Research Note

Research and Politics
October-December 2017: 1–7
© The Author(s) 2017
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/2053168017732231
journals.sagepub.com/home/rap
SAGE

Moving forward with time series analysis

Peter K. Enns[1], Nathan J. Kelly[2], Takaaki Masaki[3] and Patrick C. Wohlfarth[4]

Abstract

In a recent *Research and Politics* article, we showed that for many types of time series data, concerns about spurious relationships can be overcome by following standard procedures associated with cointegration tests and the general error correction model (GECM). Matthew Lebo and Patrick Kraft (LK) incorrectly argue that our recommended approach will lead researchers to identify false (i.e., spurious) relationships. In this article, we show how LK's response is incorrect or misleading in multiple ways. Most importantly, when we correct their simulations, their results reinforce our previous findings, highlighting the utility of the GECM when estimated and interpreted correctly.

Keywords

Cointegration, general error correction model, error correction model, time series

Introduction

We are grateful for the opportunity to continue the dialogue about appropriate applications of the general error correction model (GECM) with Matthew Lebo and his coauthors. Although this discussion has been underway for several years now,[1] our first article on the topic followed a *Political Analysis* time series symposium, where Grant and Lebo (2016) (GL) and Lebo and Grant (2016) (LG) argued that the GECM is rarely (if ever) appropriate with political data. Like many time series researchers, much of their concern stemmed from the potential for estimating spurious relationships. Our article – Enns et al. (2016) (EKMW) – showed that GL were far too skeptical of the ongoing utility of the GECM. When *Y* contains a unit root, when *Y* is bounded and contains a unit root, when *Y* is stationary, or when *Y* is near-integrated (i.e., $\rho \geq 0.90$), LG's concerns about spurious relationships are easily overcome by following standard procedures associated with cointegration tests and the GECM.[2] Specifically, to conclude that cointegration exists with a GECM, researchers should: (1) conduct statistical tests to confirm that *Y* and *X* contain unit roots (our simulations used augmented Dickey–Fuller (ADF) tests);[3] (2) confirm that the error correction model (ECM) parameter (associated with Y_{t-1}) is statistically significant using appropriate MacKinnon critical values; and (3) confirm that the coefficient for the lag of *X* is statistically significant.[4] We showed that if all three of these conditions are met, the Type-I error rate for the estimated relationship between *X* and *Y* falls at, or below, the standard 5% threshold.

Much of our evidence relied on GL's own simulation results. Lebo and Kraft (2017) (LK) now conduct new simulations in an effort to show that our approach will routinely lead researchers to identify false (i.e., spurious) relationships. They also conduct simulations which suggest that the negative bias on the error correction parameter is much more severe than we report. A careful look at LK's response, however, shows that it does *not* undermine our conclusions and that it is easy to reconcile the seemingly disparate recommendations. In fact, had LK followed exactly our recommended procedure, their simulation results would have looked extremely different and would, in fact, support our conclusions.

[1]Cornell University, Ithaca, NY, USA
[2]University of Tennessee, Knoxville, TN, USA
[3]College of William & Mary, Williamsburg, VA, USA
[4]University of Maryland, College Park, MD, USA

Corresponding author:
Peter K. Enns, Cornell University, 205 White Hall, Ithaca, NY 14850, USA.
Email: peterenns@cornell.edu

Our advice does not over-produce false-positives

Based on 60,000 simulations, LK conclude that the ADF test is “drastically underpowered” to reject the null hypothesis of a unit root (Lebo and Kraft, 2017: 4). This, of course, is a well-known finding (e.g., Blough, 1992; Cochrane, 1991), and LK actually quote us making the exact same point in our article (Lebo and Kraft, 2017: 3). It is important to remember *why* we chose to use an underpowered test in our simulations. Just three lines below the sentence LK quoted, we explain: “this means we are biasing our simulations against support for the GECM since we are more likely to incorrectly conclude the series contains a unit root and thus inappropriately utilize the GECM as a test of cointegration (thereby inflating the rate of Type-I errors with those cointegration tests).” For the four types of data that we analyzed, the Type-I error rate using 0.05 *p*-values approximated the expected 5%. Using stronger unit-root tests would only reduce the rate of spurious findings. Thus, what LK present as a bug was actually a feature of our analysis.

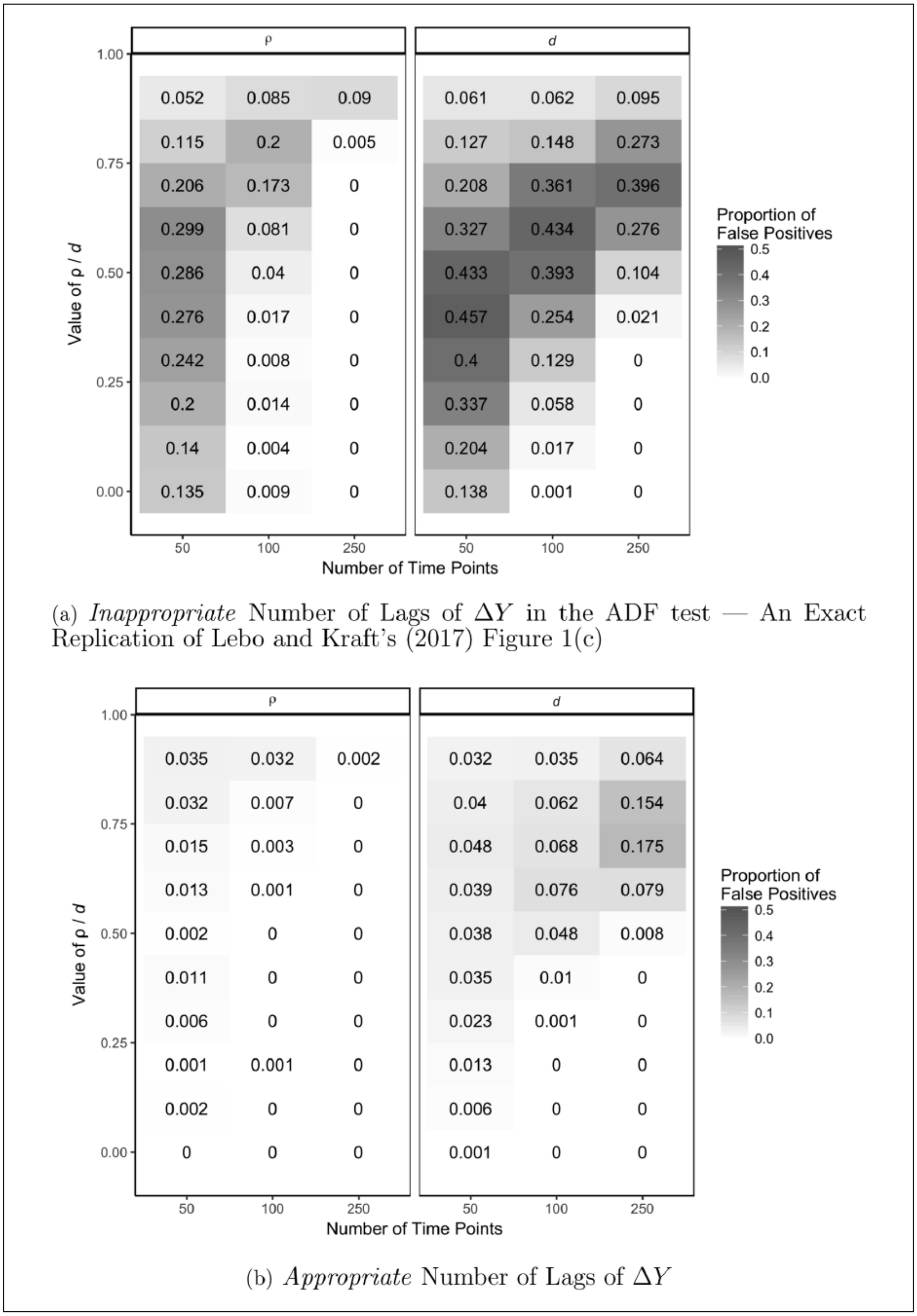
What, then, are we to make of LK’s simulation results in their Figure 1(c), which claim to follow the “exact procedures” we advocate and report false-positive rates greater than 5% in 38 out of 60 sets of simulations when *Y* is stationary or fractionally integrated? A review of LK’s approach reveals that their spurious relationships emerge because they did *not* follow our “exact procedures.”[5]

Lebo and Kraft’s first oversight results because they incorrectly used the **adf.test** function in the “tseries” R package, which includes a default number of lags for ΔY that equals $(T-1)^{\frac{1}{3}}$ in the ADF test.[6] For instance, if $T = 50$, the adf.test package includes 3 lags of differenced *Y*. While this default could be appropriate in some settings, given the data-generating processes employed by LK, we would not expect the ADF to require 3 lags to eliminate serial correlation in the residuals from the ADF test (additional lags of ΔY are typically added until the ADF test produces white noise residuals (Box-Steffensmeier et al., 2014: 134)). Including too many lags (as LK did) is problematic because, as Box-Steffensmeier et al. (2014: 135) explain, “the probability that we incorrectly diagnose a unit root increases.” In other words, by relying on the default lag length, LK risk incorrectly concluding that *Y* contains a unit root. This inappropriate diagnosis would lead to estimating the GECM when they should not, which would inflate the Type-I error rate. In our own simulations, we reproduce this result from LK in Figure 1(a). In Figure 1(b), by contrast, we replicate LK’s analysis but determine the number of lags of ΔY to include based on the specification that produces white noise residuals in the ADF regression, as suggested by Box-Steffensmeier et al. (2014). This is also the procedure we followed with our original simulations. Just by employing the ADF appropriately (which addresses LK’s first oversight), we have solved the spurious regression problem that LK report when *Y* is stationary and nearly solved the problem when *Y* is fractionally integrated.[7]

Researchers should also be aware, however, that LK skipped two other steps that are necessary to conclude that cointegration is present. First, *both X* and *Y* should be tested for a unit root before utilizing the ECM parameter as a test of cointegration. Second, even if both series showed evidence of a unit root and the ECM parameter was significant with the MacKinnon critical values, the estimated coefficient on X_{t-1} should also be significant (using traditional critical values) before concluding that there is a long-term relationship between *X* and *Y*. Consistent with Enns et al. (2016), when we follow all of the necessary steps, the false-positive rate is at or below 0.05 in every set of simulations reported above except for two (where the false-positive rate is 0.06 and 0.08) (see Appendix, Figure A1).

Instead of concerns about potential spurious relationships, LK’s Figures 1 (a) and (b) focus on the point estimate of the error correction parameter. On one hand, we want to be careful not to place too much emphasis on this point estimate. Researchers are typically most interested in whether a relationship exists between *X* and *Y*. Although the rate of error correction can be informative, this is generally not the quantity of primary interest upon which tests of substantive theories critically depend. However, even if not of primary interest, we feel that researchers should be made aware that LK’s results again reflect a fundamental error and are thus misleading.

Recall that LK estimated a bivariate GECM with two unrelated series with varying data-generating processes. LK’s Figures 1(a) and 1(b) plot the mean value of the ECM parameter ($\hat{\alpha}_1^*$) on the X-axis. LK argue this value should be equal to zero indicating no cointegrating relationship (Lebo and Kraft, 2017: 4). Unfortunately, this logic is flawed because it depends on an improper application of the GECM. The problem arises because LK report the mean value of the ECM parameter for all series they generated – *even the ones where they rejected the null of a unit root using the ADF test*. These estimates should never have been interpreted as ECM parameters because the failure to accept the null hypothesis of a unit root in *Y* means the data fail to satisfy the *first* requirement of evidence of cointegration with a GECM. As LK explain, “α_1^* is not a cointegration test” with stationary time series (Lebo and Kraft, 2017: 3). By treating the estimates of α_1^* as ECM parameters – even when they reject the null of a unit root – LK bias their estimates in a negative direction. The left side of their Figure 1(a) (Lebo and Kraft, 2017: 4) can be used to illustrate the severity of this bias. When $T = 250$ and $\rho = 0.0$, *none* of the simulations accept the null hypothesis of a unit root, which means that *none* of the estimates of α_1^* should be interpreted as an ECM parameter. Instead, because they incorrectly treat all of the estimates of α_1^* as ECM parameters, LK report that the mean value of the ECM parameter is approximately -1.0. This result is wrong and it misrepresents our recommendations and the performance of the GECM.[8]



(a) *Inappropriate* Number of Lags of ΔY in the ADF test — An Exact Replication of Lebo and Kraft's (2017) Figure 1(c)

(b) *Appropriate* Number of Lags of ΔY

Figure 1. The proportion of false-positives when the augmented Dickey–Fuller test includes the *Inappropriate* vs. *Appropriate* number of lags of ΔY.

Note: As explained in the text, these false-positive rates ignore two other necessary steps to implement the general error correction model.

Some further points of clarification

We were surprised by the statement, "Enns et al. provide no justification for expanding when these [critical] values should be used — to NI [near integrated] data, FI [fractionally integrated] data, or any other type. Yes, Enns et al.'s advice prevents some spurious findings but that does not mean they are the correct critical values" (Lebo and Kraft, 2017: 6). This statement is misleading for three reasons. First, LK suggest we had no justification for the critical values we used, but as we explained in our original article, the simulation results we presented for near integrated data come directly from GL's tables G.1–G.5. In other words, we relied on results that they reported based on MacKinnon critical values. Second, we did offer a theoretical justification for using these critical values (see, especially, Enns et al., 2016: 6–7, 9, and note 21). Third, we did not simply show that using these critical values "prevents some spurious findings." We showed that the false-positive rate was approximately 5 percent or less with these values.

We also disagree with LK's suggestion that truly cointegrated series will mimic Stock and Watson's (2011) textbook example of cointegration (reported in LK's Figure 2). To highlight the importance of the Stock and Watson example, LK reference Lebo's previous work, stating: "Error correction between variables is a very close relationship that should be obvious in a simple glance at the data" (Lebo and Grant, 2016: 22). We are strong proponents for the utility of plotting time series. However, identifying a single textbook example of cointegration and using it as the benchmark for future analyses is overly simplistic. LK could have just as easily pointed to Enders's (2014) textbook example of three cointegrated series (shown in Figure 2), which appears more similar to the Kelly and Enns data shown in LK's Figure 4. But, relying on Enders's figure would be equally problematic. The problem, of course, is that the choice of figure – as well as subjective assessments comparing applied data to the chosen figure – involves substantial researcher discretion. To avoid this subjectivity, we conducted simulations which show that we would expect to falsely reject the true null hypothesis only about 5% of the time if researchers use the procedure we highlight. Researchers should definitely plot their data, but they should also use systematic statistical tests to evaluate whether cointegration exists.

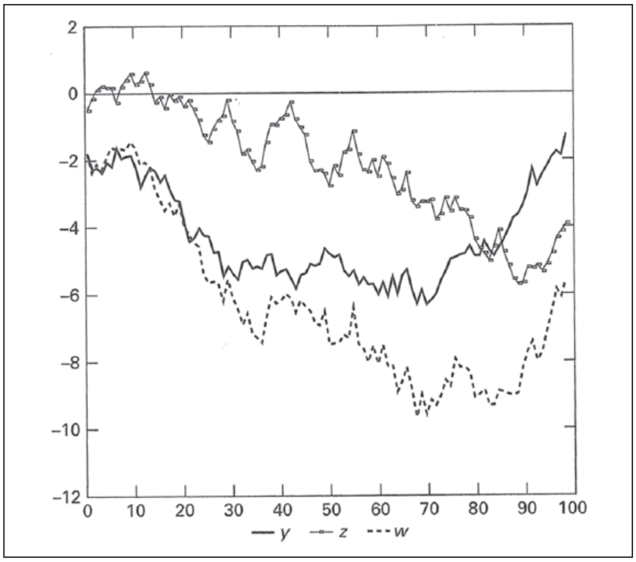


Figure 2. Enders's (2014) Figure 6.2 from *Applied Economic Time Series* (2014) (reprinted with permission from John Wiley & Sons.).

Advancing methodological debates

We have always been eager to advance our methodological understanding, even when it requires us to reconsider our previous work (e.g., Enns et al., 2014, 2016). However, our experience with this exchange suggests some general insights about how to engage usefully and constructively within a methodological debate.

First, even when the primary debate is a methodological one, existing substantive theory and research should be engaged and treated seriously. For example, not only has a sizeable literature explored – and found – a relationship between public opinion and Supreme Court decisions (e.g., Enns and Wohlfarth, 2013; Epstein and Martin, 2011; Flemming and Wood, 1997; Link, 1995; McGuire and Stimson, 2004; Mishler and Sheehan, 1993, 1996), GL found such a relationship using *our* data and *their* preferred fractional integration (FI) methods (Grant and Lebo, 2016: 23). These results should be acknowledged when critiquing literature on this topic. Similarly, any critique of research on the relationship between inequality and support for redistribution should acknowledge formal (Shayo, 2009), experimental (Trump, forthcoming), cross-national (Cavaillé and Trump, 2015), and other time series (Luttig, 2013) analyses that are consistent with the argument being critiqued. Methodological discussions and conclusions can be improved by paying attention to existing substantive literature, theoretical arguments, and related analyses.

Second, to advance the methods literature, it is most helpful to build a positive case for a new method, or a broader application of an existing one. In this instance, we think it would be extremely beneficial to make a positive case for the FI techniques advocated by GL and LG. We would be very interested in further incorporating FI techniques into our research, but as we pointed out in our previous article, we believe three aspects of FI still need to be tackled. First, concerns with estimating the FI parameter, *d*, with short time series must be addressed. Second, LG's "practical guide" to estimating *d* ignores the many choices involved and the fact that estimates can be highly sensitive to these choices.[9] Finally, our past work has shown that there is reason to question whether the three-step fractional error correction model (FECM) approach that GL recommend can reliably detect true relationships in the data (see also Enns and Wlezien, 2017). Validating FI methods in a variety of contexts and offering a realistic guide for implementation would provide an important service to the discipline.

The heart of time series methodology involves balancing the many tradeoffs inherent in applied modeling to minimize errors and avoid incorrect inferences when testing substantive theory. Although we share Lebo and his coauthors' concern that research continues to be published in top political science journals that uses the GECM incorrectly because MacKinnon critical values are ignored, we have shown that after correcting the errors in Lebo and Kraft (2017), their simulations reaffirm the conclusions of Enns et al. (2016). While care must certainly be taken, a fairly straightforward procedure can protect applied time-series researchers against false-positives when attempting to estimate relationships with many types of data that are common in social science research.

Declaration of Conflicting Interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. This exchange dates back to 2013 when three of us shared a conference paper with Matthew Lebo (see, Enns et al., (2014)). That paper showed that much political science research (including some of our own) incorrectly interpreted De Boef and Keele (2008) to imply that the general error correction model (GECM) was more flexible than it is and we emphasized that using the correct MacKinnon critical values was an important part of cointegration tests when estimating the GECM with nonstationary series (also see Enns et al., 2016). Those insights remain essential points of agreement that Lebo and Kraft (2017) identify.
2. We also discussed fractionally integrated series, but that discussion focused on: (1) explaining why Lebo and Grant's (LG's) conclusions seemed to (but did not actually) contradict Esarey (2016); and (2) showing readers that estimating the *d* parameter with fractional integration techniques is much more complicated than LG acknowledge.
3. As described below, we used the augmented Dickey–Fuller test because it offered a conservative test in the context of our simulations, but considering multiple unit root tests is often advised (e.g., Box-Steffensmeier et al., 2014).
4. In a multivariate setting with more than one X, the situation becomes somewhat more complicated, but our past simulations show that with two I(1) predictors, one that is cointegrated with Y and the other that is unrelated to Y, the false-positive rate is 5%-6% when T={30, 60, 100} (Enns et al., 2014).
5. We should clarify that we originally did not analyze the types of fractionally integrated (FI) series that Lebo and Kraft (LK) analyze in their Figure 1(c). Esarey (2016) considered FI series where *d* ranged from 0 to 0.45 and we simply offered an explanation for why Lebo and Grant obtained different results when analyzing the same FI series. Thus, LK's analysis of FI series where *d* ranges from 0.5 to 1.0 departs significantly from our article. As we show below, however, our new simulations indicate that a proper application of the general error correction model would not excessively produce false-positives.
6. Following this calculation, adf.test drops the decimal and retains the integer.
7. Although other methods for selecting the lag length in augmented Dickey–Fuller tests exist (e.g., Agunloye et al., 2013; Cavaliere et al., 2015), selecting the lag length based on white noise residuals (e.g., Box-Steffensmeier et al., 2014: 134), eliminates the spurious regression concern when the general error correction model is appropriately estimated. Figure A2, in the Appendix, shows that the results are virtually identical if we use the Breusch–Godfrey test for serial correlation to determine the number of lags.
8. Although Lebo and Kraft's results cannot offer accurate information about the bias in the error correction model (ECM) parameter, we show in Enns et al. (2016) that there are situations when the ECM parameter is biased in a negative direction and this bias should be taken seriously, particularly as this bias would lead researchers to conclude faster rates of error correction than in fact exist. Fortunately, we found that this bias tends to be small, it affects estimates of the long-run multiplier in a conservative direction, and it decreases as T increases.
9. Instead of clarifying the choices involved in estimating *d*, Lebo and Kraft (LK's) discussion of Casillas et al. (2011) ignores alternative augmented Dickey–Fuller (ADF), autoregressive fractionally integrated moving average (ARFIMA), and autoregressive integrated moving average (ARIMA) tests that support the conclusion that two of the series analyzed by Casillas et al. (2011) contain unit roots. In seeming contrast to Grant and Lebo's recommendation that, "decisions should be made based on rigorous testing of the data in hand using unit-root tests and direct estimates of the fractional integration parameter" (Grant and Lebo, 2016: 72), LK simply assert that these variables are "very unlikely to contain unit roots" because they are "computed anew each year based on the Court's decisions" (Lebo and Kraft, 2017: 8).

Carnegie Corporation of New York Grant

This publication was made possible (in part) by a grant from Carnegie Corporation of New York. The statements made and views expressed are solely the responsibility of the author.

References

Agunloye OK, Arnab R and Shangodoyin DK (2013) A new criterion for lag-length selection in unit root tests. *American Journal of Theoretical and Applied Statistics* 2(6): 293–298.

Blough SR (1992) The relationship between power and level for generic unit root tests in finite samples. *Journal of Applied Econometrics* 7(3): 295–308.

Box-Steffensmeier JM, JR, Hitt MP, et al. (2014) *Time Series Analysis for the Social Sciences*. New York, NY: Cambridge University Press.

Casillas CJ, Enns PK and Wohlfarth PC (2011) How public opinion constrains the U.S. Supreme Court. *American Journal of Political Science* 55(1): 74–88.

Cavaillé C and Trump K-S (2015) The two facets of social policy preferences. *Journal of Politics* 77(1): 146–160.

Cavaliere G, Phillips PCB, Smeekes S, et al. (2015) Lag length selection for unit root tests in the presence of nonstationary volatility. *Econometric Reviews* 34(4): 512–536.

Cochrane JH (1991) A critique of the application of unit root tests. *Journal of Economic Dynamics and Control* 15(2): 275–284.

De Boef S and Keele L (2008) Taking time seriously. *American Journal of Political Science* 52(1): 184–200.

Enders W (2014) *Applied Econometric Time Series*. 4th edition. Chichester, UK: John Wiley & Sons.

Enns PK and Wlezien C (2017) Understanding equation balance in time series regression." *The Political Methodologist*. Available at: https://thepoliticalmethodologist.com/2017/06/23/understanding-equation-balance-in-time-series-regression/ (accessed 7 October 2017).

Enns PK and Wohlfarth PC (2013) The swing justice. *Journal of Politics* 75(4): 1089–1107.

Enns PK, Kelly NJ, Masaki T, et al. (2016) Don't jettison the general error correction model just yet: A practical guide to avoiding spurious regression with the GECM. *Research and Politics* 3(2): 1–13.

Enns PK, Masaki T and Kelly N (2014) Time series analysis and spurious regression: An error correction. Paper presented at the Annual Meeting of the Southern Political Science Association, New Orleans, 9–11 January 2014. Available at: http://takaakimasaki.com/wp-content/uploads/2014/08/EnnsMasakiKelly_ECM_9.25.14.pdf (accessed 7 October 2017).

Epstein L and Martin AD (2011) Does public opinion Influence the Supreme Court? Possibly yes (but we're not sure why). *University of Pennsylvania Journal of Constitutional Law* 13(2): 263–281.

Esarey J (2016) Fractionally lntegrated data and the Autodistributed Lag Model: Results from a simulation study. *Political Analysis* 24(1): 42–49.

Flemming RB and Wood BD (1997) The public and the Supreme Court: Individual justice responsiveness to American policy moods. *American Journal of Political Science* 41(2): 468–498.

Grant T and Lebo MJ (2016) Error correction methods with political time series. *Political Analysis* 24(1): 3–30.

Lebo MJ and Grant T (2016) Equation balance and dynamic political modeling. *Political Analysis* 24(1): 69–82.

Lebo MJ and Kraft PW (2017) The general error correction model in practice. *Research and Politics* 4(2). E-print online before publication. DOI: https://doi.org/10.1177/2053168017713059.

Link MW (1995) "Tracking public mood in the Supreme Court: Cross-time analyses of criminal procedure and civil rights cases. *Political Research Quarterly* 48(1): 61–78.

Luttig M (2013) The structure of inequality and Americans' attitudes toward redistribution. *Public Opinion Quarterly* 77(3): 811–821.

McGuire KT and Stimson JA (2004) The least dangerous branch revisited: New evidence on Supreme Court responsiveness to public preferences. *Journal of Politics* 66(4): 1018–1035.

Mishler W and Sheehan RS (1993) The Supreme Court as a counterma-joritarian institution? The impact of public opinion on Supreme Court decisions. *American Political Science Review* 87(1): 87–101.

Mishler W and Sheehan RS (1996) Public opinion, the attitudinal model, and Supreme Court decision making: A micro-analytic perspective. *Journal of Politics* 58(1): 169–200.

Shayo M (2009) A model of social identity with an application to political economy: Nation, class, and redistribution. *American Political Science Review* 103(2): 147–174.

Stock JH and Watson MW (2011) *Introduction to Econometrics*. 3rd edition. Boston, MA: Addison-Wesley.

Trump K-S (forthcoming) "Income Inequality Influences Perceptions of Legitimate Income Differences." *British Journal of Political Science*. doi:10.1017/S0007123416000326.

Appendix. Rate of false positives when the general error correction model (GECM) is estimated correctly

The following figure shows that when the appropriate lag lengths are used in the augmented Dickey–Fuller (ADF) test and the GECM is implemented correctly, the rate of false-positives (i.e., the rate of finding a statistically significant effect of X_{t-1} after testing for both integration and cointegration) in the data Lebo and Kraft (LK) analyze is below 0.05 in every case except for two (where the false-positive rate is 0.08 and 0.06).

The results reported in Figures 1(b) and A1 selected the number of lags to include in the ADF test based on the portmanteau (Q) test for white noise residuals. As Figure A2 shows, if we selected the number of lags based on a Breusch–Godfrey test for serial correlation, virtually the same results emerge. Both tests indicate that LK's decision to rely on a default of 3 lags for the ADF test was not appropriate for these simulations.

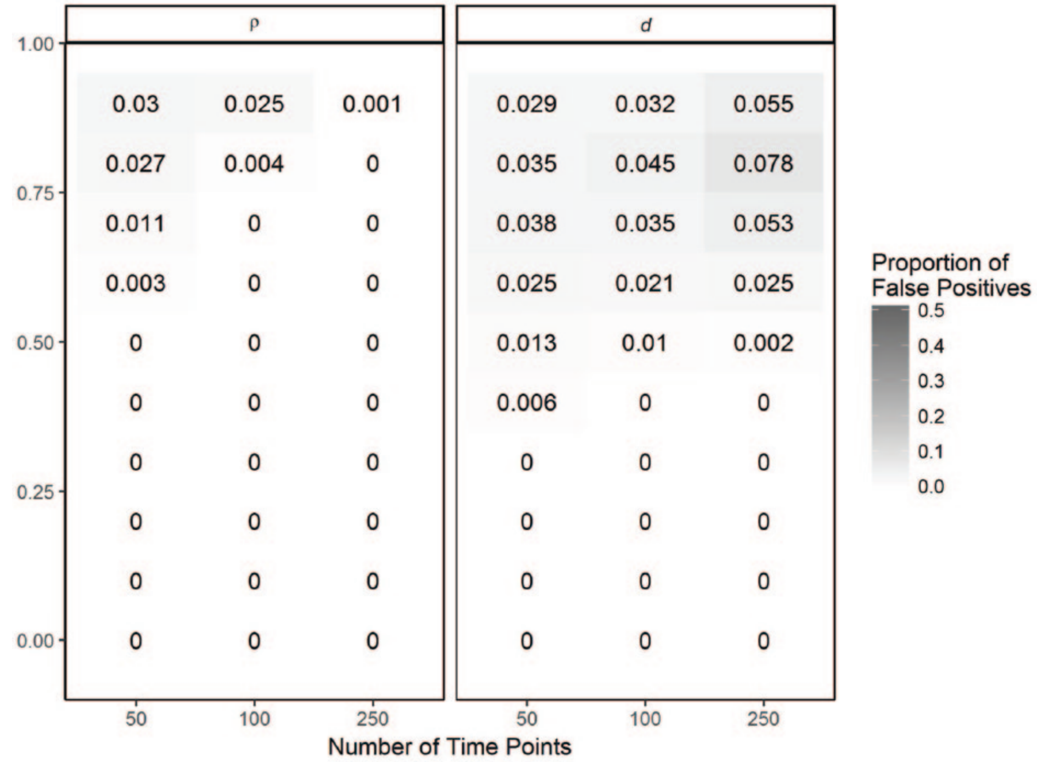


Figure A1. Rate of false-positives for X_{t-1} when the general error correction model is estimated correctly.

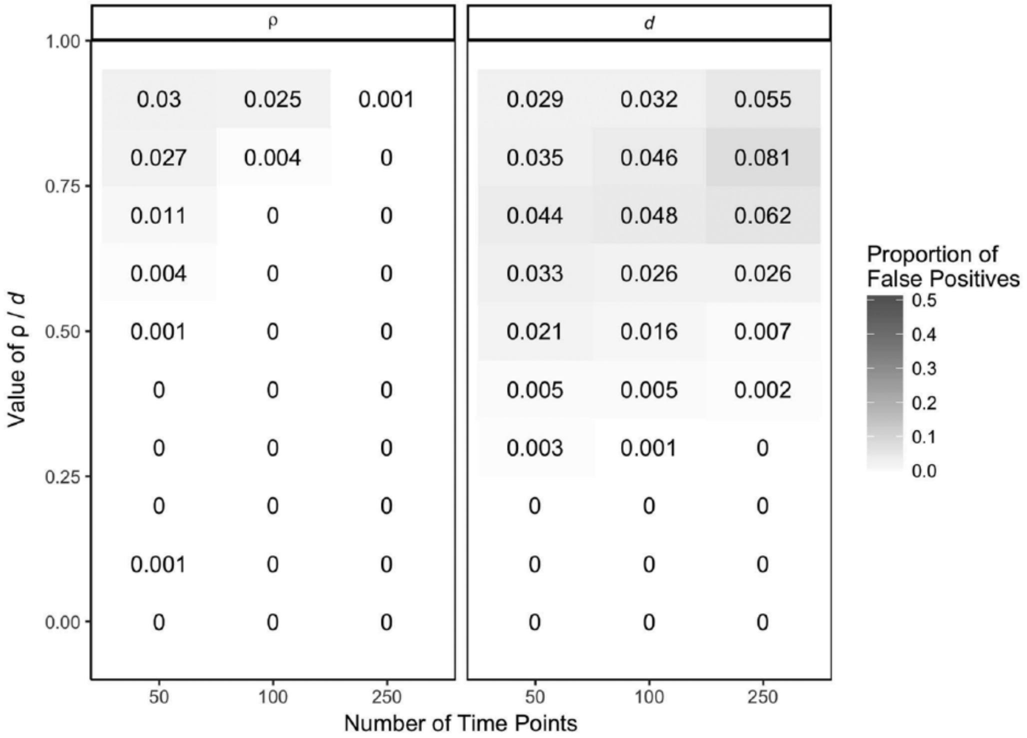


Figure A2. Rate of false-positives for X_{t-1} when the general error correction model is estimated correctly (lag length for the augmented Dickey–Fuller test selected based on a Breusch–Godfrey test for serial correlation.).