Electoral districts as samples from the national population[*]

Chris Hanretty
Royal Holloway, University of London
chris.hanretty@rhul.ac.uk

ABSTRACT The relationship between national and district-level electoral is a foundational topic in the study of elections. I argue that selected district-level electoral outcomes can be modelled as though they were the outcome of a process involving random sampling without replacement. For seat-level outcomes, district magnitude is analogous to sample size. For vote-level outcomes, I argue that the role of sample size is played by a pseudo-magnitude, or district magnitude plus a constant. My theory gives point predictions and variances for the effective and 'raw' numbers of seat-winning parties, and the effective number of vote-winning parties. I estimate linear and nonlinear Bayesian regressions to test this theory using CLEA data. The principal coefficients in these regressions are within one to three percent of their predicted values; predictions from this theory are between 2% and 45% more accurate than predictions from the leading account of district-level electoral outcomes.

1 Introduction

Almost all democratic electoral systems divide their territory into electoral districts. Systems which don't do this—like Israel and the Netherlands—are as well-known as they are rare. The ubiquity of electoral districts makes it important to understand the relationship between national-level outcomes and district-level outcomes. This relationship can be understood in two ways: "bottom-up" and "top-down" (Barceló and Muraoka 2018, 44). For those who study the relationship "bottom up", party politics starts with proto-parties forming in each district, developing their appeal locally, and then engaging in a process of aggregation (Chhibber and Kollman 1998) or linkage (Moenius and Kasuya 2004), forming (or failing to form) national parties

[*] I would like to thank Jack Bailey for giving me his code to clean CLEA data, and for continuing discussions about the logical quantitative models. I would also like to acknowledge the use of two different large language models, ChatGPT v4.1 and Google Gemini, for help with Taylor expansions, results which I have verified myself.

from local slates (Cox 1999; Ferree, Powell, and Scheiner 2014). Bottom-up approaches leave ample room for small-p politics and contingency: proto-parties might fail to bridge cultural or linguistic divides, or struggle with registration requirements (Birnir 2004). Yet there is also room for simple maxims which locate constraints at the district level, like Cox's "M+1" rule (Cox 1997, 99).

Others view the relationship between national and district outcomes "top-down". For these scholars, "all politics is national" (Shugart and Taagepera 2017, ch. 10), and what happens in any given district is severely constrained not just by the district magnitude but by the carrying capacity of the system as a whole. Famously, Shugart and Taagepera (2017) claim that national *and* district-level outcomes can be modelled as a function of the *seat product*, or the average (mean) district magnitude times the assembly size. There is less room for politics in top-down approaches: key institutional features are taken to be so constraining that culture, language and ideology all play a residual role (Shugart and Taagepera 2017, 118–22).

There is no reason in principle why scientific progress can't be made both bottom-up and top-down. To use a metaphor favoured by Derek Parfitt, we can climb the mountain from different sides. But progress must be made, because much rides on our understanding. The relationship between national and district outcomes is related to the concept of *party system nationalisation* ("the territorial homogeneity of support of all parties included in the [party] system" (Bochsler 2010, 155)) and the more specific concept of *party system inflation*, or the degree to which the (effective) number of parties nationally is greater than the (effective) number of parties at some more local level (Chhibber and Kollman 1998; Cox 1999; Moenius and Kasuya 2004). These concepts in turn matter for public policy outcomes like the provision of healthcare (Hicken, Kollman, and Simmons 2016) and the composition of public spending (Lago-Peñas and Lago-Peñas 2009; Crisp, Olivella, and Potter 2013). If we don't understand the relationship between national and district outcomes, we can't tailor electoral systems to produce specific levels of party system nationalisation. If we can't achieve specific levels of party system nationalisation, we may be forgoing improvements in the quality of public service provision.

The goal of this article is to set out a particular "top-down" theory of district-level outcomes. This theory takes national level outcomes as given, and predicts effective numbers of seat and vote-winning parties given information on district magnitude and assembly size. The theory is based on the intellectual conceit that district-level outcomes can be viewed as "samples" from a national "population". Once we adopt this perspective, we can use a range of tools designed to analyse the behaviour of samples drawn without replacement. I show that this theory provides a better fit to the data than Shugart and Taagepera's embeddedness coefficient, the leading top-down account of district outcomes.

This theory has four chief theoretical virtues. First, the theory has limited substantive content, and relies on the barest of details surrounding weakly proportional methods of seat allocation in simple electoral systems. This means that the theory does not rely on assumptions about mass or elite behaviour or rationality. Second, the theory is parsimonious: the model of the district-level effective number of seat-winning parties has no free parameters; the model of the effective number of vote-winning parties just one. Third, the theory is cumulative, building on the work of Shugart and Taagepera (2017) at national level: for although the model takes national level outcomes as given, these national level outcomes can be replaced with predictions from the Seat Product Model. Fourth, the theory generates not just expected values of district outcomes *but also* variances of those outcomes, taking dispersion seriously (Braumoeller 2006).

In the next section, I narrow down the outcomes that I am interested in, and give an overview of the Seat Product Model (SPM, Shugart and Taagepera (2017)). That model, and the accompanying notation, provide the framework for a theory section that proceeds outcome-by-outcome, starting with the expected effective number of seat-winning parties and the *variance* of that quantity, before moving on to the raw number of seat-winning parties and the effective number of *vote-winning* parties. I describe how I build upon CLEA data (Kollman et al. 2024), before testing the predictions of my theory in a series of heteroskedastic linear and nonlinear Bayesian regressions. I conclude by considering the limitations of such a substantively minimal approach.

2 Background

National and district-level electoral outcomes are of many different kinds, and not all of these fall within my scope. Here I am only interested in — and can only provide an account of — the relationship between national and district-level outcomes considered at a single point in time. That is, I do not consider the *dynamics* of party (system) nationalization (Morgenstern, Swindle, and Castagnola 2009), and the outcomes I am interested in are *levels*, rather than *changes*. Even restricting our scope to research on the *static* relationship between national and district-level outcomes we find different electoral outcomes have been studied. A minority of researchers believe that the most important (district-level) outcome is whether a party is present or absent in a district, and that talk of the relationship between national and district level outcomes is really an invitation to explore patterns of contestation and candidacy (Lago and Montero 2014). For most researchers — and all researchers who use the concept of party system inflation — the key outcomes are effective numbers (Laakso and Taagepera 1979). Here I shall use the notation introduced by Shugart and Taagepera (2017) to refer to these effective numbers at national and district level. The term N_V gives the effective number of vote-winning parties at national level. The same term with a prime mark attached, N_V', gives the corresponding quantity at the district level. Similarly, N_S and N_S' give the national and district-level number of seat-winning parties. The (effective) number of seat-winning parties is constrained by the district magnitude m.

Those who write on party system inflation use effective numbers to operationalize that concept. For Chhibber and Kollman (1998), the difference between the national and average district-level effective numbers is a measure of party system inflation or alternately of "problems of party aggregation". Cox (1999) suggested dividing this difference by the national effective number; Moenius and Kasuya (2004) suggested (in my view correctly) that it makes more sense to divide by a weighted average of the district-level effective numbers. The use of effective numbers to calculate measures of party system inflation is therefore standard, and there seems nothing objectionable to their use to aid description.

The problem is that measures of party system inflation, although they may be suited to description, are chronically unsuited to inferential analysis. More than a century ago Karl Pearson warned of the risk of finding spurious correlations when analysing variables involving a ratio (Pearson 1897). More recently we have learned that the analysis of outcome variables involving differences does not permit causal inference except under specific circumstances (Tennant et al. 2022). Party system inflation scores, because they are based on a difference divided by one of the two constituent terms, are vulnerable to both these criticisms. They are especially vulnerable when authors fail to control for the national level effective number. Consequently, the conclusions from many published regression equations analysing the "determinants of party system inflation" (Simón 2013; Golosov 2015; Simón and Guinjoan 2018) are in my view unsafe.

By criticising multiple regression models estimated on observational data I do not mean to pave the way for some new design that does permit causal identification of the links between national and district-level outcomes. My claim is more radical: I argue that we *cannot* achieve causal identification when modelling the relationship between national and district level outcomes, because we are not dealing with individually manipulable terms, but with a part-whole relationship. Because all votes must be cast in some district, it is impossible to manipulate an effective number at national level without also manipulating effective numbers in one or more districts. Similarly, it is impossible to alter an effective number at district level without affecting the corresponding effective number at national level, even though the effect may be negligible. Because these quantities are not separately manipulable, we must seek a non-causal explanation of the link between national and district-level outcomes (Woodward 2018).

Non-causal explanation is a characteristic feature of the logical quantitative models pioneered by Rein Taagepera (Taagepera 2008). A logical quantitative model makes numeric predictions based partly or wholly on what values are logically possible, given the values of other variables in the model. This is often coupled with invocation of the principle of insufficient reason, according to which our best guess given a lower and an upper bound is an arithmetic or geometric mean of these bounds (Taagepera 2008, ch. 9). In electoral studies, paying attention to these logical bounds has been unreasonably successful, and so whilst a procedure of "guessing somewhere between the lowest and highest possible numbers" might not have much to recommend it *a priori*, it does come recommended by its empirical track record.

The most famous logical quantitative model is the Seat Product Model (SPM) (Shugart and Taagepera 2017), which gives expectations for both national and district level outcomes, and which respects the relationship between part and whole. It is a "top-down" account, in that it first derives expectations at the national level, before working out the district outcomes that are consistent with these national-level figures. In the SPM, a special role is played by the "raw" number of seat-winning parties nationally, N_{S0}. This quantity is equal to the seat product, or mean district magnitude times assembly size, raised to the power of one quarter:

$$N_{S0} \approx (MS)^{\frac{1}{4}}$$

The effective number of seat-winning parties nationally is equal to the raw number, raised to the power of two-thirds (Shugart and Taagepera 2017, 105–8):

$$N_S \approx N_{S0}^{\frac{2}{3}} = (MS)^{\frac{1}{6}}$$

The number of "pertinent" vote-winning parties, N_{V0}, is equal to the number of seat-winning parties plus one. The effective number of vote-winning parties N_V is equal to the number of pertinent vote-winning parties raised to the power of two-thirds, mirroring the relationship between N_S and N_{S0}.

These equations govern national level outcomes, but the SPM generates district level predictions. These are created by use of an "embeddedness coefficient" k, which reflects not just the magnitude of each district m, but also its magnitude in relation to assembly size S. Thus, $N'_S = m^{2k/3}$, $N'_V = \left[N_S'^{3k/4} + 1\right]^{4k/3}$.

If the Seat Product Model can predict effective numbers at both national and district level, and thereby predict levels of party system inflation, where is the need for a new "top-down" approach? One answer to this question lies in the embeddedness coefficient k, which the authors themselves describe as "baffling" (p. 161), which requires "strenuous" reasoning to explain (p. 174), and which is based ultimately on empirical fit (ibid.). The full formula is

$$k = 0.5 + 0.2067 \log \frac{S}{m} \cdot m^{1/4}$$

Parts of this equation follow on from the rest of the Seat Product Model. The initial "0.5" ensures that when there is a single district, the district-level predictions for the single district match the national-level predictions for that same single district. The term $\log \frac{S}{M}$ means that the embeddedness coefficient collapses to just 0.5 when there is a single national district, since $M = S$, and $\log \frac{S}{S} = \log 1 = 0$. However, the two remaining terms (the coefficient 0.2067 and the exponent on M of one quarter) are empirically determined. Whilst the specific values of these terms might work for the elections studied by Shugart and Taagepera (2017), they may not work for other elections. If we can "do more with less" – that is, if we can improve our predictions of district-level outcomes whilst reducing our reliance on empirically determined parameters – then that would be a meaningful improvement on the state of the art.

3 Theory

In this section, I argue that district level outcomes are related to national level outcomes in the same way that samples and sample statistics are related to populations and population statistics. The details of the relationship depend on whether the outcome being modelled concerns votes or seats. For quantities concerning seats, I argue that district magnitude is exactly analogous to sample size. Thus, the effective number of seat-winning parties in a district is equal to what we would expect if we took a sample of m legislators (with party affiliations) from a population of S legislators, drawing randomly without replacement. Because the diversity of a sample cannot consistently be greater than the population from which it is drawn, this district-level expectation is lower than the national level effective number; *how much lower* will depend on the district magnitude. For quantities concerning votes, I argue that we can construct a *pseudo-magnitude*: a quantity greater than the district magnitude, and which acts analogously to sample size. In particular, I'll argue that the pseudo-magnitude m^* is an additive function of the actual district magnitude. This means that rather than thinking of votes as individual counts, we should think of them in batches, with each district of magnitude m having a certain number of slots ($m^* \gg m$) for those batches.

I analogise districts to samples because (random) sampling provides the most well developed intellectual toolset for thinking about the part-whole relationship. "Random sampling" is a very specific type of part-whole relationship, but it is also the type of sampling that requires the fewest additional assumptions about the nature of the part-whole relationship. If the sampling relationship were stratified or involved some kind of rejection stage, we would have to specify the details of that stratification or rejection process. Thus, although the formulae I offer give specific quantitative predictions, they are consequences of a natural idea which proceeds from ignorance: "what if the allocation of legislators (or batches of votes) to districts was random and shorn of any structure".

Thinking of districts as samples is an analogy, and analogies are never perfect. Here are three points of disagreement. First, "district magnitudes" are almost always small relative to "sample sizes" in social science research. We almost never try to learn from a sample of size one, but there are many examples of districts with magnitude one. Of course, in social sciences the object is to infer properites of the population given properties of the sample, whereas here we are performing inference "in reverse". Second, districts are large compared to the populations of which they are part. The average district in my data makes up 1% of its legislature. Compare that with the average sample size for opinion polls (1,000) compared to the median country population of almost six million. Third, districts are also almost always defined by geography, whereas most social scientific samples try to avoid geographic patterning. The third difference is the most significant, and one which I ignore. Although geography surely matters for *which* parties are most competitive in each seat, it may not matter for measures of diversity or concentration like the effective number. I return to this point in the conclusion.

These are the *theoretical* reasons for thinking of districts as though they were samples. The practical pay-off comes from the equations that I introduce in the following subsections. These equations give expressions for:

- the expected value and variance of the district-level effective number of seat-winning parties (N'_S)
- the number of distinct parties that win seats in a district (N'_{S0})
- the expected value and variance of the district-level effective number of vote-winning parties (N'_V)

When describing any theory – even a substantively minimal theory like this one – it is important to describe the scope conditions which restrict the theory's application. My theory applies only to "simple" electoral systems, or systems in which a voter expresses a single nominal preference between parties, and where the allocation of seats within an electoral district depends entirely on votes cast in that district (Shugart and Taagepera 2017, 31–34). This excludes systems with multiple tiers, whether these are visible to voters (mixed-member sytems) or largely invisible (remainder pooling systems). It also excludes preferential voting systems, although these tend to behave as though they were simple. The theory also applies only to systems that are "weakly proportional" in the sense of Theil (1969), or "nonmajoritarian" (Shugart and Taagepera 2017). Finally, this excludes block vote systems, but includes systems with single member districts, since these can be seen as limiting cases of closed party list proportional representation. In all of these systems, draws of seats are not independent either within district (block vote) or across districts. Again, I return to this point in the conclusion.

3.1 The expected value of N'_S

Our goal in this subsection is to find the expected value of the district-level effective number of seat-winning parties, N'_S under random sampling without replacement. As a matter of definition, the effective number is defined as the reciprocal of the sum of squared shares. District level seat shares are defined as the number of seats won by a party divided by the district magnitude. I use lower-case n_i to refer to the seats won by the ith party, and lower-case m to refer to district magnitude. The effective number of seat-winning parties at district level can therefore be written as:

$$N'_S = \frac{1}{\sum_i \left(\frac{n_i}{m}\right)^2} \tag{1}$$

It will often be easier to work with the inverse of the effective number. In ecology, this is known as the Simpson index (Simpson 1949); in economics, as the Herfindahl-Hirschmann index. Following notational conventions in ecology, I shall use λ to refer to this inverse. I'll use the prime mark to indicate when this is a district-level quantity, and subscripts to indicate whether this quantity is calculated using vote or seat shares. Inverting Equation 1 and simplifying therefore gives us

$$\lambda'_S = \frac{1}{m^2} \sum_i n_i^2 \tag{2}$$

where district magnitude m is a constant and where the number of seats won by any party n_i is a random variable.

At this point I introduce an indicator variable which has a value of one if a specified party wins a specified seat

$$I_{i,t} = \begin{cases} 1 & \text{if party } i \text{ wins seat } t \\ 0 & \text{in all other cases} \end{cases}$$

I rewrite the sum of squared seat tallies using this indicator variable:

$$\sum_i n_i^2 = \sum_i \left(\sum_{t=1}^{m} I_{i,t} \right)^2 \tag{3}$$

where the summation inside the brackets sums over the different seats from one to m. Rather than asking directly, "how many seats did party i win", we ask of each seat in turn "did party i win this seat?"

We can expand $\left(\sum a_i\right)^2$ to $\sum a_i^2 + \sum_{i \neq j} a_i a_j$ or as $\sum a_i^2 + 2\sum_{i<j} a_i a_j$. I'll use this second form. Although this second form imposes some ordering on the seats, this ordering is only there for reasons of book-keeping. I am not making any substantive assumptions about the order of election of different parties. Equation 3 therefore becomes

$$\sum n_i^2 = \sum \left(\sum_{t=1}^{m} I_{t,i}^2 + 2 \sum_{1 \leq t < u \leq m} I_{t,i} I_{u,i} \right)$$

for arbitrary seat indices t and u. Because $I_{t,i}$ is an indicator variable, and because multiplying an indicator variable by itself returns the original value, this equation simplifies to

$$\sum n_i^2 = \sum_i \left(\sum_{t=1}^{m} I_{t,i} + 2 \sum_{1 \leq t < u \leq m} I_{t,i} I_{u,i} \right)$$

At this point we separate out the summation and make use of the fact that the sum of the indicator variables $I_{t,i}$, summing over all parties and all seats in the district, must just equal the district magnitude. We are left with

$$\sum n_i^2 = m + \sum_{i=1} 2 \sum_{1 \leq t < u \leq m} I_{t,i} I_{u,i} \tag{4}$$

Here I introduce a second indicator variable, J_{tu}. This variable has a value of one if the same party won seats t and u. This indicator variable is defined formally as being the sum of several different products of indicator variables:

$$J_{tu} = \sum_i I_{i,t} I_{i,u}$$

This second indicator allows me to rearrange and simplify Equation 4 as follows:

$$\begin{aligned}
\sum n_i^2 &= m + \sum_{i=1} 2 \sum_{1 \leq t < u \leq m} I_{t,i} I_{u,i} \\
&= m + 2 \sum_{i=1} \sum_{1 \leq t < u \leq m} I_{t,i} I_{u,i} && \text{(bring multiplicative factor to the front)} \\
&= m + 2 \sum_{1 \leq t < u \leq m} \sum_{i=1} I_{t,i} I_{u,i} && \text{(change order of summation)} \\
&= m + 2 \sum_{1 \leq t < u \leq m} J_{t,u} && \text{(from definition of } J_{t,u}\text{)}
\end{aligned} \tag{5}$$

We can now insert Equation 5 into Equation 2 to give a expression for the inverse of the effective number of seat-winning parties:

$$
\begin{aligned}
\lambda'_S &= \sum_i \left(\frac{n_i}{m}\right)^2 \\
&= \frac{1}{m^2}\sum_i n_i^2 \\
&= \frac{1}{m^2}\left(m + 2\sum_{1\le t<u\le m} J_{tu}\right) \\
&= \frac{1}{m} + \frac{2}{m^2}\sum_{1\le t<u\le m} J_{tu}
\end{aligned}
\tag{6}
$$

This new expression only helps us insofar as we can work with the probability of the same party winning seats t and u, $\mathbb{E}\left[J_{tu}\right]$. Here we must once again rely on the national seat shares $\mathbf{p}$. For any given party, the probability of it winning two seats under simple random sampling without replacement is equal to the probability of it winning one seat, times the probability of it winning another seat, or p_i^2. Since we wish to calculate the probability of *any party* winning both seats t and u, we sum probabilities across parties. Because it involves a summation of these probabilities (squared shares) across parties, the expectation of J_{tu} is therefore just the inverse of the national effective number of seat-winning parties (N_S), which I'll denote as λ_S:

$$
\mathbb{E}\left[J_{tu}\right] = \lambda_S \tag{7}
$$

This means that each J_{tu} can be replaced by λ_S, but we must still work out the number of ordered pairs of seats t and u there are. The total number of pairs of seats is district magnitude times district magnitude minus one; the total number of distinct pairs (i.e., ignoring order) is half this. We must therefore add $\frac{m(m-1)}{2}$ copies of λ_S. We can therefore rearrange as follows:

$$
\begin{aligned}
\lambda'_S &= \frac{1}{m} + \frac{2}{m^2}\sum_{1\le t<u\le m} J_{tu} \\
&= \frac{1}{m} + \frac{2}{m^2}\frac{m(m-1)}{2}\lambda_S && \text{From definition of } J_{tu} \text{ and the no. of distinct pairs} \\
&= \frac{1}{m} + \frac{m(m-1)}{m^2}\lambda_S && \text{cancelling the factor of two} \\
&= \frac{1}{m} + \frac{(m-1)\lambda_S}{m} && \text{cancelling the factor of } m \\
&= \frac{1+\lambda_S(m-1)}{m} && \text{combining fractions} \\
&= \frac{\lambda_S m}{m} + \frac{1-\lambda_S}{m} && \text{multiplying out; splitting fractions} \\
\lambda'_S &= \lambda_S + \frac{1-\lambda_S}{m}
\end{aligned}
\tag{8}
$$

Equation 8 ignores the fact that we are drawing from a population *without replacement*. Our expression therefore has to be modified to include a finite population correction. Here, the finite population correction is equal to the assembly size minus the district magnitude, divided by the assembly size minus one. When assembly size S is very large and m very small, this finite population correction approaches one. When there is one district ($S = m$), the finite population correction will equal zero, and Equation 8 will collapse to give $\lambda'_S = \lambda_S$.

We therefore rewrite including the finite population correction:

$$\lambda'_S = \lambda_S + \frac{S-m}{S-1} \cdot \frac{1-\lambda_S}{m} \tag{9}$$

Equation 9 gives us an expression for the expectation $\mathbb{E}[\lambda'_S]$. From Jensen's inequality, we know that $\mathbb{E}[\frac{1}{x}] \geq \frac{1}{\mathbb{E}[x]}$. The effective number of seat-winning parties will therefore be slightly greater than one divided by the expected value of λ'_S. However, we can use the inverse of Equation 9 to give a first-order approximation of N'_S. In practice, the degree of approximation error is negligible. After inverting and re-arranging, we are left with the equation

$$N'_S \approx \frac{mN_S(S-1)}{mS + SN_S - S - mN_S} \tag{10}$$

This completes the derivation of the expectation of the district level effective number of seat-winning parties.

3.2 The variance of N'_S

We now wish to express the variance of N'_S. Our approach will be similar to the approach taken to calculate the expectation: we will calculate the variance of $\lambda'_S = \frac{1}{N'_S}$, which we will use to approximate the variance of N'_S. We will also proceed assuming sampling with replacement, and only latterly introduce a finite population correction.

Given the expression for λ'_S given in Equation 6, we can write the variance of λ'_S as

$$\begin{aligned}
\text{Var}\,(\lambda'_S) &= \text{Var}\,(\frac{1}{m} + \frac{2}{m^2}\sum_{1\leq t<u\leq m} J_{tu}) \\
&= \text{Var}\,(\frac{2}{m^2}\sum_{1\leq t<u\leq m} J_{tu}) && \text{(since Var } X + a = \text{Var } X \text{ for constant } a) \\
&= \frac{4}{m^4}\,\text{Var}\,(\sum_{1\leq t<u\leq m} J_{tu}) && \text{(since Var } Xa = a^2 \text{ Var } X \text{ for constant } a)
\end{aligned} \tag{11}$$

Calculating the variance of the inverse of the district-level effective number of seat-winning parties therefore turns out to involve calculating the variance of an indicator variable which tracks whether the same party won two specified seats.

Just as before, we can expand $\text{Var}\,(\sum_i a_i)$ to $\sum_{i=1} \text{Var}(a_i) + \sum_{i\neq j} \text{Cov}(a_i, a_j)$ or $\sum_{i=1} \text{Var}(a_i) + 2\sum_{i<j} \text{Cov}(a_i, a_j)$. I use this second form.

$$\text{Var}\,(\sum_{1\leq t<u\leq m} J_{tu}) = \sum_{1\leq t<u\leq m} \text{Var}\, J_{t,u} + 2 \sum_{\substack{1\leq t<s\leq m\\ 1\leq u<v\leq m\\ (u,v)>(t,s)}} \text{Cov}\,(J_{t,u}, J_{v,w}) \quad (12)$$

The second summation runs over pairs of seats – for example, seats one and two, and seats one and three. The notation $(u, v) > (t, s)$ indicates that each pair-of-pairs is counted once.

Consider first Var(J_{tu}). We know that J_{tu} is an binary (indicator) variable, and that therefore its variance is just its expectation, times one minus its expectation. From Equation 7, we know that the expectation of J_{tu} is λ_S. We therefore have

$$\text{Var}\,(J_{tu}) = \lambda_S(1 - \lambda_S)$$

We also know that there are $\frac{m(m-1)}{2}$ distinct pairs of seats, and so the first sum in Equation 12 is equal to

$$\frac{m(m-1)}{2}\lambda_S(1 - \lambda_S) \quad (13)$$

Now consider the covariances between the indicator variables for seats t, u, v, w, arranged in distinct pairs. We must consider two cases:

- **case one:** the two pairs of seats have no index in common. For example, $t = 1, u = 2$, but $v = 3, w = 4$. In these cases the covariance is zero.
- **case two:** the two pairs of seats have an index in common. For example, $t = 1, u = 2$, but $v = 2, w = 3$. In these cases, we must calculate the covariance.

By the definition of covariance,

$$\text{Cov}\,(J_{tu}, J_{uv}) = \mathbb{E}\,(J_{tu} \cdot J_{uv}) - \mathbb{E}\,(J_{tu}) \cdot \mathbb{E}\,(J_{uv})$$

To form an intuition about $\mathbb{E}\,(J_{tu} \cdot J_{uv})$, recall that J_{tu} is just an indicator that has the value of one if a party wins seat t and seat u. Because there is an overlap between these two indicators arising from the common index u, $J_{tu}J_{uv}$ will be equal to one when the same party wins all three seats. When we were considering the chances of a single party winning two seats, we multiplied its shares. Similarly here the chances of a single party i sweeping three seats are p_i^3. The chances of *any party* sweeping three seats are $\sum_i p_i^3$. I'll denote this quantity as R. R can be calculated from the vector of shares **p** or approximated using N_S.

The remaining term $\mathbb{E}\,(J_{tu}) \cdot \mathbb{E}\,(J_{uv})$ is simpler, because we know that the expectation is just equal to λ_S. This means that

$$\text{Cov}\,(J_{tu}, J_{uv}) = R - \lambda_S^2$$

We then have to work out how many pairs of pairs which share one index (trios) there are there are. We can do this by first picking the index to be shared, and then enumerating the number of other pairings. There are m ways of picking the first index, which leaves $m - 1$ other indices to be joined with in a triplet. There are $\binom{m-1}{2}$ ways of forming pairings between the $m-1$ indices left over. The total number of trios is therefore $\frac{m(m-1)(m-2)}{2}$.

$$\text{Var}(\sum_{1\leq t<u\leq m} J_{tu}) = \underbrace{\sum_{1\leq t<s\leq m} \text{Var}(J_{ts})}_{\text{sum of individual variances}} \;+\; \underbrace{2 \sum_{\substack{1\leq t<s\leq m\\ 1\leq u<v\leq m\\ (u,v)>(t,s)}} \text{Cov}(J_{ts},\, J_{uv})}_{\text{sum of covariances between distinct pairs}}$$

$$= \frac{m(m-1)}{2}\,(\lambda_S - \lambda_S^2) \;+\; m(m-1)(m-2)\,(R - \lambda_S^2). \tag{14}$$

The expression for the overall variance of λ'_S, combining Equation 11 and Equation 14, is

$$\text{Var}\,(\lambda'_S) = \frac{4}{m^4}\left[\frac{m(m-1)}{2}(\lambda_S - \lambda_S^2) + m(m-1)(m-2)(R - \lambda_S^2)\right] \tag{15}$$

This expression gives us the variance of λ'_S, but we want the variance of N'_S. To do this, we write out the expression for the variance of N'_S

$$\text{Var}\,(N'_S) = \mathbb{E}\,[N'_S - \mathbb{E}\,[N'_S]^2]$$

and substitute in $h(x) = 1/x$

$$\text{Var}\,(N'_S) = \mathbb{E}\,[h(\lambda'_S) - \mathbb{E}\,[h(\lambda'_S)]^2]$$

We will turn $h(\lambda'_S)$ into a linear approximation by using a Taylor expansion around the mean value of λ'_S, denoted by μ:

$$h(\lambda'_S) = h(\mu) + h'(\mu)(\lambda'_S - \mu)$$

We approximate N'_S using this

$$N'_S \approx h(\mu) + h'(\mu)(\lambda'_S - \mu)$$

and re-calculate the variance using this approximation:

$$\begin{aligned}
\text{Var}\,(N'_S) &\approx \text{Var}\,(h(\mu) + h'(\mu)(\lambda'_S - \mu)) \\
&= \text{Var}\,(h'(\mu)(\lambda'_S - \mu)) && \text{(since Var } X + a = \text{Var } X \text{ for constant } a) \\
&= h'(\mu)^2\,\text{Var}\,((\lambda'_S - \mu)) && \text{(since Var } aX = a^2 \text{ Var } X \text{ for constant } a) \\
&= h'(\mu)^2\,\text{Var}\,\lambda'_S && \text{(from the definition of variance)}
\end{aligned}$$

Since $h(x) = 1/x = x^{-1}$, $h'(x) = -x^{-2} = \frac{-1}{x^2}$, and $h'(x)^2 = \frac{1}{x^4}$. This means that

$$\text{Var}\,(N'_S) \approx \frac{1}{\mu^4}\,\text{Var}\,(\lambda'_S) \tag{16}$$

This completes the derivation of the variance of the district-level effective number of seat-winning parties.

3.3 The number of seat-winning parties N'_{S0}

If we had access to the full vector of seat shares $\mathbf{p}$, it would be easy to calculate the expected number of seat-winning parties under sampling with replacement. We would proceed party-by-party. For each party, we would ask: what is the probability that this party does *not* win a seat, given m draws? This would equal one minus their share, raised to the power m:

$$Pr(n_i = 0) = (1 - p_i)^m$$

This implies that the probability they win at least one seat is $1-(1-p_i)^m$. The expected number of parties is equal to the sum of the probabilities for each party.

$$\mathbb{E}\left(N'_{S0}\right) = \sum_{i=1}^{N_{S0}} (1 - (1 - p_i)^m) \tag{17}$$

However, the Seat Product Model does not give us an expectation for the vector of seat shares $\mathbf{p}$, only summary characteristics like the effective number and the length of the vector. We must therefore approximate.

One way of proceeding is to adopt distributional assumptions. Suppose that the vector of seat shares $\mathbf{n}$ follows a Dirichlet-Multinomial distribution, where the Dirichlet distribution is a symmetric distribution with the concentration parameter θ chosen so that the expected number of effective components is close to N_S. Per Huillet and Paroissin (2009), this means that

$$\theta = \frac{N_S - 1}{N_{S0} - N_S}$$

With the Dirichlet-Multinomial distribution with symmetric concentration parameter θ and number of draws m, the expected number of classes turns out to be

$$E[N'_{S0}] = N_{S0}\left(1 - \frac{B(\theta, (N_{S0} - 1)\theta + m)}{B(\theta, (N_{S0} - 1)\theta)}\right) \tag{18}$$

where $B(.)$ is the beta function. Proof of this is given in the appendix.

3.4 Worked examples

Suppose we take a system with an assembly size of 270 divided into twenty-seven districts each of ten seats. For reference purposes: this is a system with a slightly larger seat product than Portugal ($MS = 2700$ compared to Portuguese $MS = 2405$).

Per the Seat Product Model, this system would have an effective number of seat-winning parties of $(270 \times 10)^{(1/6)} = 3.73$, and a raw number of seat-winning parties equal to $(270 \times 10)^{(1/4)} = 7.2$, which we can round to seven. This system would have an effective number of vote-winning parties equal to 4.07 $(= (7.2 + 1)^{(2/3)})$.

Per Equation 9, the *inverse* effective number of seat-winning parties in each (equally large) district is

$$\begin{aligned}\lambda'_S &= \frac{1}{3.73} + \frac{270-10}{270-1} \cdot \frac{1-\frac{1}{3.73}}{10} \\ &= 0.34\end{aligned}$$

We therefore approximate N'_S as 2.94 ($= \frac{1}{0.34}$).

The section on the variance of N'_S gave an expression which relied on $R := \sum_i p^3$. Although in a real election it would be possible to calculate R explicitly, here I approximate R. I know that R has to be greater than λ_S^2, since $\lambda_S := \sum p_i^2$ and $R := \sum p_i^3$. I also know that R has to be less than $p_{max} \cdot \lambda_S$. Here, following the Seat Product Model, I know that the largest seat share is approximately equal to $\frac{1}{\sqrt{7.2}} = 0.373$. This means that R is somewhere between $0.34^2 = 0.1156$ and $0.373 \cdot 0.34 = 0.127$. Suppose for ease of calculation that $R = 0.12$.

Per Equation 15, the variance of λ'_S is:

$$\begin{aligned}\text{Var}\left(\lambda'_S\right) &= \frac{4}{10^4}\left[\frac{10(10-1)}{2}(0.34 - 0.34^2) + 10(10-1)(10-2)(0.12-0.1156)\right] \\ &= 0.0053064\end{aligned}$$

However, per Equation 16 the variance of N'_S depends on this figure and the mean prediction.

$$\begin{aligned}\text{Var}\left(N'_S\right) &= \frac{1}{0.34^4} 0.0053064 \\ &= 0.397\end{aligned}$$

If we were to formulate approximate 95% confidence intervals as $N'_S \pm 1.96\sigma$, then this would imply that our estimates of N'_S might be off by up to 1.2 units in this case. These confidence intervals are approximate because they rely on a normal approximation, and do not incorporate any information about lower bounds or possible values of N_S given discrete seat counts. These confidence intervals may therefore have poor coverage properties.

3.5 The effective number of vote-winning parties

In this final subsection I consider the effective number of vote-winning parties. Here we cannot rely on the district magnitude: this would imply that there could only be one vote-winning party in a district with magnitude one. Instead I rely on a pseudo-magnitude, m^*. We know (from the example of the single member districts) that the pseudo-magnitude must be larger than the magnitude, but what functional form might link actual magnitude to the pseudo-magnitude?

Consider the two simplest forms: an additive form $m^* = a + m$ and a multiplicative form $m^* = a \times m$. Assume, for the sake of argument, that the Seat Product Model is true at national level, and that we are considering the case where $m = 1$, and where $S = 270$. Per the SPM:

- the raw number of seat-winning parties nationally is $MS^{\frac{1}{4}} = 4.05$;
- the number of "pertinent" vote-winning parties is $4.05 + 1 = 5.05$;

- the effective number of vote-winning parties is $5.05^{\frac{2}{3}} = 2.94$.

We know that the effective number of vote-winning parties at district level can scarcely be as low as one, but how much higher should it be? Suppose, again for the sake of argument, that the effective number of vote-winning parties at district level was two and one thirds. We might motivate this on the basis of the M + 1 rule (Cox 1997), making generous allowance for non-Duvergerian equilibria. What would this imply for the parameter values in the functional forms above?

Let us assume that turnout in each district is exactly proportional to the district's share of total seats, and that therefore $T_i = m/S$. We can amend the previous equation Equation 9 to give the following expression for λ'_V.

$$\lambda'_V = \lambda_V + (1 - T) \cdot \frac{1 - \lambda_V}{m^*} \tag{19}$$

This gives as a first order approximation the expression

$$N'_V = \frac{m^* - Tm^* + T + N_V - 1}{m^* N_V} \tag{20}$$

We now insert $N_V = 2.94$, $N'_V = 2.33$ and $T = \frac{1}{270}$ into Equation 20 and solve for m^*. This gives a figure of approximately 7.4. If instead we say that N'_V will be two, then $m^* \approx 4.1$. Alternately, if $N'_V = 2.5$, then $m^* \approx 11$.

This implies that under the additive form, the parameter α might be $7.4 - 1$, but that it might be as low as $4.1 - 1$ or as high as $11 - 1$. For the multiplicative form, it implies that the parameter α might be 7.4, but might be between $[4.1, 11]$.

If we are willing to assume a particular value for m^*, we can calculate a variance for N'_V. If we adopt a value of 7.4, then we can calculate an expression for the variance of λ'_V using Equation 15. We can then convert this to an expression for the variance of N'_V using Equation 16. To use Equation 15, we must make an assumption regarding the term $R := \sum p_i^3$. Here I assume that R is equal to the geometric mean of the lower bound (λ_V^2) and the upper bound $(v_{max} \cdot \lambda_V)$. I know from the seat product model that $v_{max} \approx \sqrt{\frac{1}{N_{V0}}} = 44.5$. This gives a value of $R \approx 0.132$.

Per Equation 15, the variance of λ'_V is:

$$\begin{aligned}
\text{Var}\left(\lambda'_V\right) &= \frac{4}{{m^*}^4}\left[\frac{m^*(m^*-1)}{2}(\lambda_V - \lambda_V^2) + m^*(m^*-1)(m^*-2)(R - \lambda_V^2)\right] \\
\text{Var}\left(\lambda'_V\right) &= \frac{4}{7.4^4}\left[\frac{7.4(7.4-1)}{2}(0.34 - 0.34^2) + 7.4(7.4-1)(7.4-2)(0.132 - 0.34^2)\right] \\
&= 0.012683
\end{aligned}$$

However, per Equation 16 the variance of N'_S depends on this figure and the mean prediction for λ'_V.

$$\begin{aligned}
\text{Var}\left(N'_V\right) &= \frac{1}{0.429^4} 0.012683 \\
&= 0.3738055
\end{aligned}$$

If we were to formulate approximate 95% confidence intervals as $N'_V \pm 1.96\sigma$, then this would imply that our estimates of N'_S might be off by up to 1.2 units in this case, similar to the worked example for seat-winning parties.

Now consider the case where $m = 2$. We now have a new set of national effective numbers from the Seat Product Model. In particular we expect $N_V \approx 3.23$. On the basis of the reasoning above, we might also expect $R \approx 0.111$. Our expectations for N'_V now depend on the functional form we adopt.

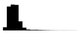
Suppose we adopt an additive functional form, and that $\alpha = 6.4$ as before. Then $m^* = m + 6.4 = 8.4$, and we would expect a value of N'_V approximately equal to 2.56. Alternately, suppose we adopt the multiplicative functional form, and that $\alpha = 7.4$ as before. Then $m^* = m \cdot 7.4 = 14.8$, and we would expect a value of N'_V approximately equal to 2.81.

It is very hard to say on the basis of these *expected values* which functional form is more plausible. However, we can also appeal to the *variance* implied by each of these pseudo-magnitudes. If we adopt the additive functional form, then the implied confidence interval runs from 1.26 to 3.86. If instead we adopt the multiplicative form, then the implied confidence interval runs from 1.7 to 3.92.

This second confidence interval implied by the multiplicative functional form is too high, in the sense that it excludes a logically possible outcome. I can show this by calculating the possible outcomes for *seat-winning* parties. When $m = 2$, $S = 270$, then (from the Seat Product Model) $N_S \approx 2.54$, and (from Equation 10) $N'_S \approx 1.44$. This is our best guess as to the expected value of N'_S, but with only two values either the effective number of seat-winning parties is one, or it is two. Our expectation implies that 56% of the time we are in a situation where one party wins both seats, and 44% of the time we are in the situation where two parties split the seats. When one party wins both seats, there is no useful lower bound on the effective number of vote-winning parties: it could be as low as one. But when two parties split the seats, then the first party cannot be more than two times larger than the second placed party, otherwise it would have been entitled to both seats under any proportional allocation mechanism. This means that the lowest possible value of N'_V is $\frac{1}{(\frac{2}{3})^2+(\frac{1}{3})^2}$, or 1.8. The *overall* lower bound, averaging over situations where one party wins both seats and situations where two party split the seats, is therefore $0.56 \cdot 1 + 0.44 \cdot 1.8 = 1.352$. But the confidence interval described above for the multiplicative functional form effectively *rules out* this logically possible outcome. It seems therefore that any multiplicative form which "works" for the $m = 1$ case does not also work for the $m = 2$ case.

This means that if we were forced to choose between the additive and multiplicative functional forms, we would have reasons to prefer the additive form. Any multiplicative parameter which gets us "enough" vote-winning parties in the single member district case yields "too many" vote-winning parties for larger district magnitudes. Although we could consider two or three-parameter functional forms, it is even harder to form expectations for the parameters in models using such forms. Parsimony and Cromwell's rule (never assign zero probability to something logically possible) both suggest that a simple additive form for the pseudo-magnitude could work well.

Table 1: Summary statistics for effective numbers at district and national level, together with summary statistics for district magnitude and assembly size.

	Unique	Missing Pct.	Mean	SD	Min	Median	Max	Histogram
N'_S	351	0	1.1	0.5	1.0	1.0	8.7	
N'_V	27176	0	2.5	0.8	1.0	2.4	10.4	
N_S	252	0	2.3	0.6	1.0	2.2	9.7	
N_V	265	0	3.0	0.9	1.4	3.0	10.9	
District magnitude	57	0	1.7	7.4	1.0	1.0	450.0	
Population share p	19422	28	0.0	0.1	0.0	0.0	1.0	
Assembly size S	88	0	377.1	209.1	4.0	308.0	659.0	

4 Data

To test my theory I use data from the Constituency Level Elections Archive (Kollman et al. 2024). I combine this district-level information with election-level information from the Democratic Electoral Systems (DES) dataset, version 5.0 (Bormann and Golder 2013). I combine these two datasets to ensure that the data respects the scope restrictions set out above and that the district-level data is free of major errors. I use the DES variable `tier1_formula` to exclude complex systems (two round systems, ranked systems, mixed systems, or systems which use remainder pooling) and block vote systems. I use the DES variables `enep` and `enpp` to remove elections where the effective number of vote- or seat-winning parties, *calculated on the basis of CLEA data*, is ten percent greater or smaller than the figure reported in the DES data. Finally, I use the DES variable `seats` to remove elections where the assembly size, *calculated on the basis of CLEA data*, is more than one seat away from the figure reported in the DES data. I do this assuming that the DES data is a better guide to election level statistics than aggregating CLEA data. I have additionally excluded a small number of district results where those results come from districts which use plurality rules to elect more than one member per seat (mostly Canadian ridings in the early C20th). In total I have information on 27269 in 265.

Table 1 provides descriptive statistics for the effective numbers at national level, and also for the range of district magnitudes, assembly sizes, and population shares. The district magnitude variable is highly skewed, with most observations equal to one but some very large district magnitudes (which often come from districts spanning the entire territory, as with the Ukrainian election of 2006, where $m = 450$).

5 Analysis

This section follows the structure of the theory section, testing expectations for the effective number of seat-winning parties, the raw number, and the effective number of vote-winning parties. A concluding subsection compares the accuracy of this theory compared to the Seat Product Model.

5.1 Analysis of the effective number of seat-winning parties at district level

Our expectation for the effective number of seat-winning parties at district level is given by Equation 10. This equation has no free parameters, and so it is possible to display the predicted value against the actual value graphically. This is done in Figure 1. The non-parametric smooth bulges upwards at a predicted value of $\hat{N'_S} \approx 4$, before falling below the 1:1 line in part as a result of districts in the 2019 Belgian election.

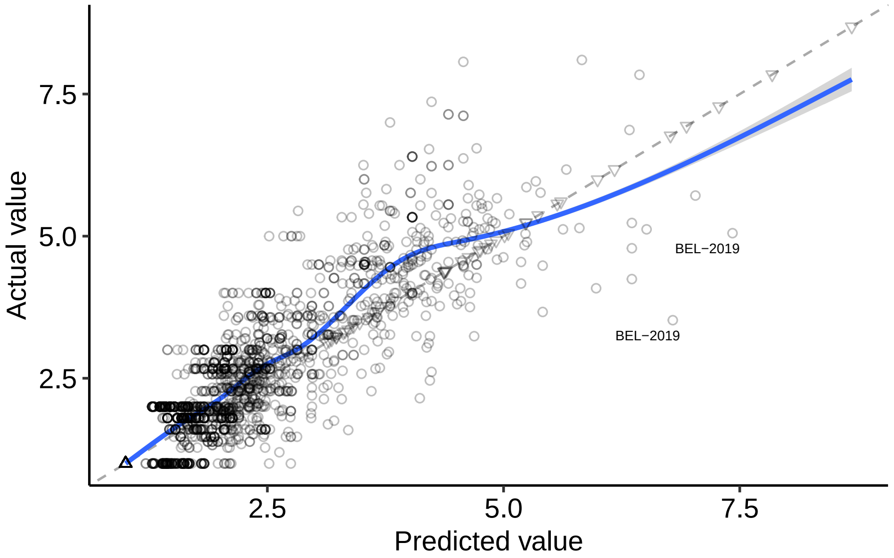


Figure 1: Scatter-plots of the district level effective number of seat-winning parties against predicted values. Dashed line shows the line $y = x$; solid blue line shows the best fitting local smooth. Plotted points show district magnitude is equal to one ( ); is equal to assembly size ( ) or is between 1 and S ( ). Selected outliers are labelled with ISO3166 country code and year.

Figure 1 shows the association between the predicted value and the actual value, but falls short of a formal test of the theory. To test the theory, I regress the actual value on the predicted value together with a modelled residual standard error. That is, I model

$$N'_S \sim N(\alpha + \beta \cdot \hat{N'_S} + \eta_j, \gamma \cdot \sqrt{\mathrm{Var}\left(\mathrm{N'_S}\right)}) \tag{21}$$

where $\hat{N'_S}$ comes from Equation 10, where $\mathrm{Var}(N'_S)$ comes from Equation 16, and where η_j is an election specific random intercept. The election specific random intercept is included because district results are nested within elections and as a result are not independent. The theory predicts that the intercept α will be approximately equal to zero; that the coefficient β will be approximately equal to one, and that the variance multiplier γ will be approximately equal to one. The theory makes approximate predictions because the derivation given in the theory section relies on Taylor approximations rather than exact identities. Because the variance of N'_S is zero when

Table 2: Model of N'_S

	(1)
α: Intercept, expected value of 0	0.154
	$[0.044, 0.266]$
β: Slope, expected value of 1	1.010
	$[0.969, 1.050]$
γ: Multiplicative term of predicted variance, expected value of 1	0.917
	$[0.890, 0.943]$
σ_η: Standard deviation of election random intercepts	0.225
	$[0.171, 0.287]$
N	1697
RMSE	0.64

$m = 1$ or $m = S$, the model is estimated only on cases where $1 < m < S$.[1] I estimate the model using Bayesian methods and provide a full discussion of prior specifications in the appendix.

Table 2 gives the result of the regression together with the root mean square error (RMSE) as a measure of model fit. The table shows that the slope coefficient is close and indistinguishable from its predicted value of 1. The intercept is slightly higher than, and significantly different to, the expected value of zero. This may reflect either approximation error, or higher-than-expected values at low values of m. For example: the binomial electoral system used in Chile between 1989 and 2013 rarely saw either electoral alliance "shut out", partly due to carefully design districts (Polga-Hecimovich and Siavelis 2015), and so delivered effective numbers of seat-winning parties very close to two. Finally, the multiplier term on the predicted variance whilst close to its predicted value is somewhat smaller. This may reflect the use of election random intercepts, which soak up some of the variation between elections but may also capture some of the predicted residual variation.

5.2 Analysis of the raw number of seat-winning parties at district level

We can test the implications for the raw number of seat-winning parties at district level in the same way. We can estimate a regression where we predict the raw number using the predicted value taken from an equation – in this case, Equation 18. In this case, I leave the residual variance unmodelled.

[1] Dropping these observations makes this a conservative test. It would be easy to run a regression on all data points, find that the intercept and slope were close to zero and one respectively, and declare victory – but this victory could be determined entirely by a series of easy wins where we predict that a district of magnitude one has an effective number equal to one. Whilst theories must avoiding predicting impossible things, successful predictions of this kind do not provide support for the theory.

Table 3: Model of N'_{S0}

	(1)
α: Intercept, expected value of 0	-0.063
	$[-0.201, 0.075]$
β: Slope, expected value of 1	1.025
	$[0.995, 1.055]$
σ_η: Standard deviation of election random intercepts	0.408
	$[0.342, 0.485]$
N	1697
RMSE	0.67

Table 4: Model of N'_V

	(Additive)
α: Coefficient	9.866
	$[9.345, 10.401]$
γ: Multiplicative term of predicted variance, expected value of 1	0.706
	$[0.695, 0.718]$
σ_η: Standard deviation of election random intercepts	0.234
	$[0.210, 0.261]$
N	19 591
LOOIC	21 050.3
RMSE	0.45

Because this model tests predicted against actual values, we expect an intercept of close to zero and a slope of close to one. In both cases the 95% credible interval encompasses these values. The coefficient value in particular within three percent of the theoretically predicted value of unity. Somewhat surprisingly, the theory seems to work better for the raw number of parties than it does for the effective number.

5.3 Analysis of the effective number of vote-winning parties

The final model estimates the district-level effective number of seat-winning parties. The regression equation is the same regression equation as before, *except that* the magnitude m in Equation 10 and Equation 16 is replaced with a pseudo-magnitude equal to the magnitude plus the parameter α.

The best fitting value of the parameter α is close to ten. This suggests that a single member district has a pseudo-magnitude of roughly eleven, close to the upper end of the range considered in Section 3.5.

Because this model relies on an estimated parameter, we cannot produce a plot exactly analogous to Figure 1, but I do show a plot of fitted values from Equation 20, with a specified α value of ten. The plot shows first that the predicted values capture well the average effective number, except that the best fitting local smooth drifts south at the extreme end of the scale where the equation predicts far higher effective numbers for Belgian districts than in fact are the case.

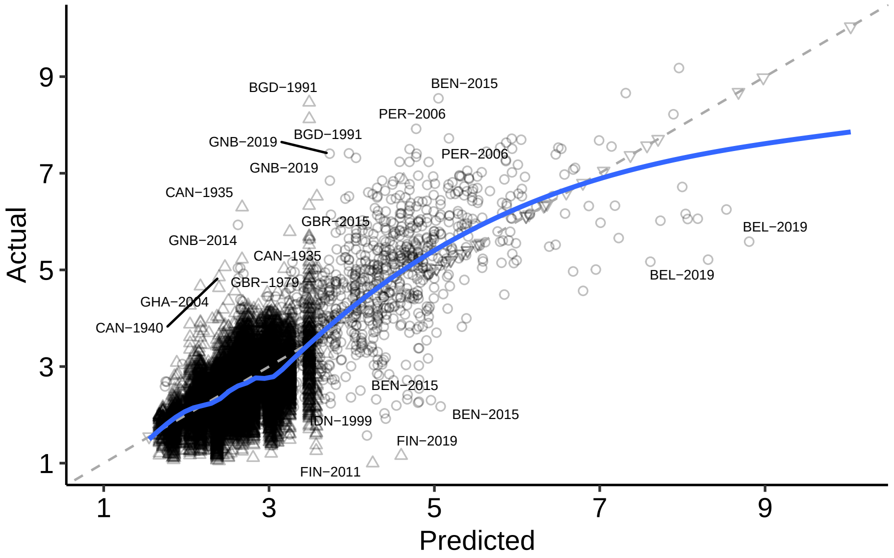


Figure 2: Scatter-plots of predicted district level effective number of vote-winning parties against actual values. Dashed line shows the line $y = x$; solid blue line shows the best fitting local smooth. Plotted points show district magnitude is equal to one ( ); is equal to assembly size ( ) or is between 1 and S ( ). Selected outliers are labelled with ISO3166 country code and year.

The points labelled in the graph (all points with an absolute standardized residual of greater than five) show the importance of political factors. Broadly there are three kinds of outliers – outliers where the effective number is depressed for linguistic reasons; outliers where the effective number is inflated by failures of coordination, and outliers with result from unpredictable but countervailing districts. Considering cases where the effective number is depressed: it would be a rich model indeed which was able to explain the low effective number of parties in the Åland islands, where different regionalist parties have won up to 89% of the vote. Similarly the effective number of vote-winning parties in Belgian districts reflects the fact that Belgian elections are not really Belgian elections but separate Flemish, Walloon and Brussels elections. Cases where the effective number is inflated by failures of coordination include elections after a period of non-democratic rule, such as the Bangladeshi elections of 1991 or the Guinnea-Bissau elections of 2014. Finally, elections which supply outliers either side of the line include the Beninese election of 2015.

5.4 Comparison to embeddedness and other rules

How do predictions from the equations listed above compare to predictions from the Seat Product Model, the leading top-down account of district-level electoral outcomes? Table 5 gives the root mean squared error (RMSE) for predictions from sampling theory and the SPM for three different outcomes:

Table 5: Root mean squared error (RMSE) for predictions from sampling theory and the Seat Product Model. 'PRE' = proportional reduction in error, or the difference between the RMSE from sampling theory and the RMSE from the Seat Product Model, divided by the RMSE from the Seat Product Model. See text for description of embeddedness coefficient k and other equations. "

	Sampling theory		Seat Product Model		
Quantity	Equation	RMSE	Equation	RMSE	PRE
N'_S	$\frac{mN_S(S-1)}{mS+SN_S-S-mN_S}$	0.667	$m^{2k/3}$	1.02	0.3461
N'_S	$\frac{m\hat{N}_S(S-1)}{mS+S\hat{N}_S-S-m\hat{N}_S}$	1.002	$m^{2k/3}$	1.02	0.0176
N'_{S0}	$N_{S0}\left(1-\frac{B(\theta,(N_{S0}-1)\theta+m)}{B(\theta,(N_{S0}-1)\theta)}\right)$	0.84	$\sqrt{m}$	1.522	0.4481
N'_V	$\frac{(m+10)N_S(S-1)}{(m+10)S+SN_S-S-(m+10)N_S}$	0.501	$({N'_S}^{3/4k}+1)^{4k/3}$	0.556	0.0989

N'_S, N'_{S0} and N'_V, together with the equations used to generate those predictions.

The first row of the table shows error on the predicted effective number of seat-winning parties at district level, *for those districts with magnitude greater than one*. The prediction from sampling theory is substantially more accurate than the prediction from the SPM, with a proportional reduction in error of 35%.

The comparison in the first row of the table is unfair. The Seat Product Model predicts national-level quantities and then predicts district-level quantities on the basis of those predictions. The prediction from sampling theory helps itself to a known quantity in N_S. The second row of the table therefore repeats this comparison, but replacing N_S with the prediction from the Seat Product Model. The predictions from sampling theory are still better, but only 2% better.

The third row of the table compares error on the prediction for the number of seat-winning parties. The prediction from the Seat Product Model relies solely on the district magnitude, and does not include any adjustment for embeddedness. Here our prediction is substantially more accurate (PRE of 45%).

The final row of the table compares errors for the district-level effective number of vote-winning parties. Sampling theory makes its prediction based on district magnitude and national level effective numbers; the SPM makes its prediction based on the (actual, not predicted) district-level effective number of seat-winning parties. Despite the fact that both theories are able to use district-level quantities, the predictions from sampling theory are 10% more accurate. Sampling theory, with modifications for quantities concerning votes, therefore offers predictions which have a clear rather than "baffling" basis and which are empirically more accurate.

6 Extension to party system inflation

Now that we have a way to formulate an expectation for effective numbers of seat- and vote-winning parties at district level, we can turn to the issue of party system inflation. In this section I show exactly how party system inflation depends on the average district magnitude and the variance of district magnitudes. I begin with seat-level party system inflation, since the expression for district-level effective numbers of seat-winning parties is simpler and involves no estimated parameters. I then turn to party system inflation at the level of votes, which is where almost all scholarly attention has been directed.

Moenius and Kasuya (2004) give the following formulae for party system inflation I:

$$I_S := \frac{N_S - \bar{N}'_S}{\bar{N}'_S} \cdot 100 \qquad\qquad I_V := \frac{N_V - \bar{N}'_V}{\bar{N}'_V} \cdot 100$$

which can also be expressed in the following more convenient form, which omits the conversion to percentage points:

$$I_S := \frac{N_S}{\bar{N}'_S} - 1 \qquad\qquad I_V := \frac{N_V}{\bar{N}'_V} - 1 \tag{22}$$

In this equation, the averages $\bar{N}'_S$ and $\bar{N}'_V$ are weighted arithmetic means with weights equal to the district's share of total seats or share of total votes V:

$$\bar{N}'_S := \sum_i \frac{m_i}{S} N'_{S(i)} \quad \bar{N}'_V := \sum_i \frac{V_i}{V} N'_{V(i)} \tag{23}$$

This measure of party system inflation is very closely related to what ecologists refer to as "beta-diversity", or the ratio between global diversity of an ecosystem or a meta-community and the diversity in individual patches or communities (Leinster 2021, ch. 8).

We can re-express the average district-level effective number $\bar{N}'_S$ as follows:

$$\bar{N}'_S = \sum_{i=1}^{E} \frac{m_i}{S} \left(\frac{m_i N_S (S-1)}{m_i S + S N_S - S - m_i N_S} \right)$$

where E is the number of electoral districts. We bring a number of terms outside of the summation to the front:

$$\bar{N}'_S = \frac{N_S(S-1)}{S} \sum_{i=1}^{E} \frac{m_i^2}{m_i S + S N_S - S - m_i N_S}$$

and substitute this into the expression for party system inflation to give:

$$I = \frac{N_S}{\frac{N_S(S-1)}{S} \sum_i^E \frac{m_i^2}{m_i S + S N_S - S - m_i n}} - 1$$

which simplifies to:

$$I_S = \frac{S}{(S-1)\sum_i^E \frac{m_i^2}{m_i S + SN_S - S - m_i n}} - 1$$

I'll use Σ to refer to the summation in the denominator, and describe the summand as a function

$$g(m) = \frac{m^2}{mS + SN_S - S - mN_S} \tag{24}$$

We can calculate the Taylor expansion of this function around the mean district magnitude $\bar{m}$. The first three terms of the Taylor expansion around $\bar{m}$ are given by:

$$g(m_i) \approx g(\bar{m}) + g'(m_i - \bar{m}) + \frac{1}{2}g''(\bar{m})(m_i - \bar{m})^2$$

The summation is therefore

$$\Sigma = \sum_i g(m_i) \approx \sum_i \left[g(\bar{m}) + g'(m_i - \bar{m}) + \frac{1}{2}g''(\bar{m})(m_i - \bar{m})^2 \right] \tag{25}$$

$$\approx E \cdot g(\bar{m}) + g'(\bar{m}) \sum_i (m_i - \bar{m}) + \frac{1}{2}g''(\bar{m}) \sum_i (m_i - \bar{m})^2 \tag{26}$$

However, we know that $\sum_i (m_i - \bar{m})$ must equal zero, and the sum of squared differences from the mean is equal to E times the variance of district magnitude. We can therefore rewrite the above approximation as

$$\Sigma \approx E \cdot g(\bar{m}) + \frac{1}{2}g''(\bar{m}) \cdot E \cdot \text{Var}\,(m_i)$$

Now, inflation I depends on the inverse of Σ, so as Σ gets bigger, I gets smaller. This in turn means that the relationship between variance of district magnitudes and inflation depends crucially on the sign of the second derivative of $g(m)$. This second derivative is

$$g''(m) = \frac{2(N_S S)^2}{(N_S(S-m) + Sm)^3}$$

We know that the numerator, because it involves two positive quantities, must be positive. In the case where there is more than one district, then $S - m_i > 0$ for all districts, and so the denominator is positive. (Obviously when there is just a single nationwide district the variance of district magnitudes is zero, and the question of party system inflation does not even arise). The sign of the second derivative is therefore positive, and so *given a constant* N_S, greater variance of district magnitudes is associated with a larger summand Σ and lower party system inflation. This provides a formal basis for the simulation studies and analysis reported in Barceló and Muraoka (2018).

The overall expression for party system inflation is:

$$I_S = \frac{S}{S-1} \frac{1}{E \cdot \left(g(\bar{m}) + \frac{1}{2}\,\text{Var}\,(m) g''(\bar{m})\right)} - 1 \tag{27}$$

where the function $g()$ is given by Equation 24. Similarly to the prediction for seat-level effective numbers, this gives an expression for seat-level party system inflation which has no free parameters.

We can give an expression for vote level party system inflation I_V by repeating these same steps. That is, we write out the expression for the population-share-weighted mean of district-level quantities:

$$\bar{N_V'} = \sum_{i=1}^{E} p_i \left(\frac{m_i^* N_V}{m_i^* + p_i + N_V - p_i N_V - 1} \right)$$

and bring the effective number of vote-winning parties out to the front before substituting into the equation for vote-level party system inflation to give:

$$I_V = \frac{1}{\sum_{i=1}^{E} p_i \left(\frac{m_i^*}{m_i^* + p_i + N_V - p_i N_V - 1} \right)} - 1 \tag{28}$$

Note that Equation 28 is different from the expression for seat-level party system inflation, because it depends on two variables measured at the level of the constituency: district magnitude m_i and district population share p_i. These two quantities are strongly related: more populous constituencies generally elect more representatives. However, the link is not exact: moderate levels of legislative malapportionment are common (Samuels and Snyder 2001; Kamahara, Wada, and Kasuya 2021), and legislative malapportionment can be a response to the same kinds of demands for especial representation that can give rise to party system inflation. We cannot therefore elide p_i and m_i, and we cannot therefore use the same techniques which allowed us to eliminate the sum when working with seat-level quantities. Our predictions for vote-level party system inflation must be based on aggregating predictions at the district level.

Figure 3 shows two scatter-plots which give predicted values of party inflation (horizontal axis) against actual levels of party inflation (vertical axis) for seat-winning parties (left panel) and vote-winning parties (right panel). The predicted values of party system inflation at seat level are taken by predicting the value of N_S' using Equation 10 for each district and calculating the size-weighted arithmetic mean. The predicted values of party system inflation are created by using Equation 20 with $m^* = m + 10$. This means that these predicted values are purely theoretical, and do not benefit from the slightly different coefficient values shown in Table 2 or Table 4, or the random intercepts used in those same models.

The left panel shows that party system inflation at seat level is as we would expect given the predicted values. There are two elections which appear as outliers – the Beninese election of 1991, and the 2019 Belgian election. The Beninese election was the first multiparty election, and was marked by strong differences in support for Nicéphore Soglo's UTRD between the north and south. The Belgian election was, like many Belgian elections marked by strong differences between Flemish and Walloon areas. The overall fit of the ordinary least squares regression line is driven by a large number of systems which use single member systems, where party system inflation is just equal to the overall effective number of seat-winning parties.

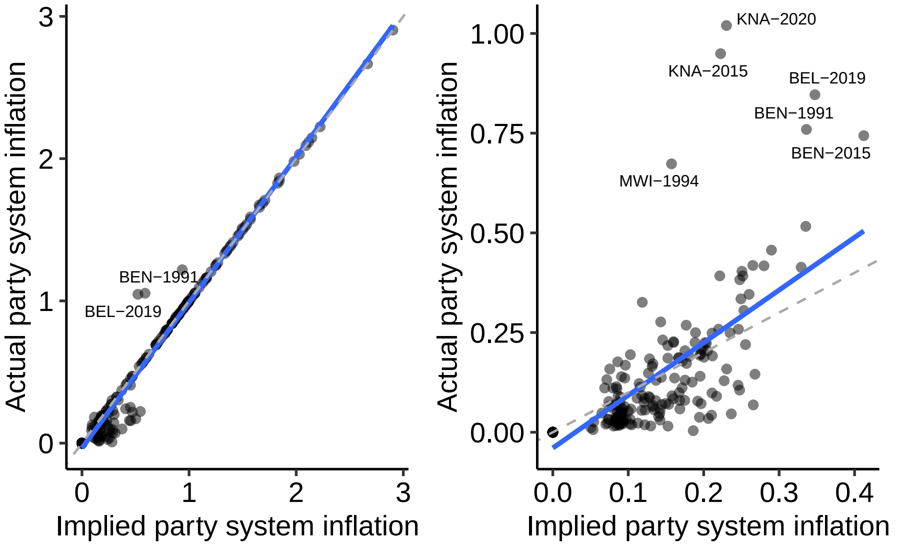


Figure 3: Scatter-plots of the levels of party system inflation implied by the models of district-level effective numbers, and actual levels. Panel (a) shows seat-level party system inflation; panel (b) shows vote-level party system inflation. Dashed line shows the line $y = x$. Observations with a Cook's distance greater than one are labelled by country code and year.

The right panel shows that party system inflation at vote level is broadly similar to what we would expect given the predicted values although here the instances of under-prediction are more numerous. These include Beninese and Belgian elections, but also two elections in St. Kitts and Nevis, where the "Team Unity" alliance won both elections but campaigned separately in St. Kitts (People's Action Movement) and Nevis (Concerned Citizen's Movement). If "Team Unity" is treated as a single party, the coefficient on the ordinary least squares regression line falls much closer to one.

These plots have been created by taking the point prediction for N'_S or N'_V in each district and aggregating over districts to get $\bar{N}'_V$ and $\bar{N}'_S$. It would be possible (and preferable) to simulate $\bar{N}'_V$ and $\bar{N}'_S$ by simulation, repeatedly drawing from assembly seats or quantized bundles of votes without replacement. Such a procedure would make for more realistic predictions in cases where the prediction for N'_V is highly variable, and or where there is substantial variance in district magnitudes.

7 Conclusion

In this article, I've argued that to explain the number of seat-winning parties at district level, we don't need to point to invoke theories of party linkage or coordination – we just need to point to the national level outcome and some well understood consequences of sampling without replacement when the sample size is limited. I've argued that the idea of districts as analogous to samples from a national population can also help us explain district-level effective numbers of vote-winning parties, if we are prepared to accept a "pseudo-magnitude" equal to the actual district magnitude plus ten.

Viewing districts as though they were samples from a national population is novel to political science but not to other disciplines. In ecology, many aspects of diversity can be explained by sampling. How we react to this finding is up to us. McGill (2011), discussing the situation in ecology, imagined that many of his colleagues would "find it disappointing that... cherished biodiversity patterns are explained by sampling... [and aspects] unique to ecology vs. those aspects that are probably quite general and apply to colored marbles, atoms, and any other entities that can be imagined to be sampled" (McGill 2011, 481–82). More optimistically, we can regard this finding as establishing a baseline useful in evaluating whether party systems are more nationalized than one would expect given effective numbers at national and district level.

My argument only applies to those simple electoral systems where seats are independent within and between districts. For complex electoral systems, the link between district magnitude and district-level effective numbers is likely to be different. Here I set out the four main families of complex electoral system, together with an indication of how they break the simple link described above.

First, in some systems seats are not allocated solely on the basis of district votes because parties must meet a national electoral threshold. Where this legal threshold is high relative to the natural threshold implied by district magnitude, it may shape voters' decision-making. The relationship between district magnitude and the effective number of vote-winning parties at district level would be attenuated: voters is large-magnitude, low implicit threshold districts would behave more like voters in small magnitude districts. The relationship would be attenuated more the more demanding the threshold and the bigger the district. It might be possible to incorporate the legal threshold into another pseudo-magnitude, as proposed by Bochsler, Hänni, and Grofman (2024), but this introduces an additional level of complexity and it is not clear that voters respond to district magnitudes and legal thresholds in the same way.

Second, in mixed systems seats are allocated in multiple (separate or linked) tiers. In mixed systems, there is little interest in explaining district outcomes in the proportional tier, for this tier usually operates with a single national district. For the nominal tier, for systems which allow split votes between tiers, it would in principle be possible to explain vote-level outcomes provided that the national effective number N_V is calculated only on the basis of votes cast in that tier, but I have chosen not to pursue that route here.

Third, in some systems voters express multiple preferences, making it difficult to calculate any kind of effective number of vote-winning parties, unless it be on the basis of first preferences. It may be that systems which use rank preference orderings work like party-list PR, or that the first round in a two-round system works like single member district plurality when one conditions on the national-level effective number of vote winning parties, but this is conjecture.

Finally, there are some seats where seats are not allocated solely on the basis of votes cast in that district, but where the additional element of remainder pooling is generally opaque to the voter and small in proportion to the total number of seats. The smaller the number of levelling up seats, the more these complex systems would approximate the simple systems I study here.

Earlier I suggested that random sampling without replacement, while it might account for concentration of party support, might not account for cooccurrence or covariance between shares for specific parties. This proves to be the case. Additional analyses reported in the appendix show that in a substantial minority of elections studied, the party-by-party products of party seat shares are significantly different from what one would expect under random sampling alone. Future research must therefore resolve the paradox of why concentration appears as-if random when competition and cooccurrence are so evidently patterned.

References

Barceló, Joan, and Taishi Muraoka. 2018. "The Effect of Variance in District Magnitude on Party System Inflation." *Electoral Studies* 54: 44–55.

Birnir, Jóhanna Kristín. 2004. "Stabilizing Party Systems and Excluding Segments of Society?: The Effects of Formation Costs on New Party Foundation in Latin America." *Studies in Comparative International Development* 39 (3): 3–27.

Bochsler, Daniel. 2010. "Measuring Party Nationalisation: A New Gini-Based Indicator That Corrects for the Number of Units." *Electoral Studies* 29 (1): 155–68.

Bochsler, Daniel, Miriam Hänni, and Bernard Grofman. 2024. "How Proportional Are Electoral Systems? A Universal Measure of Electoral Rules." *Electoral Studies* 87: 102713.

Bormann, Nils-Christian, and Matt Golder. 2013. "Democratic Electoral Systems Around the World, 1946–2011." *Electoral Studies* 32 (2): 360–69.

Braumoeller, Bear F. 2006. "Explaining Variance; or, Stuck in a Moment We Can't Get Out Of." *Political Analysis* 14 (3): 268–90.

Chhibber, Pradeep, and Ken Kollman. 1998. "Party Aggregation and the Number of Parties in India and the United States." *American Political Science Review* 92 (2): 329–42.

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

———. 1999. "Electoral Rules and Electoral Coordination." *Annual Review of Political Science* 2 (1): 145–61.

Crisp, Brian F, Santiago Olivella, and Joshua D Potter. 2013. "Party-System Nationalization and the Scope of Public Policy: The Importance of Cross-District Constituency Similarity." *Comparative Political Studies* 46 (4): 431–56.

Ferree, Karen E, G Bingham Powell, and Ethan Scheiner. 2014. "Context, Electoral Rules, and Party Systems." *Annual Review of Political Science* 17 (1): 421–39.

Golosov, Grigorii V. 2015. "Factors of Party System Fragmentation: A Cross-National Study." *Australian Journal of Political Science* 50 (1): 42–60.

Hicken, Allen, Ken Kollman, and Joel W Simmons. 2016. "Party System Nationalization and the Provision of Public Health Services." *Political Science Research and Methods* 4 (3): 573–94.

Huillet, Thierry, and Christian Paroissin. 2009. "Sampling from Dirichlet Partitions: Estimating the Number of Species." *Environmetrics: The Official Journal of the International Environmetrics Society* 20 (7): 853–76.

Kamahara, Yuta, Junichiro Wada, and Yuko Kasuya. 2021. "Malapportionment in Space and Time: Decompose It!" *Electoral Studies* 71: 102301.

Kollman, Ken, Allen Hicken, Daniele Caramani, David Backer, and David Lublin. 2024. "Constituency-Level Elections Archive." Ann Arbor, MI: Center for Political Studies, University of Michigan.

Laakso, Markku, and Rein Taagepera. 1979. "'Effective' Number of Parties: A Measure with Application to West Europe." *Comparative Political Studies* 12 (1): 3–27.

Lago, Ignacio, and José Ramón Montero. 2014. "Defining and Measuring Party System Nationalization." *European Political Science Review* 6 (2): 191–211.

Lago-Peñas, Ignacio, and Santiago Lago-Peñas. 2009. "Does the Nationalization of Party Systems Affect the Composition of Public Spending?" *Economics of Governance* 10 (1): 85–98.

Leinster, Tom. 2021. *Entropy and Diversity: The Axiomatic Approach*. Cambridge university press.

McGill, Brian J. 2011. "Linking Biodiversity Patterns by Autocorrelated Random Sampling." *American Journal of Botany* 98 (3): 481–502.

Moenius, Johannes, and Yuko Kasuya. 2004. "Measuring Party Linkage Across Districts: Some Party System Inflation Indices and Their Properties." *Party Politics* 10 (5): 543–64.

Morgenstern, Scott, Stephen M Swindle, and Andrea Castagnola. 2009. "Party Nationalization and Institutions." *The Journal of Politics* 71 (4): 1322–41.

Pearson, Karl. 1897. "Mathematical Contributions to the Theory of Evolution.—on a Form of Spurious Correlation Which May Arise When Indices Are Used in the Measurement of Organs." *Proceedings of the Royal Society of London* 60 (359-367): 489–98.

Polga-Hecimovich, John, and Peter M Siavelis. 2015. "Here's the Bias! A (Re-) Reassessment of the Chilean Electoral System." *Electoral Studies* 40: 268–79.

Samuels, David, and Richard Snyder. 2001. "The Value of a Vote: Malapportionment in Comparative Perspective." *British Journal of Political Science* 31 (4): 651–71.

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

Simón, Pablo. 2013. "The Combined Impact of Decentralisation and Personalism on the Nationalisation of Party Systems." *Political Studies* 61 (1_suppl): 24–44.

Simón, Pablo, and Marc Guinjoan. 2018. "The Short-Term and Long-Term Effects of Institutional Reforms on Party System Nationalization." *Comparative European Politics* 16 (5): 762–82.

Simpson, Edward H. 1949. "Measurement of Diversity." *Nature* 163 (4148): 688–88.

Taagepera, Rein. 2008. *Making Social Sciences More Scientific: The Need for Predictive Models*. OUP Oxford.

Tennant, Peter WG, Kellyn F Arnold, George TH Ellison, and Mark S Gilthorpe. 2022. "Analyses of 'Change Scores' Do Not Estimate Causal

Effects in Observational Data." *International Journal of Epidemiology* 51 (5): 1604–15.

Theil, Henri. 1969. "The Desired Political Entropy." *American Political Science Review* 63 (2): 521–25.

Woodward, James. 2018. "Some Varieties of Non-Causal Explanation." In *Explanation Beyond Causation: Philosophical Perspectives on Non-Causal Explanations*, edited by Alexander Reutlinger and Juha Saatsi. Oxford University Press.

Appendix to ‘Electoral districts as samples from the national population’

Table of contents

Election years included

Table 1: Country-years included.

Country	Years
Albania	2017
Antigua and Barbuda	1984; 1989; 1994; 1999; 2004; 2009; 2014; 2018

Continued on next page

Table 1: Country-years included. (Continued)

Country	Years
Bahamas	2002; 2007; 2012; 2017
Bangladesh	1991; 2001
Barbados	1971; 1976; 1981; 1986; 1994; 1999; 2003; 2008; 2013
Belgium	2019
Belize	1984; 1989; 1993; 1998; 2003; 2008; 2012; 2015; 2020
Benin	1991; 2015
Bhutan	2008; 2013
Botswana	1969; 1974; 1979; 1984; 1989; 1994; 1999; 2004; 2009; 2014
Bulgaria	1991; 1994; 1997; 2001; 2005; 2013; 2014; 2017
Burundi	2020
Canada	1925; 1926; 1930; 1935; 1940; 1945; 1949; 1953; 1957; 1958; 1962; 1963; 1965; 1968; 1972; 1974; 1979; 1980; 1984; 1988; 1993; 1997; 2000; 2004; 2006; 2008; 2011; 2015; 2019
Cape Verde	1995; 2011; 2016
Costa Rica	2010; 2014
Cyprus	1981
Czechia	2010; 2013; 2017
Dominica	2009; 2014; 2019
Dominican Republic	2002
Finland	1999; 2003; 2007; 2011; 2015; 2019
Gambia	2017
Ghana	2004; 2012
Grenada	2018
Guinea-Bissau	1994; 2014; 2019
Indonesia	1999; 2004
Israel	1949; 1951; 1955; 1959; 1961; 1969; 1973; 1977; 1981; 1984; 1988; 1992; 1996; 1999; 2003; 2006; 2009; 2013; 2015; 2019 (Apr, Sep); 2020
Jamaica	2002; 2007; 2011; 2016
Kyrgyzstan	2020
Latvia	1995; 1998; 2002; 2006; 2010; 2011; 2014

Continued on next page

Table 1: Country-years included. (Continued)

Country	Years
Luxembourg	2004; 2018
Malawi	1994
Malaysia	1974; 1978; 1982; 1986; 1990; 1995
Moldova	1994; 2005; 2009 (Apr, Jul); 2010
Montenegro	2012; 2016
Namibia	1994; 1999; 2004; 2009; 2014
New Zealand	1946; 1949; 1951; 1954; 1957; 1960; 1963; 1966; 1969; 1972; 1975; 1978; 1981; 1987; 1990; 1993
North Macedonia	2014; 2020
Pakistan	1977
Paraguay	2013
Peru	1980; 1990; 2006; 2011
Philippines	1965
Poland	2015; 2019
Portugal	2015; 2019
San Marino	2012; 2019
Serbia	2007; 2008; 2012; 2014; 2016; 2020
Sierra Leone	2018
Slovakia	1998; 2006; 2010; 2012; 2020
Solomon Islands	1984
St. Kitts and Nevis	2015; 2020
St. Lucia	2006; 2011; 2016
St. Vincent and Grenadines	2001; 2005; 2010; 2015; 2020
Suriname	2015
Timor-Leste	2007; 2017
Trinidad and Tobago	1966; 1971; 1976; 1981; 1986; 2015; 2020
Turkey	1983; 1995; 1999; 2002; 2007; 2011; 2015 (Jun, Nov)
Ukraine	2006; 2007
United Kingdom	1964; 1966; 1970; 1974 (Feb, Oct); 1979; 1992; 1997; 2001; 2005; 2010; 2015; 2019

Continued on next page

Table 1: Country-years included. (Continued)

Country	Years
United States	1996; 2006; 2008; 2010; 2012; 2014; 2016; 2018; 2020

Election years not included

The data used in this article is a small subset of the CLEA data. In part this is because elections were removed because of inconsistencies in the data; mostly, however, it is because of the need to meet the scope conditions (elections held under simple electoral systems in democracies).

The 2024 CLEA release includes information on 2092 elections. After removing elections for data inconsistencies at this stage (no seat information, no vote information, district magnitude missing, etc.,), I was left with information on 2046 elections.

I then joined the CLEA data together with the Bormann and Golder dataset. This dataset includes information on 1826 democratic parliamentary elections. When I restrict the number to simple systems with values for N_S and N_V, the number drops to 1085. All but one of these systems is also featured in data, meaning that after the merge I had information on 1084 elections.

When I then filter to elections where the tally of seats won (calculated on the basis of CLEA data) is within one seat of the total number of seats (as reported in the Bormann/Golder data), I lose 689 elections, and drop down to 395 elections. When I further remove cases where the value of N_V or N_S (calculated on the basis of CLEA data) is more than 10% away from the value reported in the Bormann/Golder data, I drop down to 293 elections. I drop a further 10 elections where the electoral formula used to allocate seats is not present.

Finally, in the process of aggregating the data to the district level, I lose a further eighteen elections because of infinite values of N_V' or N_S' (i.e., where the reported vote tallies were also zero).

Derivation of the expression for the expected raw number of seat-winning parties

In the main article I claimed that the expected raw number of seat-winning parties is equal to:

$$E[N_{S0}'] = N_{S0}\left(1 - \frac{B(\theta, (N_{S0}-1)\theta + m)}{B(\theta, (N_{S0}-1)\theta)}\right) \tag{1}$$

where $B(.)$ is the beta function, m is district magnitude, and θ is a symmetric Dirichlet concentration parameter, equal to $\frac{N_S - 1}{N_{S0} - N_S}$.

We can write out the number of as seat-winning parties as the sum of different indicator variables. Let I_j be an indicator variable which has a value of one if party j wins at least one seat. Then the expected number of seat-winning parties, $\mathbb{E}[N'_{S0}]$, is equal to $\sum_j^{N_{S0}} \mathbb{E}[I_j]$.

Because we have assumed that this Dirichlet-Multinomial distribution comes from a symmetric Dirichlet, where all components have the same size *in expectation*, then we can rewrite the expected number of seat-winning parties as the product of the probability that a randomly chosen party wins at least one seat, multiplied by the number of parties. That is, $\mathbb{E}[N'_{S0}] = N_{S0} \cdot \mathbb{E}[I_1]$.

It turns out to be easier to work with the corresponding probability, that party j wins no seats. This is equal to the probability that party j fails wins a given seat (i.e., $1 - p$), raised to the power of the number of seats.

We therefore want to evaluate the integral of this expression with respect to the probability p, which in turn depends on the Dirichlet distribution. That is, we want to know

$$\int_0^1 (1-u)^m f(u) du$$

where u depends on the Dirichlet.

The marginal distributions of any Dirichlet distribution is a Beta distribution with parameters $\theta, (k-1)\theta$, where k is the number of categories (in our case, the national number of seat-winning parties. The marginal density is given by:

$$f(u) = \frac{1}{B(\theta, (N_{S0} - 1))} u^{\theta - 1} (1-u)^{(k-1)\theta - 1}$$

We substitute this marginal density into the integral to give

$$\int_0^1 (1-u)^m f(u) d(u) \frac{1}{B(\theta, (N_{S0} - 1))} u^{\theta - 1} (1-u)^{(k-1)\theta - 1} du$$

We then pull terms not involving u out to the front, and combine powers of $(1-u)$ to give

$$\frac{1}{B(\theta, (N_{S0} - 1))} \int_0^1 u^{\theta - 1} (1-u)^{(N_{S0} - 1)\theta - 1 + m} du$$

From the definition of the Beta function,

$$B(a, b) = \int_0^1 t^{a-1} (1-t)^{b-1} \, dt.$$

which allows us to replace the integral. The probability of a focal party winning at least one seat is therefore

$$\frac{B(\theta, (N_{S0} - 1)\theta + m)}{B(\theta, (N_{S0} - 1)\theta)}.$$

and the probability of all parties that win seats nationally (N_{S0}) parties *failing* to win seats is equal to N_{S0} times one minus that probability

$$N_{S0}\left(1-\frac{B(\theta,(N_{S0}-1)\theta+m)}{B(\theta,(N_{S0}-1)\theta)}\right).$$

which is the claim we started out with.

Prior specifications

Effective number of seat-winning parties

The model of the effective number of seat-winning parties at district level has four parameters: the intercept α, the coefficient β, the coefficient on the predicted standard deviation, γ, and the standard deviation of the election random intercepts η_j. The expected values for these parameters are zero for the intercept, and one for the two coefficients. But what are reasonable priors for these parameters?

For the coefficients, I adopt a standard normal prior. This is a generic weakly informative prior which works in this application. The predicted effective number of seat-winning parties is on the same scale as the actual effective number of seat-winning parties. When modelling variables which are on the same scale in a bivariate regression, the slope must be between -1 and +1. The standard normal prior places 68% probability on values between these two extremes, and so this prior allows for values that are "extreme" given the fact that the outcome and main predictor variable are on the same scale. I adopt the same prior for the coefficient on the predicted variance for similar reasons.

For the intercept, I also adopt a standard normal prior. It's true that the intercept could take on large positive values: if, for example, the coefficient on the predicted value is exactly zero, then the intercept would have to adjust to match the average (mean) value of the outcome variable. For districts with magnitude greater than one, that mean value is close to 2.7. A standard normal prior places very low probability on values this extreme. At the same time, it still allows for the prediction to be off by a large amount in the context of prediction. With this prior, the prediction for low district magnitudes could be off by two whole units, which would imply catastrophic errors.

Now consider the standard deviation of the country-year intercepts. Under a worst case scenario, where the predicted value $\hat{N_S'}$ explains nothing about the actual values of N_S', then the standard deviation of the country intercepts might approach the standard deviation of the outcome itself. For districts with magnitude greater than one, this standard deviation is roughly one and a quarter. One prior which places a small (5% probability) on values as extreme as 1.25 is the exponential distribution with rate parameter 2.4.

The prior distributions are therefore:

$$\alpha,\beta,\gamma \overset{\text{iid}}{\sim} N(0,1)$$

$$\sigma(\eta) \sim \text{Exp}(2.4)$$

Raw number of seat-winning parties

The model of the raw number of seat-winning parties at district level has a slope with expected value of one and an intercept with expected value of zero. In this respect it is similar to the model of the effective number of seat-winning parties. I therefore adopt the same prior distributions for the parameters α, β and γ.

For the prior on the standard deviation of the country-year intercepts, I adopt the same reasoning as before. Under a worst case scenario, where the predicted value $\widehat{N'_{S0}}$ explains nothing about the actual values of N'_{S0}, then the standard deviation of the country intercepts might approach the standard deviation of the outcome itself. For districts with magnitude greater than one, this standard deviation is roughly two. One prior which places a small (5% probability) on values as extreme as two is the exponential distribution with rate parameter 1.5.

Effective number of vote-winning parties

The model of the effective number of vote-winning parties is different insofar as it features an additive parameter, α. In the body of the text, I suggests that this parameter might range between 4 and 11. One prior which places large probability (80%) on values in this range is a N(7.5, 2.75).

Co-occurrence

One open question is whether some kind of sampling model can capture co-occurrence of parties. Here I suggest that it cannot. I base my argument on the products of party seat shares in each constituency. Consider the matrix z where $z_{i,j} = s_i s_j$, where s_i and s_j are the seat shares of two parties i and j. Under random sampling, we know how to create the expectation for each cell in matrix z: it is simply equal to the product of the national seat shares p_i and p_j. For each pair of parties, we can calculate the disparity between the actual product $s_i s_j$ and the expectation $p_i p_j$. Sometimes the actual product will be greater; some times smaller. To create a matrix analogous to a chi-square statistic, we can express the squared difference as a ratio of the expected value, and sum this over pairs of parties:

$$\text{X} = \sum_{i,j} \frac{((s_i s_j) - (p_i p_j))^2}{(p_i p_j)}$$

If we repeatedly draw samples from the national distribution of seats, we can calculate when the observed value of this statistic is different from what one would expect on chance alone.

For example: in Belgium in 2019, the value of this statistic in the actual data, summing over multiple districts, was 903. The average value of the statistic over

500 simulations was 65.4. In no simulation was the value of the statistic greater than the observed value of 903. The probability that the products of seat shares could have come from random sampling is therefore less than one in the number of simulations.

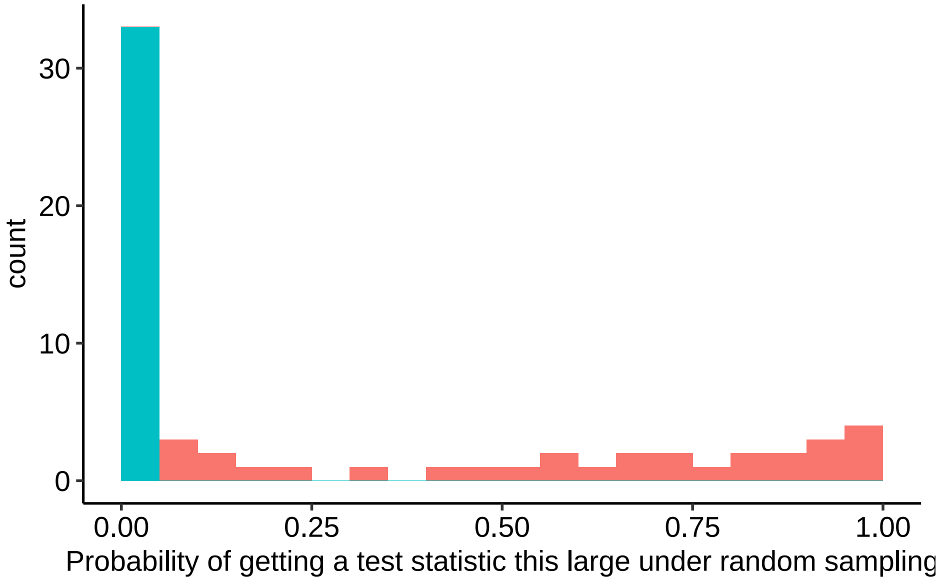


Figure 1: Histogram of the probability that test statistics in 129 elections held in districts with magnitudes greater than one

Figure 1 gives a histogram of the probability of getting a test statistic as large as the observed value under random sampling for 129 elections held in systems where the district magnitude was greater than one and less than the assembly size. 33 elections in 64 – roughly half – have patterns that depart from what we would expect under simple random sampling. The elections with the most extreme test statistics are all elections where there was strong regional pattern, including the Belgian election of 2019, the Benin election of 2015, and several different Bulgarian elections.