies, but also across elections, but only to a single national pattern. The model describes the data well–at least 5% of the votes was cast neither by constant nor by swing voters. The overall proportion of votes cast by constant voters 90% and by swing voters 5%. The swing weights are 1.18 for Conservatives and 0.93 for Labour. The fit of the uniform swing model is similar. Only at least 5% of the votes come neither from constant voters (92%) nor from those, who switch according to the 'uniform' swing (3%).

The constant voting only model fits the data well, reflecting the fact that under the British electoral system even small changes in the overall support can produce considerable turnovers in the Commons. The proportions of votes by constant, swing, and other voters under both swing models are close to each other. This is not surprising, as the models have a similar number of parameters (Table 5.12). Importantly, the proportional swing model has two parameters for the swing component, but the uniform swing model only one. Accordingly, although both residual fractions round to 5%, their difference rounds to 1% in favor of the proportional swing. This is further illustrated by the summaries of constituency residual shares in Table 5.13.

How well is each constituency described by the models, and how much it contributes to their misfit, is indicated by the constituency residuals. The best and worst described constituencies are shown in Figure 5.5 and those that contribute the least and most to the aggregate residuals in Figure 5.6. One constituency makes four appearances in the figures,

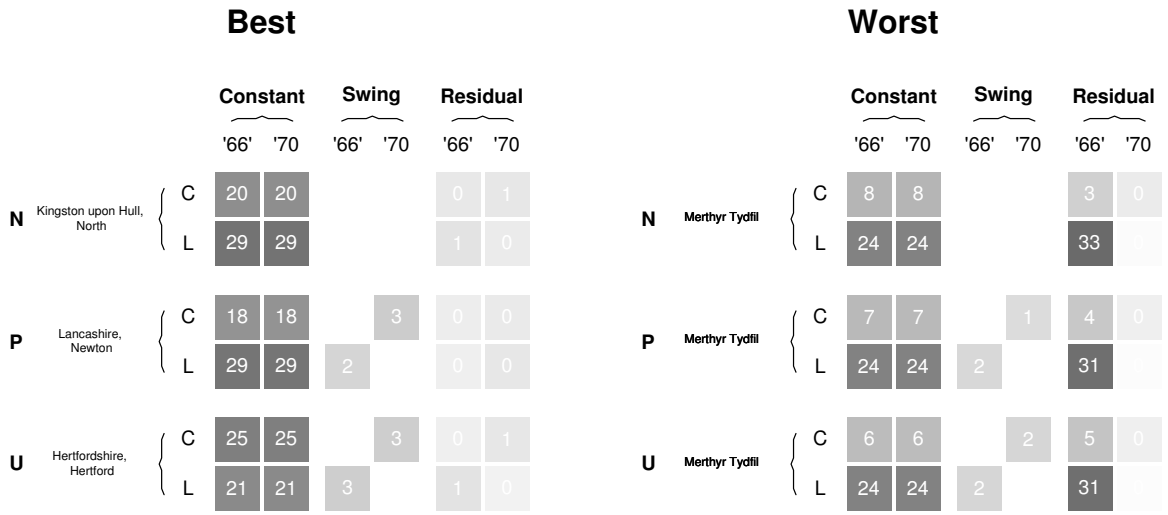


Figure 5.5: Shares of constant, swing, and residual votes in four constituencies with the best and worst fits in terms of residual fraction. Values in rounded percents of constituency two-party votes. Best fits are in constituencies Kingston upon Hull, North (88,725 votes, 2% residual), Lancashire, Newton (119,482 votes, 1% residual), and Hertfordshire, Hertford (127,228, 2% residual). Worst fits are in constituency Merthyr Tydfill (38,222 votes, 36, 35, and 35% residuals). N–no swing; P–proportional swing; U–uniform swing. Data source: Ansolabehere et al. (2014).

always among the worst described–the Welsh Merthyr Tydfill. No other constituency gets close to it in misfit under any of the three models (see Figure C.2 in the Appendix). As its residual votes show, its misfit its due to the 1966 Labor votes, of which there are more than twice as many as any of the three models would have. The 1970 elections in this constituency were fairly unusual. The veteran Labour incumbent SO Davies split with his Constituency Labour Party, and defended his seat against its official candidate as an independent (BBC, 2005).

Another pattern shown in Figure 5.5 is that the best described constituencies are those with small and moderate margins between the major parties. However, as shown in Figure 5.7, the residual fractions are only moderately correlated with two party margins.

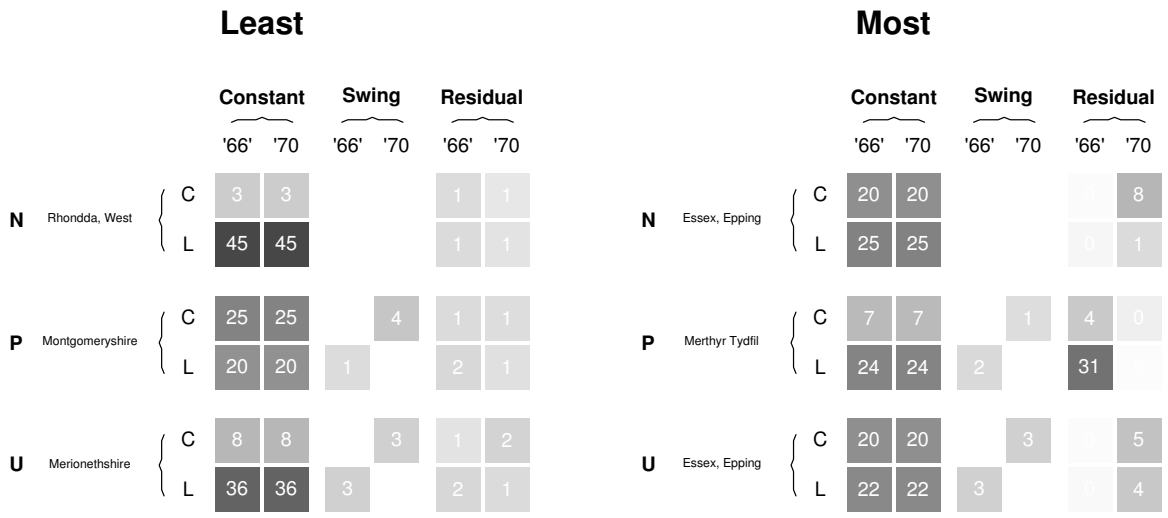


Figure 5.6: Shares of constant, swing, and residual votes in five constituencies with the least and most contributions to the aggregate residuals. Values in rounded percents of constituency two-party votes. Best fits are in constituencies Rhonda, West (41,404 votes, 4% residual), Montgomeryshire (25,901 votes, 6% residual), and Merionethshire (23,402, 6% residual). Worst fits are in Essex, Epping (154,975 votes, twice 9% residual), and Merthyr Tydfill (38,222 votes, 35% residuals). N–no swing; P–proportional swing; U–uniform swing. Data source: Ansolabehere et al. (2014).

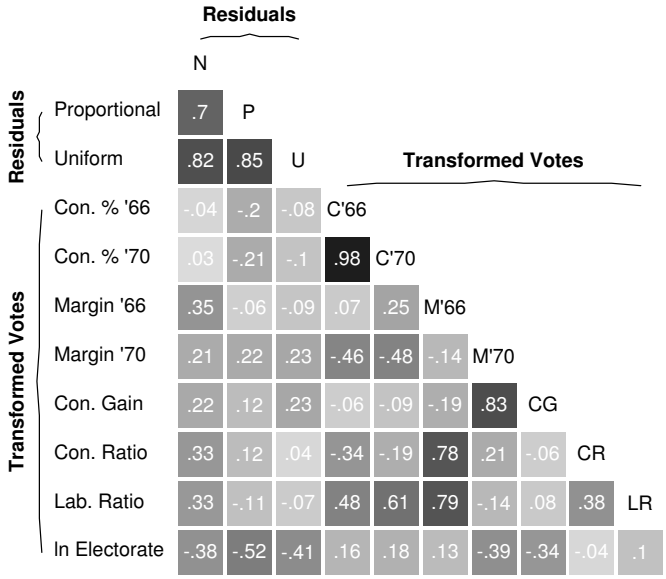


Figure 5.7: Pearson correlations of residuals and transformed votes (N=614). Data source: Ansolabehere et al. (2014).

5.6 Conclusion

Voter transitions are often estimated from aggregate data. The existing methods offer various trade-offs between how informative the generated quantities and how easy and testable the underlying assumptions are. This chapter proposed a new method for the analysis of aggregate data in settings where voter transitions are of interest. The method is based on the mixture index of fit, and rests on splitting the observations into those described by a restrictive model, such as uniform swing or constant voting, and those that are not, while minimizing the share of the latter. The method is highly flexible, and applies to any positive numbers of parties and elections. In this chapter, the method has been applied to several restrictive models of cross-election voting behavior: fully constant voting (always sticking with the same party), batch- and office-constant voting (sticking to a party only within a cycle, or for an office), and to proportional and uniform swing.

Chapter 6

Roll Call Analysis with the Mixture Index of Fit

This chapter introduces the mixture index of fit to roll call analysis. The index provides a general framework applicable to a variety of problems from this domain and instead of replacing the existing methods enhances them. In this chapter, it is applied to measure partisan and other kinds of group voting, and evaluate and substantively interpret ideal point models. The applications are illustrated with congressional roll calls related to the Civil Rights Act of 1964.

6.1 Introduction

Roll call votes are widely used in study of legislative politics, and typically analyzed statistically. The generated quantities summarize the votes on the level of individual legislators or parties, and are substantively interpreted as 'ideal points,' or measures of characteristics such as party unity or cohesion. This chapter introduces the mixture index

of fit (Rudas et al., 1994) to roll call analysis, to provide a flexible general framework that enhances the existing methods rather than replacing them. The approach rests on decomposing the observations, such as legislators or their votes, into two classes–one described by the model of interest, and one that is not–while minimizing the size of the latter class. The chapter provides a set of tools which can be applied to tasks such as measurement of partisanship and party cohesion, or dimensionality selection in ideal point estimation. The methods are presented here within the context of legislative voting, with focus on parties, individual legislators, and their geographically defined groupings, but can be used in any context where the same statistical models are applied to analyze voting behavior.

To demonstrate the methods, roll calls on the Civil Rights Act of 1964 (CRA, or the Act) introduced in Section 7.2.1 are used throughout the chapter. The mixture index of fit is introduced in Section 6.3 by way of defining new measures of partisan, or more generally, group voting in legislatures. Section 6.4 presents new measures of party cohesion and unity from the ideal points perspective by applying the mixture index to preference models commonly used in sensory research. In Section 6.5 this approach is extended to the related and more flexible models from the domain of Item Response Theory (IRT), which are used to estimate the legislator ideal points. The use of the index to evaluate ideal point estimates by testing for Differential Item Functioning (DIF) is introduced in Section 6.6. The chapter concludes by discussing some additional possible applications of the mixture index in the study of legislative politics.

6.2 Data: The Civil Rights Act of 1964 in the U.S. Congress

To illustrate the introduced methods, this chapter uses the U.S. Congress roll calls related to the Civil Rights Act of 1964 available at `www.voteview.com` (Poole and Rosenthal, 2000). Three reasons motivate this choice. First, it is one of the most widely known legislative battles (see e.g. Rodriguez and Weingast, 2003; Jeong et al., 2009). Second, the support and opposition to the Act cut across party lines. Third, the opposition was particularly strong among congressmen from the South. Thus, it provides a good example for investigation of cohesiveness of alternative legislator groupings.

Throughout the chapter, both 'ayes' are 'yeas' ar labeled as 'ayes,' so that their one-letter abbreviations correspond to *a*ffirmative and *n*egative. To simplify the exposition, absences or abstaining from voting while present are excluded, as well as are paired ayes and nays. To the same aim, announced ayes are merged with ayes, as are announced nays with nays. In some analyses, the legislators are grouped by whether their state was in the Confederacy,[1] and in some others based on their U.S. Census Regions or Divisions (see Table D.1 in the Appendix).

The bulk of the legislative battle took place in the Senate, where a total of 120 roll calls related to the Act were called, as opposed to the three in the House (Table 6.1). The subset of the data used the most in the illustrations are the six out of the 120 Senate roll calls related to the Act, in which all 100 senators either cast or announced an aye or nay. The roll calls are described in Table 6.2, and shown in Figure 6.1. All the 120 CRA-related Senate roll calls are shown in Figure D.1 in the Appendix.

[1]The following 11 states are classified as former CSA members: Alabama, Arkansas, Georgia, Florida, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee, Texas, and Virginia.

Table 6.1: Three House roll calls related to the Civil Rights Act of 1964. Ayes and Nays exclude the announced ones. Quotes from the Voteview descriptions in inverted commas. All dates in 1964. Data source: Poole and Rosenthal (2000).

No.	Date	By	Ayes	Nays	Description
127	Feb. 8	Albert, D-OK	220	175	'Albert motion that the House adjourn until Monday, Feb. 10, rather than move immediately to considerations of final titles of the bill and a vote on passage.'
128	Feb. 10		290	130	Passage of H.R. 7152
182	July 2		289	126	'H.R. Civil Rights Act of 1964. Adoption of a resolution (H. Res. 789) providing for House approval of the bill as amended by the Senate.'

Table 6.2: Six Senate roll calls related to the Civil Rights Act of 1964. Ayes and Nays exclude the announced ones. Quotes from the Voteview descriptions in inverted commas. All dates in 1964. Data source: Poole and Rosenthal (2000).

No.	Date	By	Ayes	Nays	Description
298	June 09	Morton, R-KY	51	48	'Morton amend. to entitle a defendant to demand a trial by jury on a criminal contempt charge arising under any section of the Act except Title I, covering voting rights.'
302	June 10	Mansfield, D-MT	49	48	'Mansfield[-]Dirksen motion that the Senate invoke cloture on the Southern filibuster.'
304	June 10	Russell, D-GA	40	59	'Russell amend. to postpone the effectiveness of the public-accommodations section until Nov. 15, 1965.'
388	June 17	Ervin, D-NC	47	51	'Ervin. amend. to delete authority for one member of the Equal Opportunities Comm. to file a charge of discrimination and initiate an investigation.'
408	June 19	Gore, D-TN	25	74	'Gore motion to recommit to the Judiciary Comm. w[ith] instructions that it report it back "forthwith" w[ith] the amend. stating that federal funds should not be w[ith]drawn from any school district unless that district had disobeyed a court order that is desegregate.'
409	June 19		73	27	Passage of H.R. 7152

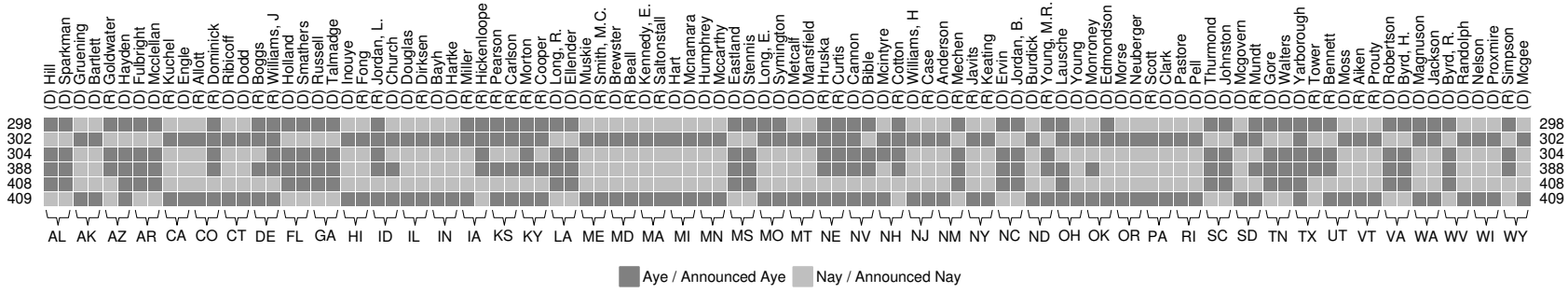


Figure 6.1: Six Senate roll calls related to the Civil Rights Act of 1964. Data source: Poole and Rosenthal (2000).

6.3 Group Cohesion and Partisan Voting

In the study of legislative politics, party cohesion or unity is typically measured from roll call data. The measures are defined for parties, but alternative groupings, such as geographic or demographic ones can be inspected with them as well. The measures can be though of as distance metrics from the observed legislative behavior to a model under which all the legislators from the same groups act the same way.

The indexes of cohesion are defined for a single party in a single roll call, and input its votes $\mathbf{v} = \{v_1, \ldots, v_O\}$, where O is the number of options, such as 'aye,' 'nay,' or 'abstain.' The oldest is the Rice (1928) index,

$$RI = \frac{|v_1 - v_2|}{\sum_o v_o},$$

defined for two options–aye and nay–only. The more recent Attinà (1990) index,

$$AI = 2 \times \frac{\max \mathbf{v}}{\sum_o v_o} - 1,$$

was defined for three options. Finally, Hix et al. (2007) rescaled the Attinà index to the unit interval,

$$HI = \frac{3}{2} \times \frac{\max \mathbf{v}}{\sum_o v_o} - \frac{1}{2}.$$

This can be generalized to any number of options as

$$CI = \frac{1}{O-1} \times \left(O \times \frac{\max \mathbf{v}}{\sum_o v_o} - 1 \right).$$

The re-scaling assures that the index is always on the unit interval, with endpoints attached to substantively interpretable states, but makes the interpretation of the values between the endpoints somewhat less intuitive. If a one-number summary for a number of roll calls

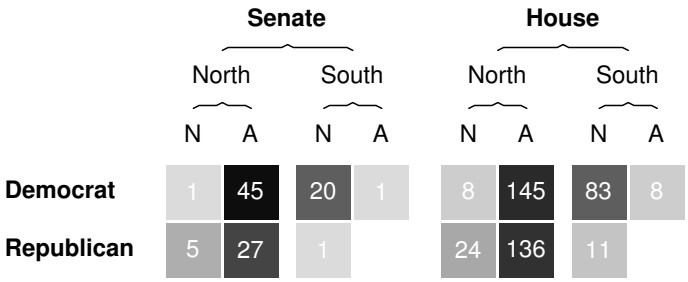


Figure 6.2: Votes on the passage of the Senate version of the Civil Rights Act of 1964 (Senate vote no. 409 and House vote no. 182). Color by proportion in the respective chambers. Abbreviations: N–Nay/Announced Nay, A–Aye/Announced Aye. Data source: Poole and Rosenthal (2000).

and/or parties is desired, the indexes are averaged over them.

A related measure is the Party Unity Score (see e.g. Lupoli, 2009), typically computed for a single legislator over a group of selected roll calls as the share of those in which the legislator votes with the majority of their party. Party values of the score are calculated by averaging the legislator scores, which in the context of a single roll call yields

$$PUS = \frac{\max \mathbf{v}}{\sum_o v_o}.$$

This can be thought of as splitting the legislators into those who follow the party line and those that do not. Furthermore, party unity can be thought of as a model under which the party membership implies only one option.

The same perspective can be applied to measure cohesion. Under perfectly incohesive voting all options from the set are equally likely within the group. Consequently, any option can occur only as many times as the least common one. This yields the Party Incohesion Score

$$PIS = O \times \frac{\min \mathbf{v}}{\sum_o v_o},$$

which is the largest possible fraction of the group evenly distributed over the options.

Applying this to cohesion, the resulting score is the complement of PIS,

$$PCS = 1 - O \times \frac{\min \mathbf{v}}{\sum_o v_o}.$$

Its value is zero if the members are equally split among the options, and one if at least one option is not chosen in the group at all, which corresponds to perfect unity if there are only two options. The PUS, PIS, and PCS always lie on the unit interval, and since they are population fractions, they are easy and intuitive to interpret, which can in some contexts render them preferable to the Rice, Attinà, and Hix et al. cohesion indexes.

The splitting of the legislators into those described by a model and those that are not can be applied to measure partisan voting as well, taking into account that it differs from party cohesion. Cohesion implies that the party's option probabilities differ. On the other hand, partisan voting means that the option probabilities vary across parties. For example, if all parties vote unanimously for the same option, they are perfectly cohesive and unified, but the party is not associated with the option. Thus, partisan voting can be defined as an association between the party and option, and non-partisan voting as its absence. In this formulation, partisan voting is understood purely descriptively, and causality is not considered. Since there is only one way how two categorical variables can be independent, but many in which they can be associated (Goodman and Kruskal, 1954; Kendall and Stuart, 1961), the non-partisan voting is the more restrictive model, and thus offers a better starting point.

From this perspective, in any roll call the votes belong to two latent classes, those that were cast independently of the parties, and those that were not. The fraction of non-partisan votes is

$$\pi_{np} = \frac{\sum_g \sum_o n_{g,o}}{\sum_g \sum_o v_{g,o}},$$

where g indexes the group (party) and o the option, n are the non-partisan votes, and r

the residual ones, and the fraction of the partisan ones is its complement. Consequently, the following representation is possible

$$\mathrm{V} = (1-\pi)\mathrm{M} + (\pi)\mathrm{R}, \tag{6.1}$$

where V is the observed distribution of votes over groups and options, M the distribution that belongs to the model of non-partisan voting, and R a residual distribution.

The residual votes by definition belong to partisan voting, and their distribution is thus unrestricted. Because of this flexibility, different decompositions according to (6.1) are possible, all of which will describe the data perfectly if π lies on $[\pi_l, 1]$, where π_l is the lowest such value that will still result in perfect fit. Consequently, the decomposition (6.1) is a special case of the mixture index of fit (Rudas et al., 1994; Clogg et al., 1995; Rudas, 2002). The mixture index of fit measures the misfit of a statistical model by the smallest fraction of the population which cannot be described by the model. Formally,

$$\pi^*(\mathrm{O}, \mathcal{M}) = \mathsf{inf}\{\pi \colon \mathrm{O} = (1-\pi)\mathrm{M} + \pi\mathrm{R},\ \pi \in [0,1],\ \mathrm{M} \in \mathcal{M},\ \mathrm{R}\ \mathsf{unspecified}\}, \tag{6.2}$$

where O is the observed distribution, M an element from the model ($\mathcal{M}$), and R an unspecified residual distribution. Among its appealing features are reliance on assumptions that are always true, easy and intuitive interpretation, and a new kind of substantively informative residual analysis. These features will be introduced in more detail below, jointly with the applications of the index.

Above, non-partisan voting was defined as option-party independence. For any number of options and groups it corresponds to the log-linear model (see e.g. Rudas, 1998b; Agresti, 2002) of independence

$$\ln m_{g,o} = \lambda^0 + \lambda^G_g + \lambda^O_o, \tag{6.3}$$

Figure 6.3: Decomposition of the votes on the passage of the Senate version of the Civil Rights Act of 1964 in the House (vote no. 182). Color by proportion in the House. Fractional values rounded. Abbreviations: N–Nay/Announced Nay, A–Aye/Announced Aye. Data source: Poole and Rosenthal (2000).

where $m_{g,o}$ are the non-partisan votes for option o in group g, λ^0 the main term, and λ^G and λ^O the group and option terms, respectively. The decomposition (6.2) can be obtained using the EM algorithm (Dempster et al., 1977; Rudas et al., 1994).

To illustrate the use of the model (6.3), consider the votes on the passage of the Senate version of the Civil Rights Act of 1964 shown in Figure 6.2. Figure 6.3 shows on the left the classification of the representatives into a non-partisan group and a residual, partisan, one. About one in eight legislators (12%) cannot be described as voting in a non-partisan way, all Democrats opposed to the Act. Compare this with the right panel of Figure 6.3, which shows non-regional voting, under which the option is independent of the region. This model fits considerably worse, leaving more than on in five legislators (22%) unaccounted for, all Southerners voting against the passage. The contrast between the non-partisan and non-regional voting is much starker if the residuals are transformed into their rows' fractions–one in five Democrats (21%) and nine in ten Southerners (91%).

The use of log-linear models with the mixture index of fit can be extended beyond the measurement of partisanship to multiple group memberships. For the case of *p*arty and *r*egion

$$\ln m_{g,o} = \lambda^0 + \lambda^P_p + \lambda^R_r + \lambda^O_o + \lambda^{PR}_{p,r}, \tag{6.4}$$

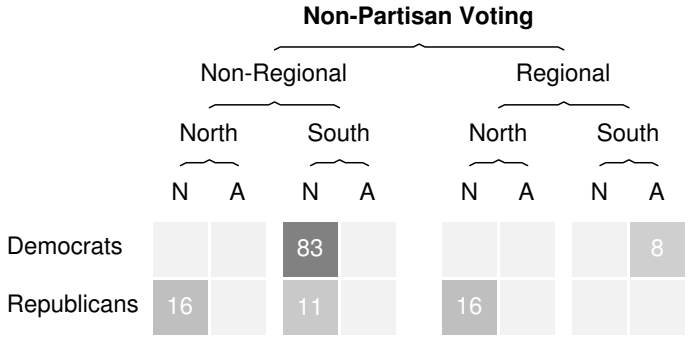


Figure 6.4: Residuals under two models of non-partisan voting. Non-regional voting on the left and regional voting on the right. Fractional values rounded. Empty cells shown as numberless squares. Data source: Poole and Rosenthal (2000).

where λ^P and λ^R are the party and region terms, respectively, and λ^{PR} their interaction, which accounts for different party shares in the regions. Here, voting is neither partisan nor regional. Partisan voting is introduced by the option-party interaction λ^{OP} and regional voting by the option-region interaction λ^{OR}.

For the House roll call on the passage of the Senate version of the Act, shown in Figure 6.2, about one in four legislators (26%) is not described by non-partisan and non-regional voting. Partisan voting is only a slightly better description, leaving more than one in five legislators (23%) out. However, under regional voting only slightly more than one in twenty legislators (6%) are out. As shown in Figure 6.4, while non-partisan and non-regional voting does not account for most opponents of the Act, regardless of their parties and regions, under non-partisan regional voting mainly the Northern Republican opponents of the Act are left out. In short, taking into account that most Southern representatives were Democrats, the opposition to the Act was clearly a regional matter and not a partisan one.

When a set of roll calls is inspected, their index values can be aggregated by taking a weighted average or simply by fitting the log-linear model

$$\ln m_{g,o} = \lambda^0 + \lambda^G_g + \lambda^O_o + \lambda^C_c + \lambda^{GC}_{g,c} + \lambda^{OC}_{o,c}\,, \tag{6.5}$$

where c indexes the roll calls, and λ^C accounts for varying numbers of votes in the roll calls. The two interactions λ^{GC} and λ^{OC} account for the fact that the group turnout and the option tallies might differ across the roll calls. In this context, the mixture index is no longer a fraction of the legislators who are not accounted for, but of their votes.

Furthermore, if the options can be placed on a known dimension, such as support for a specific policy, it is possible to inspect how much of the voting was uniformly partisan. The corresponding log-linear model is

$$\ln m_{g,o} = \lambda^0 + \lambda^G_g + \lambda^O_o + \lambda^C_c + \lambda^{GO}_{g,o} + \lambda^{GC}_{g,c} + \lambda^{OC}_{o,c}\,, \tag{6.6}$$

which contains an additional interaction, λ^{GO}, to allow partisan voting, which is restricted across the roll calls.

The use of the log-linear models (6.5) and (6.6) can be illustrated with the six Senate roll calls reported in Table 6.2 and Figure 6.1. Non-partisan voting does not describe at least one in ten (11%) of the 600 votes cast, but uniformly partisan voting is only marginally better (10%). On the other hand, taking North and South as the relevant groupings, non-regional voting leaves at least one in five (21%) of the votes out, but uniformly regional voting only one in twelve (8%). Under the uniformly regional voting, all the Southern senators are expected to vote with Thurmond, as is a comparable number of the Northern ones. In short, the six roll calls are best described as united Southerners pitted against less united Northerners.

Finally, the roll calls on the Civil Rights Act of 1964 can illustrate the relationship between party unity and partisan voting. Figure 6.5 shows the relationship between party unity and partisanship for the 120 Senate roll calls related to the Act, highlighting five roll calls that represent different patterns of unity and partisanship, which are reported in tables on the right. Non-partisan voting occurs only when there were small differences in

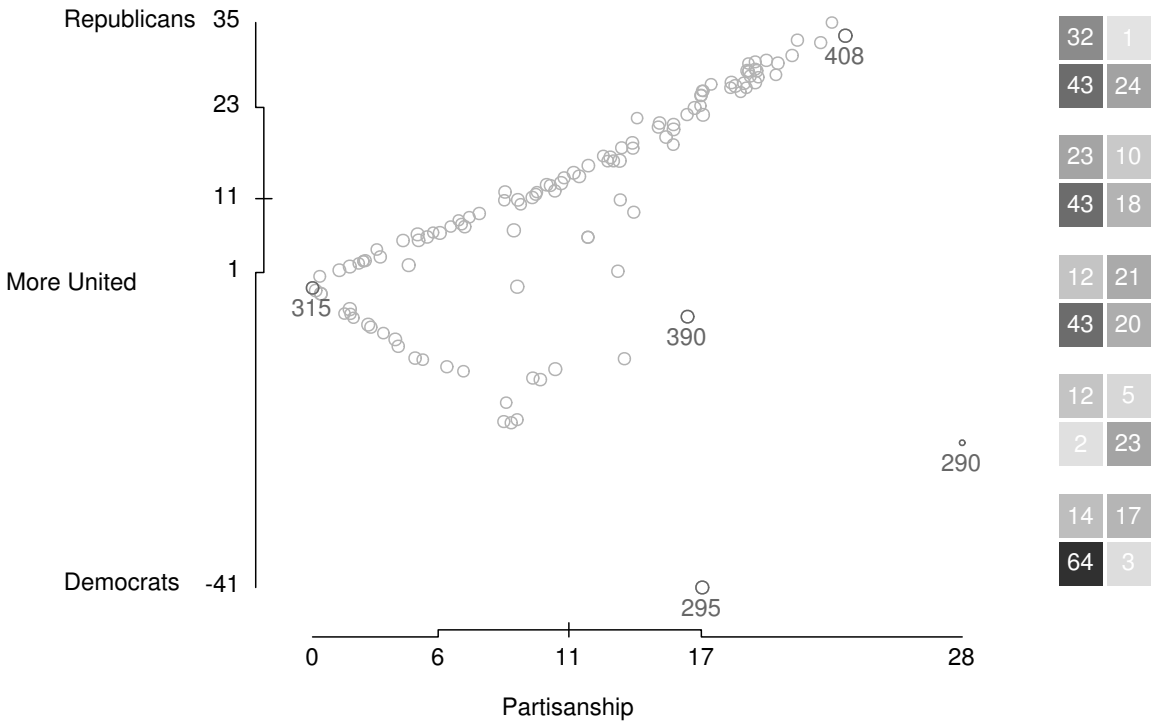


Figure 6.5: Partisanship [x-axis] and the difference in party unity [y-axis] for the 120 Senate roll calls related to the Civil Rights Act of 1964. The axes show the minima and maxima, and 25%, 50%, and 75% quantiles. Point size by the total number of votes cast in the roll call. In the tables first rows are Republicans and first columns Nays; cell color according to Senate seat shares. Data source: Poole and Rosenthal (2000).

party unity–either both parties are united behind the same option, or evenly split among them. On the other hand, partisan voting can occur if only one party is unified, or if both are unified behind opposing options. Detailed plots of the Party Unity Scores and the mixture index of fit for the model of non-partisan voting are reported in Figure D.2 in the Appendix.

6.4 Party Cohesion and Ideal Points

The previous section has operationalized partisan voting as association between party and option, and shown that this does not require the party to unify behind an option. In some

contexts, it is appealing to model legislative voting as a categorical measurement of an underlying position in a continuous space, and subject to stochastic noise. Typically, this is done to extract the 'ideal points' of political actors (see e.g. Clinton, 2012; Armstrong et al., 2014). In terms of legislators' ideal points, a party can be cohesive even if its members do not always vote the same. This section shows how this operationalization of party cohesion can be measured with the mixture index. The ideal points models presented here are more restrictive relatives of the IRT models discussed in Section 6.5.

In a single roll call the distribution of votes over options $\mathbf{v}$ of a party with s seats can be modeled as

$$\mathbf{v} \sim Multinomial(s, \mathbf{p}),$$

where $\mathbf{p} = \{p_1, \ldots, p_O\}$ are the party's option probabilities. For a single roll call, the model will fit perfectly, if the observed option fractions are taken for the probabilities $\mathbf{p}$, which leaves no room for the application of the mixture index. If, however, several related roll calls take place, misfit can occur. Since in a set of roll calls in some ayes and in other nays can correspond to the same underlying policy position, the options can be recoded according to this direction. Assigning equal weight to the roll calls, for the i^{th} representative taking part in n votes the number of times she has voted for the o^{th} option is

$$v_{o,i} \sim Binomial(n, p_o). \tag{6.7}$$

The assumption of identical legislator positions can be relaxed by modeling them with a parametric probability distribution. One possible choice of the distribution is the Beta-binomial. Consequently, (6.7) can be generalized as

$$v_{o,i} \sim Beta\text{-}binomial(n, \alpha_o, \beta_o), \tag{6.8}$$

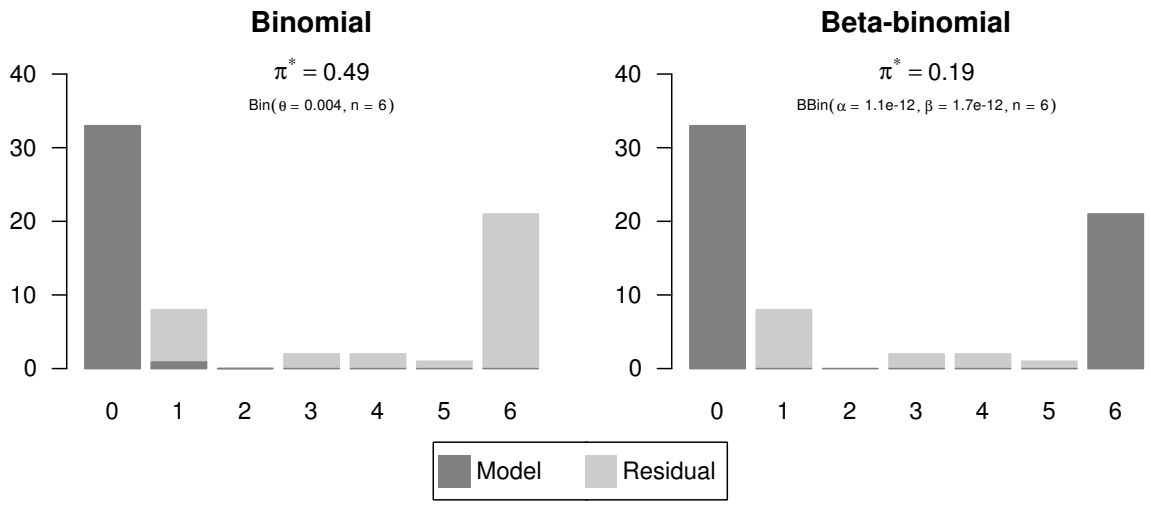


Figure 6.6: The fixed and varying preference models applied to the data on voting among the Democratic senators with S. Thurmond in the six roll described in Table 6.2 and shown in Figure 6.1. Data source: Poole and Rosenthal (2000).

where α_o and β_o are the shape parameters governing the Beta distribution that describes the positions. The shape parameters can be transformed to produce a more intuitive quantity, the mean position $\alpha_o/(\alpha_o + \beta_o)$. The mixture index can be obtained by plugging the models into (6.2).

Models of this kind are used in sensory analysis, in settings where evaluators compare stimuli known to be different in some way (Ennis and Bi, 1998; Bi, 2008). The goal is to uncover the distribution of the preferences for the stimuli in the population from which the evaluators are recruited. In the roll call application, the legislators correspond to evaluators, and roll calls on the same issue to repeated administration of the same set of stimuli.

Consider the example of the six Senate roll calls on the Civil Rights Act of 1964 described in Table 6.2 and Figure 6.1. The votes can be recoded into those supportive of the Act and those against it, using as the baseline the votes of S. Thurmond, who uncompromisingly opposed the Act. Although about one half of the 67 Democratic senators never voted for the same option as Thurmond in the six roll calls, the rest did at least once. Applying

(6.7) with the mixture index of fit to this data shows that only one half of the senators can be described by the same underlying position, even if we allow for stochastic factors to influence their votes (Figure 6.6, left). Allowing their positions to vary, as in (6.8) yields a much better description–only one in five Democratic senators are not described by the model (Figure 6.6, right). The underlying Beta distribution is sharply bi-modal, in effect splitting the senators into those in full support of the Act and those against it, with the rest out of the model. In short, neither a perfect ideal points cohesion nor a sharp polarization describe well the inspected voting records of the Democratic senators.

The method presented in this section rests on two assumptions. First, that the direction of voting can be ex-ante associated with the direction of the underlying policy dimension. Second, that the roll calls can be assigned the same weight. Both assumptions are relaxed in the following sections, and the two models (6.7) and (6.8) as special cases of the IRT models discussed in Section 6.5. Consequently, the method can be applied even with the more flexible models.

6.5 Ideal Point Estimation with IRT Models

IRT models offer a flexible approach to ideal point estimation. In most settings, the ideal point estimates produced by these models are close to the widely used ones of the NOMINATE family (Clinton et al., 2004). However, the IRT models yield themselves better to including covariates of the ideal points or modeling hierarchical structures such as group memberships, as they can be considered as hierarchical generalized linear models (De Boeck, 2008; De Boeck et al., 2011).

IRT models are less restrictive relatives of the Binomial and Beta-binomial preference models discussed in Section 6.4. Typically, the functional form in (6.7) is used, but the option probabilities are allowed to vary across individuals, and only two options are considered.

The general form is

$$v_{i,c} = Binomial(1, p_{i,c}),$$

where i indexes the legislators, c the roll calls, v an indicator variable where one means an affirmative and zero a negative vote, and p is the probability of voting affirmatively. The same can be applied to (6.7), with

$$v_{i,c} = Beta\text{-}binomial(\alpha_{i,c} = p_{i,c}\gamma,\ \beta_{i,c} = (1 - p_{i,c})\gamma),$$

where γ is a dispersion parameter, which increases the flexibility compared to the Binomial model. The probability of voting affirmatively is modeled as a function of the weight of the roll call and the legislator ideal point, and in some cases of additional parameters. The simplest such model is also known as the Rasch model (Rasch, 1980),

$$\text{logit}(p_{i,c}) = \delta_c - \theta_i$$

δ_c the roll call weight ('item difficulty'), and θ_i the legislator ideal point ('ability').

When the ideal points θ are restricted to a single value, the Rasch model is identical to the log-linear model of independence in an C-way contingency table that cross-classifies the legislators by the options they took in the roll calls (Cite). For example, for three roll calls the functional form is

$$\ln v_{c1,c2,c3} = \lambda_0 + \lambda_{o1}^{C1} + \lambda_{o2}^{C2} + \lambda_{o2}^{C2},$$

where $c1, \ldots, c3$ indexes the calls, λ_0 is the 'intercept,' and the remaining three λs are option parameters for each of the three roll calls. Unlike in the models shown in Section 6.4, this does not assume all the roll calls are of equal weight on the underlying dimension, nor

does it require to recode them to have the same substantive direction. Consequently, it can be interpreted as more general measure of party cohesion in terms of ideal points.

Applied to the six Senate roll calls (Table 6.2 and Figure 6.1), almost two thirds (65%) of the senators are not accounted for. In other words, at most about one third of the senators can be described by the same ideal point. For the Democratic senators, at most slightly more than two in five (43%) can hold the same position, which is only a slightly better fit than that shown in Figure 6.6. On the other hand, at most one third (33%) of the Republican senators can have the same ideal point. Consequently, it is interesting to inspect the fit of models which assign probability distributions to the ideal points. However, this is computationally considerably more expensive than the above models, and as such out of the scope of the present chapter.

6.6 Detecting Differential Item Functioning in Ideal Point Models

In educational testing, it is usually desirable that the chance of giving correct answers would depend only on the measured ability. If differential item functioning (DIF) occurs, equally able test takers will answer an item differently conditional on some of their demographic characteristics, such as gender or ethnicity. Thus, to assure test validity, items are tested for DIF.[2] In roll call analysis, DIF is of interest if the options are associated with a substantively meaningful dimension, which takes the place of ability. In a differentially functioning roll call legislators with similar ideal points will vote differently conditional on some other attribute of theirs. In ideal points estimation DIF detection can aid model selection and

[2]Interested readers can find a concise exposition of DIF as well as the conventional methods for its detection e.g. in Magis et al. (2010).

substantive interpretation of the extracted dimensions.

A DIF detection method based on the mixture index of fit has been proposed by Rudas and Zwick (1997) (RZ). Unlike the conventional methods it does not rely on tests of the statistical significance of deviation from the null hypothesis and is not sensitive to sample size. This is an advantage in roll call analysis, where sample sizes are typically in the low to mid hundreds and much lower than in educational testing, and conventional tests might thus lack the power to detect DIF. Furthermore, the RZ method allows through its residual analysis to identify the parts of the population affected by DIF, aiding the substantive interpretation of the findings.

The RZ method rests on two nested models of a three-way contingency table that cross-classifies the respondents according to a demographic characteristic, item response, and ability. The first model does not allow DIF, and the second restricts it to be uniform across ability levels. Under no DIF, the answers can be associated with ability, but not with the demographic characteristic of interest. This can be represented with a log-linear model, to allow for any number of demographic groups and response options. The log-linear functional form is

$$\ln y_{g,o,a} = \lambda^0 + \lambda^O_o + \lambda^G_g + \lambda^A_a + \lambda^{OA}_{o,a} + \lambda^{GA}_{g,a},$$

where g indexes the demographic groups, o options (answers), and a ability levels. Both the option and the group can be associated with the ability, due to the λ^{OA} and λ^{GA} interactions. If on the other hand DIF is uniform, the answer can be associated with the demographic characteristic, but identically across the ability levels. In log-linear terms,

$$\ln y_{g,o,a} = \lambda^0 + \lambda^O_o + \lambda^G_g + \lambda^A_a + \lambda^{OA}_{o,a} + \lambda^{GA}_{g,a} + \lambda^{GO}_{g,o},$$

where λ^{GO} is the group-answer interaction. The mixture index is applied by plugging each model separately into (6.2). The method can be applied in roll call analysis, substituting

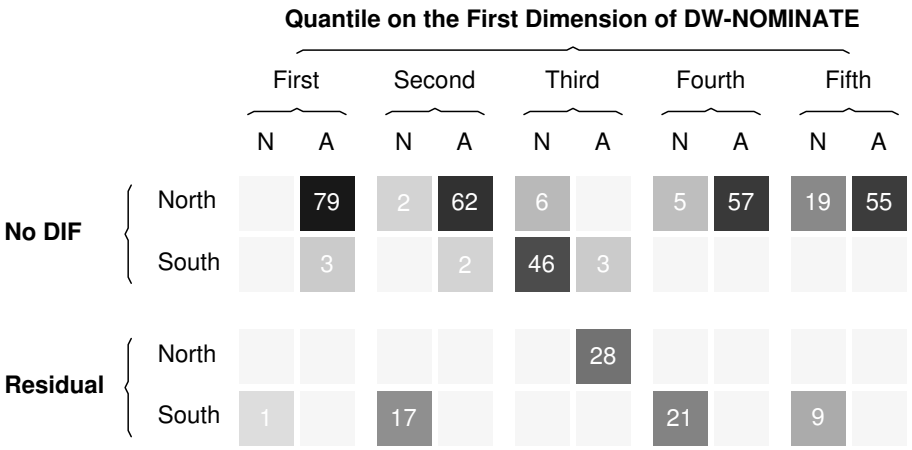


Figure 6.7: Decomposition of the votes on the passage of the Senate version of the Civil Rights Act of 1964 (House vote no. 182) into DIF-free and residual votes. N–nay; A–Aye. South–eleven former Confederate states; North–the remaining states. Data source: Poole and Rosenthal (2000).

roll calls for test items, a substantively meaningful dimension for ability, and options taken in the roll call for item responses.

The application can be demonstrated with the most widely used estimates of ideal points in U.S. Congress, DW-NOMINATE.[3] DWN places the congressmen on two dimensions, the first of which is interpreted as economic liberalism vs. conservatism. The interpretation of the second dimension varies, but from the late 1930s to mid-1970s it is interpreted as related to the civil rights of African Americans (Poole and Rosenthal, 2000). The RZ method can be used to validate the interpretations. The House vote on the passage of the Senate version of the Act (see Figure 6.2) is a particularly well-suited case, as the Act is a landmark piece of legislation on the civil rights of African Americans, and this was the final roll call on it. The DWN points of the representatives can be seen in Figure D.3 in the Appendix.

[3]The most popular methods for scaling roll calls are of the NOMINATE family, to which DW-NOMINATE belongs (Poole and Rosenthal, 1985, 1991, 2000; Poole, 2005; Carroll and Poole, 2014; Laver, 2014).

If the interpretation of the second dimension is correct, then legislators with close ideal points on it should vote the same way regardless of any grouping, including by region. The log-linear models defined above require a discretization of the ideal points. In this analysis, the representatives were sorted into five groups of 83, and to simplify the exposition, the tests of sensitivity to the grouping method are not reported here. This data is well described as free of DIF–only one in twenty (5%) of the legislators is not accounted for. In short, Southerners voted the same as Northerners did, if we take into account the second dimension ideal points, and its conventional interpretation appears fitting.

This contrasts with the first dimension. In that case, again sorting the legislators into five equally sized groups, the model of no DIF does not fit well, leaving almost one in five (18%) of the legislators out. Figure 6.7 shows the decomposition into the no DIF and residual components. In four of the five groups, it is the Southerners opposing the passage, and in the remaining the Northerners supporting it, who voted differently. This is further evidenced by the fact that uniform DIF fits practically perfectly (1% residuals). Across the first dimension, the region and the option are associated the same way, and unlike the second dimension, it does not describe the votes of both Northerners and Southerners well. To sum up, the RZ method validates the view that the vote was driven by factors associated with South and captured by the second dimension, which relate to the civil rights of African Americans.

6.7 Conclusion

This chapter presents a framework based on the mixture index of fit that extends some of the existing methods in roll call analysis. These include, but are not limited to, the measurement of group cohesion, and evaluation and substantive interpretation of ideal point models. To streamline the exposition, three issues that might be of interest were not

discussed.

First, in some contexts the researchers might be interested in obtaining interval estimates to reflect the attached uncertainty. These include the ideal point estimates, but also the values of the mixture index of fit. In any of the applications presented above, such estimates can be obtained with jackknife resampling for the index, as well as any quantities computed simultaneously with it (see Dayton, 2003, 2008).

Second, the standard practice in the literature to ignore the absences and abstaining from voting while present is followed in this chapter. This is equivalent to treating them as data missing at random, which is not always justified, and can potentially adversely affect inferences (Powell, 2015). This is by no means necessary–in addition to the many conventional techniques for missing data imputation, the generalization of the mixture index of fit to missing data (Rudas, 2005; Rudas and Verdes, 2015) can be applied.

Finally, the chapter applies the index to IRT models, but there are other alternatives in ideal point estimation. The most popular methods are of the NOMINATE family. Yet other alternatives are multiple correspondence analysis (Greenacre and Blasius, 2006) (known aslo as homogeneity analysis (De Leeuw, 2006; de Leeuw, 2007)) and optimal classification (Poole, 2000, 2005). In principle, the mixture index of fit can be applied to all of them. There are however two pragmatic reasons that hinder the application. First, for C roll calls each with two options, the cross-classification by option in each roll call has 2^C cells. Since most legislatures have hundreds or tens of seats, it is very likely that for more than a few calls many of the cells will be empty. Consequently, the tables need to flattened or smoothed, which is often done by adding a small constant to all cells (Rudas and Zwick, 1997; Rudas, 2002). The flattening procedures can affect estimation enough to contaminate substantive inferences, and thus a extensive sensitivity analysis is recommended. Second, the current software for the scaling methods does not handle data which in the contingency table representation contains fractional values. These issues can be resolved by increased

computational power and new software.

Chapter 7

The Mixture Index of Fit in Text Analysis

Text is analysed statistically across a variety of fields to a variety of aims. However, most model evaluation metrics in use focus on predictive performance, and there are no metrics for description and exploration currently in wide use. This chapter introduces the Rudas–Clogg–Lindsay mixture index of fit to statistical text analysis. The index is applied to classification as well as scaling problems. The models include the unigram model, the mixture of unigrams model, latent class analysis known in this context as probabilistic latent semantic analysis/indexing, correspondence analysis, and log-linear models for multivariate categorical associations. The demonstrations use a dataset built from ten platforms of five U.S. parties from 1996 and 2000.

7.1 Introduction

Statistical models of text are used in linguistics, computer science, psychology, and the social sciences to a variety of aims (see e.g. Roberts, 2000; Grimmer and Stewart, 2013). Despite their wide use, most metrics for their evaluation originate in computer science. The aims to which the same models are put in other domains, including the social sciences, are often different. In computer science, the focus is on models' predictive performance, and its metrics are not suited to evaluate exploratory or descriptive usefulness (Chang et al., 2009; Blei, 2012). In contrast, model evaluation techniques developed in the social sciences are focused on substantive validity (Lowe and Benoit, 2013; Grimmer and Stewart, 2013). In short, there are currently no widely-applicable metrics in use which would provide concise numeric summaries of descriptive and exploratory performance of statistical models of text.

This chapter introduces to statistical text analysis a flexible framework for model evaluation–the π^* mixture index (Rudas et al., 1994). The framework abandons the standard assumption that the model describes the whole population. Instead, the observations are split into those that are and those that are not described by the model, while minimizing the fraction of the latter, which is the model fit metric. In addition to the easy-to-interpret metric, the framework provides also a new kind of residual analysis that can inform substantive inferences. The framework can be conveniently applied to any statistical model of text, provided the observed and expected values can be represented as contingency tables of fractions without any loss of information. The secondary contribution is a new text analytic approach based on log-linear models.

The chapter proceeds as follows. Statistical analysis of text is introduced in Section 7.2 focusing on political science applications. Subsection 7.2.1 discusses a widely-used transformation of text into quantitative information for statistical analysis, and introduces a dataset built from the platforms of five U.S. political parties from the 1996 and 2000

election cycles, which is used throughout the chapter for illustrations. Subsection 7.2.2 discusses the existing model evaluation methods. The mixture index of fit is introduced in Section 7.3, and applied to five different methods in the subsequent sections. These are the unigram model in Section 7.4, the mixture of unigrams model in Section 7.5, latent class analysis–known in this context also as probabilistic latent semantic analysis or indexing–in Section 7.6, text scaling with correspondence analysis in Section 7.7, and multivariate association models in Section 7.8. The chapter concludes by discussing the limitations and potential extensions of the use of the mixture index in statistical text analysis.

7.2 Text as Data

7.2.1 From Text to Numbers

Text data is standardly organized into documents nested in collections (corpora). For statistical analysis, text has to be transformed into quantitative information. Commonly, this is achieved by treating the documents as sequences of units each of which is an instance of a category from a set thereof. The category set can be defined in a variety of ways. In the social sciences, it commonly consists of words, their sequences, word roots (lemmas or stems) (see e.g. Jurafsky and Martin, 2000), or word or word sequence types. To simplify the exposition, in this chapter the categories are called 'terms,' and only word stems are used, calling them simply 'words' for further simplification.

In most applied settings, it is assumed that the order of the units in a document can be ignored, which is known as the 'bag of words' assumption. Consequently, a corpus can be represented as a document-term matrix $\mathbf{o}$ in which each element $o_{i,j}$ is the observed frequency of j^{th} term in i^{th} document. Then, under a statistical model $\mathcal{M}$ the expected values are a document-term matrix $\mathbf{e}$ with the same number of rows and columns as $\mathbf{o}$. Alternatively, the matrices can be used transposed into a term-document matrix in which

i indexes the terms and j the documents. This can in some contexts make the matrices easier to inspect.

The illustrations in this chapter use a dataset built from the platforms of five U.S. political parties–Democrat, Green, Libertarian, Reform, and Republican–in the 1996 and 2000 election cycles. More generally, the corpus consists of ten documents, each of which belongs to one party out of five and one cycle out of two. The text was processed by first excluding numbers and the most common function words (stop-words), transforming all the text into lower case, reducing the words to their stems using the `tm` package (Feinerer et al., 2008; Feinerer and Hornik, 2014), and excluding stems that did not occur at least once in each of the source documents. The resulting dataset consists of 10,416 instances of 43 word stems. The sources of the texts are shown in Table E.1 in Appendix E, and the data is shown as a three-way contingency table in Figure 7.1.

7.2.2 Model Evaluation in Statistical Analysis of Text

A variety of statistical techniques is used to draw inferences with text data. A detailed overview focused on political science is provided by Grimmer and Stewart (2013). To summarize it, in political science the goal of statistical text analysis is usually to either place the documents in a substantively meaningful low-dimensional space (i.e., to scale it), or sort pieces of text into categories. The methods can be further classified according to the amount of input required from the analyst. Most importantly, some scaling methods require the positions of some documents, and some classification methods the character of the categories to be known ex ante. In this chapter, examples of classification methods are shown in Sections 7.5 and 7.6, and of scaling in Section 7.7. The method introduced in Section 7.8 is somewhat less straightforward to classify within the Grimmer and Stewart (2013) schema, as it rests on modeling associations between term frequencies and document attributes such as author or time of creation.

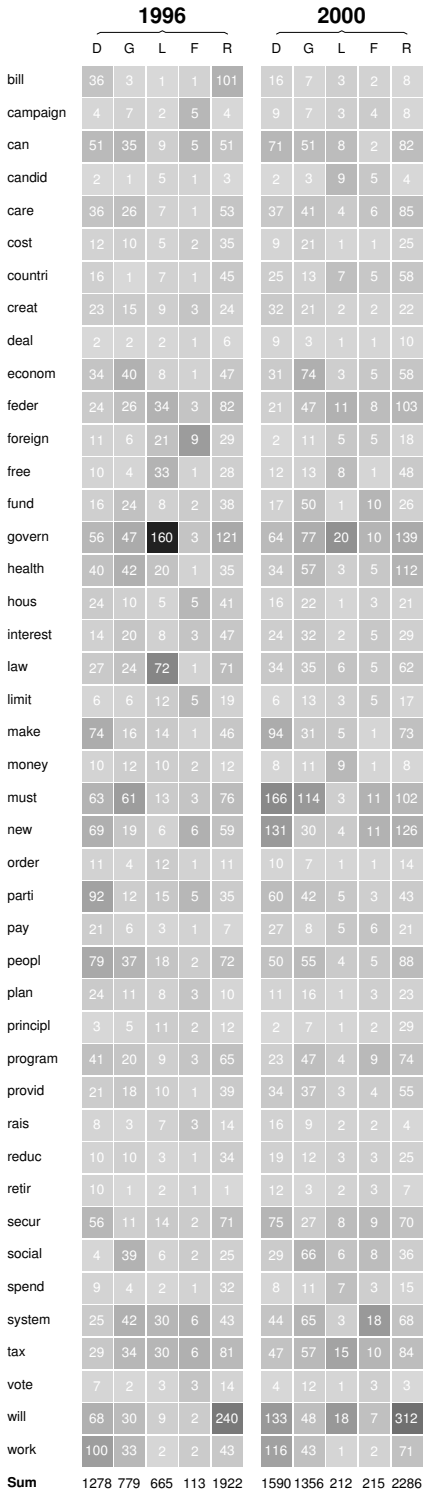


Figure 7.1: The term-document matrix used in the illustrations. N=10,416. Color by column fraction. The parties: D–Democratic, G–Green, L–Libertarian, F–Reform, R–Republican. Source: author's calculation.

Most model evaluation metrics in statistical text analysis come from computer science, reflecting the fact that some of the most popular models originate in its domain. In typical computer science applications, the priority is on prediction, and the most widely defined metric is the likelihood of a set of documents given a model estimated with a different sample of documents, which can be calculated in a variety of ways (Wallach et al., 2009; Blei, 2012). In the context of topic models, another metric of predictive performance is 'perplexity,' a function of the ratio of the held-out likelihood over the number of units (Chang et al., 2009).

In social science applications of statistical text analysis, the priority is on substantive interpretability of the generated quantities (Grimmer and Stewart, 2013). For instance, in the use of models designed to extract ideological or policy positions from text, such as Wordscores (Laver et al., 2003; Lowe, 2008) or Wordfish (Slapin and Proksch, 2008), the substantive interpretability of the uncovered spaces is the priority. Its evaluation typically relies extensively on human judgement, and can involve comparing the quantities obtained by different procedures (Grimmer and Stewart, 2013; Lowe and Benoit, 2013; Hjorth et al., 2015). In the context of topic models, procedures that aim to quantify substantive interpretability using human input have been proposed by Chang et al. (2009). Their downside is relative expensiveness.

Finally, many of the widely-used model fit metrics, such as the likelihood ratio statistic or the Akaike or the Bayesian Information Criterion can be used in some contexts. As these metrics capture different aspects of goodness of fit, and can rest on different assumptions, their choice should be driven by substantive and operational concerns. The following section introduces a widely-applicable model fit metric, the mixture index of fit (Rudas et al., 1994), to text analysis.

7.3 The Mixture Index of Fit

The π^* mixture index of fit is a general measure of model fit introduced by Rudas et al. (1994). Unlike the conventional model fit metrics, it abandons the assumption that the model describes the whole population. Instead, it assumes that the population is composed of two latent classes, one of which is described by the model and the other is not. This assumption is always true. The size of the out-of-model class is the model fit metric. Formally,

$$\pi^*(\mathrm{O}, \mathcal{M}) = \mathsf{inf}\{\pi \colon \mathrm{O} = (1-\pi)\mathrm{M} + \pi\mathrm{R},\ \pi \in [0,1],\ \mathrm{M} \in \mathcal{M},\ \mathrm{R}\ \mathsf{unspecified}\}, \tag{7.1}$$

where O is the observed distribution, M an element from the model ($\mathcal{M}$), and R an unspecified residual distribution.

In the context of document-term (or term-document) matrices the decomposition of the population under the mixture index can be formulated as

$$\mathbf{o} = (1-\pi^*)\,\mathbf{e} + \pi^*\,\mathbf{r},$$

where $\mathbf{o}$ is the observed matrix, $\mathbf{e}$ the expected matrix under the restrictive model, $\mathbf{r}$ the unrestricted residual one the mixture weights of which π^* is minimized. In the context of all contingency table models, and thus all the statistical models of text discussed in this chapter, the mixture index can be applied with the combination of the Expectation-Maximization (EM) and binary search algorithms proposed by Rudas et al. (1994). An implementation in the R language is available in the `pistar` package.

Since the mixture index rests on assumptions known to be true, it always provides validly defined residuals. Furthermore, the residuals are always fractions of the observations. Consequently, in addition to the easy-to-interpret model fit metric, the mixture index allows

a new kind of residual analysis that rests on inspecting the scaled distribution under the out-of-model class, and can inform substantive inferences (Rudas et al., 1994; Clogg et al., 1995, 1997).

7.4 Single-Topic Corpus

The most restrictive widely used statistical model of document-term matrices is document-term independence, also known as the unigram model. Substantively, it can be interpreted as the whole corpus being dedicated to the same topic. Under the model, each document is restricted to be an independent draw from a multinomial distribution (see e.g. Blei et al., 2003). Formally,

$$y_{i,j=1,\ldots,J} \sim Multinomial\left(\theta_{1,\ldots,J}, N_i\right),$$

where $\theta_{1,\ldots,J}$ are the terms probabilities and N_i is the length of the document. Consequently, the terms and the documents are restricted to be independent, which corresponds to the log-linear model

$$\ln y_{i,j} = \lambda^0 + \lambda_i^D + \lambda_j^T, \tag{7.2}$$

where λ^0 is the main terms, and λ^D and λ^T the document and term terms, respectively.

The mixture index of fit can be applied by plugging (7.2) into (7.1). In this way, the corpus is split into two parts, one of which belongs to a single topic and the other does not, while minimizing the size of the residual non-topical part. Consequently, each document is decomposed as well, and the value of the index can be calculated for it. For the U.S. platform data introduced in Subsection 7.2.1 the value of the mixture index for the unigram model is 0.49. In words, at most a half of the text belongs to a single topic.

Table 7.1 shows the document values of the index. In general, the values of the index

Table 7.1: Document values of the mixture index for the unigram model fit to the U.S. party platform corpus. Values in rounded percents.

	1996	2000
Democratic	46	48
Green	63	63
Libertarian	91	86
Reform	82	88
Republican	39	27

Table 7.2: Ten most common words (stems) in the topic of the unigram model fit under the mixture index of fit to the U.S. party platform corpus. Values in percents.

will	govern	must	peopl	new	care	can	tax	make	secur
9.6	7.8	6.1	5.3	5	4.4	4.3	4.2	3.9	3.8

tend to be larger for the shorter platforms (Reform and Libertarian). The ten most frequent words from the single topic are shown in Table 7.2. Only two of the words ('tax' and 'secur') are related to specific policy fields, and the rest are general words used in the context of outlining plans for a future administration. In contrast, among the ten words with the largest residual fractions, shown in Table 7.3, the majority belongs to words related to specific policy fields, such as foreign policy ('countri' and 'foreign') or campaign finance ('rais,' 'candid,' and 'money'). In short, while all the platforms outline policies for future administrations, the fields of these policies differ across the platforms.

7.5 Multi-Topic Corpus of Single-Topic Documents

The unigram model discussed in Section 7.4 can be substantively interpreted as the corpus containing only a single topic. This assumption can be relaxed by allowing each document

Table 7.3: Words (stems) with largest residual fractions under the unigram model fit under the mixture index to the U.S. party platform corpus. Values in rounded percents.

countri	foreign	retir	bill	social	principl	vote	rais	candid	money
90	89	89	86	86	83	82	81	75	70

to be drawn form one topic out of a set thereof. Formally,

$$y_{i,j=1,\ldots,J} \sim Multinomial\left(\sum_k \mathbf{1}(u_i = k)\,\theta_{k,j=1,\ldots,J}, N_i\right), \tag{7.3}$$

where i indexes the documents and k the topics, and u_i is the topic of the i^{th} document.

Again, the mixture index can be applied by plugging the model into (7.1). Since the model 7.3 is estimated by an EM algorithm as is the mixture index of fit, this can be computationally expensive in many applied settings. A comparatively computationally inexpensive approximation is available via k-means clustering, which minimizes the objective function

$$f(\mathbf{u}, \mathbf{p}, \boldsymbol{\mu}) = \sum_i \sum_k \mathbf{1}(u_i = k)\left(\sum_j (p_{i,j} - \mu_{k,j})^2\right), \tag{7.4}$$

where $p_{i,j}$ is the document fraction of j^{th} word in i^{th} document, and $\mu_{k,j}$ its fraction in the k^{th} category. This works as an approximation of the model (7.3) under the mixture index because if the model fits perfectly, then substituting $\boldsymbol{\mu}$ with $\boldsymbol{\theta}$ in (7.4) will also produce a perfect fit.

Substantively, the application of the index to the mixture of unigrams model can be interpreted as a decomposition of the corpus into a component that belongs to a set of topics and a residual component. In this way, terms and documents which are not described well by the model can be identified and inspected. Furthermore, the index can be used in

Table 7.4: Values of the mixture index for five mixture of unigram models fit to the U.S. party platform data. Fractions in rounded percents.

Unigrams (topics)	1	2	3	4	5
π^* index	49	46	35	21	19

Table 7.5: Document topics under five mixture of unigram models fit to the U.S. party platform data under the mixture index.

		No. of Topics				
		1	2	3	4	5
1996	Democratic	1	1	2	4	2
	Green	1	1	3	2	1
	Libertarian	1	2	1	3	4
	Reform	1	1	3	2	5
	Republican	1	1	2	1	3
2000	Democratic	1	1	2	4	2
	Green	1	1	3	2	1
	Libertarian	1	1	2	1	3
	Reform	1	1	3	2	1
	Republican	1	1	2	1	3

model selection to set the number of the topics.

The values of the index for a series of mixture of unigrams models fit to the U.S. party platform corpus are shown in Table 7.4. The first model–a single topic–is the unigram model. The addition of the second and third topic improve the fit somewhat, taking the residual fraction of the observations from about one half to about one third. The four and five topic models fit very similarly, leaving out about one in five of the observations out. Whether this fit is satisfactory or not depends on the goals of the analysis.

Table 7.5 shows the topics of the documents under the mixture index. Save for the single-topic model, the 1996 Libertarian platform always stands out with its own topic.

The other two minor parties–Green and Reform–on the other hand always belong to the same topic regardless of the cycle, with the exception of the five-topic model. The two major parties–Democratic and Republican–belong to the same topic regardless of the cycle under the two- and three-topic models, and each party separately belongs to the same topic regardless of the cycle under all the models.

7.6 Multi-Topic Documents

Section 7.5 relaxes the assumption of a single-topic corpus by allowing each document to be drawn from a single topic out of a set thereof. This can be further relaxed by allowing each document to be composed of multiple topics. One statistical procedure that allows this is latent class analysis (LCA). Under LCA, terms and documents are independent conditional on the observations' membership in latent classes. Formally,

$$p(\text{document}, \text{word}) = p(\text{document}) \sum_{\text{class}} p(\text{word}|\text{class}) p(\text{class}|\text{document}).$$

Under the Poisson sampling distribution this gives the model

$$o_{i,j} \sim Poisson\left(\sum_{k} \zeta_k \, m_{i,j,k}\right), \tag{7.5}$$

that is, the observed document-term table $\mathbf{o}$ is composed of K sub-tables $\{\mathbf{m}_1, \ldots, \mathbf{m}_K\}$ with mixing weights $\{\zeta_1, \ldots, \zeta_K\}$, and each of the sub-tables is perfectly described by the independence model (7.2).

The latent class model (see Lazarsfeld et al., 1968; Goodman, 1974; McCutcheon, 1987) corresponds to a multi-topic model (Blei et al., 2003) also known in some contexts as Probabilistic Latent Semantic Analysis (pLSA) (Hofmann, 1999a) or Probabilistic Latent Semantic Indexing (pLSI) (Hofmann, 1999b). The application of the mixture index in

Table 7.6: Class proportions for the two-class model fit under the mixture index to the 1996 U.S. party platforms. Proportions in rounded percents.

	Class		
	1st	2nd	Residual
Democratic	6	58	36
Green	21	22	57
Libertarian	91	0	9
Reform	4	19	77
Republican	18	58	24

Table 7.7: Ten most frequent terms in the two-class model fit under the mixture index to the 1996 U.S. party platform data. Terms ordered by their in-topic frequencies.

	Term									
Class	1	2	3	4	5	6	7	8	9	10
First	govern	law	feder	tax	system	health	peopl	parti	foreign	make
Second	will	must	secur	peopl	program	new	care	can	econom	work

latent class analysis has been investigated by Formann (2003b). Again, the index can be applied by plugging (7.5) into (7.1).

For simplicity, consider the five 1996 platforms out of the U.S. corpus introduced in Subsection 7.2.1. The unigram model is equivalent to the one-class model, and does not fit the data very well, leaving about nine in twenty of the observations out. The model with two restricted latent classes fits the data somewhat better, leaving out about one third of the observations. Table 7.6 shows the decomposition of the platforms under this model. Most of the Democratic and Republican platforms belongs to the second class, and most of the Libertarian platform to the first. The Green and the Reform platforms on the other hand mostly belong to the residual component. Table 7.7 shows the ten most common words under the two-topic model. The top ten stems of the second topic suggest that these

are related to an active role of the federal government. Also, at least some of the terms of the first topic are congruent with restricting the role of the federal government, which has been one of the core policies of the Libertarian Party.

7.7 Policy Positions of Documents

In some text analytic problems from the domain of political science the goal is to place the documents in a low-dimensional latent space, which is typically interpreted substantively as an ideological or policy space. The two most popular text scaling techniques in political science are Wordscores (Laver et al., 2003; Lowe, 2008) and Wordfish (Slapin and Proksch, 2008), both of which can extract only a single dimension in one pass. The former requires the analyst to supply the positions of at least two of the documents, while the latter does not. As Lowe (2013) shows, there is a variety of other methods for text scaling, all of which can be understood within a single framework, among which correspondence analysis stands out with its computational inexpensiveness.

Conventional models of policy or ideological spaces use a low number of dimensions, usually from one to three (see e.g. Laver, 2014). Applications of text scaling usually proceed by extracting a low number of dimensions, and then evaluating the substantive validity of the extracted spaces. In this process, model fit is rarely considered. Consequently, the analysis runs the risk of drawing inferences on the basis of models which ex post seem to make substantive sense, but do not fit the data well. The mixture index can be applied in this context as well. If political text is scaled with correspondence analysis, the index can help to select the dimensionality of the model.

Consider the U.S. party platform corpus described in Subsection 7.2.1. Table 7.8 shows the fit of correspondence analysis to the data up to five dimensions. From the perspective of scaling, the unigram model discussed in Section 7.4 can be considered as a zero-dimensional

Table 7.8: The mixture index of fit for correspondence analysis models of the U.S. platform corpus. Fractions in rounded percents.

	Number of Dimensions					
	0	1	2	3	4	5
π^*	49	41	30	19	12	9

model. It describes at best a half of the data. In general, adding further dimensions up to four improves the fit by about one ninth of the observations. U.S. politics of the past three decades is usually understood in terms of a single Liberal-Conservative dimension. However, the single-dimensional model does not fit very well to the analyzed corpus of platforms, leaving at least about two in five observations unaccounted for. The two-dimensional model leaves one in three observations out, and the three-dimensional model about one in five. In short, even models with numbers of dimensions exceeding those used in theories of U.S. politics do not fit particularly well to the inspected data.

Figure 7.2 shows the extracted space under the two-dimensional model. While the Democratic, Republican, and Green manifestos do not move much in the space, the Reform and Libertarian ones do. However, the space does not appear to have an intuitive substantive interpretation. Perhaps the most straightforward intuition on the first dimension is that it captures the views on the role of the federal government, since one end is occupied by the Democratic and the other by the Libertarian platforms. However, the word positions do not provide particularly strong support for this interpretation. Another intuition based on the party positions is that the second dimension represents positions on welfare and government spending, with one end occupied by the Green manifestos and the other by Republican ones. Again, the evidence provided by the word positions is limited at best. To sum up, the one-, two-, and three-dimensional models do not seem to fit the platform data well, nor do they have intuitive substantive interpretations. A possible explanation is that

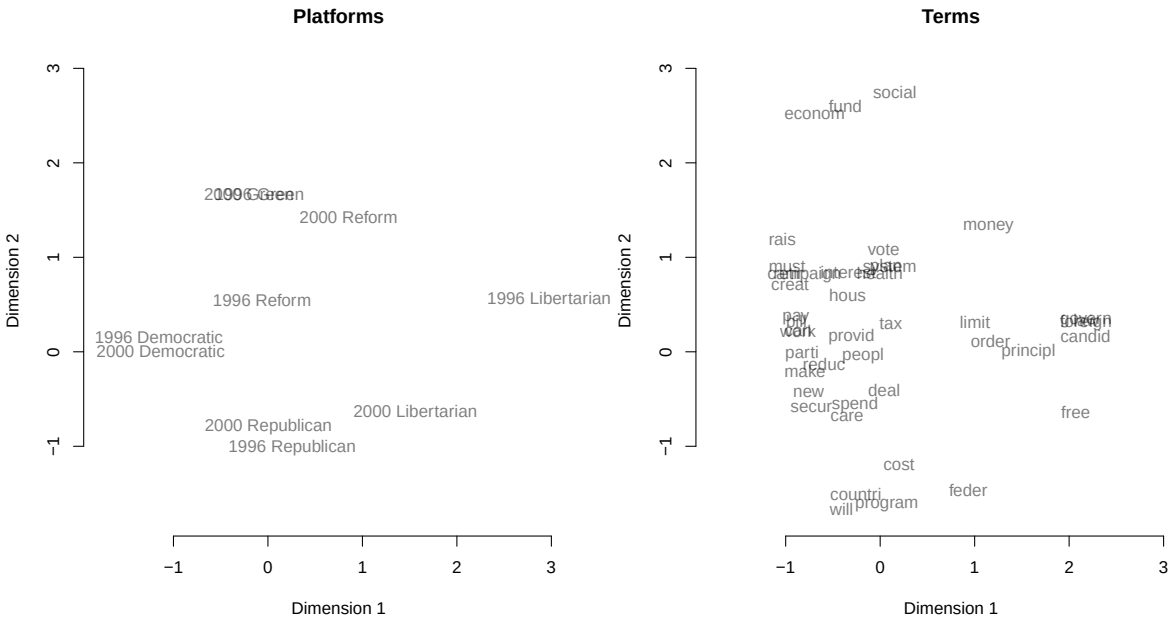


Figure 7.2: Platform and term positions extracted by two-dimensional correspondence analysis under the mixture index of fit.

since the observed document-term matrix contains only terms that occur in each of the documents, it is not informative with regards to the parties' positions, the information on which is contained in terms that are not shared by the whole corpus.

7.8 Relationships between Words, Authors, and Time

The unigram model discussed in Section 7.4 corresponds to the log-linear model of term-document independence. This section extends the use of log-linear models to settings where the documents have categorical attributes of substantive interest. The method introduced here can be applied if the attributes are either discrete, or can be discretized with acceptable information loss. The method is illustrated here with a corpus of documents by different authors and from different points in time. Consequently, the word frequencies can be recorded in a three-way contingency table, with a indexing authors, t time, and j words (terms).

With data of this type, the analyst might wish to consider questions on associations between the terms and the document attributes. In such analysis, an appealing starting point is provided by the independence model

$$\ln y_{a,t,j} = \lambda^0 + \lambda^A_a + \lambda^T_t + \lambda^W_j, \tag{7.6}$$

where λ^A, λ^T, and λ^W are the author, time, and word coefficients, respectively. The independence model describes the data well if the words are not associated with the authors and the time.

The fit of the independence model provides a baseline to which the fit of less restrictive models can be compared. Continuing with the example, four such models are

$$\begin{aligned}
\ln y_{a,t,j} &= \lambda^0 + \lambda^A_a + \lambda^T_t + \lambda^W_j + \lambda^{AT}_{a,t}, && (7.7)\\
\ln y_{a,t,j} &= \lambda^0 + \lambda^A_a + \lambda^T_t + \lambda^W_j + \lambda^{AT}_{a,t} + \lambda^{AW}_{a,j}, && (7.8)\\
\ln y_{a,t,j} &= \lambda^0 + \lambda^A_a + \lambda^T_t + \lambda^W_j + \lambda^{AT}_{a,t} + \lambda^{TW}_{t,j}, && (7.9)\\
\ln y_{a,t,j} &= \lambda^0 + \lambda^A_a + \lambda^T_t + \lambda^W_j + \lambda^{AT}_{a,t} + \lambda^{AW}_{a,j} + \lambda^{TW}_{t,j}, && (7.10)
\end{aligned}$$

where λ^{AT} is the author-time association that allows the authors to produce the words at different times, and λ^{AW} and λ^{TW} are the author-word and time-word associations, respectively. The only model that will fit perfectly by definition is the saturated model,

$$\ln y_{a,t,j} = \lambda^0 + \lambda^A_a + \lambda^T_t + \lambda^W_j + \lambda^{AT}_{a,t} + \lambda^{AW}_{a,j} + \lambda^{TW}_{t,j} + \lambda^{ATW}_{a,t,j},$$

where λ^{ATW} is the author-time-word association.

The mixture index of fit can be applied in model selection. Table 7.9 shows the values of the index for the five models (7.6–7.10) fit to the U.S. party platform data. The independence model can account for at most half of the words, and the fit is improved only

Table 7.9: The mixture index of fit for five log-linear models fit to the U.S. party platform data. Values of the mixture index in rounded percents.

Model	π^*
Independence	50
Word, Party-Year	49
Party-Year, Party-Word	15
Party-Year, Year-Word	42
Party-Year, Year-Word, Party-Word	10

Table 7.10: Residual fractions for the U.S. party platforms under the party-year, year-word, party-word model. Fractions in rounded percents.

	1996	2000
Democratic	14	10
Green	18	11
Libertarian	4	61
Reform	46	9
Republican	7	5

slightly by allowing the parties to produce platforms of different lengths across the cycles with the author-time interaction. Allowing further the words to be associated with the cycles improves the fit considerably more, but still leaves more than two in five words out of the model. However, when the words are also allowed to be associated with the parties, the fit improves markedly, leaving only one in ten words out. Substantively, the frequency with which a word is used in the data is related primarily to the party that authored the platform.

Further substantive insights can be provided by an inspection of the values of the index for each platform, shown in Table 7.10. In the first cycle it is the Reform and in the second the Libertarian party that has the most words unaccounted for by the model (7.9). In both cases, between two and three out of five words in the platforms are not accounted for by

the model.

7.9 Conclusion

The mixture index of fit provides an easy-to-interpret metric for model evaluation, which is especially appealing when the analysis has descriptive goals. Furthermore, it enables a new kind of residual analysis that can inform substantive interpretations. The framework can be conveniently applied to all statistical models of text for which the observed and the expected term frequency values can be represented as contingency tables. In this chapter, the use of the framework was shown on variety of classification and scaling models, as well as multivariate association models.

The only drawback of the framework is the computational expensiveness of its application in many contexts. Perhaps the most important example of this are the highly popular Latent Dirichlet Allocation and associated topic models (Blei et al., 2003; Blei, 2012). Even without applying the mixture index, their estimation is often computationally more expensive than that of any of the models to which the index was applied in this chapter. However, where computational power is available, nothing prevents the analyst to apply the mixture index to them. Just as in the case of the models discussed in this chapter, the mixture index can be applied to them using the combination of the EM algorithm and binary search introduced by Rudas et al. (1994), as implemented e.g. in the `pistar` package for the R language.

Conclusion

This thesis investigated how the Rudas–Clogg–Lindsay mixture index of fit can help to deal with two issues common in political science, unobserved heterogeneity and non-stochastic samples, as well as how log-linear models can help to handle another common issue, categorical data. Furthermore, the thesis is accompanied by an R package for the mixture index of fit named `pistar` and made available as open source software. Each method was applied in five out of the six investigated substantively motivated problems. The mixture index and the `pistar` package are discussed in Chapter 1.

In Chapter 2, the mixture index and log-linear models were applied to the detection of electoral fraud from digit distributions, also known as 'election forensics.' In conventional election forensics digit distributions are investigated by testing the statistical significance of the deviation from the null hypothesis that the data are drawn from a distribution believed to characterize fraud-free results. The goal of this procedure is to arrive at a qualitative judgement whether the observed returns are fraudulent or not. Since electoral returns are better understood as population data, and setting the significance level for the test is at best very difficult, the conventional testing framework is not appealing. The mixture index and the dissimilarity index offer an appealing alternative–if the distributional assumptions of election forensics hold, the indexes can provide an answer to the quantitative question of how much fraud did occur. Furthermore, log-linear models can be used to relax the

distributional assumptions, and compare digit distributions from multiple sets of electoral returns.

In Chapter 3, a widely used procedure for the allocation of seats over options according to the distribution of votes over the options, known as the D'Hondt method, is given an interpretation from the perspective of the mixture index. From this perspective, the D'Hondt method allocates the seats over the options by representing the largest possible share of the votes proportionally, and the remainder not at all. In addition to providing an intuitive interpretation of the D'Hondt method, the mixture representation also allows to formulate a new index of disproportionality and generalize it to settings where the votes are observed only partially. The findings apply to any seat allocation method that minimizes the quantity known as the D'Hondt index.

In Chapter 4 log-linear models and the dissimilarity index are applied to the problem known in party research as 'party nationalization.' Party nationalization is understood as uniformity of party support across the territories in a country. From this perspective, a nationalized party enjoys the same level of support across all the country's territories. Consequently, if a party system is nationalized, territory is not associated with party choice. Several measures of nationalization have been proposed in the literature both for parties and their systems. Chapter 4 proposes a new index, which has a more intuitive interpretation, and is related to the well known Pedersen volatility index as well as to the D index of residential segregation. The index is a special case of the dissimilarity index. With log-linear models the index can be used to inspect not only the territorial, but also the spatial variability of the vote using the dissimilarity index. Finally, the dissimilarity index can be applied to mixture models to inspect the diversity of local patterns of competition.

Chapter 5 developed from the mixture index minimum mixture measures of voter transitions in aggregate data, as well as applied the index to the problem of 'swing.' The minimum mixture measures of voter transitions rest on splitting the votes into those cast by

voters who are constant in the sense that in the inspected elections they always picked the same option (e.g. party). If the researcher wishes to take into account some other attributes of the elections, these can be achieved with log-linear models. For instance, where batches of elections for multiple offices take place simultaneously, batch-only and office-only constancy can be represented with log-linear models. This approach was extended to restrictive models of swing, specifically the 'uniform' and 'proportional' swing, the first of which can be represented by a linear and the second by a log-linear model.

Chapter 6 applied the mixture index to a variety of problems in the analysis of roll call data in the study of legislative politics. The index was used to formulate a new measure of partisan voting in legislatures. The measure rests on decomposing the observed votes into those cast independently of the party and others, while maximizing the fraction of the former. It can be interpreted as the largest possible fraction of non-partisan votes. The measure extends to any other legislator groupings of interest, such as states in federal systems, and any other bodies where roll call voting takes place, such as executive cabinets or supreme courts. If more than one grouping is of interest, as e.g. in federal systems where each member of the legislature can be classified by party and state, the index can be applied with log-linear models to evaluate which member characteristics best describe the observed vote. The index is also applied to item response models which are used in the study of legislative politics in vote scaling to extract the ideological or policy positions of the legislators. The index can be used to evaluate the fit of the item response models as well as to detect the presence of differential item functioning in ideal point models.

In Chapter 7, the mixture index was applied to model criticism in statistical text analysis. Although statistical models of text are used across a variety of domains, the existing widely-used measures of fit originate in computer science, and focus on predictive performance of the model. However, in many applications, including social science ones, description and exploration have priority over prediction. The mixture index provides

an easy-to-interpret alternative that captures the descriptive performance of the model. Furthermore, unlike the conventional methods it does not assume that the model describes the whole corpus, nor does it require to assume that the observed documents are a stochastic sample from a larger population. The main drawback to the application of the index in text analysis, which in the present is significant enough to prevent its wide adoption, are the increased computational demands it creates.

Across the chapters, several avenues for further research were identified. The most common ones are related to the computational demands imposed by the procedures for the computation of the index. This issue was pronounced the most in Chapters 6 and 7, due to the fact that some of the models used in roll call and text analysis are computationally demanding already in their existing forms, and the application of the index to them further increases these demands. Asides from the development of new, computationally less demanding procedures several other developments can help to overcome this challenge. The first is a computationally more efficient implementation of the existing procedures. The version of the `pistar` package used in this thesis is implemented in pure R, which has the advantage of greater transparency and easier maintenance of the code. Computational expensiveness can be decreased by re-implementing selected parts of the code in C++. The computational demands can be decreased also by re-implementing some of the existing procedures for various widely-used model families in a computationally more efficient manner in the `pistar` package.

Another avenue for further research is the development of practically applicable procedures for the computation of the index in multivariate continuous settings. Finally, in some contexts the investigators might want to take into account extra-data information on the fraction of the population described by the model. In this case, a Bayesian take on the index, under which the extra-data information would be expressed as informative priors, could serve to this aim.

Appendix A

Appendix to 'Latent Class and Log-Linear Election Forensics'

A.1 The Benford Distribution

Table A.1: Distributions of numerals under Benford's law for the first nine positions. Source: author's calculation.

	Position								
Numeral	1^{st}	2^{nd}	3^{rd}	4^{th}	5^{th}	6^{th}	7^{th}	8^{th}	9^{th}
0		.119679269	.101784365	.100176147	.100017592	.100001759	.100000176	.100000018	.100000002
1	.301029996	.113890103	.101375977	.100136888	.100013681	.100001368	.100000137	.100000014	.100000001
2	.176091259	.108821499	.100972198	.100097673	.100009771	.100000977	.100000098	.10000001	.100000001
3	.124938737	.10432956	.100572932	.1000585	.100005862	.100000586	.100000059	.100000006	.10000000
4	.096910013	.100308202	.100178088	.100019371	.100001953	.100000195	.10000002	.100000002	.1
5	.079181246	.096677236	.099787576	.099980285	.099998044	.099999805	.09999998	.099999998	.1
6	.06694679	.093374736	.09940131	.099941242	.099994136	.099999414	.099999941	.099999994	.099999999
7	.057991947	.090351989	.099019207	.099902241	.099990228	.099999023	.099999902	.09999999	.099999999
8	.051152522	.087570054	.098641184	.099863284	.099986321	.099998632	.099999863	.099999986	.099999999
9	.045757491	.084997352	.098267164	.099824369	.099982414	.099998241	.099999824	.099999982	.099999998

A.2 Simulations

Figures A.1 and A.2 report the results for π^* and Δ, respectively. For the same pairs of means and standard deviations, the values of each index are strongly correlated across the simulated processes, with Pearson's $\rho > 0.88$ (N=400) for all pairs excluding Mebane's (2006a) two mechanisms and $\rho > 0.45$ (N=169) for all possible pairs (shown in Figure A.3).

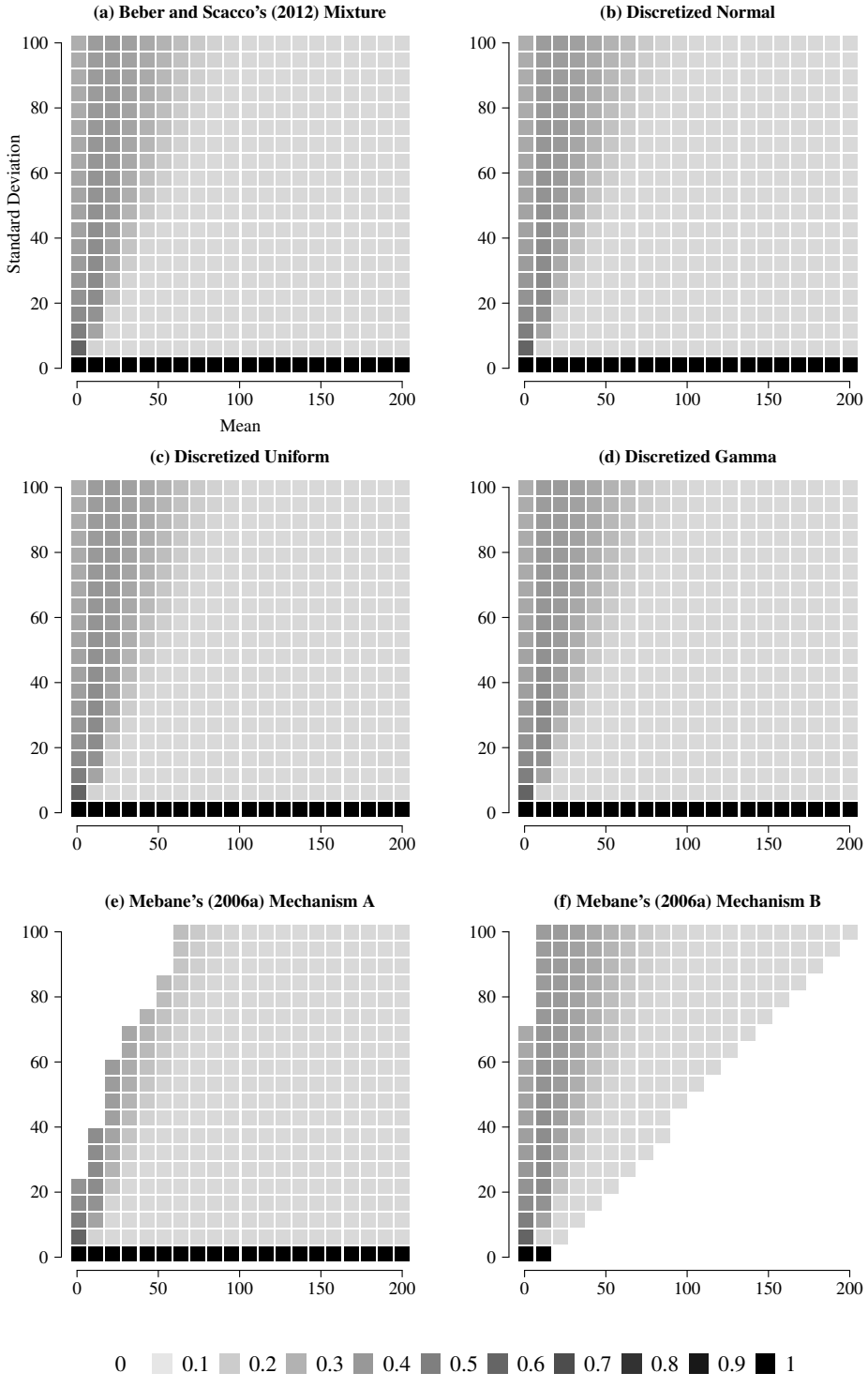


Figure A.1: Mean values of the π^* mixture index of fit for the model of uniformity for Beber and Scacco's (2012) simulated scenarios. Two of the six mechanisms are Mebane's (2006a). For each combination of mechanism, mean, and standard deviation 1,000 samples of 1,000 numbers were simulated.

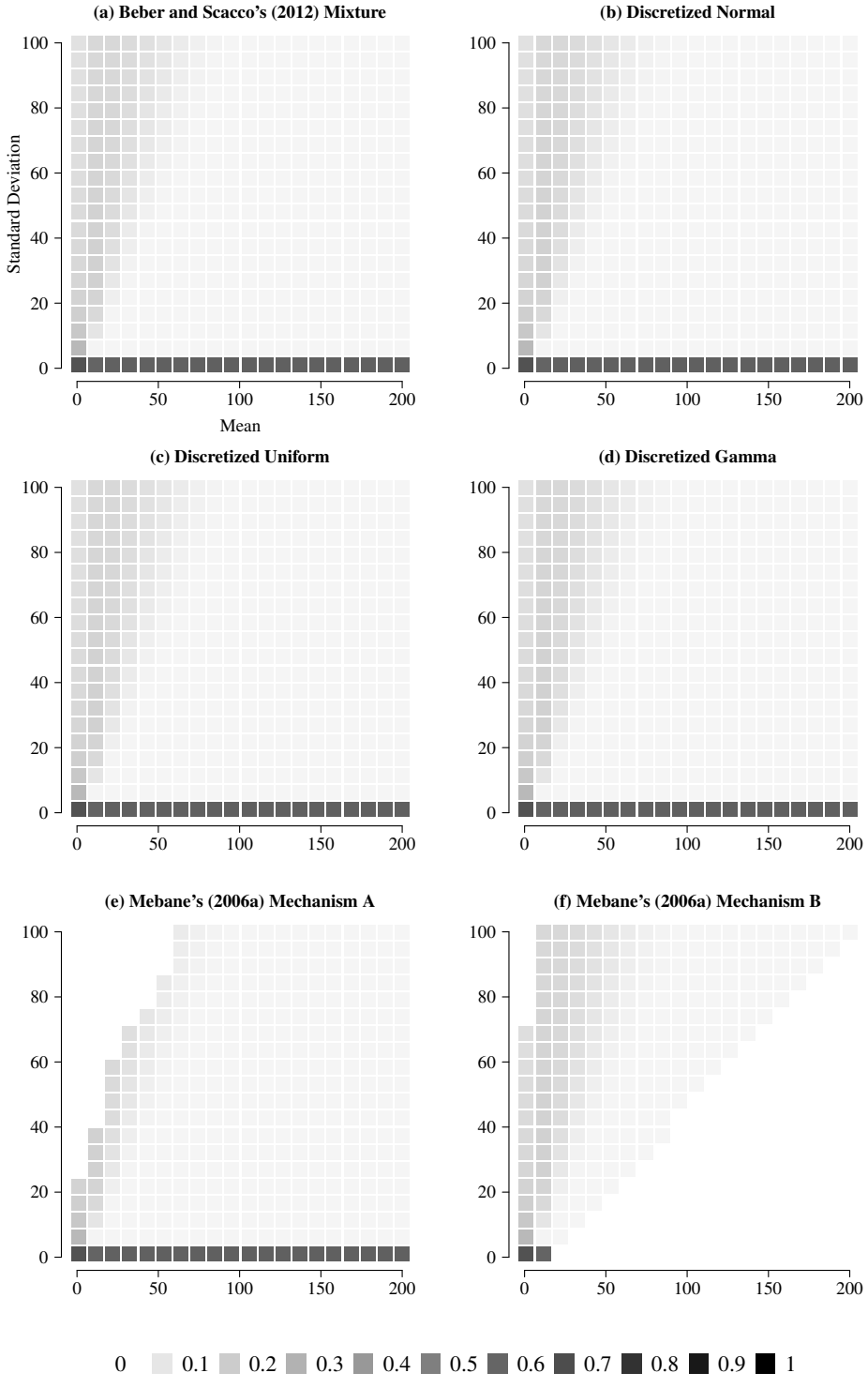


Figure A.2: Mean values of the Δ dissimilarity index for the model of uniformity for Beber and Scacco's (2012) simulated scenarios. Two of the six mechanisms are Mebane's (2006a). For each combination of mechanism, mean, and standard deviation 1,000 samples of 1,000 numbers were simulated.

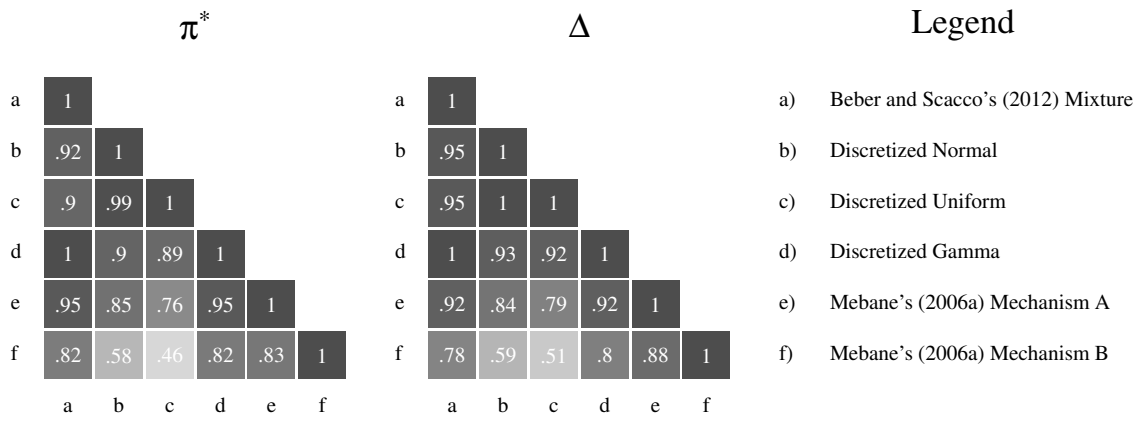


Figure A.3: Pearson correlations for the two indexes for pairs of the six mechanisms simulated by Beber and Scacco (2012). Mechanisms (e) and (f) devised by Mebane (2006a). For pairs of the first four scenarios, N=400, for the rest N=169.

A.3 Empirical Demonstration

A.3.1 Sweden

Table A.2 reports the last digits classified by numeral, party, and result from the Swedish parliamentary elections of 2002 for the two parties with the largest national vote totals.

Table A.2: Last digits of ward-level vote counts with three or more digits in the 2002 Swedish parliamentary elections for two parties with the largest national vote counts. 5,963 wards inspected, 13 wards where the two parties tied excluded. Source: author's calculation. N=8,940. Data source: Beber and Scacco (2012).

	SAP		MSP	
	Won	Lost	Won	Lost
0	538	54	39	285
1	492	57	43	292
2	533	62	33	305
3	527	61	37	292
4	526	57	41	305
5	496	52	46	268
6	528	62	33	278
7	488	55	35	294
8	514	43	34	270
9	518	46	35	266

As reported in Table A.3, both NHST and latent class methods indicate that uniformity describes all inspected subsets of numerals well. Under the strong distributional assumption this is evidence of no fraud.

Table A.3: Evaluation of uniformity for ten subsets of last digits from the Swedish data. Fraction sizes in %. Reference distributions of tests statistics obtained with one million simulations.

	N	χ^2	p	π^*	Δ
SAP Victory	5,160	5.63	0.78	5	1
SAP Loss	549	6.87	0.65	22	4
SAP	5,709	6.75	0.66	5	2
MSP Victory	376	4.85	0.85	12	5
MSP Loss	2,855	6.66	0.67	7	2
MSP	3,231	6.32	0.71	7	2
Victory	5,536	4.50	0.88	6	1
Loss	3,404	10.15	0.34	8	2
Registered	5,962	7.75	0.56	9	1
Total	5,959	3.54	0.94	4	1

Under the relaxed distributional assumption the numeral subsets can be compared using

a series of log-linear models that represent different processes by lifting a different set of restrictions. The independence model allows the number of last digits to vary across parties and ward result, and some numerals to be more common than others. The second model further allows for one party to win/lose more wards than the other. Under this model the same probability distribution describes the numerals of both parties wards or only those won/lost by the party. This corresponds to equal kind and amount of fraud for the inspected subsets. The remaining three models further allow numeral probabilities to vary across parties, ward results, and both, respectively, corresponding to different amounts of fraud across the inspected subsets.

Table A.4: Fit of the five log-linear models to the Swedish data. N=8,940. Fraction sizes in %. Jackknifed confidence intervals.

	χ^2	df	p	π^*	95% ci		Δ	95% ci	
Independence	5,439.38	28	<0.01	34	(27,	41)	36	(36,	37)
Party-Result, Numeral	15.98	27	0.95	4	(1,	6)	1	(1,	2)
Party-Result, Numeral-Party	11.00	18	0.89	2	(0,	4)	1	(0,	1)
Party-Result, Numeral-Result	9.53	18	0.95	2	(0,	6)	1	(0,	1)
Party-Result, Numeral-Party, Numeral-Result	4.51	9	0.87	1	(0,	3)	1	(0,	1)

As shown in Table A.4, the assessment is similar under NHST and the latent class approach. Unsurprisingly, since the two parties won different numbers of wards, the independence model fits badly. The second model delivers a near perfect fit with a π^* of 4% and Δ of 1%, and is not rejected by the χ^2 test of fit at the 5% level. Lifting further assumptions is left with little room to improve the fit.

A.3.2 Nigeria

Table A.5 reports the data–last digits of numbers with three or more digits in polling station returns from the Plateau state for the two electorally strongest parties–by numeral, party,

and polling station result, excluding the seven tied polling stations.

Table A.5: Last digits of polling station level vote counts with three or more digits in the 2003 Nigerian presidential elections in the Plateau state for two parties with largest vote counts. Seven polling stations where the two parties were tied are excluded. N=3,045. Source: author's calculation. Data source: Beber and Scacco (2012).

	ANPP		PDP	
	Won	Lost	Won	Lost
0	78	30	195	27
1	59	35	171	27
2	58	36	183	25
3	88	38	198	27
4	52	32	200	23
5	50	26	169	21
6	58	25	170	17
7	69	32	189	21
8	86	25	178	28
9	78	23	170	28

Fit of uniformity to several subsets of the data is reported in Table A.6. Several sets of digits depart noticeably from uniformity, including some of the parties' numbers as well as numerals from the counts of registered and total voters. Uniformity is rejected for ANPP's returns where it won by the χ^2 test at the 5% significance level. The π^* and Δ distances from uniformity are similar for votes for ANPP where it won, lost, or both, and for all votes for PDP as well as for votes where it lost.

Table A.6: Evaluation of uniformity for ten subsets of last digits from the Nigerian data. Fraction sizes in %. Jackknifed confidence intervals. Reference distributions of tests statistics obtained with one million simulations.

	N	χ^2	p	π^*	95% ci		Δ	95% ci	
ANPP won	676	26.40	<0.01	26	(6,	46)	9	(5,	13)
ANPP lost	302	8.20	0.52	24	(0,	54)	7	(2,	13)
ANPP	978	20.53	0.02	22	(6,	39)	6	(3,	9)
PDP won	1,823	7.64	0.57	7	(0,	21)	3	(1,	5)
PDP lost	244	5.18	0.82	30	(0,	62)	6	(1,	12)
PDP	2,067	8.06	0.53	10	(0,	22)	3	(0,	5)
Lost	546	8.14	0.52	23	(1,	45)	5	(1,	9)
Won	2,499	16.08	0.07	12	(1,	23)	3	(1,	5)
Registered	2,565	113.54	<0.01	17	(7,	28)	7	(5,	8)
Total	2,546	292.28	<0.01	21	(10,	31)	10	(9,	12)

Table A.7: Fit of the five log-linear models to the Nigerian data. N=3,045. Fraction sizes in %. Jackknifed confidence intervals.

	χ^2	df	p	π^*	95% ci		Δ	95% ci	
Independence	193.53	28	<0.01	16	(11,	20)	8	(7,	10)
Party-Result, Numeral	28.03	27	0.41	10	(4,	15)	4	(2,	5)
Party-Result, Numeral-Party	18.87	18	0.40	5	(2,	9)	2	(1,	4)
Party-Result, Numeral-Result	23.11	18	0.19	7	(3,	10)	3	(2,	5)
Party-Result, Numeral-Party, Numeral-Result	13.26	9	0.15	2	(0,	4)	2	(1,	3)

Under the relaxed distribution assumption the data can be inspected by the same log-linear models as in the Swedish case. Table A.7 reports their fit. The independence model does not fit well, again due to the different numbers of territories carried by the parties. The second model fits markedly better, especially in terms of χ^2 and Δ, and only the two least restrictive models fit near perfectly according to both latent class indexes. The modeled and residual frequencies under Δ are shown in Figure A.4. Substantively, the extent of contamination by fraud appears roughly similar across all the inspected subsets. The subset that most stands out are PDP's numbers from wards carried by it, which might

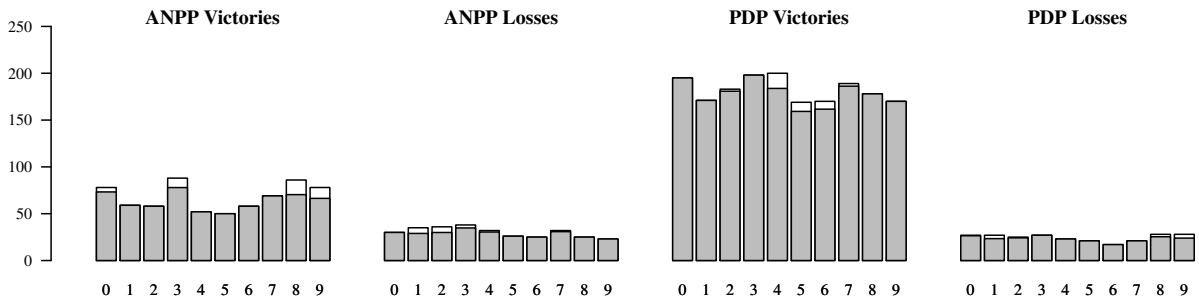


Figure A.4: Model fit under Δ of the second log-linear model (result-party, numeral) to the Nigerian data. Observations that do not need to be reallocated in grey, residuals in white.

be considerably more or less contaminated then the other subsets.

Appendix B

Appendix to 'Analysis of Electoral Support with the Dissimilarity Index'

B.1 Data Description

The CLEA dataset (Kollman et al., 2014) in its version from 12th August 2014 contains 1781 sets of constituency-level results from 132 countries. First, I have excluded all constituencies which were uncontested and/or had missing data on at least one party. From this data all the 1495 sets with data on more than one party with a positive national vote count were used. The resulting set contains elections from 119 countries, the earliest being from 1789 and the latest from 2013, and contains vote counts for 19518 party-election combinations.

Table B.1: List of 1495 elections from 119 countries used for comparing the measures. Number of elections in parenthses. Source: Kollman et al. (2014).

Albania	(4) 2001–2013	Estonia	(6) 1992–2011	Nicaragua	(5) 1990–2011
Andorra	(4) 1997–2011	Finland	(35) 1907–2007	Nigeria	(1) 2003
Angola	(2) 2008–2012	France	(7) 1978–2002	Norway	(34) 1882–2009
Anguilla	(6) 1989–2010	Gambia	(2) 1997–2007	Pakistan	(1) 2002
Antigua and Barbuda	(13) 1951–2009	Georgia	(1) 2012	Paraguay	(3) 1998–2008
Argentina	(14) 1983–2009	Germany	(38) 1871–2005	Peru	(8) 1963–2011
Armenia	(2) 2007–2012	Ghana	(3) 2000–2008	Philippines	(2) 1998–2007
Australia	(34) 1901–1984	Greece	(22) 1926–2000	Poland	(5) 1991–2005
Austria	(25) 1919–2008	Grenada	(15) 1951–2013	Portugal	(14) 1975–2011
Azerbaijan	(1) 2010	Guatemala	(7) 1984–2011	Puerto Rico	(5) 1992–2008
Bahamas	(4) 1997–2012	Guinea	(1) 2013	Romania	(4) 1990–2000
Bangladesh	(3) 1991–2008	Guinea-Bissau	(2) 1994–2004	Russian Federation	(3) 2003–2011
Barbados	(11) 1966–2013	Guyana	(6) 1953–2006	Saint Kitts and Nevis	(13) 1952–2010
Belgium	(61) 1847–1995	Honduras	(7) 1980–2005	Saint Lucia	(14) 1951–2006
Belize	(14) 1954–2012	Hungary	(6) 1990–2010	Samoa	(1) 2011
Benin	(3) 1991–2011	Iceland	(26) 1916–1995	Seychelles	(1) 2007
Bermuda	(12) 1963–2012	India	(13) 1977–2009	Singapore	(11) 1963–2006
Bhutan	(2) 2008–2013	Indonesia	(2) 1999–2004	Somaliland	(1) 2005
Bolivia	(9) 1979–2009	Iraq	(1) 2010	South Africa	(1) 2009
Bosnia and Herzegovina	(1) 2006	Ireland	(26) 1922–1997	Spain	(8) 1977–2004
Botswana	(9) 1969–2009	Italy	(16) 1919–1996	Sri Lanka	(12) 1952–2010
Brazil	(14) 1945–2010	Jamaica	(14) 1944–2002	St. Vincent and the G.	(16) 1951–2010
British Virgin Islands	(3) 2003–2011	Japan	(21) 1947–2012	Suriname	(3) 2000–2010
Bulgaria	(6) 1991–2009	Kenya	(5) 1961–2013	Sweden	(31) 1911–2006
Cambodia	(1) 2008	Korea	(16) 1958–2012	Switzerland	(45) 1848–1995
Cameroon	(2) 1997–2002	Kosovo	(2) 2007–2010	Taiwan	(6) 1986–2004
Canada	(40) 1867–2011	Latvia	(3) 1998–2006	Tanzania	(1) 2005
Cape Verde	(4) 1995–2011	Lesotho	(7) 1965–2012	Thailand	(8) 1969–1992
Cayman Islands	(2) 2005–2009	Liberia	(1) 2005	Togo	(2) 2007–2013
Colombia	(4) 1998–2010	Liechtenstein	(21) 1945–2013	Trinidad and Tobago	(7) 1991–2010
Costa Rica	(15) 1953–2010	Luxembourg	(18) 1919–1994	Turkey	(16) 1950–2011
Croatia	(1) 2007	Macedonia	(4) 2002–2011	Turks and Caicos Islands	(2) 2007–2012
Cyprus	(3) 2001–2011	Malawi	(2) 1999–2004	UK	(38) 1832–2010
Czech Republic	(6) 1990–2006	Malaysia	(1) 2013	Ukraine	(1) 1998
Denmark	(69) 1849–2011	Mauritius	(9) 1967–2005	Uruguay	(11) 1954–2009
Dominica	(3) 1995–2005	Mexico	(8) 1991–2012	US	(284) 1789–2012
Dominican Republic	(11) 1962–2006	Mozambique	(3) 1999–2009	US	(1) 1980
Ecuador	(9) 1979–2013	Nepal	(1) 2008	Zambia	(5) 1968–2006
El Salvador	(7) 1994–2012	Netherlands	(36) 1888–2012	Zimbabwe	(2) 2005–2013
Equatorial Guinea	(1) 1993	New Zealand	(20) 1946–2011		

Table B.2: Numbers of constituencies and options covered by elections in the Belgian data. Data source: Kollman et al. (2014).

	'46	'49	'50	'54	'58	'61	'65	'68	'71	'74	'77	'78	'81	'85	'87	'91	'95
Constituencies	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	20
Options	8	7	7	8	8	7	7	10	9	10	10	11	13	13	13	14	13

Appendix C

Appendix to 'Minimum Mixture Models of Voter Transitions in Aggregate Data'

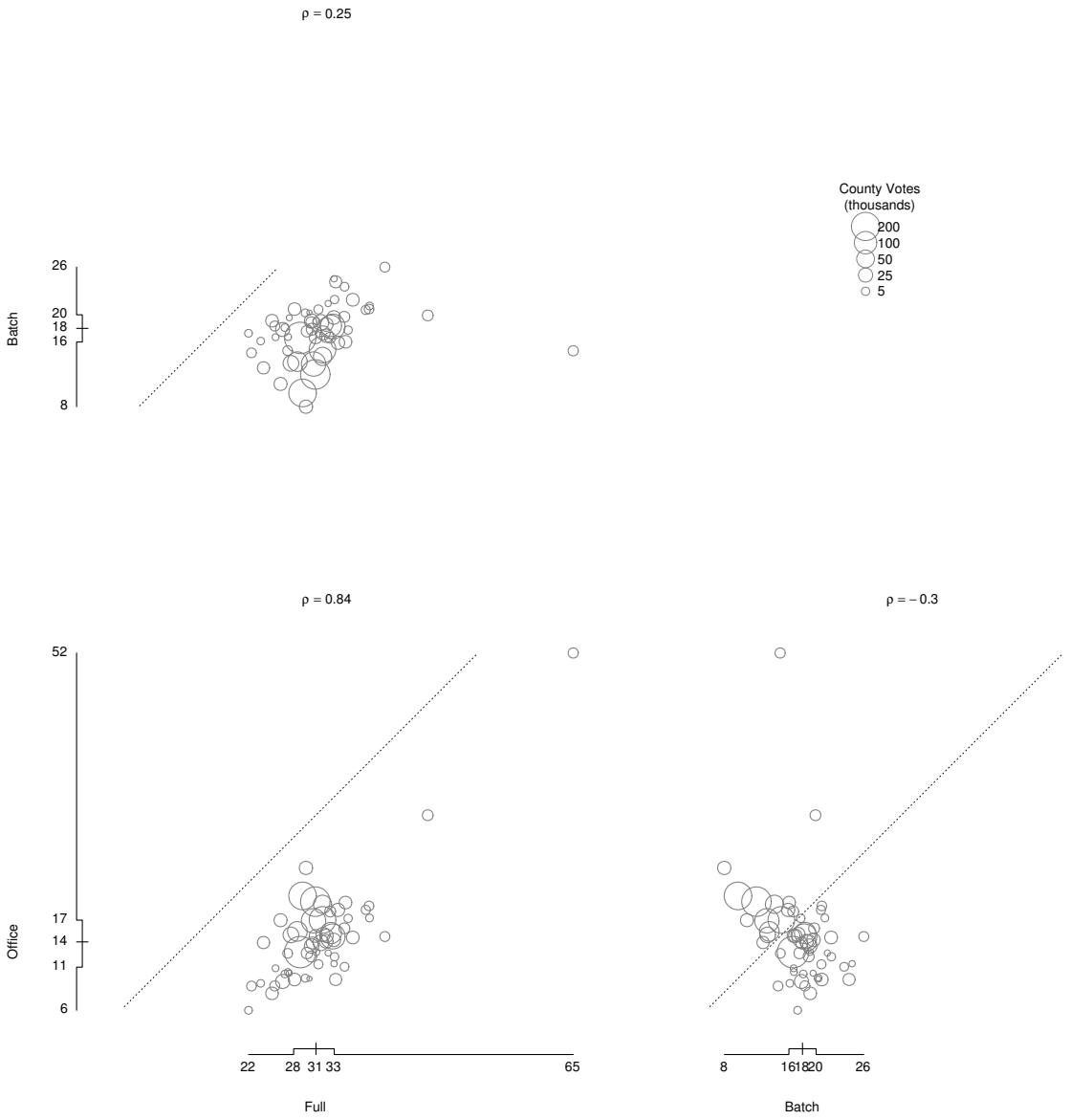


Figure C.1: Distribution of residual fractions across counties under the three models of constant voting fit to the Montana 2004–2008 data aggregated by county. Fractions in rounded percents. Size by county number of votes. Data source: Ansolabehere et al. (2014).

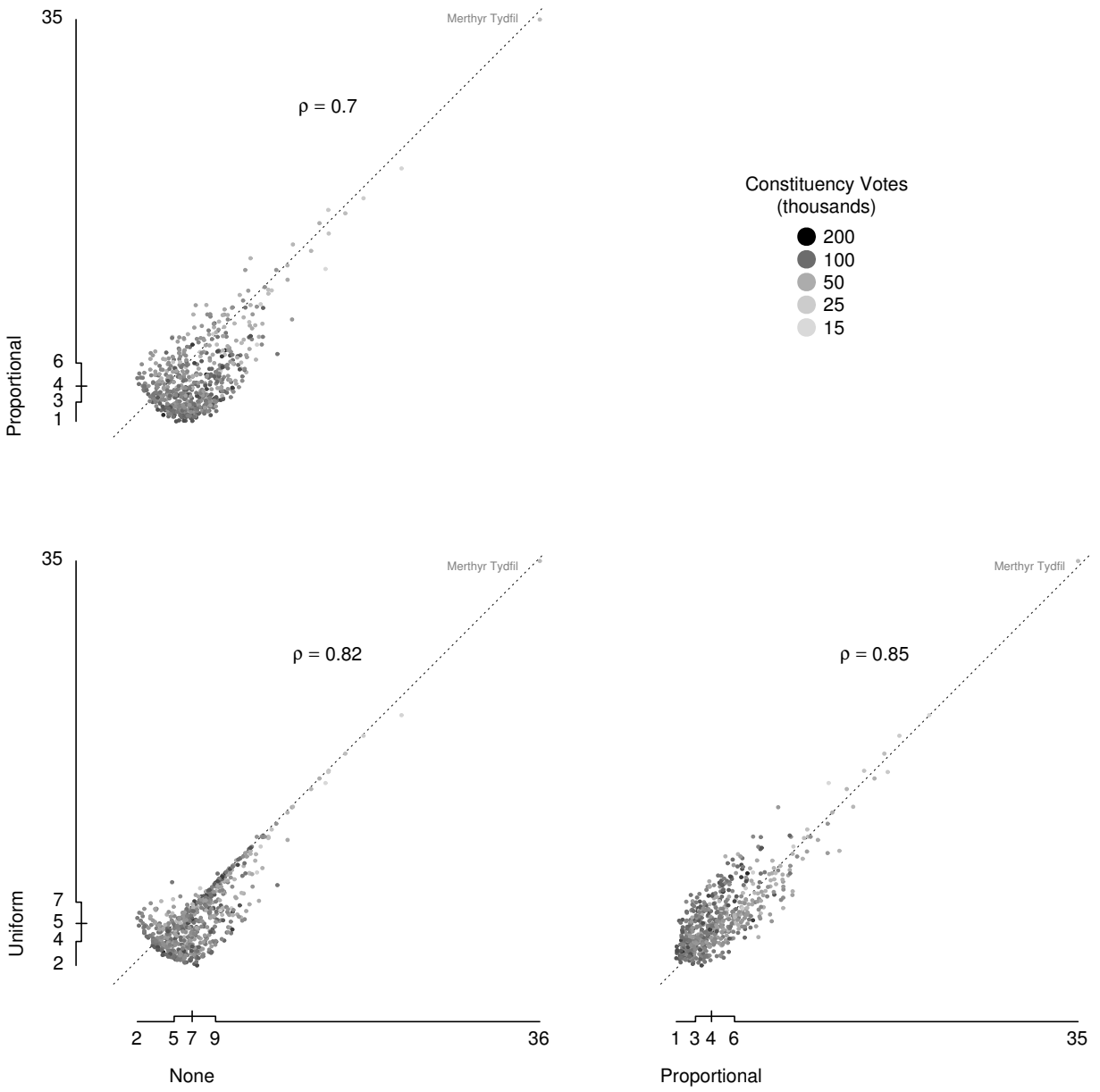


Figure C.2: Distribution of residual fractions across constituencies under the three models (no, uniform, and proportional swing) fit to the UK 1966–1970 data. Fractions in rounded percents. Color by constituency number of votes. Data sources: Kollman et al. (2014); Kimber (2015).

Appendix D

Appendix to 'Roll Call Analysis with the Mixture Index of Fit'

Table D.1: U.S. Census Regions (cite)

Region	Division	States
1. Northeast	1. New England	Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, and Vermont
	2. Mid-Atlantic	New Jersey, New York, and Pennsylvania
2. Midwest	3. East North Central	Illinois, Indiana, Michigan, Ohio, and Wisconsin
	4. West North Central	Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota, and South Dakota
3. South	5. South Atlantic	Delaware, Florida, Georgia, Maryland, North Carolina, South Carolina, Virginia, Washington D.C., and West Virginia
	6. East South Central	Alabama, Kentucky, Mississippi, and Tennessee
	7. West South Central	Arkansas, Louisiana, Oklahoma, and Texas
4. West	8. Mountain	Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah, and Wyoming
	9. Pacific	Alaska, California, Hawaii, Oregon, and Washington

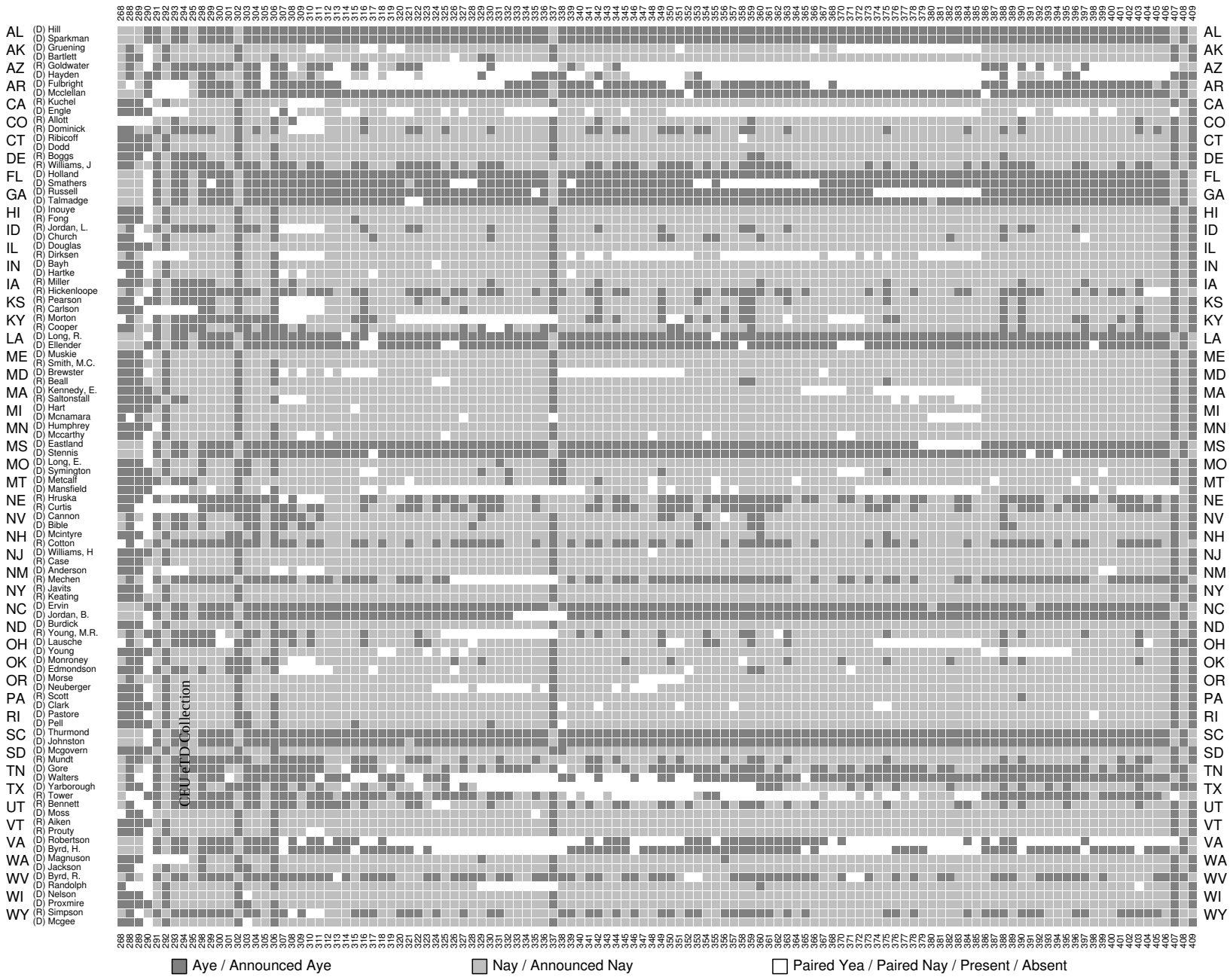


Figure D.1: 120 Senate roll calls related to the Civil Rights Act of 1964. Data source: Poole and Rosenthal (2000).

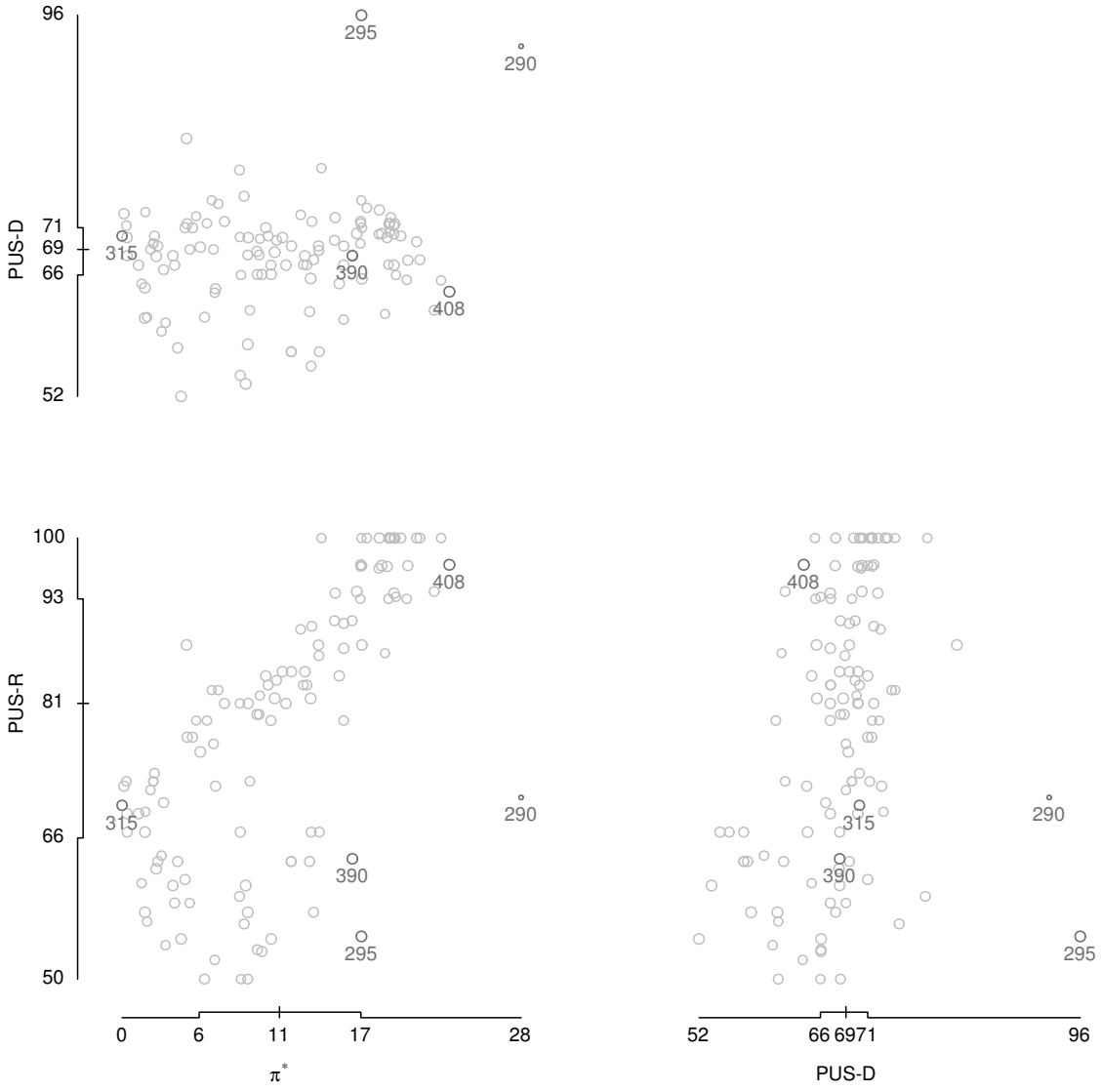


Figure D.2: Scatterplots of the mixture index of fit for party-option independence, and Party Unity Scores for Democrats and Republicans for the 120 Senate roll calls (see Figure D.1). Point size by the number of senators casting or announcing an aye or nay. Data source: Poole and Rosenthal (2000).

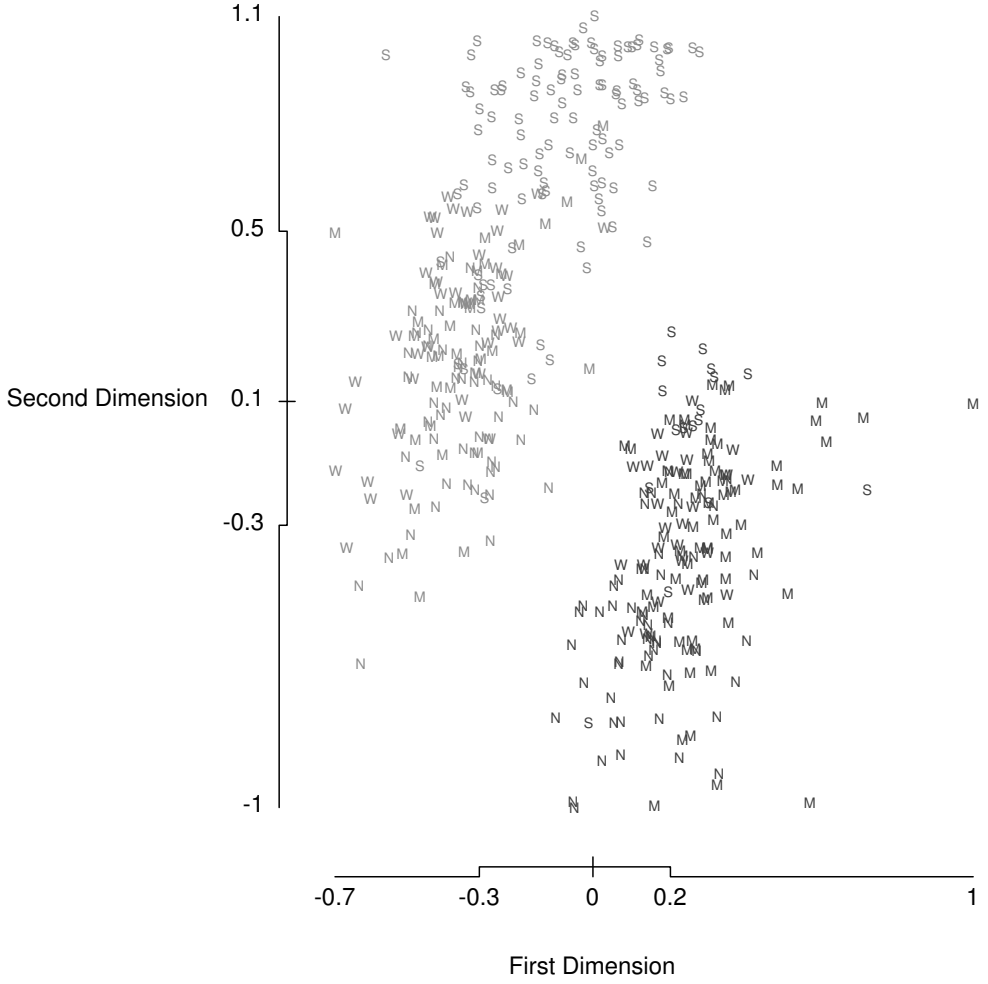


Figure D.3: Positions of the representatives who cast or announced an aye or nay on the passage of the Civil Rights Act of 1964 on the two dimensions of DW-NOMINATE. Democrats in lighter type and Republicans in darker type. Letters according to the U.S. Census regions (see Table D.1). Data source: Poole and Rosenthal (2000).

Appendix E

Appendix to 'The Mixture Index of Fit in Text Analysis'

Table E.1: Sources of the ten U.S. party platforms.

Party	Cycle	URL
Democratic	1996	http://www.presidency.ucsb.edu/ws/index.php?pid=29611
	2000	http://www.presidency.ucsb.edu/ws/index.php?pid=29612
Green	1996	http://janda.org/politxts/PartyPlatforms/OtherParties/green.996.html
	2000	http://janda.org/politxts/PartyPlatforms/OtherParties/green.2000.html
Libertarian	1996	http://janda.org/politxts/PartyPlatforms/OtherParties/libertar.996.html
	2000	http://janda.org/politxts/PartyPlatforms/OtherParties/libertar.2000.html
Reform	1996	http://janda.org/politxts/PartyPlatforms/OtherParties/reform.996.html
	2000	http://janda.org/politxts/PartyPlatforms/OtherParties/reform.2000.html
Republican	1996	http://www.presidency.ucsb.edu/ws/index.php?pid=25848
	2000	http://www.presidency.ucsb.edu/ws/index.php?pid=25849

Bibliography

Achen, C. H. and W. P. Shively (1995). *Cross-Level Inference*. University of Chicago Press.

Agresti, A. (2002). *Categorical Data Analysis, 2nd ed.* John Wiley & Sons.

Alemán, E. and M. Kellam (2008). The nationalization of electoral change in the Americas. *Electoral Studies 27*(2), 193–212.

Alvarez, R. M., L. R. Atkeson, and T. E. Hall (2012). *Evaluating Elections: A Handbook of Methods and Standards*. Cambridge University Press.

Alvarez, R. M., T. E. Hall, and S. D. Hyde (2009). *Election Fraud: Detecting and Deterring Electoral Manipulation*. Brookings Institution Press.

Ansolabehere, S., M. Palmer, and A. Lee (2014). Precinct-level election data.

Armstrong, D. A., R. Bakker, R. Carroll, C. Hare, K. T. Poole, H. Rosenthal, et al. (2014). *Analyzing Spatial Models of Choice and Judgment with R*. CRC Press.

Attinà, F. (1990). The voting behaviour of the European Parliament members and the problem of the Europarties. *European Journal of Political Research 18*(5), 557–579.

Baker, S. G. (1994). The multinomial-Poisson transformation. *The Statistician 43*(4), 495–504.

Balinski, M. L. and H. P. Young (1982). *Fair Representation: Meeting the Ideal of One Man, one Vote*. Yale University Press.

Bartels, L. M. (1998). Electoral continuity and change, 1868–1996. *Electoral Studies 17*(3), 301–326.

BBC (2005). Rebel history lesson for new MP. `http://news.bbc.co.uk/go/pr/fr/-/2/hi/uk_news/politics/vote_2005/wales/4526753.stm`.

Beber, B. and A. Scacco (2011). Replication data for: What the numbers say: A digit-based test for election fraud. `http://hdl.handle.net/1902.1/17151`. Harvard Dataverse, V2.

Beber, B. and A. Scacco (2012). What the numbers say: A digit-based test for election fraud. *Political Analysis 20*(2), 211–234.

Benaglia, T., D. Chauveau, D. Hunter, and D. Young (2009). mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software 32*(6), 1–29.

Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 551–572.

Berg, S. (1988). Spatial influence on voter transitions in Swedish elections: An application of Johnston's maximum entropy method. *Electoral Studies 7*(3), 233–250.

Bi, J. (2008). *Sensory Discrimination Tests and Measurements: Statistical Principles, Procedures and Tables*. John Wiley & Sons.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM 55*(4), 77–84.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent Dirichlet allocation. *The Journal of Machine Learning Research 3*, 993–1022.

Bochsler, D. (2010). Measuring party nationalisation: A new Gini-based indicator that corrects for the number of units. *Electoral Studies 29*(1), 155–168.

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association 71*(356), 791–799.

Brent, R. P. (1973). *Algorithms for Minimization Without Derivatives*. Courier Corporation.

Breunig, C. and A. Goerres (2011). Searching for electoral irregularities in an established democracy: Applying Benford's law tests to Bundestag elections in unified Germany. *Electoral Studies 30*(3), 534–545.

Butler, D. and E. Stokes, Donald (1969). *Political Change in Britain*. Macmillan.

Buttorf, G. (2008). Detecting fraud in America's gilded age. Unpublished manuscript, University of Iowa.

Calvo, E. and J. Rodden (2015). The Achilles heel of plurality systems: Geography and representation in multiparty democracies. *American Journal of Political Science*.

Cantú, F. and S. M. Saiegh (2011). Fraudulent democracy? An analysis of Argentina's infamous decade using supervised machine learning. *Political Analysis 19*(4), 409–433.

Caramani, D. (2000). *Elections in Western Europe Since 1815: Electoral Results by Constituencies*. Macmillan.

Caramani, D. (2004). *The Nationalization of Politics: The Formation of National Electorates and Party Systems in Western Europe*. Cambridge University Press.

Carroll, R. and K. Poole (2014). Roll call analysis and the study of legislatures. *The Oxford Handbook of Legislative Studies*, 103–124.

Chang, J., S. Gerrish, C. Wang, J. L. Boyd-Graber, and D. M. Blei (2009). Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems*, pp. 288–296.

Chhibber, P. and K. Kollman (1998). Party aggregation and the number of parties in India and the United States. *American Political Science Review*, 329–342.

Cleave, N., P. J. Brown, and C. D. Payne (1995). Evaluation of methods for ecological inference. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 55–72.

Clinton, J., S. Jackman, and D. Rivers (2004). The statistical analysis of roll call data. *American Political Science Review 98*(02), 355–370.

Clinton, J. D. (2012). Using roll call estimates to test models of politics. *Annual Review of Political Science 15*, 79–99.

Clogg, C., T. Rudas, and L. Xi (1995). A new index of structure for the analysis of models for mobility tables and other cross-classifications. *Sociological Methodology*, 197–222.

Clogg, C. C., T. Rudas, and S. Matthews (1997). Analysis of contingency tables using graphical displays based on the mixture index of fit. In J. Blasius and M. Greenacre (Eds.), *Visualization of Categorical Data*, pp. 425–439. Academic Press.

Cox, G. (1999). Electoral rules and electoral coordination. *Annual Review of Political Science 2*(1), 145–161.

Dayton, C. M. (2003). Applications and computational strategies for the two-point mixture index of fit. *British Journal of Mathematical and Statistical Psychology 56*(1), 1–13.

Dayton, C. M. (2008). Applications and extensions of the two-point mixture index of model fit. In G. R. Hancock and K. M. Samuelsen (Eds.), *Advances in Latent Variable Mixture Models*, pp. 299–316. IAP.

De Boeck, P. (2008). Random item IRT models. *Psychometrika 73*(4), 533–559.

De Boeck, P., M. Bakker, R. Zwitser, M. Nivard, A. Hofman, F. Tuerlinckx, I. Partchev, et al. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software 39*(12), 1–28.

De Leeuw, J. (2006). Principal component analysis of binary data by iterated singular value decomposition. *Computational Statistics & Data Analysis 50*(1), 21–39.

de Leeuw, J. (2007). Principal component analysis of senate voting patterns. In S. S. Sawilowsky (Ed.), *Real Data Analysis*, pp. 405–410.

Deckert, J., M. Myagkov, and P. C. Ordeshook (2011). Benford's law and the detection of election fraud. *Political Analysis 19*(3), 245–268.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological) 39*(1), 1–38.

Ducan, O. D. and B. Ducan (1955). A methodological analysis of segregation indices. *American Sociological Review 20*, 200–217.

Duncan, O. D. and B. Davis (1953). An alternative to ecological correlation. *American Sociological Review 18*(6), 665–666.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*, Volume 38. SIAM.

Ennis, D. M. and J. Bi (1998). The Beta-binomial model: Accounting for inter-trial variation in replicated difference and preference tests. *Journal of Sensory Studies 13*(4), 389–412.

Ersson, S., K. Janda, and J.-E. Lane (1985). Ecology of party strength in Western Europe a regional analysis. *Comparative Political Studies 18*(2), 170–205.

Feinerer, I. and K. Hornik (2014). tm: Text mining package. *R package version 0.5-7.1*.

Feinerer, I., D. Meyer, and K. Hornik (2008). Text mining infrastructure in R. *Journal of Statistical Software 25*(5), 1–54.

Formann, A. K. (2000). Rater agreement and the generalized Rudas–Clogg–Lindsay index of fit. *Statistics in Medicine 19*(14), 1881–1888.

Formann, A. K. (2003a). Latent class model diagnosis from a frequentist point of view. *Biometrics 59*(1), 189–196.

Formann, A. K. (2003b). Latent class model diagnostics–a review and some proposals. *Computational Statistics & Data Analysis 41*(3), 549–559.

Formann, A. K. (2006). Testing the Rasch model by means of the mixture fit index. *British Journal of Mathematical and Statistical Psychology 59*(1), 89–95.

Freedman, D. A. (1999). Ecological inference and the ecological fallacy. In N. J. Smelser and P. B. Baltes (Eds.), *International Encyclopedia of the Social & Behavioral Sciences, Vol. 6*, pp. 4027–4030.

Freedman, D. A., S. P. Klein, J. Sacks, C. A. Smyth, and C. G. Everett (1991). Ecological regression and voting rights. *Evaluation Review 15*(6), 673–711.

Gallagher, M. (1991). Proportionality, disproportionality and electoral systems. *Electoral studies 10*(1), 33–51.

Gelman, A., D. K. Park, S. Ansolabehere, P. N. Price, and L. C. Minnite (2001). Models, assumptions and model checking in ecological regressions. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 164*(1), 101–118.

Giles, D. E. (2007). Benford's law and naturally occurring prices in certain ebay auctions. *Applied Economics Letters 14*(3), 157–161.

Gini, C. (1914). Di una misura della dissomiglianza tra due gruppi di quantità e delle sue applicazioni allo studio delle relazione statistiche. *Atti del Reale Instituto Veneto di Scienze, Lettere ed Arti (Series 8) 74*, 185–213.

Glynn, A. N. and J. Wakefield (2010). Ecological inference in the social sciences. *Statistical Methodology 7*(3), 307–322.

Golosov, G. V. (2014). Party system nationalization the problems of measurement with an application to federal states. *Party Politics*, Advance access.

Golosov, G. V. and E. Ponarin (1999). Regional bases of party politics: a measure and its implications for the study of party system consolidation in new democracies.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika 61*(2), 215–231.

Goodman, L. A. and W. H. Kruskal (1954). Measures of association for cross classifications*. *Journal of the American Statistical Association 49*(268), 732–764.

Goodman, L. A. and W. H. Kruskal (1959). Measures of association for cross classifications. II: Further discussion and references. *Journal of the American Statistical Association 54*(285), 123–163.

Greenacre, M. and J. Blasius (2006). *Multiple Correspondence Analysis and Related Methods*. CRC Press.

Grego, J. (2010). R code for a set of functions and sample session demonstrating Clogg, Lindsay, and Rudas's method for mixing an independent table and fully nonparametric table. `http://www.stat.sc.edu/~grego/courses/stat770/CLR.txt`.

Greiner, J. D. and K. M. Quinn (2009). R×C ecological inference: bounds, correlations, flexibility and transparency of assumptions. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 172*(1), 67–81.

Grimmer, J. and B. M. Stewart (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 1–31.

Healy, K. J. (2015). 2015 UK election data. `https://github.com/kjhealy/uk-elections/`.

Hernández, J. M., V. J. Rubio, J. Revuelta, and J. Santacreu (2006). A procedure for estimating intrasubject behavior consistency. *Educational and Psychological Measurement 66*(3), 417–434.

Hidalgo, F. D. (2010). Digital democratization: Suffrage expansion and the decline of political machines in Brazil. *Manuscript, Department of Political Science, University of California at Berkeley*.

Hill, T. P. (1995). A statistical derivation of the significant-digit law. *Statistical Science*, 354–363.

Hix, S., A. G. Noury, and G. Roland (2007). *Democratic Politics in the European Parliament*. Cambridge University Press.

Hjorth, F., R. Klemmensen, S. Hobolt, M. E. Hansen, and P. Kurrild-Klitgaard (2015). Computers, coders, and voters: Comparing automated methods for estimating party positions. *Research & Politics 2*(2), 2053168015580476.

Hofmann, T. (1999a). Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, pp. 289–296. Morgan Kaufmann Publishers Inc.

Hofmann, T. (1999b). Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 50–57. ACM.

Imai, K. and D. Tingley (2012). A statistical method for empirical testing of competing theories. *American Journal of Political Science 56*(1), 218–236.

Ispány, M. and E. Verdes (2014). On the robustness of mixture index of fit. *Journal of Mathematical Sciences 200*(4), 432–440.

Jeong, G.-H., G. J. Miller, and I. Sened (2009). Closing the deal: negotiating civil rights legislation. *American Political Science Review 103*(04), 588–606.

Jiménez, R. and M. Hidalgo (2014). Forensic analysis of Venezuelan elections during the Chávez presidency. *PloS ONE 9*(6), e100884.

Johnston, R. (1980). Federal and provincial voting: contemporary patterns and historical evolution. In D. J. Elkins and R. Simeon (Eds.), *Small Worlds: Provinces and Parties in Canadian Political Life*, pp. 131–78. Toronto: Methuen.

Johnston, R. and A. Hay (1982). On the parameters of uniform swing in single-member constituency electoral systems. *Environment and Planning A 14*(1), 61–74.

Johnston, R. and C. Pattie (2000). Ecological inference and entropy-maximizing: An alternative estimation procedure for split-ticket voting. *Political Analysis 8*(4), 333–345.

Johnston, R. and C. Pattie (2003). Evaluating an entropy-maximizing solution to the ecological inference problem: Split-ticket voting in New Zealand, 1999. *Geographical Analysis 35*(1), 1–23.

Johnston, R. J. and A. Hay (1983). Voter transition probability estimates: An entropy-maximizing approach. *European Journal of Political Research 11*(1), 93–98.

Jones, M. P. and S. Mainwaring (2003). The nationalization of parties and party systems: An empirical measure and an application to the Americas. *Party Politics 9*(2), 139–166.

Judge, G. and L. Schechter (2009). Detecting problems in survey data using Benford's law. *Journal of Human Resources 44*(1), 1–24.

Jurafsky, D. and J. H. Martin (2000). *Speech and Language Processing*. Pearson.

Kamary, K., K. Mengersen, C. P. Robert, and J. Rousseau (2014). Testing hypotheses via a mixture estimation model. *arXiv preprint arXiv:1412.2044*.

Kawato, S. (1987). Nationalization and partisan realignment in congressional elections. *American Political Science Review 81*(04), 1235–1250.

Kendall, M. G. and A. Stuart (1961). *The Advanced Theory of Statistics. Vol. 2: Inference and Relationship*. Griffin.

Kimber, R. (2015). Political science resources: UK general elections. `http://www.politicsresources.net/election.htm#british`.

Knott, M. (2005). A measure of independence for a multivariate normal distribution and some connections with factor analysis. *Journal of Multivariate Analysis 96*(2), 374–383.

Kollman, K., A. Hicken, D. Caramani, D. Backer, and D. Lublin (2014). Constituency-level elections archive. `http://www.electiondataarchive.org/cite.html`.

Lago, I. and J. R. Montero (2014). Defining and measuring party system nationalization. *European Political Science Review 6*(02), 191–211.

Laver, M. (2014). Measuring policy positions in political space. *Annual Review of Political Science 17*, 207–223.

Laver, M., K. Benoit, and J. Garry (2003). Extracting policy positions from political texts using words as data. *American Political Science Review 97*(02), 311–331.

Lazarsfeld, P. F., N. W. Henry, and T. W. Anderson (1968). *Latent Structure Analysis*. Houghton Mifflin Boston.

Lee, A. (1988). The persistence of difference: electoral change in Cornwall. In *Political Studies Association Conference, Plymouth*.

Leemann, L. and D. Bochsler (2014). A systematic approach to study electoral fraud. *Electoral Studies 35*, 33–47.

Leemis, L. M., B. W. Schmeiser, and D. L. Evans (2000). Survival distributions satisfying Benford's law. *The American Statistician 54*(4), 236–241.

Loosemore, J. and V. J. Hanby (1971). The theoretical limits of maximum distortion: some analytic expressions for electoral systems. *British Journal of Political Science 1*(04), 467–477.

Lowe, W. (2008). Understanding Wordscores. *Political Analysis 16*(4), 356–371.

Lowe, W. (2013). There's (basically) only one way to do it. *Available at SSRN 2318543*.

Lowe, W. and K. Benoit (2013). Validating estimates of latent traits from textual data using human judgment as a benchmark. *Political Analysis 21*(3), 298–313.

Lupoli, J. B. (2009). Party unity score. In L. Sabato and H. R. Ernst (Eds.), *Encyclopedia of American political parties and elections*, pp. 268. Infobase Publishing.

Magis, D., S. Béland, F. Tuerlinckx, and P. De Boeck (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods 42*(3), 847–862.

Massey, D. S. and N. A. Denton (1988). The dimensions of residential segregation. *Social Forces 67*(2), 281–315.

McCutcheon, A. L. (1987). *Latent Class Analysis*. Number 64. Sage.

McLean, I. (1973). The problem of proportionate swing. *Political Studies 21*(1), 57–63.

Mebane, W. R. (2006a). Election forensics: the second-digit Benford's law test and recent American presidential elections. In *Election Fraud Conference*.

Mebane, W. R. (2006b). Election forensics: Vote counts and Benford's law. In *Summer Meeting of the Political Methodology Society, UC-Davis, July*.

Mebane, W. R. (2007). Election forensics: Statistical interventions in election controversies. In *Annual Meeting of the American Political Science Association*.

Mebane, W. R. (2008). Election forensics: Outlier and digit tests in America and Russia. In *American Electoral Process conference, Center for the Study of Democratic Politics, Princeton University*.

Mebane, W. R. (2010a). Election fraud or strategic voting? Can second-digit tests tell the difference? In *Summer Meeting of the Political Methodology Society, University of Iowa*.

Mebane, W. R. (2010b). Fraud in the 2009 presidential election in Iran? *Chance 23*(1), 6–15.

Mebane, W. R. (2011). Comment on "Benford's law and the detection of election fraud". *Political Analysis 19*(3), 269–272.

Mebane, W. R. and K. Kalinin (2009). Comparative election fraud detection. In *Annual Meeting of the American Political Science Association*.

Medzihorsky, J. (2015a). Election fraud: A latent class framework for digit-based tests. *Political Analysis forthcomming*.

Medzihorsky, J. (2015b). pistar: Rudas, Clogg and Lindsay mixture index of fit. `https://github.com/jmedzihorsky/pistar`. R package version 0.5.2.5.

Medzihorsky, J. (2015c). Replication data for: Election fraud: A latent class framework for digit-based tests. `http://dx.doi.org/10.7910/DVN/1FYXUJ`. Harvard Dataverse, V2.

Miller, W. (1972). Measures of electoral change using aggregate data. *Journal of the Royal Statistical Society. Series A (General)*, 122–142.

Moenius, J. and Y. Kasuya (2004). Measuring party linkage across districts some party system inflation indices and their properties. *Party Politics 10*(5), 543–564.

Monroe, B. L. (1994). Disproportionality and malapportionment: Measuring electoral inequity. *Electoral Studies 13*(2), 132–149.

Morgenstern, S., J. Polga-Hecimovich, and P. M. Siavelis (2014). Seven imperatives for improving the measurement of party nationalization with evidence from Chile. *Electoral Studies 33*, 186–199.

Morgenstern, S. and R. F. Potthoff (2005). The components of elections: district heterogeneity, district-time effects, and volatility. *Electoral Studies 24*(1), 17–40.

Morgenstern, S., S. M. Swindle, and A. Castagnola (2009). Party nationalization and institutions. *The Journal of Politics 71*(04), 1322–1341.

Mustillo, T. and S. A. Mustillo (2012). Party nationalization in a multilevel context: Where's the variance? *Electoral Studies 31*(2), 422–433.

Nelder, J. A. and R. Mead (1965). A simplex method for function minimization. *The Computer Journal 7*(4), 308–313.

Newcomb, S. (1881). Note on the frequency of use of the different digits in natural numbers. *American Journal of Mathematics 4*(1), 39–40.

Nickerson, R. S. (2002). The production and perception of randomness. *Psychological Review 109*(2), 330.

Nohlen, D. (2005). *Elections in the Americas: A Data Handbook, Volume 2*. Oxford: Oxford University Press.

Norris, P., R. W. Frank, and F. M. i Coma (2014). *Advancing Electoral Integrity*. Oxford University Press.

Pedersen, M. N. (1979). The dynamics of European party systems: changing patterns of electoral volatility. *European Journal of Political Research 7*(1), 1–26.

Pennisi, A. (1998). Disproportionality indexes and robustness of proportional allocation methods. *Electoral Studies 17*(1), 3–19.

Pericchi, L. and D. Torres (2011). Quick anomaly detection by the Newcomb–Benford law, with applications to electoral processes data from the USA, Puerto Rico and Venezuela. *Statistical Science 26*(4), 502–516.

Poole, K. T. (2000). Nonparametric unfolding of binary choice data. *Political Analysis 8*(3), 211–237.

Poole, K. T. (2005). *Spatial Models of Parliamentary Voting*. Cambridge University Press.

Poole, K. T. and H. Rosenthal (1985). A spatial model for legislative roll call analysis. *American Journal of Political Science*, 357–384.

Poole, K. T. and H. Rosenthal (1991). Patterns of Congressional voting. *American Journal of Political Science*, 228–278.

Poole, K. T. and H. Rosenthal (2000). *Congress: A Political-Economic History of Roll Call Voting*. Oxford University Press.

Powell, E. N. (2015). Pure position-taking in the US House of Representatives. *Unpublished manuscript*.

Powell, E. N. and J. A. Tucker (2014). Revisiting electoral volatility in post-communist countries: New data, new results and new approaches. *British Journal of Political Science 44*(01), 123–147.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rasch, G. (1980). *Probabilistic Models for Some Intelligence and Attainment Tests*. University of Chicago Press.

Revuelta, J. (2008). Estimating the π^* goodness of fit index for finite mixtures of item response models. *British Journal of Mathematical and Statistical Psychology 61*(1), 93–113.

Rice, S. A. (1928). *Quantitative Methods in Politics*. Knopf.

Roberts, C. W. (2000). A conceptual framework for quantitative text analysis. *Quality and Quantity 34*(3), 259–274.

Rodriguez, D. B. and B. R. Weingast (2003). The positive political theory of legislative history: New perspectives on the 1964 Civil Rights Act and its interpretation. *University of Pennsylvania Law Review*, 1417–1542.

Rose, R. and D. W. Urwin (1975). *Regional Differentiation and Political Unity in Western Nations*. Sage.

Rudas, T. (1998a). The mixture index of fit. In A. Ferligoj (Ed.), *Advances in Methodology, Data Analysis, and Statistics*, pp. 15–22. FDV.

Rudas, T. (1998b). *Odds Ratios in the Analysis of Contingency Tables*. Sage.

Rudas, T. (1999). The mixture index of fit and minimax regression. *Metrika 50*(2), 163–172.

Rudas, T. (2002). A latent class approach to measuring the fit of a statistical model. In J. A. Hagenaars and A. L. McCutcheon (Eds.), *Applied Latent Class Analysis*, pp. 345–365. Cambridge University Press.

Rudas, T. (2005). Mixture models of missing data. *Quality & Quantity 39*(1), 19–36.

Rudas, T., C. Clogg, and B. Lindsay (1994). A new index of fit based on mixture methods for the analysis of contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological) 56*(4), 623–639.

Rudas, T. and E. Verdes (2015). Model based analysis of incomplete data using the mixture index of fit. In G. R. Hancock and G. B. Macready (Eds.), *Advances in Latent Class Analysis: A Festschrift in Honor of C. Mitchell Dayton*. Information Age Publishing.

Rudas, T. and R. Zwick (1997). Estimating the importance of differential item functioning. *Journal of Educational and Behavioral Statistics 22*(1), 31–45.

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons.

Shively, W. P. (1991). A general extension of the method of bounds, with special application to studies of electoral transition. *Historical Methods: A Journal of Quantitative and Interdisciplinary History 24*(2), 81–94.

Slapin, J. B. and S.-O. Proksch (2008). A scaling model for estimating time-series party positions from texts. *American Journal of Political Science 52*(3), 705–722.

Stokes, D. E. (1965). A variance components model of political effects. In J. Claunch (Ed.), *Mathematical Applications in Political science*, pp. 61–85. Dallas: The Arnold Foundation.

Stokes, D. E. (1967). Parties and the nationalization of electoral forces. In W. Chambers and W. Burnham (Eds.), *The American Party Systems: Stages of Political Development*, pp. 182–202. OUP.

Taagepera, R. and B. Grofman (2003). Mapping the indices of seats–votes disproportionality and inter-election volatility. *Party Politics 9*(6), 659–677.

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

Tam Cho, W. K. and B. J. Gaines (2007). Breaking the (Benford) law: Statistical fraud detection in campaign finance. *The American Statistician 61*(3), 218–223.

Tractenberg, R. E., F. Yumoto, P. S. Aisen, J. A. Kaye, and R. J. Mislevy (2012). Using the guttman scale to define and estimate measurement error in items over time: the case of cognitive decline and the meaning of points lost. *PloS one 7*(2).

Verdes, E. (2002). Appendix: Algorithms to compute the mixture index of fit π^*. In J. A. Hagenaars and A. L. McCutcheon (Eds.), *Applied Latent Class Analysis*, pp. 363–365. Cambridge University Press.

Verdes, E. and T. Rudas (2003). The π^* index as a new alternative for assessing goodness of fit of logistic regression. In Y. Haitovsky and Y. Ritov (Eds.), *Foundations of Statistical Inference*, pp. 167–177. Springer.

Wallach, H. M., I. Murray, R. Salakhutdinov, and D. Mimno (2009). Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 1105–1112. ACM.

Wittenberg, J. (2006). *Crucibles of Political Loyalty: Church Institutions and Electoral Continuity in Hungary*. Cambridge University Press.

Wittenberg, J. (2013). How similar are they? Rethinking electoral congruence. *Quality & Quantity 47*(3), 1687–1701.

Xi, L. (1996). *The Mixture Index of Fit*. Ph. D. thesis, Department of Statistics, The Pennsylvania State University.

Xi, L. and B. G. Lindsay (1996). A note on calculating the π^* index of fit for the analysis of contingency tables. *Sociological Methods & Research 25*(2), 248–259.

Ziliak, S. T. and D. N. McCloskey (2008). *The Cult of Statistical Significance: How the Standard Error Costs us Jobs, Justice, and Lives*. University of Michigan Press.