Political Analysis (2016) 24:291–306
doi:10.1093/pan/mpw011

Erratum for Keele, Linn, and Webb (2016)

Luke Keele
Department of Political Science, Pennsylvania State University, State College, PA 16802
Email: ljk20.psu.edu

Suzanna Linn
Department of Political Science, Pennsylvania State University, State College, PA 16802
Email: slinn@la.psu.edu

Clayton McLaughlin Webb
Department of Political Science, University of Kansas, Lawrence, KS 66049
Email: webb767@ku.edu

The originally published version of Keele, Linn, and Webb (2016) that appeared in *Political Analysis* 24(1) was an early version of the manuscript that was mistakenly submitted as the final version. The below version is the correct version of the authors' article. The mistake is the authors' and should not in any way be attributed to *Political Analysis* or Oxford University Press.

Treating Time with All Due Seriousness

Luke Keele
Department of Political Science, Pennsylvania State University, State College, PA 16802
Email: ljk20.psu.edu

Suzanna Linn
Department of Political Science, Pennsylvania State University, State College, PA 16802
Email: slinn@la.psu.edu

Clayton McLaughlin Webb
Department of Political Science, Texas A&M University, College Station, TX 77843
Email: webb767@tamu.edu

Time series techniques see widespread use in political science. In De Boef and Keele (2008) we outlined a set of statistical methods for stationary data. Those methods have come to be widely used. Grant and Lebo (2015) contend that one of the methods we discussed, the error correction model, should generally not be used with political data. They argue that the error correction model leads to both interpretational and inferential mistakes by applied analysts. While we agree with their statements about equation balance, we show that the error correction model leads to the same inferences as the autoregressive distributed lag model when the data are stationary. We also demonstrate that careful use of an error correction model can help diagnose model misspecification when the equation is unbalanced. Such techniques are useful since pretesting for integration and fractional integration is often a highly uncertain process, which we demonstrate through a simulation exercise. We also highlight two related but often ignored complications in time series: low power and overfitting. We argue that the statistical tests used in time series analyses have little power to detect differences in many of the sample sizes typical in political science. Moreover, given small sample sizes, many analysts overfit their time series models. We argue that the results in the Grant and Lebo replications stem from inadequate sample sizes that make it difficult to conclusively use any time series model.*

In 2008 we wrote "Taking Time Seriously" (TTS), which considered issues in the estimation and interpretation of time series models for stationary data (De Boef and Keele 2008). We showed that the generalized error correction model (GECM) is a linear transformation of the well known autoregressive distributed lag model (ADL) (see also Beck 1991; Bannerjee et al. 1993). We claimed the isomorphism of the two models ensures that the GECM is appropriate for stationary data, as well as integrated and jointly cointegrated data. We *did not* suggest that the GECM is appropriate for non-stationary data that is not cointegrated or that the GECM can be used for any and all time series data. We also emphasized that researchers should focus on a quantity known as the long run multiplier (LRM), which is the cumulative effect of a covariate on the outcome. Finally, we urged the analyst to think carefully about the correspondence between theory and empirical tests. We stand by these claims and recommendations.

In "Error Correction Methods with Political Time Series", Grant and Lebo (GL) challenge some of the arguments made in TTS related to the GECM (Grant and Lebo 2015). They argue that the GECM should only be used in "rare instances." They argue that the parameterization of the GECM leads to interpretational mistakes. They also claim that the GECM leads to spurious

*The originally published version of Keele, Linn and Webb (2016) that appeared in *Political Analysis* 24(1) was an early version of the manuscript that was mistakenly submitted as the final version. This version is the correct version of the authors' article. The mistake is the authors' and should not in anyway be attributed to *Political Analysis* or Oxford University Press.

https://doi.org/10.1093/pan/mpw011 Published online by Cambridge University Press

inferences at higher rates than alternative time series regression models. While we agree that the GECM can lead to spurious inferences in some contexts, we argue that the GECM does not lead to spurious inferences at a higher rate than other regression models. Our goal in this paper is to offer additional clarity about when the GECM is appropriate, how the GECM should be interpreted, and how the GECM relates to the ADL.

Of course, we do not completely disagree with the points made in GL. First, we agree that balance matters. We agree that the GECM is appropriate for integrated and jointly cointegrated data and that it is inappropriate for a) integrated but not cointegrated data (GL case 1), b) bounded unit root processes that are not cointegrated (GL case 2), c) fractionally integrated (and cointegrated) data (GL case 4), and d) explosive time series (GL case 5). However, issues of balance are not confined to GECMs. When the equation is unbalanced one cannot apply a GECM or *any* other time series regression model. Below we outline specific combinations of acceptable equation balance.

Of course, there are many places where we disagree. First, GL note that stationary data usually guarantees a significant error correction parameter and claim that "this is problematic when common practice is to use α_1 to speak to error correction and re-equilibration between variables" (16). We argue that this statement indicates a misunderstanding of the concept of equilibration between time series by conflating a parameter that describes the speed of adjustment — the error correction rate – with one that assesses the existence and nature of the long run relationship between them — the long run multiplier. Below we clarify the meaning of equilibration in the stationary, integrated, and fractionally integrated cases.

GL also claim that the GECM and the ADL produce inconsistent inferences. While they concede that the GECM is a linear parameterization of the ADL, they argue that the same inferences do not follow from both models. They go on to suggest that this problem is exacerbated when data are near-integrated. We use simulations to show that the inferences are the same from both models and that any inferential errors are avoided if analysts focus on the LRM.

Third, GL also claim that stationary times series as well as integrated and cointegrated time series are rare in political science. Instead they suggest many time series are bounded unit roots or fractionally integrated. While we believe many time series are stationary, we are somewhat agnostic on many of these claims given the weak power of statistical tests and the uncertainty associated with estimates of long run behavior when sample sizes are relatively small. We demonstrate below that drawing inferences about the existence and extent of fractional integration is problematic in sample sizes typically seen in political science; particularly in the applications they criticize.[1] Given this result we are less optimistic than GL about the need for, or efficacy of, fractional cointegration to model relationships in political time series data. Finally, we highlight two related but often ignored complications in time series that explain much of the problem with the applied work criticized by GL: low power of our statistical tests and overfitting. The statistical tests used to diagnose the properties of time series have weak power to detect differences in many of the sample sizes typical in political science. Moreover, we argue that many analysts overfit their time series models. Overfitting occurs when a statistical model describes random error or noise instead of the underlying relationship. Given the short length of many time series and the surfeit of parameters used in many models, we argue that overfitting is a very real danger. In fact, we argue that the ills diagnosed in the empirical applications of GL mostly stem from small sample sizes. Many of the samples are so small that we should be hesitant to make bold claims based on *any* time series model.

The remainder of the paper proceeds as follows. In the next section we discuss univariate dynamics and the importance of the concept of a long run equilibrium for modeling time series relationships. In part two we highlight the areas where we disagree with GL, and defend the utility of the GECM for the analysis of stationary data. Part three discusses issues relevant to pretesting and balance, and part four examines the power of common statistical tests and the problem of overfitting. We end with some suggestions for applied analysis of time series data.

[1] While longer time series are quite common, as in the analysis of financial data or event data, many applications of the GECM have relatively small sample sizes.

https://doi.org/10.1093/pan/mpw011 Published online by Cambridge University Press

1 Modeling Time Series Relationships

Fundamental to time series analysis is the concept of equilibrium. Regardless of the dynamic properties of single time series, we are using statistical models to make statements about how relationships between variables change over time and then stop changing, i.e. equilibrate. However, the concept of equilibrium depends on the type of time series data. The modeling strategy we adopt to test hypotheses about relationships depends on the univariate dynamics of our data and the joint properties of linear combinations of the data. In this section, we review stationary, integrated, and fractionally integrated time series and the form of equilibrium relationship that may be modeled with each of the three types of time series data.

1.1 *Stationary Time Series*

A weakly stationary time series is one for whom the mean, variance, and covariance are time invariant.[2] There are many forms of stationary time series. The most common of these can be represented by a stochastic difference equation with autoregressive and moving average features:

$$Y_t = a + \phi_1 Y_{t-1} + \ldots + \phi_t Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q}$$

where ε_t is an uncorrelated process with mean zero.

The autoregressive moving average model (ARMA(p,q)) is stationary if the roots of the equation lie outside the unit circle. For this condition to be met in an AR(p) model, for example, the sum of the ϕ coefficients must be strictly less than one. In the common AR(1) model, the rate of decay of shocks is given by ϕ_1 and the effects of shocks decay geometrically. The autocorrelation and partial autocorrelation functions can be used to identify the form of the process and estimation of the model provides information to diagnose whether the series is stationary.

When a time series is judged to be stationary, the most common method of analysis relies on linear regression models. In TTS we outlined a set of regression-based models that may be applied to stationary data and weakly exogenous time series. Specifically we focused on the the generalized error correction model (GECM) and the autoregressive distributed lag (ADL) model. The (bivariate) ADL model is given by:

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \beta_0 X_t + \beta_1 X_{t-1} + \varepsilon_t. \tag{1}$$

A simple linear transformation shows that the model in Equation 1 is exactly equivalent to the GECM model given by:

$$\Delta Y_t = \alpha_0 + \alpha_1^* Y_{t-1} + \beta_0^* \Delta X_t + \beta_1^* X_{t-1} + \varepsilon_t. \tag{2}$$

In TTS, we spoke at length at about long and short run effects, equilibration between X_t and Y_t, and the rate at which equilibration occurs. What does equilibration mean when data are stationary? We did not address this question directly in TTS, but with the benefit of hindsight we think it is worth clarifying this point. We argue that the primary use of time series regression models when the data is stationary is to estimate causal effects: what is the causal effect of X_t on Y_t. In all five of the applied examples in GL, the authors are fundamentally interested in causal effects.

In a time series setting, one can think of causal effects as the same subject being given a treatment at a point in time, where the treatment is as-if randomly applied to that time point. For example, take the case of presidential approval and economic expectations. In the dynamic context, we are using data to ask what is the causal effect of a surprise change in economic expectations on presidential approval. Since the data are collected over time it is possible to estimate a dynamic causal effect. That is, the treatment effect may not occur all at once, but Y_t may continue responding to a treatment over future time periods. Thus, we can trace the temporal path of the effect of the treatment on the outcome.

[2] A strictly stationary process is one for which all moments of the joint distribution function are constant over time. Weak stationarity is, however, all that's needed for estimating the regression models detailed below.

In the time since TTS was published, the assumptions needed to identify such dynamic causal effects have been clarified. Specifically, the value of X_t must be as-if randomly assigned. See Keele (2015) for a discussion of causal identification in this setting and in many other settings. Whether the identification assumption holds in many cases is questionable. However, subject to identification assumptions, one can estimate dynamic causal effects.

Since Y_t may respond to a surprise change in X_t across future time periods, in order to calculate the total causal effect we need to calculate the cumulative effect of X_t on Y_t. This cumulative or total causal effect of X_t on Y_t distributed across all future time periods is given by the long run multiplier (LRM). In TTS, one of our main arguments was that researchers should primarily be interested in the long run multiplier precisely because it gives the total causal effect of a change in X_t on Y_t. We argued that researchers often understated their effects, since they did not calculate the total effect in the form of the LRM.

Equilibration in the stationary case thus occurs when the causal effect of X_t on Y_t has dissipated and Y_t is no longer increasing or decreasing in relation to a surprise change in X_t. In the time series context, we can also calculate how long it takes for the total effect to accumulate. It is this rate of effect that is the error correction rate. This rate itself tells us nothing about the strength of the relationship between X_t and Y_T. For this we must turn to the LRM. In the ADL, the LRM is given by $\frac{\beta_0+\beta_1}{1-\alpha_1}$, and the error correction rate is given by $\alpha_1 - 1$. In the GECM, the LRM is $(-\frac{\beta_1^*}{\alpha_1^*})$ and the error correction rate is given by α_1^*. The equivalence of the two models ensures that the LRM and the error corrections rate will be the same in both models.

1.2 *Integrated and Jointly Cointegrated Time Series*

An integrated time series is memoryless: the value of Y_{t+1} depends only on the current value and not on the sequence of events that preceded it. As such, the effects of shocks to an integrated time series do not decay but are instead cumulated. We can write a simple I(1) series to illustrate[3]:

$$Y_t = Y_{t-1} + \varepsilon_t^3$$

Recursive substitution for Y_{t-1} reveals that Y_t is the sum of past shocks:

$$Y_t = \Sigma_{k=0}^{\infty}\varepsilon_{t-k}.$$

The cumulation of past shocks ensures that the mean, variance, and covariance depend on time. In particular the variance of the unit root process is theoretically infinite. A number of statistical tests have been developed to test the null that a process contains a unit root. The (augmented) Dickey Fuller test and the Phillips-Perron test are the most common of these. Notably, however, these tests have weak power against local alternatives such as strongly autoregressive (near-integrated) processes and stationary processes with structural breaks (De Boef and Granato 1997).

In contrast to the case where all variables are stationary, if pre-testing leads the analyst to conclude the data are individually integrated and jointly cointegrated, the long run relationship *must* be captured in an error correction model. The existence of a long run relationship among integrated variables implies cointegration and a valid error correction representation. In turn, cointegration among integrated variables implies a long run relationship that can be captured in an error correction model. Typically, textbooks lay out the Engle-Granger two step method or the Johansen reduced rank regression model for estimating the long and short run dynamics, but other options are also available (see Bannerjee et al. 1993).

Given that an integrated series is memoryless, it makes little sense to think about dynamic causal effects in the same way as stationary data. Here, the concept of equilibrium is defined such that the behavior of some Y_t, is tied to some X_t such that we expect Y_t to settle down — to cease to change — if X_t settles down, all else equal. In the integrated context, equilibrium focuses on the fact that X_t

[3] If the process contains a constant, then the process is said to be a unit root with drift. The process can also contain a deterministic time trend.

https://doi.org/10.1093/pan/mpw011 Published online by Cambridge University Press

and Y_t tend not to stray from each other. Political scientists typically adopt the Engle-Granger two step method for estimation. We maintain that this part of time series analysis is the least controversial, although debate exists over how often this scenario occurs in practice.

1.3 *Fractionally Integrated and (Fractionally) Cointegrated Time Series*

An integrated time series is memoryless. Exogenous shocks do not wear off. For a stationary series, exogenous shocks change the level of the series, but the series then returns to its mean level. A key question is the rate at which the series returns to its mean. If we assume Y_t follows an AR(1) process, we can use the following model:

$$Y_t = a + \phi Y_{t-1} + \varepsilon_t. \tag{3}$$

Here the model assumes that shocks decay at a geometric rate given by ϕ. Under a model of fractional integration, shocks decay at a much slower, hyperbolic rate. Specifically the model of fractional white noise (ARFIMA(0,d,0)) can be represented by an infinite-order autoregressive model:

$$Y_t = \sum_{k=0}^{\infty} \pi_k Y_{t-k} + \varepsilon_t \tag{4}$$

where the weights are obtained from the binomial expansion such that for a given lag k, the weights are given by ck^{d-1} where c is a constant. When tests of the null hypothesis that the data are stationary or unit roots are both rejected, a time series may be fractionally integrated. A significant estimate of d from a fractionally integrated process is often treated as evidence the series is fractionally integrated.

If data are fractionally integrated and jointly (fractionally) cointegrated, the concept of equilibrium is identical to that of stationary data. Recall that the key difference between stationary data and FI data is the rate of decay or, here, the rate at which the dynamic causal effect accumulates. When data are fractionally integrated, the total effect accumulates at the slower hyperbolic rate rather than at the geometric rate in stationary data. While we can estimate the LRM, the coefficients in these models are largely uninterpretable in their raw form: the fractionally differenced dependent variable does not have a natural interpretation such that linking estimates from fractionally cointegrated models back to our theory is difficult. However, impulse response functions can be calculated from these models to interpret the nature of the effects.

2 Points of Debate - The GECM, Stationary Data, and Long Run Effects

In this section, we note that the inferential problems highlighted by GL are not really problems since they focus on the wrong quantity in the GECM. GL's main criticism of the GECM is that it has poor inferential properties. They make two arguments. First, they argue the GECM and the ADL produce inconsistent inferences. They estimate an ADL and a GECM model using data from a study analyzing public support for the supreme court to buttress this claim. The results are presented in table five (17).

The first part of their argument focuses on short run effects. They fail to reject the null $\alpha_1 = 0$ in the ADL model but reject ($\alpha_1^* = 0$) in the GECM model. They argue that these results are inconsistent because the ADL suggests a static model while the GECM does not. This interpretation is incorrect. The fact that the coefficient on lagged Y_t is small demonstrates that the series is not autoregressive conditional on X_t. As a consequence, the error correction rate ($\alpha_1^* = 1 - \alpha_1$) will be large relative to the (same) standard error because Y_t returns quickly to its long run mean when shocked. A small α_1 in the ADL will mean a α_1^* closer to one in the GECM. These coefficients are two ways of representing the same information. Both coefficients tell us that Y_t has little to no dynamic component.

The second element of their argument deals with long run relationships. One must calculate the LRM to test hypotheses about the total effect. As we noted above, the LRM is typically of primary

https://doi.org/10.1093/pan/mpw011 Published online by Cambridge University Press

interest. This quantity can be calculated from either the ADL or GECM given the isomorphism between the two models. GL argue that the ADL and GECM produce inconsistent inferences about the long run relationship between congressional approval and supreme court approval because the coefficients for lagged court approval (α_1) and lagged congressional approval (β_1) are insignificant in the ADL while the error correction rate (α_1^*) and lagged congressional approval (β_1^*) coefficients are significant in the GECM. Based on these differences, they conclude the LRM is significant for the GECM but not significant for the ADL.

This interpretation is also incorrect. The LRM can be statistically significant even if individual terms in the regression model are not. The statistical significance of any single term in the GECM or the ADL is of little consequence for assessing long run effects. While one can calculate the LRM using information from the ADL or the GECM, the standard error for the LRM cannot be calculated directly from either model. A standard error for the LRM must be estimated separately to assess the LRM's significance. While this standard error cannot be estimated directly, an asymptotic approximation for the variance of the LRM can be calculated using the Bewley method (De Boef and Keele 2008, p. 192). The Bewley method is simply a useful way to implement the Delta method. Grant and Lebo arrive at the conclusion that the ADL and GECM produce different inferences because they try to assess the significance of the total effect without calculating the LRM or its standard error.

GL's second argument against the inferential properties of the GECM is that it performs poorly with strongly autoregressive or near integrated data. They argue that the ADL should be preferred over the GECM because the GECM is more likely to produce spurious results. They attempt to illustrate the differences between the ADL and the GECM in table six (19). The analyses presented in table six are problematic. First, the differences reported in table six are specious. The inconsistencies they report exist because they are not testing comparable amounts of information. $\alpha_1^* = 1 - \alpha$. $\beta_1^* = \beta_0 + \beta_1$. GL are comparing β_1 in the ADL against β_1^* and α_1^* in the GECM. The rejection rates are different because the coefficients contain different amounts of information. Also, GL don't actually estimate the LRMs or their standard errors. If they had, they would find that the LRMs from both models are the same. The same standard error is used to assess the significance of the LRM regardless of what information one uses to calculate it. Hence, rejection rates don't differ when one makes the correct comparisons.

Table six highlights another problem with GL's analysis. They analyze the efficacy of the GECM in the least ideal conditions. The sample sizes are small ($T = 60$) and many of the series are near integrated. These conditions aren't just bad for the GECM, they are bad for all regression models. We conducted a series of Monte Carlo experiments to illustrate this point. We simulated two unrelated AR(1) series X_t and Y_t with a range of autoregressive parameters $\phi_{X,Y} = 0.50, 0.70, 0.90, 0.95, 0.99$. We estimated GECMs, calculated LRMs, and estimated standard errors for the LRMs for each pair of unrelated series. We simulated each pair of series 1,000 times for two sample sizes – $T = 100$ and $T = 250$. We calculated rejection rates and the average biases for the LRMs. The series are unrelated so the population $LRM = 0$. See the appendix for details on the simulations and for additional results with larger sample sizes. The results from these experiments are presented in Table 1.

There are two values in each cell. The values to the left of the vertical lines are the average biases of the LRMs calculated for each combination of autoregressive parameters. The values to the right of the vertical lines are the rejection rates for the LRMs. The autoregressive parameters for X are displayed across the horizontal dimension of the table. The autoregressive parameters for Y are displayed along the vertical dimension. The top panel shows the results for $T = 100$. The bottom panels shows the results for $T = 250$.

Looking at the top panel, the results for $T = 100$ show excessive rejection rates when the series are near integrated. The rejection rates are around twice the acceptable rate when $\phi_x \geq .9$ and $\phi_y \geq .9$. The rejection rates are highest when both autoregressive parameters are equal to .99. This would seem support to GL's pessimism about the utility of the GECM with near integrated series, but there are a number of important caveats that need to be considered.

First, the ADL and the GECM produce the same results. Two models were run for each set of simulated series — an ADL and a GECM. We performed joint F-tests $\beta_0 + \beta_1 = 0$ for the ADLs,

https://doi.org/10.1093/pan/mpw011 Published online by Cambridge University Press

Table 1 Average bias and rejection rates for LRM when X_t and Y_t are unrelated

<table>
<tr><td>ϕ_y</td><td>ϕ_x</td><td></td><td></td><td></td><td></td></tr>
<tr><td>T = *100*</td><td>*0.50*</td><td>*0.70*</td><td>*0.90*</td><td>*0.95*</td><td>*0.99*</td></tr>
<tr><td>0.50</td><td>−0.000 | 0.047</td><td>0.002 | 0.063</td><td>−0.001 | 0.061</td><td>0.001 | 0.068</td><td>0.001 | 0.087</td></tr>
<tr><td>0.70</td><td>0.023 | 0.057</td><td>−0.004 | 0.058</td><td>−0.002 | 0.094</td><td>−0.003 | 0.087</td><td>−0.000 | 0.097</td></tr>
<tr><td>0.90</td><td>0.011 | 0.036</td><td>0.025 | 0.052</td><td>−0.000 | 0.095</td><td>−0.000 | 0.100</td><td>−0.002 | 0.132</td></tr>
<tr><td>0.95</td><td>−0.138 | 0.021</td><td>0.029 | 0.035</td><td>−0.002 | 0.089</td><td>0.055 | 0.107</td><td>0.042 | 0.144</td></tr>
<tr><td>0.99</td><td>−0.202 | 0.023</td><td>2.856 | 0.026</td><td>2.716 | 0.088</td><td>−0.209 | 0.106</td><td>0.573 | 0.154</td></tr>
<tr><td>$T = 250$</td><td>0.50</td><td>0.70</td><td>0.90</td><td>0.95</td><td>0.99</td></tr>
<tr><td>0.50</td><td>0.002 | 0.046</td><td>−0.004 | 0.051</td><td>0.000 | 0.074</td><td>0.000 | 0.061</td><td>−0.000 | 0.053</td></tr>
<tr><td>0.70</td><td>−0.008 | 0.041</td><td>0.010 | 0.058</td><td>−0.005 | 0.066</td><td>0.001 | 0.074</td><td>−0.002 | 0.069</td></tr>
<tr><td>0.90</td><td>−0.002 | 0.031</td><td>−0.005 | 0.039</td><td>0.003 | 0.066</td><td>0.008 | 0.076</td><td>0.003 | 0.090</td></tr>
<tr><td>0.95</td><td>−0.036 | 0.027</td><td>−0.032 | 0.034</td><td>−0.006 | 0.057</td><td>0.023 | 0.088</td><td>−0.004 | 0.109</td></tr>
<tr><td>0.99</td><td>0.094 | 0.004</td><td>0.598 | 0.013</td><td>0.286 | 0.028</td><td>−0.006 | 0.071</td><td>0.122 | 0.120</td></tr>
</table>

The data generating processes are given by $Y_t = \phi_y Y_{t-1} + e_{1t}$; $X_t = \phi_x X_{t-1} + e_{2t}$; and e_{1t}, $e_{2t} \sim IN(0, 1)$. Estimates and standard errors are from the Bewley ECM (see appendix). Values are the (average bias for LRM | rejection rate LRM). Results are for 1,000 simulations.

we performed *t*-tests $\beta_1^* = 0$ for the GECMs, and we compared the inferences from the models. LRMs were calculated based on each set of results. The LRMs from the different models were identical and the *t* and F-tests produced similar inferences in all cases. Again, the ADL and GECM are algebraically equivalent and the coefficients contain the same information. Hence, any situation where the GECM appears to be performing poorly in Table 1, the ADL performed poorly as well.

Second, many of the rejection rates fall to acceptable levels when the sample sizes change from $T = 100$ to $T = 250$. This is not surprising given that the inference relies on an asymptotic approximation. Asymptotic approximations such as those used to estimate the standard error for the LRM are useful, but of course limited when sample sizes are small. Most models perform poorly with near integrated data when sample sizes are small. GL use $T = 60$. We would expect the results to be worse under these conditions; not because there is a problem with the GECM but because these conditions are not ideal for any time series regression model. In the appendix, we report additional simulations results with sample sizes of 500 and 1000. Under these sample sizes the results improve in every case.

Third, even in these extreme conditions there is cause for optimism. For certain types of data and sample sizes the risk that one falsely concludes a long run relationship exists when one does not is higher than 5%. However, the average bias in these cases is typically quite low. One would find a relationship but the substantive interpretation of the relationship would be very small. This establishes a set of conditions analysts can look for to avoid mistakes. If one or more series are highly autoregressive, the sample size is small ($T < 200$), the LRM is statistically significant, and the long run substantive effect is minimal analysts should be wary of spurious inferences. If possible, more data should be collected. If not, analysts must decide whether the results are theoretically plausible and be aware of the potential for inferential errors.

There are a number of important conclusions one can distill from this discussion. First, the long run multiplier is the quantity of interest. One can calculate the long run multiplier using information from either the ADL or the GECM, but one must estimate the standard error of the LRM separately to assess its significance. Moreover this standard error is based on an asymptotic approximation that may be poor when the sample size is small. Second, the GECM and the ADL are algebraically equivalent and produce the same results as long as the analyst is careful to correctly interpret each part of the model and the quantities of interest. Finally, analysts can run into problems when analyzing near integrated time series if their sample sizes are too small. This is as true for the GECM as it is for the ADL and most other time series regression models.

3 Complexity in the Statistical Analysis of Time Series

Next, we draw out some specific points about the analysis of time series data that we think are implied by GL but are not clarified.

3.1 *The Role of Pretesting*

GL argue strongly for the use of pretesting and prewhitening. The link between univariate dynamics and the models we use to test the relationships between variables suggests the importance of pretesting — of determining the dynamics of individual series — before we begin to test hypotheses about how X influences Y. Pretesting time series to determine their univariate dynamics is straightforward in the textbook sense: test the properties of the dependent variable and then use the appropriate model. For example, test for a unit root, and if the series appears to be integrated, then use a model for integrated data. We agree, there is little doubt that analysts should be pretesting their data. However, for pretesting to be effective, the analysts must understand the role of equation balance. We turn to this question next. Many of the issues that GL highlight arise from equation balance as opposed to whether the outcome is stationary or not.

3.2 *The Importance of Equation Balance*

The question of equation balance is an important and yet often overlooked aspect of time series modeling. No regression model is appropriate when the orders of integration are mixed because no long run relationship can exist when the equation is unbalanced. The intuition is simple. Stable/stochastically bounded variables cannot cause (or be caused by) the path of a stochastically unbounded variable. The time series must eventually diverge by larger and larger amounts.[4] Instead, the data must be transformed to ensure the left and righthand sides of the model are of equal orders of integration.

Grant and Lebo spend extensive time demonstrating that the GECM performs poorly when equations are unbalanced. Specifically they consider explosive dependent variables and integrated or stationary regressors, integrated dependent variables and stationary regressors, and stationary dependent variables and integrated regressors. In each case, they show that tests of the true null that the series are unrelated reject too frequently. These results emphasize the inapplicability of error correction models in these cases, but for the same reason they condemn regression models generally. No regression model will produce reliable inferences when the order of integration on the left and righthand side of our equation are different such that no long run relationship exists between the regressand and regressors.

3.3 *Pretesting Amibguity*

If pretesting were so simple, it is unlikely that the literature on testing for unit roots would be so large. At issue are the weak power of statistical tests, the disjuncture between the theoretically infinite variance of unit root processes and the limited variance of bounded time series, and the lack of sufficient theory predicting unit roots. In addition, a time series may contain structural breaks, that when ignored, may make a stationary series appear integrated or fractionally integrated. Similarly, a process may be conditionally heteroscedastic, a feature of the data that confounds statistical tests for stationarity and the estimation of the fractional differencing parameter.

In fact, GL advocate for additional pretesting. GL suggest that much of the data in political science are characterized by fractional integration. They reject the notion that most political data is stationary. Strictly speaking, the distinction between FI and stationary time series is an empirical one that can be arbitrated with data. The difficulty is that the data often do not cooperate as we explore next.

[4] As GL note, hypothesis tests on the model coefficients will not follow standard distribution theory in this case, but that point is trivial given that the regression model is nonsensical.

https://doi.org/10.1093/pan/mpw011 Published online by Cambridge University Press

The general ARFIMA(p, d, q) model is given by:

$$\left(1-\sum_{i=1}^{p}\phi_i L^i\right)(1-L)^d Y_t=\left(1+\sum_{i=1}^{q}\theta_i L^i\right)\varepsilon_t \tag{5}$$

where p refers to the number of autoregressive parameters, ϕ, q refers to the number of moving average parameters, θ, and d is the fractional differencing parameter. Fractional integration would appear to offer a complete framework for thinking about time series. It can accommodate short run dynamics in the form of autoregressive and moving average parameters and long run dynamics in a fractional differencing parameter, d.[5]

Of course, in order to use this model, we must be able to reliably estimate the d parameter that describes the level of fractional integration, possibly along with ϕ and θ. A considerable body of research suggests this may not be the case in a wide variety of circumstances. In particular when samples are small to medium in size, and when the process includes short run dynamics, particularly of unknown order, estimation of d can be highly uncertain. The Stata manual on the arfima command, for example, warns against fitting a 3 parameter ARFIMA model with in an empirical example with 372 observations, saying this is a very complex dynamic model (StataCorp. 2013). A three parameter model is a model with an AR parameter, an MA parameter, and a d parameter. No independent variables are included in this model.

Many authors point to cases in which tests suggest data are fractionally integrated when they are not (Diebold and Inoue 2001; Engle and Smith 1999; Granger and Hyung 1999). The presence of outliers or structural breaks can produce time series that mimic ARFIMA processes, as can time series that are simple non-linear transformations of underlying short memory variables. Bhardwaj and Norman (2003) show that even absent these concerns, spurious long memory often arises in a number of statistical tests of short memory. They also show that standard short memory tests will provide evidence for long-memory even in cases where predictions from a number of ARFIMA model estimators of d fare worse than those from the more standard AR, MA, ARMA and related models (Bhardwaj and Swanson 2006). In fact, Granger (1999) notes that ARFIMA models may well fall into an "empty box" because these models have stochastic properties that do not mimic the properties of much of the data to which they have been, or are likely to be, applied. Bhardwaj and Swanson (2006) show that in some circumstances ARFIMA models offer superior predictions to alternative models about half the time, but only when sample sizes are large and forecast horizons long.

Here we illustrate the difficulty of drawing inferences about the existence and degree of fractional integration using simulations. We analyze the properties of the exact maximum likelihood estimate (Sowell 1992), the default estimator in Stata and the popular `R` package `ARFIMA`, and that recommended by Lebo, Walker and D Clarke (2000) and Veenstra (2013).[6]

We simulate the ARFIMA process given in equation 5 for samples of size 50, 100, 250, 500, 1000, and 1500. We allow for a range of dynamics, including ARFIMA(0,d,0), ARFIMA(1,d,0), ARFIMA(0,d,1) and ARFIMA(1,d,1) processes. The autoregressive parameter, ϕ, is set to 0.60, the moving average parameter, θ, is set to 0.60 in the AR and MA models, respectively, while $\phi = 0.50$ and $\theta = 0.30$ in the combined ARMA models.[7] In the simulations, d takes on the values 0 (no fractional integration), 0.20, 0.40, 0.45, and 0.80. In the latter case, the data is integer differenced before simulation and estimation so that d = -0.20 in the transformed data. We estimate the ARFIMA process under the optimal, but unrealistic assumption that the order of the short run

[5] See Baillie (1996) for a survey of methods for long memory data.

[6] While the most commonly used estimators are asymptotically equivalent, their performance can differ markedly in small to medium size samples. Other estimators often used are the Whittle likelihood (Robinson 1995) and the modified profile likelihood (An and Bloomfield 1993). Evidence suggests the exact MLE is not a panacea (Hauser 1999). Specifically, the modified profile likelihood dominates the exact MLE, which is biased downward, in small samples, especially when long and short run dynamics both characterize the data generating process.

[7] These values are chosen somewhat arbitrarily and allow for the possibility that the short run dynamics in the model are reasonably large. We see no reason to assume that all the dynamics associated with a process are long term.

https://doi.org/10.1093/pan/mpw011 Published online by Cambridge University Press

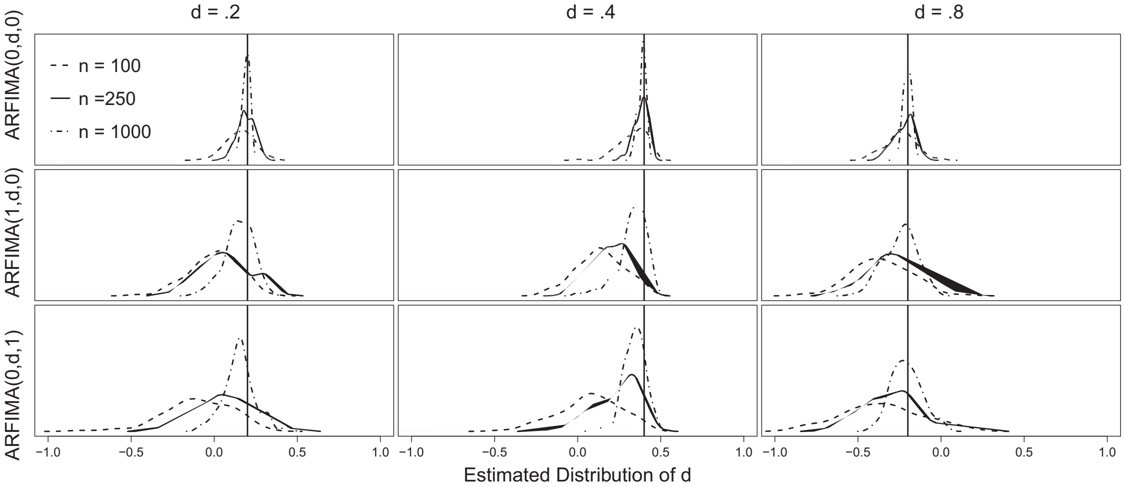

Fig. 1 Distributions of Estimates for *d*. Each panel shows the distribution of the exact maximum likelihood estimates of *d* from the simulations for samples of size $T = 100$ (dashed line), $T = 250$ (solid line), and $T = 1, 000$ (dotted line). The solid vertical line in each plot represents the true value of *d*. Details of the simulations are given in the text.

dynamics is known.[8] The difficulty of selecting the right model of short run dynamics further complicates the estimation of *d*; uncertainty over the proper short run dynamic model increases our uncertainty over the estimate of *d* and thus our confidence that the selected model mimics the data generating process.

The results from the simulations are presented in Figure 1. For the sake of clarity, we only present a subset of the results.[9] The rows show the results from the ARFIMA(0,d,0), ARFIMA(1,d,0), and ARFIMA(0,d,1) models. The value of the fractional differencing parameter (*d*) varies across the columns of the array. Results are presented for the fractional parameters $d = .2$, $d = .4$, and $d = .8$. Three sample sizes are presented: $T = 100$ (dashed line), $T = 250$ (solid line), and $T = 1, 000$ (dotted line). The solid vertical line in each plot represents the true value of *d*.

There is a considerable amount of uncertainty in the estimates of *d*. Consistent with (Hauser 1999), the estimator produces downwardly biased estimates of *d* across all models. In particular, performance is poor when $T = 100$ and $T = 250$ and is worse when the data generating process contains short-run dynamics. The estimator has particular difficulty distinguishing long run from short run dynamics. In some cases estimates range across almost all possible values of *d*. For $d = .2$ and $T = 100$ estimates range from –.18 to .36 in the ARFIMA(0,d,0) model. This range increases to [-.46,.33] in the ARFIMA(1,d,0) model and to [-.80,.30] in the ARFIMA(0,d,1) model.[10] This uncertainty may lead to misdiagnosis of *d* in small to medium samples. It is clearly inconsistent with the statement in GL that one needs at least 64 observations to reliably estimate the *d* parameter. That number appears to be too low relative to our simulations.

This poor performance may also lead to overdiagnosis of fractional integration, causing analysts to fractionally difference short memory data. Table 2 summarizes a series of simulations that illustrate this point. Columns one and two show the models and sample sizes. Column three

[8] The sample mean is used as the estimate of the true mean (which is zero). The log likelihood is given by:

$$\ell(y|\hat{\eta}) = -1/2\Big[Tlog(2\pi) + log|\hat{V}| + (y - X\hat{\beta})'\hat{V}^{-1}(y - X\hat{\beta})\Big] \qquad (6)$$

where V is the variance-covariance matrix. See Sowell (1992) for details. The models are estimated with the number of starting values set to twice the number of estimated parameters (other than the constant). The AIC is used to select the estimate when the likelihood surface has multiple modes.

[9] The remaining results are summarized in the appendix.

[10] The ARFIMA(1,d,1) results reported in the appendix show that estimates of *d* continue to deteriorate as models become more complex. Even with large samples $T = 1, 000$ and $T = 1, 500$ MLE produces very poor estimates of *d*.

https://doi.org/10.1093/pan/mpw011 Published online by Cambridge University Press

Table 2 Estimation of d

Model	*t*	*Mean*	*Rejection Rate*
	100	–.032	11 %
ARMA(0,0)	250	–.017	9 %
	1,000	–.004	13 %
	100	–.199	32 %
ARMA(1,0)	250	–.121	34 %
	1,000	–.073	21 %
	100	–.227	16 %
ARMA(0,1)	250	–.132	34 %
	1,000	–.019	12 %
	100	–.530	69 %
ARMA(1,1)	250	–.316	53 %
	1,000	–.056	27 %

Column 3 gives the mean exact maximum likelihood estimate of d for different sample sizes and different data generating processes. Column 4 reports the rejection rate on the true null hypothesis $d=0$.

shows the average estimates of d for each model-sample combination, and column four shows the rate at which each of the models produced estimates of d reliably (95%) different from zero when $d=0$.

The results presented in Table 2 are consistent with the simulations presented in Figure 1. The estimates are negatively biased, this bias is larger in small samples. The quality of the estimates deteriorate further when the data generating process contains short run dynamics, which is at least a plausible alternative hypothesis. The percentages in column four show that the risk of incorrectly rejecting the null that $d=0$ is unacceptably high across all of the models, and is particularly pronounced in the more complex models. One commits type-I error more than one third of the time in the ARMA(1,0) and ARMA(0,1) models when $T=250$, and more than half the time in the ARMA(1,1) model with the same sample size. This is a concern since samples of 250 observations or less are common in political science.

So should analysts do pretesting? Of course. Besides testing for whether data are stationary or have some level of integration, analysts should also test for structural breaks and for stochastic volatility using ARCH methods. It is always unwise to ignore the information that may be gleaned from pretesting. However, it is also clear that pretesting is no panacea. As we have noted above, it can be difficult to exactly know whether series are stationary or integrated. De Boef and Granato (1999) and De Boef (2001) highlight the difficulties that can arise in pretesting.

Given the difficulties with pretesting univariate time series to infer the true dynamics of many time series processes, it is worth noting that estimates from a GECM (and ADL) can provide some diagnostic evidence about equation balance in some cases. It is well known that estimates from a GECM provide evidence about the nature of the long run relationship specified. It is less well appreciated that model estimates can tell us something about the appropriateness of the GECM and the likelihood that the equation is balanced.

Estimated error correction rates may take on a range of values in practice. See Table 3. Typically, error correction rates lie between 0 and -1.0. Estimated error correction coefficients nearer to -1.0 imply that the effect of X_t accumulates quickly, while those closer to 0 imply a slower accumulation. Error correction rates may also lie strictly between -1.0 and -2.0. Just as with negative autocorrelation, in this scenario the approach to equilibrium is oscillating, as Y_t corrects more than 100% of the equilibrium error in the succeeding period but will slowly return to equilibrium as the overcorrection lessens after each time period. This situation is, however, very rare. If an analyst estimates an error correction rate in this range, he should consider whether such a scenario makes sense or whether some form of misspecification is likely driving the result.

Error correction coefficients outside this range or close to the bounds are often a sign of model misspecification. A positive error correction rate indicates a lack of stability in the model. The model does not converge to a long run equilibrium. The implied coefficient on lagged Y_t in the ADL

https://doi.org/10.1093/pan/mpw011 Published online by Cambridge University Press

Table 3 Error Correction Rates and Long Run Equilibria

α_I^*	α_I	*Diagnosis*
$0 > \alpha_1^* > -1.0$	$0 < \alpha_1 < 1.0$	Steady return to long run equilibrium.
$-2.0 < \alpha_1^* < -1.0$	$-1.0 < \alpha_1 < 0$	Oscillating return to long run equilibrium.
$\alpha_1^* > 0$	$\alpha_1 > 1.0$	Y is explosive, no long run equilibrium exists.
$\alpha_1^* = 0$	$\alpha_1 = 1.0$	Y is integrated, no long run equilibrium exists.
$\alpha_1^* < -2.0$	$\alpha_1 < -1.0$	Y is explosive, no long run equilibrium exists.

is greater than 1.0 ($\alpha_1^* + 1.0 > 1.0$). Here it is immediately obvious that the Y_t process is explosive and no long run equilibrium exists. Positive estimates of α_1^* likely occur because the equation is imbalanced but may occur because the equation properly specified contains unmodeled dynamics, likely a structural break (or breaks). If the time series are all integrated, a second, unmodeled cointegrating relationship may exist, producing a positive error correction rate.

Consider an estimated error correction rate in the GECM equal to -1.0. Such an estimate implies Y_t adjusts immediately and completely to any shocks in X_t and thus all the dynamic effects of X_t translate to a new value of Y_t immediately (at whatever lag they enter they model).[11] In other words, the causal effect of X_t on Y_t is not dynamic. Y_t is white noise. Estimation of the ADL would present the analyst with the corroborating evidence that the coefficient on lagged Y_t is 0. In this case, although the data are stationary, neither a GECM nor ADL model should be specified. If the error correction coefficient is less than -2.0, the underlying Y_t series is explosive. Such a scenario could also arise if a negatively autocorrelated Y_t contains a structural break, if the data contain ARCH effects, or if some other form of misspecification exists.

Finally, an error correction coefficient equal to 0 implies that Y adjusts so slowly to changes in X_t that it does not ever reach an equilibrium. This signals misspecification of a different sort. It could occur because Y_t is a unit root process and not cointegrated with X_t but may also indicate unmodeled dynamics in Y_t due to a structural break or simply that the time series was not observed long enough to witness a return to equilibrium. Therefore, when analysts use a GECM or ADL, they should take note of estimated values of α_1^* or α_1 that are close to the bounds or exceed the bounds implied by the the model, as this is evidence of a misspecified dynamic model.

4 Tests, Power, and Overfitting

GL criticize a set of applications of the GECM. We agree that there are problems with these examples, but we disagree that these problems exist because the authors used the GECM. In at least one application the authors note the lack of balance in the model (Volscho and Kelly 2012). As we noted above, this rules out an equilibrium relationship.

The broader problem with these applications is the lengths of the time series available for analysis. Table 4 lists the lengths of the time series in the five applications criticized by Grant and Lebo. The longest time series in these applications is 60 time periods. The small sample sizes contribute to three problems. First, it makes it difficult to diagnose whether these time series are stationary, fractionally integrated, or integrated. Second, diagnostic tests have weak power in small samples. Third, there is a very real danger of over fitting in each of these examples.

In some sense, we have already made the first point above. GL claim that the solution in these applications is the use of FI techniques. Perhaps. However, diagnosis of FI will be difficult in every case since the number of cases is lower than what GL recommend for estimation of d (64) and considerably lower than our simulations suggest necessary to accurately diagnose fractional integration.

Next, once a time series regression model is fit, the residuals should be tested for signs of temporal dependency. When model residuals are auto-correlated, this is a clear sign of incorrectly

[11] If none of the independent variables in the model are significant, unmodeled shocks are immediately incorporated into future values of Y.

https://doi.org/10.1093/pan/mpw011 Published online by Cambridge University Press

modeled dynamics. The analyst should go back to the drawing board. The basic difficulty is that both types of tests have little power given the length of the time series in these applications. There is simulation evidence on this point in the literature.

Keele and Kelly (2006) compared the performance of lagged dependent variable (LDV) models to alternative ARMA specifications. One of their conclusions was that problems with LDV models could be detected through testing the residuals for autocorrelation. They then conducted a series of simulations to understand the power of such tests. That is, they sought to understand how long a time series needed to be before one could reliably detect autocorrelation in the residuals of regression models with LDVs. The results are instructive given that they found one needed sample sizes of between 250 and 500 observations before these tests had much power. The five articles replicated by Grant and Lebo do not meet this threshold. Hence, one cannot expect to have much power to detect autocorrelation in the residuals.

Overfitting is another problem that complicates a critical assessment of problems cited by GL. Overfitting occurs when a statistical model describes random error or noise instead of the underlying relationship. In the statistics literature, the rule of thumb is that one should fit one parameter for each 10 observations when the data are independent and identically distributed (IID) (Babyak 2004). When there is not enough information in the data, the model can be tuned to fit random patterns in the data instead of the conditional expectation that is generally of interest in applied statistical analysis. The likelihood of finding spurious relationships is quite high when models are overfit (Babyak 2004).

Let's consider the possibility of overfitting in the applied examples in Grant and Lebo. For Volscho and Kelly (2012), $N = 60$ and $k = 10$. If we apply the rule of thumb, that would imply a maximum of 6 parameters if the data were IID. Another way to think about their model is that it is equivalent to fitting 10 separate models with a single predictor each with a sample size of 6. However, the rule above assumes we have IID data. With time series data, there is considerably less information present in the same size sample. This means the rule of thumb for time series data understates the possibility of overfitting. It is quite possible that many of the results in those models could be a function of overfitting. In fact, we believe that many of the issues that arise in the Grant and Lebo re-analysis are a function of overfitting, where the data are being fit to different random patterns, and thus the results are unstable.

In general, the things that can be learned from a small sample of data is quite limited. This is something time series analysts need to take seriously. When sample sizes are small, overfitting is possible and diagnostic tests have little power to detect violations of basic assumptions. The conclusion we should draw is that time series analysts need to use great caution and provide limited interpretations of their results when sample sizes are small.

5 Discussion

Applied time series analysis depends on the diagnosis and classification of time series and the selection of appropriate models. Given the sample sizes we see in many applied examples, this can be problematic at both the pretesting and diagnostic phases of the analysis. While there are cases in political science where there are long time series, there are many cases where sample sizes remain limited. If the analyst has solid theoretical and empirical evidence that the data are stationary, one can use autoregressive distributed lag models, error correction models, or appropriately restricted versions of these regression models. Following De Boef and Keele (2008), one can use a general to specific modeling strategy to determine which restrictions, if any, are appropriate and use the results to calculate other quantities of interest.

If one or more time series is judged to be integrated, alternative models may be necessary. If one finds that one series is integrated and the remaining series are stationary or finds that two series are integrated but not integrated of the same order, one should transform the integrated variables in order to ensure that the equation is balanced and use an ADL or GECM on the transformed data. If one finds that two series are integrated and integrated of the same order, one should test whether the series are cointegrated. If the series are cointegrated, cointegration techniques are appropriate. One can apply either the Engle-Granger two step method or Johansen reduced rank regression

https://doi.org/10.1093/pan/mpw011 Published online by Cambridge University Press

Table 4 Comparison of Observations to Parameters in Grant and Lebo Replications

Article	*Time Periods*	*Number of Parameters*
Casillas, Enns, Wohlfarth	45	7
Ura and Ellis	36	11
Sanchez et al.	60	11
Kelly and Enns	54	8
Volscho and Kelly	60	10

Table 5 Univariate Dynamics and Modeling Equilibrium Relationships

Y	X	*Modeling Strategy*
Stationary	Stationary	ADL/GECM or restricted versions thereof.
Stationary	Integrated	First difference X and estimate ADL/GECM.
Stationary	Fractionally Integrated	Fractionally difference X and estimate ADL/GECM.
Integrated	Stationary	First difference Y and estimate ADL/GECM
Integrated	Integrated	If jointly stationary, estimate error correction model; otherwise estimate model in first differences.
Integrated	Fractionally Integrated	Fractionally difference X and first difference Y to ensure stationarity and estimate the short run relationship between the differenced series.
Fractionally Integrated	Stationary	Fractionally difference Y and estimate ADL/GECM.
Fractionally Integrated	Fractionally Integrated	If jointly of lower order of integration, estimate FECM; otherwise fractionally difference both variables and estimate ADL/GECM
Fractionally Integrated	Integrated	First difference X and fractionally difference Y to ensure stationarity and estimate the short run relationship between the differenced series.

model to estimate the cointegrating relationship. Of course, other stationary variables can be included in these models. These variables will not be part of the cointegrating relationship but can impinge on the relationship. If the series are not cointegrated, the variables can be transformed and conventional models can be applied.

Analysts may find that some series are fractionally integrated. If two series are fractionally integrated but not fractionally integrated of the same order, the series should be fractionally differenced and conventional time series regression models can be applied to the fractionally differenced data. If two series are fractionally integrated of the same order, one should test whether the series are fractionally cointegrated. Like standard cointegration procedures, one only needs to use a fractional error correction model if they find that two fractionally integrated series are fractionally integrated of the same order and jointly stationary. Otherwise the variables can be fractionally differenced and the analyst can use an ADL or GECM. As we have highlighted, distinguishing between stationary and fractionally integrated series is not easy. The procedures conventionally used to identify whether series are fractionally integrated may not perform well given the sample sizes common in political science.

We present a guide for selecting a modeling strategy to estimate equilibrium relationships when pretesting leads the analyst to characterize the univariate dynamics of their data in the different ways presented in Table 5.

All of these questions are complicated when we have small samples, say less than 250 observations. Further, diagnostic tests may be of little help and overfitting is a real danger. What is the analyst to do in these cases? In these cases, analysts need to fit simpler models and be aware that inferences must be limited. Finally, in all cases time series analysts should pay close attention to

https://doi.org/10.1093/pan/mpw011 Published online by Cambridge University Press

identification conditions. Time series models are often used for causal inference, and attention must be paid to the plausibility of the identification strategy.

References

An, S and P Bloomfield. 1993. “Cox and Reid’s modification in regression models with correlated errors.” *Department of Statistics, North Carolina State University, Raleigh*.

Babyak, Michael A. 2004. “What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models.” *Psychosomatic medicine* 66(3):411–421.

Baillie, Richard T. 1996. “Long memory processes and fractional integration in econometrics.” *Journal of econometrics* 73(1):5–59.

Bannerjee, Anindya, Juan Dolado, John W. Galbraith and David F. Hendry. 1993. *Integration, Error Correction, and the Econometric Analysis of Non-Stationary Data*. Oxford: Oxford University Press.

Beck, Nathaniel. 1991. “Comparing Dynamic Specifications: The Case of Presidential Approval.” *Political Analysis* 3:27–50.

Bhardwaj, Geetesh and Norman R Swanson. 2006. “An empirical investigation of the usefulness of ARFIMA models for predicting macroeconomic and financial time series.” *Journal of Econometrics* 131(1):539–578.

Bhardwaj, Geetesh and R Norman. 2003. “Swanson “An Empirical Investigation of the Usefulness of ARFIMA Models for Predicting Macroeconomic and Financial Time Series?? working paper of.”.

De Boef, Suzanna. 2001. “Testing for Cointegrating Relationships with Near-integrated Data.” *Political Analysis* 9:78–94.

De Boef, Suzanna and Jim Granato. 1997. “Near-integrated Data and the Analysis of Political Relationship.” *American Journal of Political Science* 41(2):619–640.

———. 1999. “Testing for Cointegrating Relationships with Near-integrated Data.” *Political Analysis* 8:99–117.

De Boef, Suzanna and Luke Keele. 2008. “Taking time seriously.” *American Journal of Political Science* 52(1):184–200.

Diebold, Francis X and Atsushi Inoue. 2001. “Long memory and regime switching.” *Journal of econometrics* 105(1):131–159.

Engle, Robert F and Aaron D Smith. 1999. “Stochastic permanent breaks.” *Review of Economics and Statistics* 81(4):553–574.

Granger, Clive WJ. 1999. Aspects of research strategies for time series analysis. In *Presentation to the conference on New Developments in Time Series Economics, Yale University*.

Granger, Clive WJ and Namwon Hyung. 1999. “Occasional structural breaks and long memory.” *Department of Economics, UCSD*.

Grant, Tayler and Matt Lebo. 2015. “Error Correction Methods with Political Time Series.” *Political Analysis* Forthcoming.

Hauser, Michael A. 1999. “Maximum likelihood estimators for ARMA and ARFIMA models: A Monte Carlo study.” *Journal of Statistical Planning and Inference* 80(1):229–255.

Keele, Luke J. 2015. “The Statistics of Causal Inference.” *Political Analysis* Forthcoming.

Keele, Luke J. and Nathan J. Kelly. 2006. “Dynamic Models for Dynamic Theories: The Ins and Outs of Lagged Dependent Variables.” *Political Analysis* 14:186–205.

Keele, Luke J., Suzanna Linn and Clayton McLaughlin Webb. 2016. “Treating Time with All Due Seriousness.” *Political Analysis* 24(1):31–41.

Lebo, Matthew J, Robert W Walker and Harold D Clarke. 2000. “You must remember this: dealing with long memory in political analyses.” *Electoral Studies* 19(1):31–48.

Robinson, P.M. 1995. “Gaussian Semiparametric Estimator of Long Range Dependence.” *Annals of Statistics* 23:1630–61.

Sowell, Fallaw. 1992. “Modeling long-run behavior with the fractional ARIMA model.” *Journal of Monetary Economics* 29(2):277–302.

StataCorp. 2013. *Stata 13 Base Reference Manual*. College Station, TX: Stata Press.

Veenstra, Justin. 2013. Persistence and Anti-Persistence: Theory and Software (Thesis format: Monograph) PhD thesis Western University London.

Volscho, Thomas W and Nathan J Kelly. 2012. “The Rise of the Super-Rich Power Resources, Taxes, Financial Markets, and the Dynamics of the Top 1 Percent, 1949 to 2008.” *American Sociological Review* 77(5):679–699.

https://doi.org/10.1093/pan/mpw011 Published online by Cambridge University Press