Political Analysis (2016) 24:31–41
doi:10.1093/pan/mpv031

Treating Time with All Due Seriousness

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

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

Clayton McLaughlin Webb
Department of Political Science, University of Kansas, Lawrence KS 66045
e-mail: webb767@ku.edu

Edited by Janet Box-Steffensmeier

In this article, we highlight three points. First, we counter Grant and Lebo's claim that the error correction model (ECM) cannot be applied to stationary data. We maintain that when data are properly stationary, the ECM is an entirely appropriate model. We clarify that for a model to be properly stationary, it must be balanced. Second, we contend that while fractional integration techniques can be useful, they also have important weaknesses, especially when applied to many time series typical in political science. 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 the small sample sizes, many analysts overfit their time-series models. Overfitting occurs when a statical model describes random error or noise instead of the underlying relationship. We argue that the results in the Grant and Lebo replications could easily be a function of overfitting.

The goal of applied time-series analysts is to estimate relationships among variables whose behaviors evolve over time and use those estimates both to test hypotheses and to infer interesting features of the relationships in the short and long run. Studies of presidential approval, inequality, and macroeconomic indicators rely on the tools of time-series analysis. In a time-series analysis, the choice of the statistical model depends critically on the temporal properties of the data. In theory, the choice is clear-cut. Traditionally, the analyst faced a dichotomous choice. If the data are stationary, a number of time-series regression models can be applied in a straightforward manner. If the data are integrated but jointly stationary, cointegration techniques must be used. More recent work adds a third option: if the data are fractionally integrated but jointly of a lower order of integration, fractional cointegration methods apply. As such, an important part of a time-series analysis is the diagnosis and classification of time series. The analyst must be able to classify time series as either stationary, integrated, or fractionally integrated before more standard statistical modeling can be done.

The difficulty is that classification of time series into these three types is often not straightforward. Grant and Lebo's basic argument revolves around this choice. Essentially they argue that error correction models can only be applied to integrated outcomes, and that many

Authors' note: For comments and suggestions, we thank Neal Beck. Replication files are available at (Keele, Linn, and Clayton, 2016).

analysts have incorrectly treated fractionally integrated time series as stationary. These mistakes have led to serious inferential errors. Their final conclusion essentially amounts to a call for most political time-series models to be treated as fractionally integrated.

In this essay, we highlight three points. First, we counter Grant and Lebo's claim that the error correction model (ECM) cannot be applied to stationary data. We highlight that much of the evidence in their article is based on the use of imbalanced equations. We maintain that when data are properly stationary, the ECM is an entirely appropriate model. Next, we contend that while fractional integration techniques can be fruitfully applied, they also have important weaknesses, especially when applied to many time series typical in political science. Finally, we bring attention to two related but often ignored complications in time series: low power 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, given the small sample sizes, many analysts overfit their time-series models. Overfitting occurs when a statical 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. We end with some suggestions for applied analysis of time-series data.

1 Modeling Time-Series Relationships

In this section, we review the models appropriate for the three types of time series. The goal in any time-series analysis is to model the equilibrium relationship between X and Y, that is, to model the behavior of some Y which is tied to X over time. Here, we review how such an equilibrium relationship may be modeled with each of the three types of time-series data.

1.1 *Stationary Time Series*

When a time series is judged to be stationary, the most common method of analysis relies on linear regression models. De Boef and Keele (2008) outlined a set of regression-based models that may be applied to stationary data and weakly exogenous time series. In this case, the generalized error correction model (GECM), the autoregressive distributed lag (ADL) model, or appropriately restricted versions of these regression models are all reliable tools for inference. Each will capture the essential features of the short and long-run relationships. The (bivariate) ADL model is given by

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

The short-run effects are given by the β, the long-run effect is given by $\frac{\beta_0+\beta_1}{1-\alpha_1}$, the long-run equilibrium is calculated as $\frac{\alpha_0}{1-\alpha_1} + \frac{\beta_0+\beta_1}{1-\alpha_1} E(X)$, and the error correction rate is given by $\alpha_1 - 1$.

A simple linear transformation shows that the model in equation (1) is exactly equivalent to the GECM model given by

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

The GECM estimates changes in Y as a function of lagged values of X and Y, which capture the long-run relationship ($\frac{\beta_1^*}{\alpha_1^*}$), and changes in X, which capture short-run dynamics. The error correction rate is given by α_1^*. The equivalence of the two models ensures that the short- and long-run effects, as well as the equilibrium relationship and error correction rates, will be the same as those given for the ADL. Standard limiting distributions apply to hypothesis tests on all quantities.

Grant and Lebo maintain that the ECM has pathologies particular to it that do not exist in the ADL model. However, the long-established isomorphism between the GECM and ADL, and restricted versions of each, means the GECM and ADL offer the same information and suffer the same problems (Beck 1991; Bannerjee et al. 1993; Hendry 1995; Davidson and MacKinnon 1993). Again, we maintain that the equivalence is mathematical fact and the two models lead to the

same results under the stated assumptions. As we highlight below, problems can arise when there is overfitting or unbalanced equations, but these problems apply equally to the ADL and ECM.[1]

1.2 *Integrated and Jointly Cointegrated Time Series*

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).

Political scientists typically adopt the Engle-Granger two-step method in which the long-run relationship is estimated in step one. In step two the lagged (stationary) residuals from this equation enter the second-stage error correction model. It is easily seen that all variables in the second-stage regression are stationary in this case such that standard limiting distributions apply to all coefficients. However, as long as the dependent variable is in first differences, lagging the right-hand-side variables has the same effect as including a cointegrated set of regressors, provided they are either individually or jointly I(0). It is irrelevant whether the transformation is actually carried out in step one because any linear combination of the variables contains the same information. However, standard distribution theory applies only to test statistics on individual coefficients and any subset of coefficients that are jointly stationary (Bannerjee et al. 1993). We maintain that this part of time-series analysis is the least controversial outside of the fact that there tend to be few truly integrated time series in political science.

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. In time-series parlance, the rate at which a series returns to its mean after an exogenous shock is referred to as decay. If we assume Y_t follows an AR(1) process, we can use the following model:

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \varepsilon_t. \tag{3}$$

Here, the parameter for the lag of Y, α_1, dictates the rate of decay for any shocks to Y_t. Specifically, the model assumes that shocks decay at a geometric rate. 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. If data are fractionally integrated and jointly (fractionally) cointegrated, fractional cointegration captures the long-run relationship between the variables. While we can estimate the long-run relationship with these models, short-run effects, dynamic multipliers, long-run equilibria, and error correction coefficients are largely uninterpretable. 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.

[1]See the appendix for a summary of simulations demonstrating that estimates from the GECM have the usual desirable properties and that t-tests follow the standard t-distribution when the data are stationary. See the replication files for the simulation code (Keele, Linn, and Clayton 2016).

Thus far we have laid out the basic trichotomy of time-series analysis. Once the analyst decides which form of analysis is appropriate, he or she can simply pick from the above menu of models. As we highlight below, best practice is more complex.

2 Complexity in the Statistical Analysis of Time Series

2.1 *Unbalanced Time-Series Regressions*

Stable long-run relationships in turn imply balanced equations. Each of the three cases above involves balanced regressions. However, 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.[2] Instead, the data must be transformed to ensure the left and right-hand sides of the models are of equal orders of integration.

Much applied work estimates long-run relationships with little attention to the underlying dynamic properties of the time series (or their joint properties) and the existence of equation balance implied by the specified dynamic relationship. In fact, applied work citing De Boef and Keele (2008) has been used to justify the GECM writ large. While it is true that the GECM can be used to estimate long-run relationships between stationary series and between integrated and jointly cointegrated time series, such a strategy invites spurious inferences when the regressand and regressors are unbalanced.

Grant and Lebo spend extensive time demonstrating that the GECM performs poorly in a set of cases, most of which deal explicitly with unbalanced equations: explosive dependent variables and integrated or stationary regressors, integrated dependent variables and stationary regressors, stationary dependent variables and integrated regressors. These results emphasize the inapplicability of error correction models, but for the same reason they condemn regression models generally in these cases: No regression model will produce reliable inferences when the order of integration on the left- and right-hand side of our equation are different such that no long-run relationship exists between the regressand and regressors.

2.2 *Error Correction Rates, Balance, and Long-Run Relationships*

Estimates from an ECM provide some evidence not only about the nature of the long-run relationship specified but also about the appropriateness of the ECM and the likelihood that the equation is balanced. Recall that the error correction rate gives us the rate at which Y_t changes to restore the long-run equilibrium between it and X_t. Assume X_t and Y_t are out of equilibrium and call the equilibrium error e. If at time $t - 1$, e is positive, the value of Y_t is too high (above its equilibrium value), Y_t must adjust downward at time t. Similarly, if e is negative, Y_t must adjust upward at time t. Thus, the movement in Y_t is in the opposite direction of the disequilibrium. This implies that the error correction coefficient must be negative for the long-run equilibrium to be restored.

Estimated error correction rates may take on a range of values. See Table 1. Typically, error correction rates lie between 0 and -1.0. In this case, the long-run equilibrium is restored gradually. Estimated error correction coefficients nearer to -1.0 imply a quick return to the long-run equilibrium; those closer to 0 imply a slower return. 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

[2]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.

Table 1 Error correction rates and long-run equilibria

α_1^*	α_1	*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.

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 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 equilibrium value of Y_t immediately (at whatever lag they enter the model).[3] In other words, the Y_t process 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 an ADL model should be specified. Any equilibrium relationship is driven solely by the independent variables.

If the error correction coefficient is less than -2.0, no equilibrium could exist among the untransformed variables. The implication is that 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 Autoregressive conditional heteroskedasticity (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 shocks 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 an 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 model, as this is evidence of a misspecified dynamic model.

2.3 *Fractional Integration*

Fractional integration techniques present their own set of challenges. Above, we discussed the problem of interpretability. Here, we discuss the interwoven issues of estimation and inference of fractional dynamics in a single time series. Strictly speaking, the distinction between Fractionally Integrated (FI) and stationary time series is an empirical one that can be arbitrated with data. The difficulty is that time-series modeling is bedeviled with such questions. Integrated time series are said to be memoryless to distinguish them from either stationary or fractionally integrated time series. However, a voluminous amount of ink has been spilled on the subtleties of testing whether a

[3] If none of the independent variables in the model are significant, unmodelled shocks are immediately incorporated into future values of Y (this is Grant and Lebo's case 3).

series is integrated or not. And as we will show below, these subtleties extend to fractional integration.

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.[4]

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 within an empirical example with 372 observations, saying this is a very complex dynamic model (StataCorp., 2013). Moreover, 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 the data is fractionally integrated when it is not (Engle and Smith 1999; Granger and Hyung 1999; Diebold and Inoue 2001). 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 Swanson (2006) demonstrate 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. 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 (Bhardwaj and Swanson 2006).

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).[5]

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. 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 dynamics is

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

[5]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.

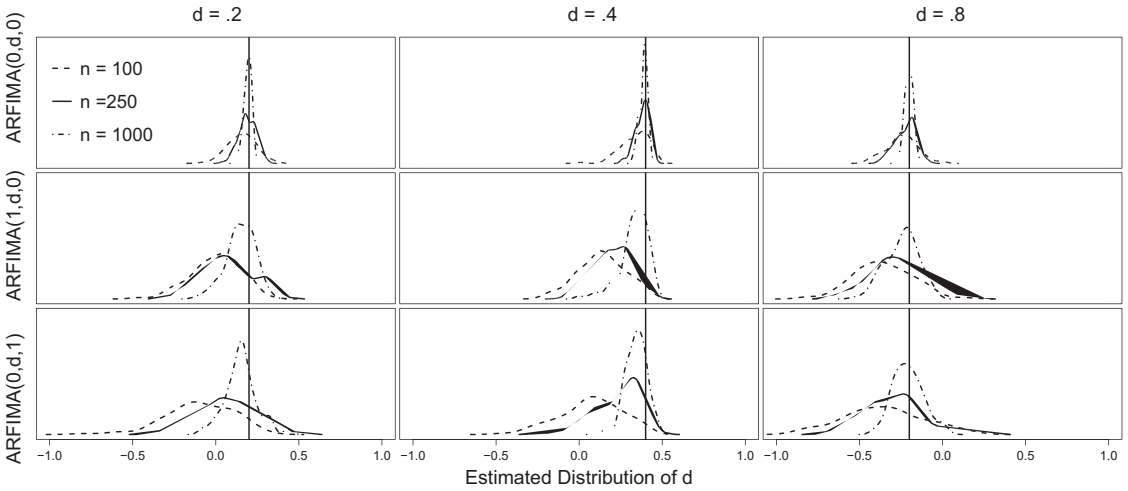

Fig. 1 Distributions of Estimates for *d*.
Each panel shows the distribution of the exact maximum likelihood estimates of *d* from the simulations for samples of size $t = 100$ (dashed line), $t = 250$ (solid line), and $t = 1000$ (dotted line). The solid vertical line in each plot represents the true value of *d*. Details of the simulations are given in the text.

known.[6] 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 Fig. 1. For the sake of clarity, we only present a subset of the results.[7] 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 $d = 0.2$, $d = 0.4$, and $d = 0.8$. Three sample sizes are presented: $t = 100$ (dashed line), $t = 250$ (solid line), and $t = 1000$ (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 $d = 0.2$ and $t = 100$, estimates range from -0.18 to 0.36 in the ARFIMA(0,d,0) model. This range increases to [-0.46, 0.33] in the ARFIMA(1,d,0) model and to [-0.80, 0.30] in the ARFIMA(0,d,1) model.[8] This uncertainty may lead to misdiagnosis of *d* in small to medium samples.

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 shows the average estimates of *d* for each model-sample combination, and column four shows

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

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

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.

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

[8]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 $t = 1000$ and $t = 1,\ 500$ MLE produces very poor estimates of *d*.

Table 2 Estimation of d

Model	*t*	*Mean*	*Rejection rate (%)*
	100	−0.032	11
ARFIMA(0,d,0)	250	−0.017	9
	1,000	−0.004	13
	100	−0.199	32
ARFIMA(1,d,0)	250	−0.121	34
	1,000	−0.073	21
	100	−0.227	16
ARFIMA(0,d,1)	250	−0.132	34
	1000	−0.019	12
	100	−0.530	69
ARFIMA(1,d,1)	250	−0.316	53
	1000	−0.056	27

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

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 Fig 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. 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 ARFIMA(1,d,0) and ARFIMA(0,d,1) models when $t=250$, and more than half the time in the ARFMA(1,d,1) model with the same sample size. This is a concern since samples of 250 observations or less are common in political science.

3 Tests, Power, and Overfitting

We end with two interrelated points: one about the power of statistical tests and one about overfitting. Selecting the correct time-series model depends on a series of statistical tests that diagnose the autoregressive properties of the data. Then, once the 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 modeled dynamics. The basic difficulty is that both types of tests have little power given the length of the typical time series in political science. Table 3 lists the length of the time series in the five applications considered by Grant and Lebo. The longest time series there is 60 time periods. This should give us pause, considering some of the simulation evidence in the literature.

Keele and Kelly (2006) compared the performance of lagged dependent variable (LDV) models as compared 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.

That implies that for all five articles replicated by Grant and Lebo, one cannot expect to have much power to detect autocorrelation in the residuals. Of course, tests for FI should generally be subject to the same constraints. Grant and Lebo note in a footnote that one needs at least 64 observations to reliably estimate the d parameter. First, that number appears to be too low relative to our simulations. More to the point, that implies that FI techniques also cannot be reliably used in any of these empirical applications. This highlights the difficulties of using time-series data with small samples. The statistical tests needed to perform critical model diagnostics have a low power.

Table 3 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

Overfitting is another problem that complicates time-series analysis in political science. 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 to the conditional expectation which 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 examplies in Grant and Lebo. For Volscho and Kelly, $N = 60$ 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. 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 issues that arise in the Grant and Lebo reanalysis are a function of overfitting, where the data are being fit to different random patterns, and thus the results are unstable.

In general, time-series analysis must take seriously that in many instances what can be learned from the data is quite limited. 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.

4 Discussion

Applied time-series analysis depends on the diagnosis and classification of time series and the selection of appropriate models. This essay highlights three points relevant to this endeavor. First, error correction models can be applied to stationary and non-stationary data alike, but the equations must be balanced. The misspecifications highlighted by Grant and Lebo apply to cases where this condition is not met. Second, we demonstrate the shortcomings of fractional integration techniques for most political science applications. Finally, we bring attention to two problems endemic to time-series analyses in political science—low power and overfitting. While most of this essay has been devoted to a discussion of when analysts should be cautious in the application of different time-series models, we would like to conclude by offering analysts advice on how to proceed.

The first step in any time-series analysis is the diagnosis of the individual time series. Analysts must classify time series as stationary, integrated, or fractionally integrated. If all the series are stationary, one can use autoregressive distributed lag and 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 and use an ADL or GECM. 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 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.

Finally, 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. 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 a GECM. As we have highlighted, distinguishing between stationary and fractionally integrated series is not easy, and the procedures conventionally used to identify whether series are fractionally integrated may not perform well given the sample sizes common in political science. Finally, analysts should take care not to overfit the data. Analysts should not fit more than one parameter per 10 observations, and may want to err on the side of not more than one parameter per 20 observations.

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):411–21.

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):539–78.

Davidson, Russell, and James G. MacKinnon. 1993. *Estimation and inference in econometrics*. New York: Oxford University Press.

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):131–59.

Engle, Robert F., and Aaron D. Smith. 1999. Stochastic permanent breaks. *Review of Economics and Statistics* 81(4):553–74.

Granger, Clive W. J. 1999. Aspects of research strategies for time series analysis. Presentation to the Conference on New Developments in Time Series Economics, Yale University.

Granger, Clive W. J., and Namwon Hyung. 1999. Occasional structural breaks and long memory. Department of Economics, UCSD.

Hauser, Michael A. 1999. Maximum likelihood estimators for ARMA and ARFIMA models: A Monte Carlo study. *Journal of Statistical Planning and Inference* 80(1):229–55.

Hendry, David F. 1995. *Dynamic econometrics*. Oxford: Oxford University Press.

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. 2016. Replication data for: Treating time With all due seriousness. http://dx.doi.org/10.7910/DVN/KD4MXV, Havard Dataverse, V1.

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.