

How Similar Are They? Rethinking Electoral Congruence

Jason Wittenberg
University of California, Berkeley
witty@berkeley.edu¹

July 3, 2008

¹Prepared for the 25th annual meeting of the Society for Political Methodology, University of Michigan, Ann Arbor, July 9-12, 2008. Thanks to Miguel de Figueiredo for research assistance, and to John Bing, Thomas Brambor, Jay Goodliffe, and Sven Wilson for detailed comments. Earlier versions of this paper were presented at seminars at UC Berkeley, BYU, and Stanford University, as well as at the 2008 annual meetings of the Southwestern and Midwest Political Science associations.

Abstract

Electoral continuity and discontinuity have been a staple of voting research for decades. Most researchers have employed Pearson's r as a measure of congruence between two electoral outcomes across a set of geographic units. This paper argues that that practice should be abandoned. The correlation coefficient is almost always the wrong measure. The paper recommends other quantities that better accord with what researchers usually mean by electoral persistence. Replications of prior studies in American and comparative politics demonstrate that the consequences of using r when it is inappropriate can be stark. In some cases what we think are continuities are actually discontinuities.

1 Introduction

Electoral continuity and discontinuity have been for decades a staple of voting research. From the study of the “Solid South” and its demise in the US to the “freezing hypothesis” in Europe and party system change in the Americas and elsewhere, scholars have employed quite sophisticated methods to illustrate subtle patterns in the evolution of electoral behavior. These analyses have informed how we think about partisanship as well the ways parties interact with their electorates and with each other. This paper addresses important ambiguities in the way electoral stability and instability are conceptualized and measured. In particular, I show that for an important class of problems, researchers conflate two very different concepts of stability, one capturing the degree of association, and the other the degree of agreement, between two electoral outcomes. These errors are compounded by the near-universal use of Pearson’s correlation coefficient as a measure of persistence. The paper argues that Pearson’s r almost never measures what researchers think it does, and should be (mostly) abandoned. It shows that what researchers usually mean by electoral persistence can be captured more accurately by either a rank order correlation coefficient such as Kendall’s Q or a measure of absolute agreement. Replications of prior studies from American and comparative politics demonstrate that the consequences of using the wrong measure can be stark. In some cases what we think are continuities are actually discontinuities. The paper concludes by discussing some further pitfalls with correlational measures and applications beyond electoral persistence.

2 The Basic Schema

It is best to begin with a generic example. Figure 1 displays a scatterplot of points, where the horizontal axis represents the fraction of the vote for a particular party (or bloc or tendency) at time i , and the vertical axis represents the vote for that same party (or bloc or tendency) at time j . With appropriate labels for the party, elections, and units, such a scenario occurs quite often in electoral research, even where a scatterplot is not literally employed. It is central to the study of political continuity and discontinuity in redemocratizing regimes in Eastern Europe, Southern Europe, and Latin America, where more often than not the elections under consideration are separated by decades. Wittenberg (2006), for example, employs this schema to explore continuities in support for the Right between pre- and post-communism across Hungarian settlements. Similarly, Linz (1980) identifies continuities and discontinuities in support for the Left and Right at the provincial level between pre- and post-Franco Spain.¹ This framework also informs some strands of research on long-term continuities in American political development. Key and Munger (1959) inspired a generation of further research on the “standing electoral decision” through their exploration of persistent voting allegiances across Indiana counties.² In all these

¹Other examples include Maravall (1982), Valenzuela and Scully (1997), and Montes, Mainwaring, and Ortega (2000).

²See also Burnham (1968), Levine (1976), and Gimpel and Schuknecht (2002).

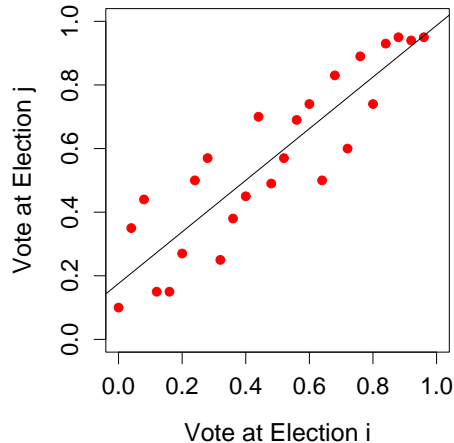


Figure 1: Generic comparison of two electoral outcomes across a set of units, with the regression line.

instances, the Pearson correlation coefficient, r , has remained the statistic of choice to characterize the relationship among the points.

King (1986) notes that although “many great things” have been attributed to r , almost none of them are true. This is particularly true of the widely-held but almost wholly erroneous idea that r measures the degree of similarity between two sets of outcomes. As Robinson (1957) recognized, it does not: Pearson’s r is a measure of linearity, not similarity. As I show in the next two sections, at best r indicates a very weak form of continuity; at worst, and almost always in practice, it is fundamentally misleading.

3 Conceptualizing Continuity: Relative Strength

At least two distinct senses of political continuity can be distilled from the literature. In this section we examine the more common but far weaker of the two, which focuses on the relative weakness and strength across districts of some party (or bloc or tendency) over time. Suppose a party performs weakly in some districts and strongly in others in some election. Suppose at some later election that the same party competes and performs weakly in some districts and strongly (say) in others. The degree of persistence is then the extent to which strong performance in the earlier election is associated with strong performance in the latter one. This conceptualization does not depend on any particular definition of “weak” and “strong” provided the researcher is comfortable comparing what counts as weakness or strength in one period with what counts as weakness or strength in another. For example, in a multiparty system winning a plurality of the district vote in both time periods would be a possible definition of strength. Another possibility would be to define weakness and strength relative to the average performance of the party across districts. A weak performance in some district is then one that falls below the average performance for that election, a strong one being above the average. In a two-party system like the U.S., the dividing line between weakness and strength

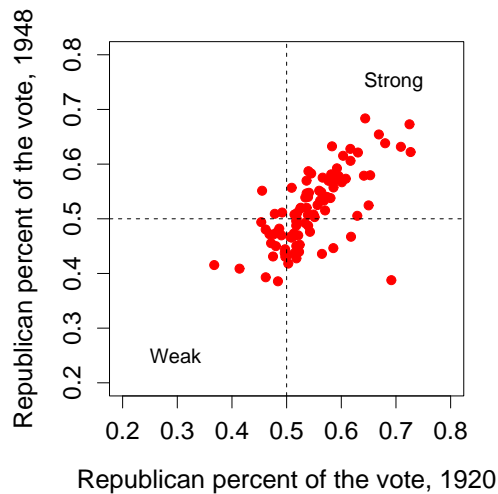


Figure 2: The traditional vote across Indiana counties (according to Key and Munger, 1959, Figure 15-2.), with “weak” and “strong” Republican counties identified. Source: Recomputed from ICPSR Dataset 1, United States Historical Electoral Returns 1824-1968.

might be 50 percent of the vote rather than a plurality.

This two-party version is precisely the setup Key and Munger (1959) employ in their classic work on the “standing electoral decision” in American politics. They investigate the persistence in voting allegiances through an analysis of electoral behavior in Indiana counties between 1868 and 1948, and note that “where the Republicans were strong in 1920, they were relatively strong in 1948; where the Democrats were weak in 1920, they were relatively weak in 1948” (pp. 283-286).

Figure 2 is a replication of this finding, with dashed horizontal and vertical lines imposed superimposed imposed to separate “weak” from “strong” electoral performances. Points in the lower left quadrant indicate counties where the Republican Party was weak in both 1920 and 1948. The upper right quadrant contains counties where the Republicans were strong in both elections. To determine the magnitude of persistence Key and Munger correlate the actual electoral outcomes and find that $r = .68$ for the Republican vote between 1920 and 1948. They interpret this, along with other evidence, as representing “the balance between two opposing party groups each with striking powers of self-perpetuation.”

Correlating the outcomes in this way is the wrong procedure for ascertaining the degree of persistence in terms of relative strength and weakness over time. The reason is that the result of such a computation will be sensitive to even small changes in party performance in a municipality, whereas the logic of relative strength implies that any weak (strong) performance should be considered equivalent to any other weak (strong) performance, regardless of whether the actual results are nearly identical or relatively far apart. For example, Tipton county, which gave the Republicans roughly 51 percent to the Republicans in both 1920 and 1948, and Steuben county,

whose support was 67 and 72 percent, respectively, both get coded as “strong” Republican counties and thus have identical outcomes from the relative strength perspective. More generally, any municipality within a given quadrant is considered equivalent to any other within the same quadrant regardless of where they happen to fall within the quadrant. Pearson’s r , however, is sensitive to within-quadrant variance.

The appropriate measure of association in a relative strength scenario is one such as Yule’s Q that respects the binary nature (weak vs. strong) of the underlying persistence concept but discounts the absolute differences among units within each category. With this notion of association perfect persistence ($Q = 1$) would obtain if all points in Figure 2 were in the lower-left and upper-right quadrants: weak Republican counties remained weak, while strong ones remained strong.³ The underlying logic is the same if outcomes are coded into more than two categories (such as “weak”, “middling”, and “strong”) as long as one uses an appropriate measure of association such as one of Kendall’s rank order coefficients.

Thus far I have argued that the correlation coefficient r is the wrong measure of persistence if persistence is understood as relative strength over time, and that the fix is easy: replace r with an appropriate rank order correlation coefficient. Yule’s $Q = .86$ for Key and Munger’s data, significantly higher than their reported $r = .68$. Key and Munger could well have claimed far more political continuity than they actually did. Yet there is a fly in the ointment, one that I suspect Key and Munger knew well: a weakness of the relative strength understanding of persistence is that it throws away much of the information in the data by discounting absolute differences among outcomes. The concomitant reduction in variance often inflates the value of the measure of association. To make matters worse, even perfect persistence is consistent with a wide variation in actual electoral outcomes since any unit within a quadrant is treated as equivalent to any other within the same quadrant. The more fundamental problem, then, is that the relative strength view provides a rather low bar for assessing persistence. Unfortunately, as we shall see in the next section, the problems with r do not go away, and arguably get worse, if we adopt an interpretation of persistence that incorporates all deviations from absolute agreement.

4 Conceptualizing Continuity: Absolute Agreement

The second understanding of continuity emphasizes the absolute agreement among outcomes across units. In this conceptualization the degree of persistence is understood as the extent to which a party in a later election exactly duplicates its performance from a prior election across some set of districts. The difference between relative strength and agreement of results is subtle but crucial. Both refer to concordances among outcomes, but the former compares rank orderings, while the latter treats even small differences as real deviations. Conceptually, agreement

³There are, in addition, degenerate cases of persistence, in which all points are either in the lower left or upper right quadrants, but not spread across both. Here the data would collapse to a single point.

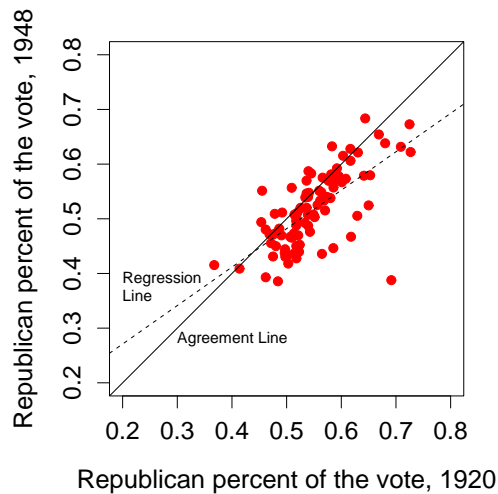


Figure 3: The traditional vote across Indiana counties (according to Key and Munger, 1959, Figure 15-2.), with lines of regression (dotted) and agreement (solid) added.

is more clear-cut than relative strength. We saw in the previous section that even conditional on a given definition of weakness and strength perfect relative strength is consistent with a wide variety of actual outcomes. Perfect agreement, by contrast, is produced in one and only one way: the latter results must be identical to the former ones for all units. The notion of absolute agreement also better accords with our intuitive notion of persistence as indicating the unchanged continuation of something from before.

Although a consideration of absolute differences would appear to augur well for the use of r , in fact it does not. The reason is that Pearson's r is a measure of linearity, not similarity: its magnitude indicates the degree of deviation around the points' regression line. Figure 3 depicts Key and Munger's data with the quadrants removed and the regression (dotted) and agreement (solid) lines added. The line of agreement is angled at 45 degrees. If the outcome in 1948 had been identical to that in 1920, then all the points would lie on this line. Any deviation from the line indicates lack of complete agreement. Note that the lines of agreement and regression coincide only where the estimated bivariate regression slope equals one and the intercept equals zero. Given how close the regression and agreement lines are to one another in this example, it is worth mentioning that the former is not statistically equivalent to the latter. The regression slope .70 with a standard error of .08, and the intercept is .13 with a standard error of .04. If rerun the analysis without the outlier in the lower right (Lake County), the slope remains different from one though the intercept does become zero.

The consequences of employing r when it is inappropriate may in fact be starker than in the Key and Munger example, where the regression and agreement lines are relatively close together, suggests. To see why consider the panels in Figure 4,

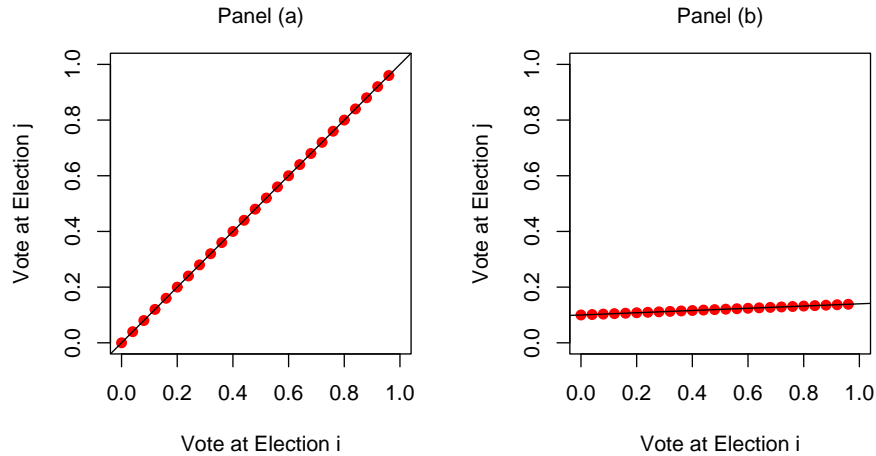


Figure 4: The Perils of r as Measure of Persistence. $r = 1$ for both panels, but panel (a) exhibits far more persistence than panel (b).

which feature two different versions of the basic schema. In each panel the electoral results across districts are perfectly correlated, yet only the points in panel (a) exhibit perfect persistence. The difference between the two lies in the intercepts and slopes of the associated regression lines. In panel (a) each outcome in election j is identical to that in election i , so that all the points lie on the 45 degree line having a slope of one and intercept of zero. In panel (b), by contrast, the regression line has slope .04 and intercept .1. For panel (b) this means that 25 percent support in a district in election i translates into 11 percent in election j , 50 percent translates into 12 percent, and 75 percent translates into 13 percent. Despite a perfect correlation among the points, there is virtually no similarity between performance in elections i and j . Clearly similar arguments hold if the points were spread about their corresponding regression lines rather than right on them.

Measuring persistence with reference to deviation about the 45 degree line of agreement rather than the points' corresponding regression line has the added benefit of enhancing our ability to compare levels of persistence across different research problems. Suppose two researchers using Pearson's r each found perfect continuity in their data. In the absence of other information there is no way to know whether the actual outcomes are closer to panel (a) or, say, panel (b), though for purposes of knowledge accumulation they would be treated as identical outcomes. In fact even more misleading situations could occur. If the points in panel (a) had been a little spread about the 45 degree line, the correlation coefficient would have dropped below one, and the result in panel (b) would have been deemed to exhibit the most similarity.

5 The Concordance Correlation Coefficient

Researchers have long been interested in assessing the degree of agreement across sets of continuous measurements, and have introduced a wide variety of pertinent indices (see Barnhart et al. (2007) for a review). I advocate the use of Lin's concordance

correlation coefficient, ρ_c (Lin 1989; 2000), a close relative of the far more widely-used (though seemingly not in political science) intraclass correlation coefficient.⁴ The principal advantage of ρ_c is that it is intuitive to understand, its statistically properties are well-understood, and the assumptions it imposes on the data are no more unreasonable than those implied in the use of the correlation coefficient. In fact, statistically speaking, ρ_c can be considered an analogue of ρ (the population Pearson correlation coefficient), but with the loss function computed with respect to the agreement rather than the regression line. Thus, like ρ , it yields a single number for the magnitude of concordance, and has a range of -1 to 1.

Following Lin (1989; 2000), assume that n independent pairs of election outcomes (v_{1i}, v_{2i}) , $i = 1, 2, \dots, n$, are taken from a bivariate population with a mean given by (μ_1, μ_2) and variance-covariance matrix $\begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}$.

Define the degree of concordance between v_1 and v_2 as the expected value of the squared perpendicular deviation from the 45 degree line (multiplied by 2):

$$E[(v_1 - v_2)^2] = (\mu_1 - \mu_2)^2 + (\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}) \quad (1)$$

To scale the index between -1 and 1 define the concordance correlation coefficient:

$$\rho_c = 1 - \frac{E[(v_1 - v_2)^2]}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2}, \quad (2)$$

where the denominator in the second term represents the expected squared deviation from the 45 degree line when v_1 and v_2 are uncorrelated, i.e., when such deviation is at its maximum.

Plugging in for $E[(v_1 - v_2)^2]$ and rearranging terms we then obtain:

$$\rho_c = \frac{2\sigma_{12}}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2} = \rho C_b, \quad (3)$$

where ρ is the Pearson correlation coefficient and C_b is a scale factor:

$$C_b = \frac{2\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2}. \quad (4)$$

ρ_c has the following desirable property: $-1 \leq -|\rho| \leq \rho_c \leq |\rho| \leq 1$. Thus, like ρ , the magnitude of ρ_c is bounded: $\rho_c = 1$ if and only if each pair of outcomes is in perfect agreement ($v_{2i} = v_{1i}$ for all i); $\rho_c = -1$ if and only if each pair is in perfect reverse agreement ($v_{2i} = -v_{1i}$ for all i). Moreover, $\rho_c = \rho$ if and only if $\sigma_1 = \sigma_2$ and $\mu_1 = \mu_2$. Taken together, these properties mean that except in the unusual case when the mean and variance of the two sets of electoral outcomes are the same, the concordance coefficient will always be smaller in absolute magnitude than the correlation coefficient. Thus every piece of research that employs correlation when it should be using concordance will be overestimating the degree of continuity.

⁴A full-text JSTOR search for the occurrence of “intra-class” and “correlation” within political science journals yielded only 44 hits. Contrast this with the search for “correlation coefficient”, which yielded 1299 hits. Accessed June 29, 2008. See Nickerson (1997) and Carrasco and Jover (2003) for the relationship between the concordance correlation coefficient and intraclass correlation coefficients.

If ρ represents the degree to which the observations deviate from the best-fit line, then C_b can be thought of as measure of how much that best-fit lines deviates from the line of absolute agreement. The smaller C_b is, the greater the deviation from the line of agreement. When $C_b = 1$, $\rho_c = \rho$, and the best fit and agreement lines coincide.

Inference is straightforward. If we plug sample values into the quantities on the right hand side of equation (3) we obtain

$$r_c = \hat{\rho}_c = \frac{2s_{12}}{s_1^2 + s_2^2 + (\bar{v}_1 - \bar{v}_2)^2}, \quad (5)$$

where s_1 , s_2 , and s_{12} are the variances and covariance of the two electoral outcomes, and \bar{v}_1 and \bar{v}_2 are the means of the two outcomes. Lin has shown that if we assume that our paired samples are drawn from a bivariate normal distribution, then r_c consistently estimates ρ_c with an asymptotic normal distribution centered at ρ_c and with variance:

$$\sigma_{r_c}^2 = \frac{1}{n-2} \left[\frac{(1-\rho^2)\rho_c^2(1-\rho_c^2)}{\rho^2} + \frac{2\rho_c^3(1-\rho_c)(\mu_1-\mu_2)^2}{\rho\sigma_1\sigma_2} - \frac{\rho_c^4(\mu_1-\mu_2)^4}{2\rho^2\sigma_1^2\sigma_2^2} \right]. \quad (6)$$

The normal approximation can be improved by implementing Fisher’s Z-transformation, yielding more realistic confidence intervals. (See Lin 1989, 2000 for details.) Lin’s concordance coefficient is implemented in the EpiCentre package for R and the concord package in Stata.

6 Replications

For Key and Munger’s data, $r = .68$, with a 95 percent confidence interval of (.56,.78), and $r_c = .61$, with a 95 percent confidence interval of (.48,.72). Thus, if Key and Munger had been interested in absolute agreement across elections, then in reporting r they overestimate the degree of Republican political continuity by only around 7 percentage points. Given the volume of subsequent research on the “standing electoral decision” it is fortunate that this early finding is robust even against the much more demanding standard of absolute agreement. In this section I replicate and apply these measures to two other analyses of political continuity: one that examines the roots of socialist strength in post-Franco Spain, and another that documents the return of the Right in post-communist Hungary.

6.1 Socialist Continuity in Spain

What accounts for the strength of the Spanish Socialist Party (PSOE) in the aftermath of the post-Franco transition to democracy? How the PSOE survived despite roughly four decades of repression and extensive economic upheavals under Franco’s dictatorship has remained a puzzle. Linz (1980) assesses the extent of continuity in voting patterns by reporting the correlation across provinces between the 1936 and 1977 vote for a series of parties and blocs, including the Socialists. He finds

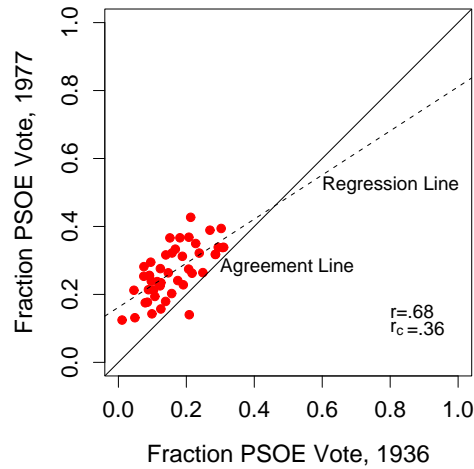


Figure 5: Support for the socialist PSOE across provinces in Spain, 1936-1977, with lines of regression (dotted) and agreement (solid) added.

overall discontinuity in voting patterns except in the case of the Socialists, where he reports $r = .6$ for the two elections. Maravall (1982: 173-174), who terms the geographic continuities in PSOE support “remarkable,” attributes this persistence to the power of political memory: “Communities, parties and families would have acted as the ‘social carriers’ of ideologies, beyond the restrictions and repressions of a non-democratic regime.”

Or did they? I attempted to replicate Linz’s and Maravall’s findings. The provincial-level electoral data for the 1977 elections to the constituent assembly are available in Caramani (2000: 83-84). The data for the last pre-Franco elections in 1936 are reported in Linz and Miguel (1977). Neither Linz nor Maravall indicate the need to account for changes in provincial borders between the two elections. I thus assume that the names of the 52 provinces, which seem not to have changed between the two elections, continue to refer to the same patches of territory.

I was not able to exactly replicate the PSOE vote correlation of $r = .6$. This is no doubt due to slightly different sample sizes and subtleties in the data. Maravall reports that he uses all 52 provinces but I could find PSOE data from 1936 for only 48 provinces. Moreover, it is not clear whether Linz or Maravall is employing “pueblo” or “capital” data, both of which are available for 1936. Nonetheless, if the tiny Spanish provinces that are enclaves in Africa, Ceuta and Melilla, are also excluded from the analysis, and the pueblo data are used, then we can compute $r = .68$, with a 95 percent confidence interval of (.48,.81). This value is nearly 10 percent higher than what they report. The regression and agreement lines are not statistically identical: the slope is .65 (.43,.87), and the intercept is .16 (.12,.20).

The data, the lines of regression and agreement, and the values of both the correlation and concordance coefficients are illustrated in Figure 5. Here we see, in contrast with the Key and Munger example, the perils of using the regression rather

than the agreement line as the baseline against which persistence is measured. First, $r_c = .36$, with a 95 percent confidence interval of (.21,.49). The magnitude of the concordance coefficient is barely more than half that of the ordinary correlation. There is thus not nearly as much regional persistence in support of the PSOE as has been commonly assumed. Though provinces close to the line of agreement certainly do exhibit continuity, this cannot be said for other areas. Second, the reasons for the lack of continuity are not that former PSOE strongholds were decimated by the Franco regime. If that were true the regression line would be flatter, and many more one-time PSOE bastions would have ended up below the line of agreement (having lost support between 1936 and 1977). Rather, PSOE discontinuity occurred because it gained strength in every region but one (Las Palmas).

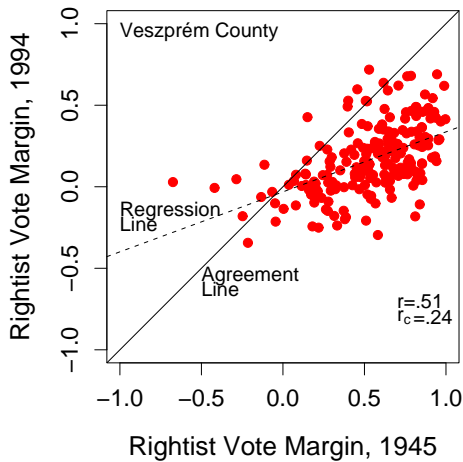
The partisan swing in favor of the PSOE between 1936 and 1977 does reveal one circumstance in which r can serve as an informative indicator of continuity. If there is an intercept shift but not a slope shift, as seems nearly the case here (and might well have actually been the case had there been more variance in the 1936 outcome), then the partisan swing can be interpreted as a uniform election-specific effect for 1977. For example, independently of any of its activities under the Franco dictatorship, the PSOE may well have received an extra bump in 1977 by virtue of voters eager to see new faces in power. If this bump were of theoretical interest to the researcher, then r_c would remain the way to go. However, if the researcher wished to remove the effect of the bump, which effectively shifts the points downward so that the intercept becomes zero, then r does become the better measure.

6.2 Rightist Continuity in Hungary

To what extent did communism transform the partisan political profile of the Hungarian electorate? Unlike in Spain, where the PSOE managed to survive the Franco dictatorship, in Hungary four decades of communism largely destroyed attachments to individual parties. What survived was more diffuse partisan loyalties to leftist and rightist parties. Wittenberg (2006) explores continuities in Hungarian support for rightist parties between pre- and post-communism by comparing municipal-level electoral outcomes from the 1945 national parliamentary elections with corresponding outcomes from the 1990, 1994, and 1998 elections. Although he discusses the differences between agreement and correlation (see pp. 71-73), in the end he assesses persistence by correlating the results for each of Hungary's 19 provinces. He notes remarkable rightist persistence between 1945 and 1994, where the correlations exceed 0.5 in three provinces: Komárom, Pest, and Veszprém (see the map on p.69). I replicated these results for these three provinces of high continuity and for Hungary as a whole.

The results appear in Figure 6, consisting of a graph on the left and a table on the right. The graph on the left, of the municipalities in the province of Veszprém, is similar to those already reported for Indiana and Spain, with lines of regression and agreement as well as the values of the correlation and concordance coefficients.⁵

⁵The Hungarian party system changed from one dominated by two blocs in 1945 to domination by three blocs in 1994. To compensate for this the victory margins of the Right over the Left are correlated rather than actual support for the Right and Left. The axes thus range from -1,



| | r | r_c | n |
|----------|------------------|------------------|------|
| Komárom | .56 (.34,.68) | .34 (.20,.47) | 65 |
| Pest | .49 (.36,.60) | .30 (.21,.39) | 154 |
| Veszprém | .51 (.41,.61) | .24 (.17,.30) | 211 |
| Hungary | .38 (.35,.42) | .21 (.19,.23) | 2851 |

Figure 6: Rightist persistence in Hungary, 1945-1994. The graph on the left illustrates the degree of persistence across municipalities in Veszprém county; the table on the right lists both correlation (r) and concordance (r_c) coefficients for three high-persistence provinces and the entire country.

The table on the right lists both r and r_c for Veszprém in addition to other high-persistence provinces and Hungary. As in the case of Spain, these results illustrate the dangers of reliance on the correlation coefficient to measure continuity. Significantly, r_c is substantially less than r in all cases. Though this does not preclude persistence in municipalities close to the agreement line, it does indicate that as a description for communities as a whole in these provinces, rightist continuity is less prevalent than previously thought.

7 Discussion

Thus far I have argued that Yule's Q and r_c are superior to Pearson's r for problems of political continuity that can be characterized by a scatterplot of units such as the basic schema of Figure 1. However, they are not silver bullets. First, like r and other correlational measures, their magnitudes are sensitive to sample heterogeneity: high values may reflect large sample variances rather than strength in the underlying relationship between the two outcomes. Though sample heterogeneity is not an issue if the purpose of the analysis is purely descriptive, it does otherwise render inference difficult (see Achen 1977; Atkinson and Nevill 1997).

Second, the approach advocated here presupposes that the quantities the researcher has chosen to compare are appropriate for the question under consideration. My argument takes no position on whether, say, Key and Munger ought to have been comparing the 1920 and 1948 Republican vote as evidence of political persistence. Such questions can be tricky. For example, for illustrating absolute agreement, com-

representing a purely leftist settlement, to 1, representing a purely rightist one.

paring electoral outcomes for particular years may not be an ideal strategy. The reason is that no two election outcomes are ever *exactly* the same across a set of districts, even if identical theoretical processes are at work in both periods. Random error will perturb the results and thus decrease the value of r_c , at least slightly. In such a case $r_c < 1$ even though by assumption there should be perfect continuity. In practice, of course, there are also almost always election-specific effects that may or may not be of interest to the researcher but enter into the computation when particular electoral outcomes are used. We saw a special case of this in the partisan swing in favor of the Socialists in Spain between 1936 and 1977. More generally, if a researcher wishes to minimize both random error and election-specific effects, one approach is to compare “normal” votes by replacing the two individual outcomes with averaged outcomes from neighboring elections.

Third, correlating units such as precincts, municipalities, or provinces that are themselves aggregations of individual voters requires careful consideration of the ecological fallacy. Suppose two elections are temporally proximate, say separated by two years, and the electoral data are available at the municipality-level. In such an instance the quantity of interest for many researchers interested in electoral persistence would be the fraction of voters who choose the same party in both elections. Perfect continuity would obtain if the party’s constituency at the second election were identical with that at the first election. The ecological fallacy tells us that r_c or any other correlational measure would serve as a very unreliable estimate of this quantity. A better strategy, assuming the absence of relevant survey data and that the same population is voting in each election, is to estimate voter transition probabilities using ecological inference techniques (see Achen and Shively 1995; King 1997). Such techniques yield direct estimates of the proportions of a party’s voters at the first election that remain loyal or defect at the second election.⁶

Although the ecological fallacy would appear to severely limit the utility of correlational measures for measuring electoral continuity, in fact it does not. Ecological inference is applicable only when the elections are proximate enough in time to render plausible the assumption that the same population is voting in both elections. However, researchers of persistence are typically interested in stability over long periods of time, often decades. This is certainly true of work on redemocratization but also quite prevalent in studies of partisan realignment in the US. The further apart in time the two elections are, the greater the transformation in the voting population due to factors such as death, the entry of new voters, and in- and out-migration, and the less reasonable the assumption that the population remains the same. Where the elections are separated by many decades there is thus no constant population whose party loyalty over time can be estimated, and correlational measures remain preferable.

⁶If the researcher were interested only in the degree to which the party duplicates its results from the prior election, regardless of mechanisms producing that duplication, then r_c would be the appropriate measure.

8 Conclusion

The procedures advocated in this paper have been introduced in the context of a comparison of two continuously-measured electoral outcomes across a set of districts. However, the measures are actually relevant for any two-way comparison where the expected relationship between the variables is one of agreement rather than linearity. For example, closely related to the idea of persistence as discussed here is the concept of electoral volatility, which is a summary of the vote changes experienced by all parties in a party system between two elections. Another example would be to measure deviation from pure proportional representation by comparing vote and parliamentary seat shares. Taagepera and Grofman (2003) evaluate 19 previously proposed indices of electoral volatility and vote-seat disproportionality, but it is unclear how well their statistical properties are understood. It might be possible to modify the concordance coefficient to serve as an alternative and statistically better-grounded general measure. A related approach would be to build on Morgenstern and Potthoff (2005), who eschew unidimensional indices of stability in favor of separately analyzing over-time and cross-district components. For an example in the case of the concordance coefficient, see Carrasco and Jover (2003).

Wand et al. (2001) and Mebane and Sekhon (2004) suggest that ecological electoral data may be especially prone to outliers, and advocate robust estimators rather than OLS to prevent these outliers from contaminating the analysis. The concordance coefficient relies on many of the same assumptions as OLS, so another avenue for further research would be to adopt a robust version that is specifically suited to the peculiarities of electoral data. King and Chinchilli (2001) and Haber and Barnhart (2008) provide an excellent starting point.

There is no reason to limit things to electoral data. Some version of the concordance coefficient could serve as an all-purpose goodness-of-fit measure. Perfect model fit obtains when the predicted values from a model equal the actual values in the data, a situation tailor-made for an agreement measure. Indeed, every argument made in this paper against r also applies to its square, the much-maligned R^2 . Vonesh, Chinchilli, and Pu (1996) have adapted the concordance coefficient to measure fit for a range of non-linear models.

Shapiro (2002: 616) argues for the value of “problematizing redescription” in the study of politics, noting that “[i]t is intrinsically worthwhile to unmask an accepted depiction as inadequate and to make a convincing case for an alternative as more apt.” I have tried to engage in such a task in this paper, which has argued that prior studies of electoral continuity have mischaracterized the degree of persistence by employing Pearson’s r rather than a more appropriate measure. At best r indicates a very weak form of continuity; at worst, and almost always in practice, it is fundamentally misleading, implying continuity when in fact there is discontinuity. In most circumstances it should be abandoned.

Bibliography

- Achen, Christopher H. 1977. "Measuring Representation: Perils of the Correlation Coefficient," *American Journal of Political Science*, Vol. 21, No. 4, November, pp. 805-815.
- Achen, Christopher H. and W. Phillips Shively. 1995. *Cross-level inference*. Chicago: University of Chicago Press.
- Atkinson, Greg and Alan Nevill. 1997. "Comment on the Use of Concordance Correlation to Assess the Agreement between Two Variables," *Biometrics*, Vol. 53, No. 2, June, pp. 775-777.
- Barnhart, Huiman X, Michael J. Haber, and Lawrence I. Lin. 2007. "An Overview on Assessing Agreement with Continuous Measurements," *Journal of Biopharmaceutical Statistics*, Vol. 17, pp. 529-569.
- Burnham, Walter Dean. 1968. "American Voting Behavior in the 1964 Election," *Midwest Journal of Political Science*, Volume XII, Number 1, February, pp. 1-40.
- Caramani, Daniele. 2000. *Elections in Western Europe Since 1815: Electoral Results by Constituencies*. London: MacMillan.
- Carrasco, Josep L. and Lluís Jover. 2003. "Estimating the Generalized Concordance Correlation Coefficient through Variance Components," *Biometrics*, Vol. 59, December, pp. 849-858.
- Gimpel, James G. and Jason E. Schuknecht, 2002. "Rethinking Political Regionalism in the American States," *State Politics and Policy Quarterly*, Volume 2, Number 4, Winter, pp. 325-352.
- Haber, Michael and Huiman X. Barnhart, 2008. "A general approach to evaluating agreement between two observers or methods of measurement from quantitative data with replicated measurements," *Statistical Methods in Medical Research*, Vol. 17, pp. 151-169.
- Key, V.O., Jr. and Frank Munger, 1959. "Social Determinism and Electoral Decision: the Case of Indiana," in Eugene Burdick and Arthur J. Broadbeck, editors. 1959. *American Voting Behavior*. Westport, CT: Greenwood Press. pp. 281-299.
- King, Gary. 1986. "How Not to Lie with Statistics: Avoiding Common Mistakes in Quantitative Political Science," *American Journal of Political Science*, Vol. 30, No. 3, August, pp. 666-687.
- King, Gary. 1997. *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton: Princeton University Press.
- King, Tonya S. and Vernon M. Chinchilli. 2001. "Robust Estimators of the Concordance Correlation Coefficient," *Journal of Biopharmaceutical Statistics*, Vol. 11, No. 3, pp. 83-105.
- Levine, Marc V. 1976. "Standing Political Decisions and Critical Realignment: The Pattern of Maryland Politics, 1872-1948," *Journal of Politics*, Volume 38, No. 2, May, pp. 292-325.
- Lin, Lawrence I-Kuei. 1989. "A Concordance Correlation Coefficient to Evaluate Reproducibility," *Biometrics*, March, Vol. 45, pp. 255-268.

- Lin, Lawrence I-Kuei. 2000. "A Note of the Concordance Correlation Coefficient," *Biometrics*, March, Vol. 56, pp. 324-325.
- Linz, Juan J. 1980. "The New Spanish Party System," in Richard Rose, editor. 1980. *Electoral Participation: A Comparative Analysis*. Beverly Hills, CA: Sage, pp. 101-189.
- Linz, Juan J. and Jesus M. De Miguel. 1977. "Hacia un análisis regional de las elecciones de 1936 en España," *Revista Española De La Opinión Pública*, No. 48 (Abril-Juno), pp. 27-68.
- Montes, J. Esteban, Scott Mainwaring, and Eugenio Ortega, 2000. "Rethinking the Chilean Party Systems," *Journal of Latin American Studies*, Volume 32, pp. 795-824.
- Maravall, José. 1982. *The Transition to Democracy in Spain*. London & Canberra: Croom Helm.
- Mebane, Walter R, Jr. and Jasjeet S. Sekhon. 2004. "Robust Estimation and Outlier Detection for Overdispersed Multinomial Models of Count Data," *American Journal of Political Science*, Vol. 48, No. 2, April, pp. 392-411.
- Morgenstern, Scott and Richard F. Pothoff, 2005. "The components of elections: district heterogeneity, district-time effects, and volatility," *Electoral Studies*, Vol. 24, pp. 17-40.
- Nickerson, Carol A. E. 1997. "A Note on 'A Concordance Correlation Coefficient to Evaluate Reproducibility,'" *Biometrics*, Vol. 53, December, pp. 1503-1507.
- Robinson, W. S. 1957. "The Statistical Measurement of Agreement," *American Sociological Review*, Vol. 22, No. 1, February, pp. 17-25.
- Shapiro, Ian. 2002. "Problems, Methods, and Theories an The Study of Politics, Or What's Wrong With Political Science and What To Do About It," *Political Theory* 30, pp. 596-619.
- Taagepera, Rein and Bernard Grofman, 2003. "Mapping the Indices of Seats-Votes Disproportionality and Inter-Election Volatility," *Party Politics*, Vol. 9, No. 6, pp. 659-677.
- Valenzuela, Samuel J. and Timothy R. Scully. 1997. "Electoral Choices and the Party System in Chile: Continuities and Changes at the Recovery of Democracy," *Comparative Political Studies*, Volume 29, No. 4, July, pp. 511-527.
- Vonesh, Edward F., Vernon M. Chinchilli, and Kewei Pu, 1996. "Goodness-of-Fit in Generalized Nonlinear Mixed-Effect Models," *Biometrics*, Vol. 52, June, pp. 572-587.
- Wand, Jonathan N., Kenneth W. Shotts, Jasjeet S. Sekhon, Walter R. Mebane, Jr., Michael C. Herron, and Henry E. Brady, 2001. "The Butterfly Did It: The Aberrant Vote for Buchanan in Palm Beach County, Florida," *American Political Science Review*, Vol. 95, No. 4, December, pp. 793-810.
- Wittenberg, Jason. 2006. *Crucibles of Political Loyalty*. Cambridge: Cambridge University Press.