PA Simulating Party Shares

Denis Cohen[1] and Chris Hanretty[2]

[1] *Mannheim Centre for European Social Research (MZES), University of Mannheim, Mannheim 68131, Germany.*
E-mail: denis.cohen@mzes.uni-mannheim.de
[2] *Department of Politics, International Relations and Philosophy, Royal Holloway, University of London, London, UK.*
E-mail: chris.hanretty@rhul.ac.uk

Abstract

We tackle the problem of simulating seat- and vote-shares for a party system of a given size. We show how these shares can be generated using unordered and ordered Dirichlet distributions. We show that a distribution with a mean vector given by the rule described in Taagepera and Allik (2006, *Electoral Studies* 25, 696–713) fits real-world data almost as well as a saturated model where there is a parameter for each rank/system size combination.

Keywords: party systems, compositional data, Dirichlet distribution

1 Introduction

Consider two scenarios:

1. A consultant is asked to advise on a country's choice of electoral system, and specifically a proposal for a legislature with 125 members ($S = 125$) and a mean district magnitude of 5 ($M = 5$). She knows that such a system is likely to feature $MS^{1/4} = 5$ seat-winning parties and that the seat share of the largest party, s_1, is $MS^{-1/8} \approx 45\%$ (Shugart and Taagepera 2017, 149), but legislators want to know the *probability* that a single party will have a majority.

2. A political scientist wishes to test a generative model of coalition formation (Golder, Golder, and Siegel 2012). She wishes to compare predicted coalition outcomes to observed outcomes in party systems with different numbers of seat-winning parties N_{S0}—but must first simulate distributions of seat shares.

In both scenarios, researchers lack ways of simulating realistic outcomes for party systems of different sizes. Although it is possible to perturb existing outcomes (Laver and Benoit 2015, 282–3), this approach cannot handle situations where researchers need to simulate party systems *ex nihilo*.

Here, we show how realistic party systems of a given size can be simulated using ordered and unordered Dirichlet distributions. The mean vector of these Dirichlet distributions is given by a formula for seat shares proposed by Taagepera and Allik (2006) which relies only on two parameters (party rank and the number of seat winning parties N_{S0}). We show that the fit of these simulations to real-world data is almost as good as a saturated model where the seat share of the *i*th ranked party in a system of size N_{S0} is given by the empirical mean for parties of that rank in systems of that size.

2 Theory

To the best of our knowledge, the only attempt to predict the size of vote- or seat-winning parties in a party system of size N_0 comes from Taagepera and Allik (2006), who suggest that the seat (vote) share of the largest party is equal to the geometric mean of two logical extremes: one extreme where the largest party wins $[100 - \epsilon]\%$ of seats (votes), where ϵ is some tiny share divided between the remaining parties, and another extreme where the largest party wins $[\frac{100}{N_0} + \epsilon]\%$ of

seats (votes), and is only fractionally larger than the remaining parties. By repeatedly appealing to the geometric mean of logical extremes, they construct the following formula for the seat (vote) share of the ith largest party:

$$s_i = \frac{1 - \sum_{j=1}^{i-1} s_j}{\sqrt{N_{S0} - i + 1}}. \tag{1}$$

Thus, for a system with five parties, the seat share of the first party is 45% (one divided by the square root of 5), the seat share of the second party is 27% (the remaining 55% divided by the square root of 4), and so on.

Taagepera and Allik ([2006](#)) also propose a second, "politically adjusted" model, which is like the model above, except that the seat share of "small parties" is half what it would be under the probabilistic model, with this surplus distributed between larger parties. Small parties are defined as parties whose rank is greater than $1/s_1$. Thus,

$$s_i' = \begin{cases} s_i + \dfrac{0.5 \sum_{i=\lceil 1/s_i \rceil}^{N_{S0}} s_i}{\sum 1(s_i \le \frac{1}{s_i})}, & i > 1/s_i, \\ 0.5 \quad s_i, & \text{otherwise.} \end{cases} \tag{2}$$

The predictions of these two models are compared visually to binned averages of election results for the nth largest parties taken from Mackie and Rose ([1997](#)). The authors conclude that the politically adjusted model fits the data better. Whether this conclusion is sound or not, these models remain deterministic. As such, they make it difficult to answer questions of the form, "what is the probability that a party system with five seat-winning parties will have a single party majority," even if we know our best guess as to the seat share of the largest party remains 45%.

3 Methods

Modeling party systems is difficult because seat and vote shares are ordered compositional data. They are *ordered* data because, since different parties compete in different countries at different times, we typically lack any way of referring to parties except by their rank within the system, and so we refer to the seat share of the first-largest party, the seat share of the second-largest party, and so on. They are *compositional* data because both seat and vote shares add up to one. Compositional data can be modeled by transforming d-dimensional compositions into a $(d-1)$-dimensional data through appropriate transforms (Aitchison [1986](#)), or by using probability distributions defined on the simplex. The Dirichlet distribution is the most common such distribution.

A Dirichlet distribution is typically governed by a vector of nonnegative concentration parameters θ. These parameters hold two different pieces of information. First, their *relative* magnitude determines the *location* of each element of the probability distribution. For instance, both three-dimensional simplexes $\mathbf{s}_A \sim \text{Dir}(\theta_A)$ with $\theta_A = \begin{bmatrix} 15 & 7.5 & 2.5 \end{bmatrix}'$ and $\mathbf{s}_B \sim \text{Dir}(\theta_B)$ with $\theta_B = \begin{bmatrix} 0.30 & 0.15 & 0.05 \end{bmatrix}'$ yield expected values of $\mathbb{E}[\mathbf{s}_A] = \mathbb{E}[\mathbf{s}_B] = \begin{bmatrix} 0.6 & 0.3 & 0.1 \end{bmatrix}'$. Second, the *absolute* magnitude determines the *scale* of the corresponding distributions: while both simplexes $\mathbf{s}_A$ and $\mathbf{s}_B$ have identical expected values, the low values of θ_B result in high dispersion, and high density near the extremes of 0 and 1 for the elements of $\mathbf{s}_B$. In contrast, the high values of θ_A result in high concentration such that there is high density around the expected values of the elements of $\mathbf{s}_A$ (see Figure [1](#)).

We therefore characterize the Dirichlet distribution in terms of a scalar concentration parameter α and a location vector of probabilities $\mathbf{p} = (p_1, p_2, \ldots, p_{N_0})$, $\sum \mathbf{p} = 1$:

$$\theta = \alpha \mathbf{p},$$
$$\mathbf{s} \sim \text{Dir}(\theta). \tag{3}$$
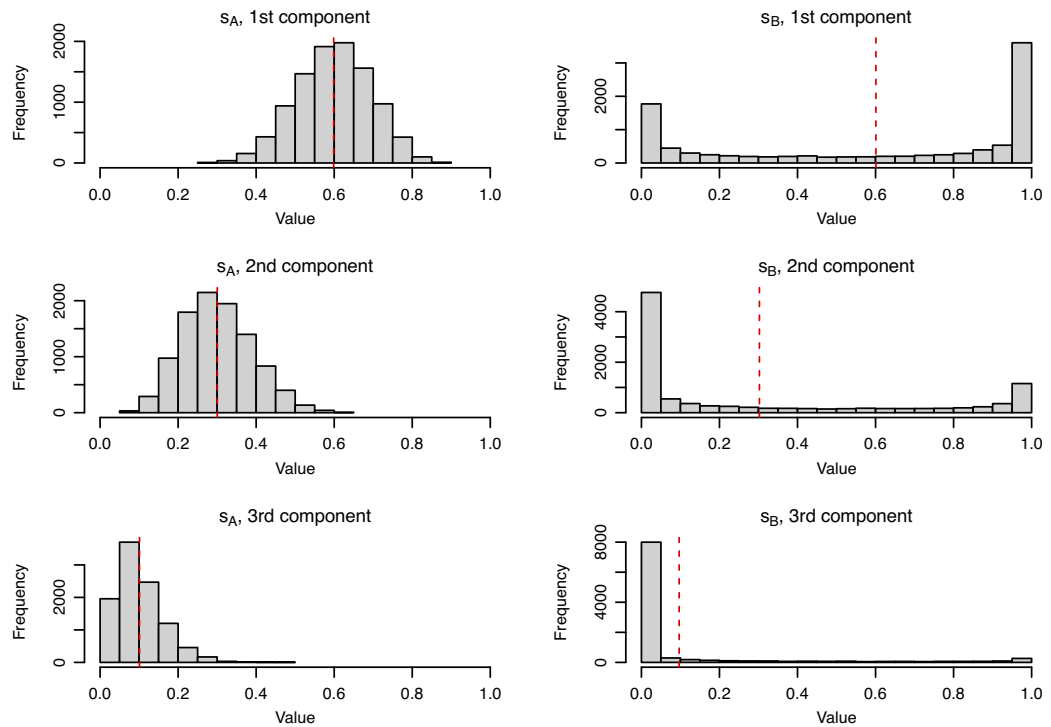
Figure 1. Marginal distributions of two Dirichlet distributions with same expected values for each component but different degrees of dispersion. Dashed red line shows mean value. $\mathbf{s_A} \sim Dir([15, 7.5, 2.5])$; $\mathbf{s_B} \sim Dir([0.3, 0.15, 0.05])$.

Parameterizing θ in terms of a product of a general concentration parameter α and a location vector **p** has attractive properties. It allows us to use past work which has formulated (deterministic) expectations regarding party seat (vote) shares **p**, while quantifying the dispersion around those expectations through the concentration parameter α, which can be estimated from real-world data.

We have described seat and vote share data as ordered data, but draws from Dirichlet distributions described by Equation (3) need not be ordered. Although Equation (1) gives us an ordered location vector (**p**), whether or not draws from this distribution will be ordered will depend on the concentration parameter α. If α is very large, draws from the distribution will more closely approximate the ordered location vector, and will in turn be more likely to be ordered. If α is small, as in our discussion above, values of all components will be more highly dispersed, and it becomes *less* likely that the resulting draws from a Dirichlet distribution $Dir(\alpha\mathbf{p})$ will be ordered.

It is possible to guarantee an ordered draw by using an ordered Dirichlet distribution (van Dorp and Mazzuchi 2004):

$$\theta^\star = \alpha\mathbf{p}^\star,$$
$$\mathbf{s} \sim \text{OrdDir}(\theta^\star),$$

(4)

where $\mathbf{p}^\star$ is an *increasing* ordered vector with length $N_0 + 1$, and values equal to the differences between successive values of $[0, \mathbf{p}, 1]$. If our value of **p** for the five-party case is [0.03, 0.08, 0.16, 0.28, 0.45], then our value of $\mathbf{p}^\star$ is [0.03, 0.05, 0.08, 0.12, 0.17, 0.55]. Phrased slightly differently, the ordered Dirichlet is the result of generating Dirichlet-distributed *differences* between party shares and taking the cumulative sum. The parameter α acts as a concentration parameter, and can be interpreted in the same way as in the standard Dirichlet distribution.

The ordered Dirichlet distribution respects the ordered property of the data, but poses practical problems. First, the ordered Dirichlet distribution requires shares to be *strictly*, not *weakly*

ordered. While vote shares in national elections are almost always strictly ordered, seat-winning parties sometimes win exactly the same number of seats. We deal with this by adding or subtracting negligible values from the seat shares of tied parties. Second, using the ordered Dirichlet means, we cannot (directly) use certain useful *analytic* properties of the standard Dirichlet distribution, such as the expression for the variance of each component s_i: $Var[s_i] = \frac{s_i(1-s_i)}{1+\alpha}$ (Aitchison 1986, 59). This may not be a problem if our sole focus is simulation. We note these problems now, and return to them later when we discuss the performance of our models.

4 Models

We fit Dirichlet and ordered Dirichlet distributions to data drawn from parliamentary elections around the world. We estimate four different models:

- The **null model**: $\mathbf{p}$ is given by the equation $\frac{1}{N_0}$, and α is estimated.[1]
- The **logical model**: $\mathbf{p}$ is given by Equation (1), and $\mathbf{p}^\star$ by taking differences, with α estimated.
- The **political model**: $\mathbf{p}$ is given by Equation (2), and $\mathbf{p}^\star$ by taking differences, with α estimated.
- The **saturated model**: $\mathbf{p}$ is estimated for each size of party system ($N_0 = 2, \dots, 20$); α is estimated.

Note that the *null model* is the only model which is not estimated using an ordered Dirichlet distribution. In the null model, all components have the same expected value, and so the *differences* between these components are equal to zero. Because Dirichlet-distributed values must be greater than zero, it is not possible to estimate an ordered Dirichlet version of the null model.

Our focus is understandably on the second and third models. The null and saturated models provide performance benchmarks, but it seems unlikely that the null model will ever capture the patterns in the data. Each model is estimated on vote- and seat-share data, for both the Dirichlet and ordered Dirichlet distributions, for a total of 14 models. We estimate these models in Stan (Stan Development Team 2022); Stan code is given in the Supplementary Material, together with further details on the generation of ordered Dirichlet deviates.

5 Evaluation metrics

We evaluate models using the following metrics:

- **Root mean squared error (RMSE)**: root mean squared error is calculated at the election level and then averaged across elections.
- **Calibration**: we calculate, for each election, the proportion of seat (vote) shares which were greater than or equal to the corresponding 5th percentile and less than or equal to the corresponding 95th percentile in the posterior distribution. We then average this across elections. Calibration ranges between 0% and 100%; values closer to 90% indicate a better model.
- **Proportional error on N_S (or N_V)**: we calculate for each simulation the effective number of simulated parties. We then subtract the actual effective number for each election. To draw meaningful comparisons across party systems with different effective numbers, we then divide this difference by the actual effective number. This quantity, expressed in percentages, ranges from −100 to +100. Values greater than zero indicate the effective number of parties was overestimated; values closer to 0 indicate a better model.
- **Proportional error on s_1 (or v_1)**: we take the share of the largest party in each simulation and subtract the actual share for each corresponding election. To enable comparison, we once again divide this difference by the share of the largest party. This quantity ranges from −100 to +100. Values greater than zero indicate the seat share of the largest party was overestimated; values closer to 0 indicate a better model.
- **Proportional error on s_2 (or v_2)**: as above, but for s_2 instead of s_1.

---

1 This is equivalent to a *symmetric* Dirichlet model.

We calculate these quantities because each taps an important aspect of party systems. RMSE is closest to an overall measure of fit to the data. Calibration is important because ours is a probabilistic model, and in order to improve on deterministic models like that proposed by Taagepera and Allik ([2006]), we need to show that the set of shares to which we assign 90% probability actually occurs 90% of the time. Proportional error on N_S is important because N_S is the key continuous property of party systems, and arguably more important than the discrete measure of party system size N_{S0}. Finally, proportional error on s_1 and s_2 is necessary to assess the claim that "political adjustments" are necessary to explain whether small parties lose a portion of the seat (or vote) share they would gain under a probabilistic model, and because the share of the largest party is arguably the *second* most important quantitative feature of a party system (Magyar [2022]).

6 Data

We estimate our models using data from ParlGov (Döring and Manow [2021]). ParlGov collects comprehensive information on electoral outcomes in a number of parliamentary and semi-presidential regimes. Information is recorded for all elections after 1945 or after full democratization, and for a limited number of countries from 1900. Parties are included if they won more than 1% of the vote or two seats or more. ParlGov covers 813 elections in 37 unique countries, far exceeding Mackie and Rose ([1997]). The raw number of seat- and vote-winning parties ranges from 2 to 20; the modal number of seat-winning parties is 5.

We use ParlGov data because its coverage of seat- and vote-shares in included elections is more complete than any other source we are aware of. ParlGov does, however, have certain limitations. Most notably, it lacks information on seat- and vote-shares in presidential regimes. It also does not cover elections in smaller parliamentary regimes located outside of Europe, such as the Westminster-model democracies in the Caribbean. We claim, however, that it would be unlikely, *when conditioning on the number of seat- or vote-winning parties,* for these systems to have very different expected seat- or vote-shares **p** (Shugart and Taagepera [2017], 187–92), or to alter substantially our parameter estimates for concentration α.[2]

7 Results

Table [1] shows evaluation metrics for models of seat shares. The null model performs poorly, with a large RMSE and an effective number of parties that is 13% too high (i.e., the model predicts more fragmentation than there really is). The logical models provide much better fit, as measured by the RMSE, and calibration that is close to nominal. The (unordered) logical model does give values of N_S which are roughly 9% too high. However, this is not due to systematically underestimating the share of the two largest components: our average estimates of s_1 and s_2 are close to zero, and the 90% credible interval encompasses zero. The political models, which might address the issue of over-estimating N_S, provide a worse fit to the data, as measured by RMSE. The fit of the logical models is impressive, with RMSE within 7% of the value for the saturated model. When comparing between logical models, the ordered Dirichlet ends up giving a less realistic picture of the effective number of parties, and has a worse fit to the data as measured by RMSE. Given the greater ease of use of the unordered Dirichlet distribution, the ordered Dirichlet does not repay its greater complexity.

Table [2] presents the same metrics for vote share. As before, the political models are worse than the logical models, and the logical models are worse than the saturated model only by a small

---

2 Section [A6] of the Supplementary Material tests this claim by modeling vote shares from 304 legislative elections in 20 presidential democracies in the Americas. We find slightly smaller estimates for α than those reported below, which results in only a minor increase in the dispersion of the simulated shares. We note that whenever researchers expect substantial heterogeneity in α, they can use the statistical models we implemented for estimating α across any subsets of elections, provided that data on seat or vote shares are available.

Table 1. Evaluation metrics for models of seat shares. RMSE measured in percentage points. Errors on N_S, s_1, s_2 are expressed in percentages of the true values [-100, +100]. Figures in square brackets are 90% credible intervals. Figures from best-performing model on each criterion (excluded the saturated model) are in bold.

Model		α	RMSE	Calibration	Error N_S	Error s_1	Error s_2
Null	Dirichlet	5.87	18.9	**89.8**	12.8	−0.151	0.108
		[5.71, 6.05]	[18.6, 19.2]		[11.0, 14.7]	[−0.541, 0.257]	[−0.317, 0.538]
Logical	Dirichlet	39.1	**7.75**	87.4	8.95	0.0242	**−0.0378**
		[38.0, 40.4]	[7.60, 7.90]		[8.03, 9.86]	[−0.141, 0.182]	[−0.198, 0.124]
	Ordered	18.5	9.30	94.8	16.9	0.0252	−0.0391
		[18.1, 19.0]	[9.11, 9.49]		[13.7, 20.2]	[−0.219, 0.263]	[−0.296, 0.217]
Political	Dirichlet	37.2	8.06	79.5	−6.03	**−0.0107**	−0.0505
		[36.2, 38.3]	[7.90, 8.22]		[−6.75, −5.31]	[−0.164, 0.146]	[−0.217, 0.112]
	Ordered	16.2	9.75	90.6	**2.16**	−0.0114	−0.0517
		[15.8, 16.6]	[9.55, 9.95]		[−0.737, 5.05]	[−0.260, 0.237]	[−0.303, 0.205]
Saturated	Dirichlet	41.9	7.24	90.6	5.00	0.0364	0.0100
		[40.6, 43.2]	[7.10, 7.39]		[3.81, 6.17]	[−0.116, 0.190]	[−0.150, 0.169]
	Ordered	16.5	12.1	87.5	−7.50	1.14	0.732
		[15.9, 17.0]	[11.7, 12.4]		[−11.8, −3.22]	[0.837, 1.45]	[0.416, 1.05]

Table 2. Evaluation metrics for models of vote shares. RMSE measured in percentage points. Errors on N_V, v_1, and v_2 are expressed in percentages of the true values [−100, +100]. Figures in square brackets are 90% credible intervals. Figures from best-performing model on each criterion (excluded the saturated model) are in bold.

Model		α	RMSE	Calibration	Error N_V	Error v_1	Error v_2
Null	Dirichlet	6.32	17.4	**90.2**	8.66	−0.0941	0.0211
		[6.16, 6.48]	[17.1, 17.7]		[6.94, 10.4]	[−0.493, 0.321]	[−0.388, 0.440]
Logical	Dirichlet	50.0	**6.34**	88.2	**4.12**	0.00891	−0.0227
		[48.5, 51.4]	[6.22, 6.46]		[3.35, 4.88]	[−0.129, 0.154]	[−0.165, 0.121]
	Ordered	21.8	7.93	95.6	9.98	**0.00644**	**−0.0203**
		[21.3, 22.4]	[7.77, 8.08]		[7.31, 12.9]	[−0.208, 0.233]	[−0.244, 0.212]
Political	Dirichlet	45.1	6.83	79.1	−11.1	−0.0240	−0.110
		[43.9, 46.3]	[6.70, 6.95]		[−11.7, −10.5]	[−0.169, 0.122]	[−0.255, 0.0362]
	Ordered	18.8	8.49	90.3	−4.66	−0.0245	−0.114
		[18.3, 19.2]	[8.32, 8.67]		[−7.06, −2.06]	[−0.254, 0.214]	[−0.356, 0.128]
Saturated	Dirichlet	59.7	5.81	90.4	3.05	0.0340	0.00369
		[57.8, 61.6]	[5.70, 5.92]		[2.04, 4.04]	[−0.0975, 0.163]	[−0.132, 0.141]
	Ordered	20.8	10.1	89.5	−9.88	0.901	0.716
		[20.2, 21.5]	[9.83, 10.5]		[−13.6, −5.74]	[0.625, 1.17]	[0.443, 0.992]

amount. Once again, the fit of the ordered Dirichlet models is inferior to the unordered models. All models save the political models over-estimate the effective number of parties, even the saturated models. Indeed, N_V is very badly under-estimated in the saturated ordered model.

8 Conclusion

Our results show that realistic looking party systems of a given size can be simulated using a (standard, unordered) Dirichlet distribution where mean seat or vote shares are given using Equation (1), and where the concentration parameter is roughly 40 (for seat shares) or 50 (for

vote shares). We can achieve similar results using an ordered Dirichlet distribution, but the ordered Dirichlet generally provides a worse fit to real-world data, and we know that the ordered Dirichlet is harder to work with than the standard Dirichlet. For these reasons, we recommend that researchers who are interested in simulating party systems use a standard Dirichlet distribution. Tools to simulate party systems of different sizes can be found in an accompanying R package sharesimulatoR and in an interactive web page.[3]

The ability to simulate party systems allows researchers to answer practical questions (provided, of course, that they know, or have expectations regarding the number of seat- or vote-winning parties). To return to the questions asked in the Introduction: a consultant who knows that the most likely number of seat-winning parties under a proposed system is 5 can use our work to show, through simulation, that the probability of a single-party majority is roughly 1 in 4. Researchers interested in coalition formation can use our work to evaluate the probability that a party system with five, seven, or nine seat-winning parties is an "open" system (per Laver and Benoit (2015), one where even the top-two parties do not have a majority). Because the number of seat- and vote-winning parties is strongly determined by the "seat product," researchers evaluating proposed electoral systems can simulate likely distributions of seat and vote shares given predicted numbers of seat- and vote-winning parties (Shugart and Taagepera 2017, 149). In our view, the questions which we can now answer with this method of simulation ("what is the probability that a single party will have a majority?" and "what is the probability that no two parties will have a majority") are simple questions which are fundamental to the operation of a party system, and which could not have been satisfactorily answered without the simulation methods given here.

Conflict of Interest
The authors are not aware of any conflicts of interest.

Data Availability Statement
Replication code for this article is available in Cohen and Hanretty (2023) at https://doi.org/10.7910/DVN/3WILXI.

Supplementary Material
For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2023.13.

References
Aitchison, J. 1986. *The Statistical Analysis of Compositional Data*. London–New York: Chapman & Hall.
Cohen, D., and C. Hanretty. 2023. "Replication Data for: Simulating Party Shares." Harvard Dataverse, V1. https://doi.org/10.7910/DVN/3WILXI
Döring, H., and P. Manow. 2021. "Parliaments and Governments Database (Parlgov): Information on Parties, Elections and Cabinets in Modern Democracies." Development Version.
Golder, M., S. N. Golder, and D. A. Siegel. 2012. "Modeling the Institutional Foundation of Parliamentary Government Formation." *Journal of Politics* 74 (2): 427–445.

---

3 R package is available at https://github.com/chrishanretty/sharesimulatoR. Interactive web page is available at https://chanret.shinyapps.io/shareSimulatoR/.

Laver, M., and K. Benoit. 2015. "The Basic Arithmetic of Legislative Decisions." *American Journal of Political Science* 59 (2): 275–291.

Mackie, T. T., and R. Rose. 1997. *A Decade of Election Results: Updating the International Almanac*. Glasgow: Centre for the Study of Public Policy, University of Strathclyde.

Magyar, Z. B. 2022. "What Makes Party Systems Different? A Principal Component Analysis of 17 Advanced Democracies 1970–2013." *Political Analysis* 30 (2): 250–268.

Shugart, M. S., and R. Taagepera. 2017. *Votes from Seats: Logical Models of Electoral Systems*. Cambridge: Cambridge University Press.

Stan Development Team. 2022. "Stan Modeling Language Users Guide and Reference Manual, V2.29."

Taagepera, R., and M. Allik. 2006. "Seat Share Distribution of Parties: Models and Empirical Patterns." *Electoral Studies* 25 (4): 696–713.

van Dorp, J. R., and T. A. Mazzuchi. 2004. "Parameter Specification of the Beta Distribution and Its Dirichlet Extensions Utilizing Quantiles." In *Handbook of Beta Distribution and Its Applications*, Statistics Textbooks and Monographs, 174, 283–318. New York: Marcel Dekker.