Economics
The Open-Access, Open-Assessment E-Journal

Vol. 12, 2018-4 | January 30, 2018 | http://dx.doi.org/10.5018/economics-ejournal.ja.2018-4

Which panel data estimator should I use?: A corrigendum and extension

Mantobaye Moundigbaye, William S. Rea, and W. Robert Reed

Abstract

This study uses Monte Carlo experiments to produce new evidence on the performance of a wide range of panel data estimators. It focuses on estimators that are readily available in statistical software packages such as Stata and Eviews, and for which the number of cross-sectional units (N) and time periods (T) are small to moderate in size. The goal is to develop practical guidelines that will enable researchers to select the best estimator for a given type of data. It extends a previous study on the subject (Reed and Ye, Which panel data estimator should I use? 2011), and modifies their recommendations. The new recommendations provide a (virtually) complete decision tree: When it comes to choosing an estimator for efficiency, it uses the size of the panel dataset (N and T) to guide the researcher to the best estimator. When it comes to choosing an estimator for hypothesis testing, it identifies one estimator as superior across all the data scenarios included in the study. An unusual finding is that researchers should use different estimators for estimating coefficients and testing hypotheses. The authors present evidence that bootstrapping allows one to use the same estimator for both.

(Replication Study)

JEL C23 C33

Keywords Panel data estimators; Monte Carlo simulation; PCSE; Parks model

Authors

Mantobaye Moundigbaye, University of Canterbury, New Zealand

William S. Rea, ✉ University of Canterbury, New Zealand, bill.rea@canterbury.ac.nz

W. Robert Reed, Department of Economics and Finance, University of Canterbury, Christchurch, New Zealand

Citation Mantobaye Moundigbaye, William S. Rea, and W. Robert Reed (2018). Which panel data estimator should I use?: A corrigendum and extension. *Economics: The Open-Access, Open-Assessment E-Journal*, 12 (2018-4): 1–31. http://dx.doi.org/10.5018/economics-ejournal.ja.2018-4

Received July 25, 2017 **Published as Economics Discussion Paper** August 28, 2017
Revised November 24, 2017 **Accepted** January 9, 2018 **Published** January 30, 2018

1 Introduction

For applied researchers using panel data, there is an abundance of possible estimators one can choose. A key issue is how one decides to handle cross-sectional dependence. There are three general approaches. One approach is to model the error-variance covariance matrix in the framework of Seemingly Unrelated Regression (SUR). Here the common estimator is Feasible Generalized Least Squares (FGLS), where the cross-sectional covariances are typically modelled parametrically. The classic reference is Parks (1967) and the corresponding data-generating process (DGP) is commonly called the Parks model.

An alternative approach is to model the cross-sectional dependencies "spatially" (Anselin, 2013; Baltagi et al., 2013; Elhorst, 2014; Bivand and Piras, 2015). This typically involves modelling the dependencies across units as a function of distance, in either a continuous or binary fashion. While this has the advantage of greatly reducing the number of parameters to be estimated, it comes at the cost of possible misspecification. Misspecification occurs if the nature of the respective cross-sectional dependencies cannot be effectively reduced to a function of distance (Corrado and Fingleton, 2012).

Another alternative is to model cross-sectional correlation as a function of time-specific common factors (Pesaran and Smith, 1995; Bai, 2003; Coakley et al., 2006; Pesaran, 2006; Eberhardt et al., 2013; Kapetanios et al., 2011). This approach has proven particularly popular in the macro panel literature (Eberhardt and Teal, 2011). While the multi-factor framework for cross-sectional correlation allows one to incorporate a number of other important issues, it also comes at the cost of possible misspecification, because it greatly reduces the number of parameters to be estimated.

Despite the existence of more recent alternatives, the Parks model continues to be relevant for applied researchers. It is the underlying statistical model for Stata's *xtgls* procedure, as well as similar procedures in other software packages such SAS, Eviews, GAUSS, RATS, Shazam, and others. However, a major problem with this model is the large number of parameters that need to be estimated. In its general form, with groupwise heteroskedasticity, group-wise specific AR(1) autocorrelation, and time-invariant cross-sectional correlation, the classic Parks model has a total of $\left(\frac{N^2+3N}{2}\right)$ unique parameters in the error variance-covariance matrix (EVCM), where N is the number of cross-sectional units.

This causes two problems. First, the FGLS estimator cannot be estimated when the number of time periods, T, is less than N, because the associated EVCM cannot be inverted (Beck and Katz, 1995). Second, even when $T \geq N$, there may be relatively few observations per EVCM parameter, causing the associated elements of the EVCM to be estimated with great imprecision. As demonstrated by Beck and Katz (1995), henceforth BK, this can cause severe underestimation of coefficient standard errors, rendering hypothesis testing useless.

To address these problems, BK proposed a modification of the full GLS-Parks estimator called Panel-Corrected Standard Errors (PCSE). PCSE preserves the (Prais-Winsten) weighting of observations for autocorrelation, but uses a sandwich estimator to incorporate cross-sectional dependence when calculating standard errors. The PCSE estimator has proven very popular, as evidenced by over 2000 citations in Web of Science. All of this has opened up a myriad of choices for applied researchers when it comes to choosing a panel data estimator.

It is in this context that Reed and Ye (2011), henceforth RY, conducted Monte Carlo experiments to test a large number of OLS and FGLS-type panel data estimators, including the estimators studied by BK. They studied panel datasets for which the number of cross-sectional units (*N*) and time periods (*T*) were small to moderate in size. Cross-sectional units ranged in size from 5 to 77; and time periods ranged from 5 to 25. RY presented three recommendations to guide researchers facing the decision of which panel data estimator to use. RY has been reasonably well-cited. At the time of this writing, RY has 27 Web of Science citations and approximately 84 Google Scholar cites, indicating interest in guidance about how to choose a panel data estimator.

There are two reasons for writing this follow-up study to RY. First, there is a mistake in the design of their experiments. In attempting to construct explanatory variables that have the properties of "real world" data, they introduced additional autocorrelation that was not present in the source datasets. As autocorrelation in the explanatory variables exacerbates the effect of autocorrelation in the error term, this should affect their analysis.

Second, in their conclusion, RY called for additional experiments to confirm their recommendations. In the Parks-type error structures used by BK and RY, there are often more than a thousand unique elements in the respective EVCM. Rather than attempting to set "plausible" values for all these parameters, RY estimate these from actual datasets, and then set these estimated values as population values for the subsequent experiments. However, because RY's experiments were based on a relatively small number of datasets, there is concern that their recommendations may not apply to other datasets. A replication of RY that extended their analysis with different datasets provides an opportunity to test the validity of their recommendations.

Our study proceeds as follows. Section 2 summarizes the experimental design and datasets used for our experiments. Section 3 demonstrates that we are able to replicate RY's main findings. Section 4 presents our results using an improved method for simulating values of the explanatory variable, and additional datasets. Section 5 concludes.

2 Experimental Design

The data generating process (DGP). The experimental design for our analysis is taken from RY. Given *N* cross-sectional units and *T* time periods, we model the following DGP,

(1) $$\boldsymbol{y} = \boldsymbol{i}\beta_0 + \boldsymbol{x}\beta_x + \boldsymbol{\varepsilon},$$

where $\boldsymbol{y}$, $\boldsymbol{i}$, $\boldsymbol{x}$, are each $(NT\times 1)$ vectors, β_0 and β_x are scalars, and $\boldsymbol{\varepsilon}$ is an $(NT\times 1)$ vector of error terms such that

(2) $$\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \boldsymbol{\Omega}_{\boldsymbol{NT}}),$$

where

(3) $$\boldsymbol{\Omega}_{\boldsymbol{NT}} = \boldsymbol{\Sigma} \otimes \boldsymbol{\Pi},$$

(4) $$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{\varepsilon,11} & \sigma_{\varepsilon,12} & \cdots & \sigma_{\varepsilon,1\mathrm{N}} \\ \sigma_{\varepsilon,21} & \sigma_{\varepsilon,22} & \cdots & \sigma_{\varepsilon,2\mathrm{N}} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{\varepsilon,\mathrm{N}1} & \sigma_{\varepsilon,\mathrm{N}2} & \cdots & \sigma_{\varepsilon,\mathrm{NN}} \end{bmatrix}, \text{ and}$$

(5) $$\boldsymbol{\Pi} = \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{T-1} \\ \rho & 1 & \rho & \cdots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{T-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \cdots & 1 \end{bmatrix}.$$

$\boldsymbol{\Omega_{NT}}$ incorporates groupwise heteroskedasticity, time-invariant cross-sectional dependence, and first-order, common autocorrelation.[1] To get realistic values for the respective EVCM elements, $\sigma_{\varepsilon,ij}$ and ρ, we estimate these parameters from actual datasets, using the same procedures that Stata and Eviews use in calculating their respective FGLS estimators.

Creation of simulated panel datasets. Table 1 lists the datasets that were employed in obtaining population parameter values for the DGPs in the Monte Carlo experiments. In order to evaluate the recommendations provided by RY, albeit with a corrected experimental design, we start with the same datasets they used. These are listed in the top panel of Table 1. However, we also use additional datasets that were not considered by RY. These are listed in the bottom panel of Table 1 ("new datasets").

The first set of experiments draw data from the Penn World Table. For a given sized panel dataset, say N=5 and T=5, we take the first N cross-sectional units and regress the log of real GDP on the ratio of government expenditures to GDP and a set of country fixed effects for the first T available time periods. We save the residuals from that regression. We then use those residuals to obtain estimates of the individual elements of the EVCM, $\hat{\sigma}_{\varepsilon,ij}$, $i,j = 1,2,\ldots,N$, and $\hat{\rho}$.[2] We then repeat that procedure for all possible samples of T contiguous years. These estimates are then averaged to obtain a "representative" EVCM,

(3') $$\overline{\overline{\boldsymbol{\Omega}}}_{\boldsymbol{NT}} = \begin{bmatrix} \overline{\hat{\sigma}}_{\varepsilon,11} & \overline{\hat{\sigma}}_{\varepsilon,12} & \cdots & \overline{\hat{\sigma}}_{\varepsilon,1\mathrm{N}} \\ \overline{\hat{\sigma}}_{\varepsilon,21} & \overline{\hat{\sigma}}_{\varepsilon,22} & \cdots & \overline{\hat{\sigma}}_{\varepsilon,2\mathrm{N}} \\ \vdots & \vdots & \ddots & \vdots \\ \overline{\hat{\sigma}}_{\varepsilon,\mathrm{N}1} & \overline{\hat{\sigma}}_{\varepsilon,\mathrm{N}2} & \cdots & \overline{\hat{\sigma}}_{\varepsilon,\mathrm{NN}} \end{bmatrix} \otimes \begin{bmatrix} 1 & \overline{\hat{\rho}} & \overline{\hat{\rho}}^2 & \cdots & \overline{\hat{\rho}}^{T-1} \\ \overline{\hat{\rho}} & 1 & \overline{\hat{\rho}} & \cdots & \overline{\hat{\rho}}^{T-2} \\ \overline{\hat{\rho}}^2 & \overline{\hat{\rho}} & 1 & \cdots & \overline{\hat{\rho}}^{T-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \overline{\hat{\rho}}^{T-1} & \overline{\hat{\rho}}^{T-2} & \overline{\hat{\rho}}^{T-3} & \cdots & 1 \end{bmatrix}.$$

---

[1] Following BK and RY, we set the AR(1), autocorrelation parameter, ρ, to be the same for all cross-sectional units.

[2] The reason we do not average cross-sectional covariances over different sets of cross-sectional units is that it would work against our goal of producing parameters that "looked like" real world data. There is no reason to believe that averaging the cross-sectional covariances of, say, (i) the US and France, (ii) the US and South Africa, and (iii) Russia and Fiji would produce anything that looked like a cross-sectional covariance from an actual pair of countries.

Table 1: Description of Datasets Used to Generate Population Parameters

Dataset	*Source*	*Dependent Variable*	*Independent Variables*	*N*	*T*
			REED AND YE (2011) DATASETS		
1	Penn World Table	Log of real GDP	Ratio of government expenditures to GDP Country fixed effects	5, 10, 20, 50, 77	5, 10, 15, 20, 25
2	Penn World Table	Real GDP growth	Ratio of government expenditures to GDP Country fixed effects	5, 10, 20, 50, 77	5, 10, 15, 20, 25
3	Reed (2008)	Log of real state PCPI	Tax Burden State fixed effects	5, 10, 20, 48	5, 10, 15, 20, 25
4	Reed (2008)	Real state PCPI growth	Tax Burden State fixed effects	5, 10, 20, 48	5, 10, 15, 20, 25
5	Penn World Table	Log of real GDP	Ratio of government expenditures to GDP Country fixed effects Time fixed effects	5, 10, 20, 50, 77	5, 10, 15, 20, 25
6	Penn World Table	Real GDP growth	Ratio of government expenditures to GDP Country fixed effects Time fixed effects	5, 10, 20, 50, 77	5, 10, 15, 20, 25
7	Reed (2008)	Log of real state PCPI	Tax Burden State fixed effects Time fixed effects	5, 10, 20, 48	5, 10, 15, 20, 25
8	Reed (2008)	Real state PCPI growth	Tax Burden State fixed effects Time fixed effects	5, 10, 20, 48	5, 10, 15, 20, 25
			NEW DATASETS		
9	Kersting & Kilby (2014)	Gross Aid Disbursement as Share of GDP (Germany Aid Allocation)	Freedom House Score Country fixed effects	5, 10, 20, 50, 77	10, 15, 20, 25
10	Casper & Tufis (2003)	Vanhanen's Democracy Index	Primary education enrolment (share of population) Country fixed effects	5, 10, 20, 50	10, 15, 20, 25
11	Biagi et al. (2012)	Crime per 100000 inhabitants	Tourists arrivals per square kilometre Country fixed effects	5, 10, 20, 50, 77	10, 15, 19

Dataset	***Source***	***Dependent Variable***	***Independent Variables***	***N***	***T***
12	Nunn & Qian (2014)	Any Conflict	US-Wheat Aid (1000 MT) Country fixed effects	5, 10, 20, 50, 77	10, 15, 20, 25
13	Kersting & Kilby (2014)	Gross Aid Disbursement as Share of GDP (Germany Aid Allocation)	Freedom House Score Country fixed effects Year fixed effects	5, 10, 20, 50, 77	10, 15, 20, 25
14	Casper & Tufis (2003)	Vanhanen's Democracy Index	Primary education enrolment (share of population) Country fixed effects Year fixed effects	5, 10, 20, 50	10, 15, 20, 25
15	Biagi et al. (2012)	Crime per 100000 inhabitants	Tourists arrivals per square kilometre Country fixed effects Year fixed effects	5, 10, 20, 50, 77	10, 15, 19
16	Nunn & Qian (2014)	Any Conflict	US-Wheat Aid (1000 MT) Country fixed effects Year fixed effects	5, 10, 20, 50, 77	10, 15, 20, 25

To obtain a representative vector of $\boldsymbol{x}$ values, we randomly select one contiguous, T-year period.[3] Let these values be given by $\widetilde{\boldsymbol{x}}$. Then for given values of β_0 and β_x, we generate simulated $\widetilde{\boldsymbol{y}}$ values from the following DGP:

(6) $$\widetilde{\boldsymbol{y}} = \boldsymbol{i}\beta_0 + \widetilde{\boldsymbol{x}}\beta_x + \widetilde{\boldsymbol{\varepsilon}},$$

where $\widetilde{\boldsymbol{\varepsilon}}$ consists of simulated, normally distributed error terms having mean 0 and an EVCM equal to $\overline{\overline{\boldsymbol{\Omega}}}_{\boldsymbol{NT}}$. The vector of $\widetilde{\boldsymbol{y}}$ and $\widetilde{\boldsymbol{x}}$ values are then used to obtain estimates of β_x for each of the estimators under study.

This procedure was followed for each of the N and T values listed in Table 1, and each of the respective datasets.[4] Note that each N and T pair produces a unique set of $\hat{\sigma}_{\varepsilon,ij}$, $i,j = 1,2,\ldots,N$, and $\hat{\rho}$ values (and thus unique EVCM), as well as unique $\widetilde{\boldsymbol{x}}$ values. Accordingly, each of the original sixteen datasets becomes the parent for anywhere from 15 to 25 artificial datasets, depending on the number of possible (N,T) combinations. These "offspring" datasets, besides having different sizes, also have different characteristics. For example, a cross-country dataset that has level of income as its dependent variable and that includes the world's largest economies such as the US, China, Germany, Japan, and the UK, will have very different heteroskedasticity characteristics than a dataset that omits these countries. Further, cross-country dependencies will vary greatly depending on the specific countries that are included.

The datasets listed in Table 1 are quite diverse. In particular, the new datasets listed in the bottom panel are distinctly different from the original RY datasets. The original RY datasets used dependent variables that were income-based, either cross-country/GDP values (level and growth) or US state/PCPI values (level and growth). In contrast, the dependent variables for the new datasets are (i) international aid (Datasets 9 and 13), (ii) a democracy index (Datasets 10 and 14); (iii) crime per capita (Datasets 11 and 15), and (iv) a binary variable indicating conflict (Datasets 12 and 16). And not just the dependent variables, but the explanatory variables are very different. This should produce a wide variety of artificial panel datasets having very different EVCMs.

---

[3] RY made a mistake in their experimental design by averaging the $\boldsymbol{X}$ values. This introduced excessive autocorrelation in $\widetilde{\boldsymbol{x}}$. When the error terms are serially correlated, the serial correlation in the regressor affects the variance of its OLS coefficient estimator variance. The following relationship connects the variance of OLS slope estimator characterised by first order serial correlation of both the error term and the regressor, $Var(\hat{\beta}_{AR(1)})$, on the one hand, and that of the usual OLS slope estimator, $Var(\hat{\beta}_{OLS})$, on the other (see Gujarati 2004, p 452): $Var(\hat{\beta}_{AR(1)}) = Var(\hat{\beta}_{OLS})\left(\frac{1+r\rho}{1-r\rho}\right)$, where r and ρ denote the first order serial correlation coefficients of the regressor and the error term respectively. Thus, exaggerating the serial correlation in the regressor worsens the bias in the estimated coefficient standard error. We note that Beck and Katz (1995) made a related error on the other side in their Monte Carlo experiments by generating x_{it} values that were "random draws from a zero-mean normal distribution" (BK, page 638). By ignoring the role of autocorrelation in the explanatory variable, they diminished the problems caused by autocorrelation. This was pointed out in a replication study by Reed and Webb (2010).

[4] The maximum N and T values listed in Table 1 are often less than the size of the panel dataset in the original dataset. For example, the original Dataset 1 used by RY contained data on 97 countries for 40 years (1961-2000). However, data issues, usually caused by problems with the Cholesky decomposition function in creating simulated error terms, forced us to limit the sizes of some of the panel datasets. For the same reason, the actual number of datasets we were able to create is less than the total possible combinations from pairing all possible N and T values in the table.

Table 2: List and Description of Panel Data Estimators to Be Studied

Estimator	***Procedure***	***Assumed Error Structure***
1	OLS-1A	IID
2	OLS-1B	Robust heteroskedasticity
3	OLS-1C	Robust heteroskedasticity + Robust autocorrelation
4	OLS-1D	Robust heteroskedasticity + Robust cross-sectional dependence
5	FGLS-1A	Groupwise heteroskedasticity
6	FGLS-2	Groupwise heteroskedasticity + autocorrelation
7	FGLS-3 (Parks)	Groupwise heteroskedasticity + autocorrelation + cross-sectional dependence
8	FGLS-4 (PCSE)	Groupwise heteroskedasticity + autocorrelation + cross-sectional dependence
9	FGLS-1B	Weight = Groupwise heteroskedasticity Var-Cov = Robust heteroskedasticity + Robust cross-sectional dependence
10	FGLS-1C	Weight = Groupwise heteroscedasticity Var-Cov = Robust heteroskedasticity + Robust autocorrelation
11	FGLS-1D	Weight = Groupwise heteroskedasticity Var-Cov = Robust heteroskedasticity

Note: Interpretation of the numbering and lettering of the procedures is given in Section 2 in the text. Further details about the estimator is given in the Appendix.

The estimators. Through this gauntlet of diverse data environments we run the respective estimators. These are identified in Table 2. These are the same estimators studied by RY. All of the estimators correspond to a particular Stata or Eviews panel data estimator.[5] Each estimator is a special case of the following:

(7) $$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{W}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{W}^{-1}\boldsymbol{y}$$

(8) $$Var(\widehat{\boldsymbol{\beta}}) = \left(\boldsymbol{X}'\boldsymbol{W}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{W}^{-1}\widehat{\boldsymbol{\Omega}}\boldsymbol{W}^{-1}\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{W}^{-1}\boldsymbol{X}\right)^{-1}$$

where $\boldsymbol{X} = [\boldsymbol{i} \quad \widetilde{\boldsymbol{x}}]$, $\widehat{\boldsymbol{\beta}} = [\hat{\beta}_0 \quad \hat{\beta}_x]$, $\boldsymbol{W}$ is the "weighting" matrix, and $\widehat{\boldsymbol{\Omega}}$ is the estimated EVCM.[6] For example, in the case of OLS with an assumed IID error structure (Estimator 1), $\boldsymbol{W} = \boldsymbol{I}$ and $\widehat{\boldsymbol{\Omega}} = \hat{\sigma}^2\boldsymbol{I}$. In the case of Estimator 5 (FGLS-1A), $\boldsymbol{W} = \widehat{\boldsymbol{\Omega}}$, where $\widehat{\boldsymbol{\Omega}}$ is the diagonal matrix with group-specific variances on the main diagonal. Estimator 9 (FGLS-1B) has the same weighting matrix $\boldsymbol{W}$, and thus produces an identical estimate, $\widehat{\boldsymbol{\beta}}$, but estimates $\boldsymbol{\Omega}$ using a robust estimator that clusters on time period, and thus produces different standard errors than Estimator 5.

Table 2 employs the notation that estimators with the same weighting matrix $\boldsymbol{W}$ have the same number index. Estimators with different $\widehat{\boldsymbol{\Omega}}$ matrices have different letter indices. So all the FGLS-1 estimators use the same weighting matrix (based on groupwise heteroskedasticity), but FGLS-1A calculates different standard errors than FGLS-1B, FGLS-1C, and FGLS-1D.

Estimators 1, 7, and 8 are particularly worth noting. Estimator 1 is conventional (pooled) OLS.[7] This will serve as the benchmark estimator against which the other estimators will be compared. Estimator 7 is the Parks estimator. It is asymptotically efficient, but requires $T \geq N$.[8] Estimator 8 is BK's PCSE estimator which has become a popular substitute for the Parks estimator because of its claimed finite sample advantages.

Performance measures. The experiments compare the respective panel data estimators on two dimensions, efficiency and accuracy in hypothesis testing. An experiment consists of 1000 replications, where each replication draws a unique panel data sample simulated from a common DGP, corresponding to given "offspring" dataset. For each experiment and each estimator, we calculate an *EFFICIENCY* value defined by,

(9) $$EFFICIENCY_{Estimator} = 100 \cdot \frac{\sqrt{\sum_{r=1}^{R}\left(\hat{\beta}_{Estimator}^{(r)} - \beta_x\right)^2}}{\sqrt{\sum_{r=1}^{R}\left(\hat{\beta}_{OLS}^{(r)} - \beta_x\right)^2}},$$

where β_x is the true value of the slope coefficient, and $\hat{\beta}_{OLS}^{(r)}$ and $\hat{\beta}_{Estimator}^{(r)}$ are the estimated values of β_x in a given replication *r* as estimated by OLS and the estimator that is being

[5] The Appendix lists the specific commands in Stata or Eviews that correspond to each estimator.

[6] Note that $\widehat{\boldsymbol{\Omega}} \neq \overline{\overline{\boldsymbol{\Omega}}}_{NT}$. $\overline{\overline{\boldsymbol{\Omega}}}_{NT}$ is the population EVCM used in the DGP to generate the simulated $\widetilde{\boldsymbol{y}}$ and $\widetilde{\boldsymbol{x}}$ data. $\widehat{\boldsymbol{\Omega}}$ is the EVCM estimated from residuals generated by regressing $\widetilde{\boldsymbol{y}}$ on $\widetilde{\boldsymbol{x}}$.

[7] Note that the DGP does not contain fixed effects, so we omit fixed effects estimators from the choice set.

[8] It is possible to estimate the full Parks model in Stata when $T < N$. This is made possible through the use of a generalized inverse function in Stata that allows one to invert matrices that are not full rank. However, our own investigations indicate that the resulting estimators do not perform well.

compared to OLS, respectively. Smaller values indicate a more efficient estimator. OLS is defined to have an *EFFICIENCY* value of 100. Estimators with *EFFICIENCY* values less than 100 are thus more efficient than OLS for datasets having the given characteristics.

To measure accuracy in hypothesis testing, we calculate two measures. The first is the coverage rate, *Coverage*, defined as the percent of 95% confidence intervals around $\hat{\beta}_x$ that include the true value of β_x. We also calculate the absolute value of the difference between 95% and the coverage rate, |95 – *Coverage*|. Estimators for which |95 – *Coverage*| is closest to zero are judged to be superior with respect to accuracy in hypothesis testing.

As seen in Table 2, estimators 5, 9, 10, and 11 all share the same weighting matrix, $\boldsymbol{W}$, weighting solely on (groupwise) heteroskedasticity. As a result, these estimators will produce identical coefficient estimates $\widehat{\boldsymbol{\beta}}$ when using the same data (cf. Equation 7). Thus, in comparing estimators on the dimension of efficiency, we treat these estimators as one and refer to Estimator 5/9/10/11. When it comes to assessing their accuracy in hypothesis testing, they will be treated separately because they produce different estimates of $Var(\widehat{\boldsymbol{\beta}})$ (cf. Equation 8).

RY's three recommendations. Based on their analysis of the performances of the eleven estimators in Table 3, RY provide three recommendations.[9]

1. When the primary concern is efficiency and *T/N* ≥ *1.50*, use *Estimator 7*.
2. When the primary concern is efficiency, *T/N* < *1*, and *Heteroskedasticity* > *1.67,* use either *Estimator 5* or *Estimator 6.*
3. When the primary concern is constructing accurate confidence intervals and *Autocorrelation* < *0.30,* use either *Estimator 8* or *Estimator 4.*

These recommendations are designed as guides for applied researchers, mapping observed/measurable characteristics of the data – such as the ratio of time periods to units, or the degree of heteroskedasticity or autocorrelation – to the choice of a "best" estimator.

Two things are noteworthy in this regard. First, the recommendations have "gaps." For example, when choosing estimators on the basis of efficiency, there is a recommendation for cases where *T/N* ≥ 1.50 and *T/N* < *1,* but nothing for 1 ≤ *T/N* < 1.50. And when it comes to selecting an estimator based on accurate confidence intervals, and hence preferred for hypothesis testing, there is no recommendation when *Autocorrelation* ≥ 0.30. The reason for these gaps is that RY could not identify a consistently best estimator for these data situations.

Also noteworthy is the fact that RY recommend different estimators depending on whether one's primary interest is efficiency or accuracy in hypothesis testing. While this is unusual, it is not contradictory. The expression for $Var(\widehat{\boldsymbol{\beta}})$ in Equation (8) does not have finite sample validity. The substitution of $\widehat{\boldsymbol{\Omega}}$ for $\boldsymbol{\Omega}$ is justified on the basis of the "analogy principle" (Manski, 1988). While correct asymptotically – assuming the respective estimates of the EVCM elements are consistent – it may be a better or worse substitute in finite samples for some estimators versus others depending on the specifics of the deviation between $\widehat{\boldsymbol{\Omega}}$ and $\boldsymbol{\Omega}$. Further, because $\widehat{\boldsymbol{\Omega}}$ factors differently into Equations (7) and (8), it is possible that this deviation affects an estimator's relative performance in hypothesis testing more or less than its relative performance in efficiency.

[9] In order to make their recommendations easier to understand, we have replaced their terminology with the nomenclature from this paper. The substituted terms are italicized.

To summarize, RY's recommendations provide a potentially useful guide to applied researchers facing a choice of panel data estimators. However, their recommendations are incomplete, and they have the unusual feature of advising different estimators for coefficient efficiency and accuracy in hypothesis testing. While their analysis introduced additional autocorrelation in the simulated values of the explanatory variables, it's not clear to what extent this affected their results. Our analysis attempts to see whether correcting this mistake alters their recommendations, and if it does, whether the new recommendations are robust when these recommendations are applied to entirely new datasets.

Table 3: Replication of Table 3 in Reed and Ye (2011)

	Reed and Ye (2011)			
	Average EFFICIENCY		***Percent of experiments where estimator is more efficient than OLS***	
	$N \leq T$	$N > T$	$N \leq T$	$N > T$
Estimator 5/9/10/11	95.2	82.9	58.8	84.4
Estimator 6	95.1	83.1	71.3	79.7
Estimator 7	73.9	---	96.3	--
Estimator 8	100.8	101.0	62.5	51.6
	Replication			
	Average EFFICIENCY		***Percent of experiments where estimator is more efficient than OLS***	
	$N \leq T$	$N > T$	$N \leq T$	$N > T$
Estimator 5/9/10/11	95.2	82.9	58.8	84.4
Estimator 6	95.1	82.6	71.3	81.3
Estimator 7	73.7	--	96.3	--
Estimator 8	100.8	101.0	63.8	57.8

3 Replication

Reproducing the RY's original results. The first step in our analysis is to confirm that we are able to reproduce RY's findings. As we had access to their computer code, this was straightforward. The top panel of Table 3 copies the values from Table 3 in RY and reports two measures of efficiency for the different panel data estimators: (i) average *EFFICIENCY*, and (ii) the percent of experiments where the estimator is more efficient than OLS. The experimental results are reported separately for datasets having $N \leq T$ and $N > T$. We recall that lower values indicate greater efficiency.

The bottom panel of Table 3 reports our replication of RY's results. We obtain very similar results. Note that our replication is unable to exactly reproduce their results. Because the datasets are randomly generated, differences will necessarily arise due to sampling error. However, as the results are averaged over 1000 replications for each experiment, these differences are expected to be relatively small. Indeed, that is the case.

Table 4 repeats the replication exercise, this time focussing on two measures of accuracy in hypothesis testing: (i) coverage rates, and (ii) the absolute value of the difference between 95 and the coverage rate. Once again, the top panel copies the results from RY (see Table 5 in RY). The bottom panel reports our replication. While there are differences, they are, again, relatively small.

Having confirmed that we are able to reproduce RY's key results for efficiency and accuracy in hypothesis testing, we next turn to correcting the mistake in RY and re-doing their analysis using both the datasets that they used, and extending their analysis to a new set of datasets.

Correlation in the explanatory variable. As discussed above, RY's original procedure introduced excessive autocorrelation in the explanatory variable. Table 5 illustrates the extent of the problem using Dataset 1 (see Table 1). There are a total of 25 different *NT* combinations, (N=5/T=5, N=5/T=10, … , N=77/T=20, N=77/T=25). For each of these *NT* combinations, we simulated 1000 datasets, first using RY's original method, then using the corrected method.

RY's original method is best explained via example. Let the values of the explanatory variable x for a given cross-sectional unit i from the parent "real" dataset be given by (x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}, x_{i6}, … , x_{iT}). To create the corresponding values of x for a simulated panel dataset having size $T=5$, RY take all possible, contiguous 5-year periods: (x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}), (x_{i2}, x_{i3}, x_{i4}, x_{i5}, x_{i6}), (x_{i3}, x_{i4}, x_{i5}, x_{i6}, x_{i7}), … ($x_{i,T-4}$, $x_{i,T-3}$, $x_{i,T-2}$, $x_{i,T-1}$, x_{iT}). $\tilde{x}_1$ is calculated as the average of the first element across all possible sets of contiguous, 5-year periods, $\tilde{x}_2$ is the average of the second element across all 5-year sets, and so on. Note that the averages $\tilde{x}_1$ and $\tilde{x}_2$ will share a large number of underlying x values, so that a regression of $\tilde{x}_t$ on $\tilde{x}_{t-1}$ will produce a high degree of "autocorrelation" in excess of any autocorrelation that may exist in the underlying values of x. The corrected method randomly selects one 5-year period of x values from the set of all possible 5-year sets and thus avoids this manufactured, spurious autocorrelation.

Table 5 compares the average, estimated AR(1) coefficient for the explanatory variable in each of the simulated, *NT* datasets generated from Dataset 1 using both RY"s method and the corrected method. For example, for simulated datasets having dimension N=5,T=5, RY's method produces explanatory variables having an average, estimated AR(1) parameter equal to 0.979. In contrast, randomly selected, 5-year contiguous periods have an average, estimated AR(1) coefficient of 0.602. Across all simulated datasets, RY's method produces an average autocorrelation of 0.969, with minimum and maximum values of 0.926 and 0.995. In contrast, the corrected method produces an average autocorrelation coefficient of 0.635, with minimum and maximum values of 0.204 and 0.833. The next section reports the results of our analysis using the corrected method for generating values of the explanatory variable. We also extend RY's analysis by expanding the set of datasets used in our simulations.

Table 4: Replication of Table 5 in Reed and Ye (2011)

	Reed and Ye (2011)			
	$N \leq T$		$N > T$	
	COVERAGE	***Absolute value of (95-COVERAGE) over all experiments***	***COVERAGE***	***Absolute value of (95-COVERAGE) over all experiments***
Estimator 1	73.6	21.9	74.2	21.9
Estimator 2	73.7	21.8	77.9	18.8
Estimator 3	83.5	11.6	91.8	3.9
Estimator 4	72.7	22.5	74	21.3
Estimator 5	69.8	25.6	72.6	22.9
Estimator 6	86.4	9.3	88.8	7.2
Estimator 7	43.3	51.7	--	--
Estimator 8	87.8	7.2	88.1	6.9
Estimator 9	66.1	28.9	65.4	29.6
Estimator 10	68.1	26.9	80.1	14.9
Estimator 11	69.5	25.9	72.4	23.2
	Replication			
	$N \leq T$		$N > T$	
	COVERAGE	***Absolute value of (95-COVERAGE) over all experiments***	***COVERAGE***	***Absolute value of (95-COVERAGE) over all experiments***
Estimator 1	73.6	21.9	75.7	20.5
Estimator 2	73.7	21.8	79.3	17.5
Estimator 3	83.5	11.6	92.7	3.0
Estimator 4	72.7	22.5	75.8	19.6
Estimator 5	69.8	25.6	74.1	21.4
Estimator 6	86.4	9.3	90.2	5.5
Estimator 7	43.3	51.7	--	--
Estimator 8	87.8	7.2	89.1	5.9
Estimator 9	66.1	28.9	66.7	28.3
Estimator 10	68.1	26.9	81.5	13.5
Estimator 11	69.5	25.9	73.9	21.7

Table 5: Example of Excessive Autocorrelation in Reed and Ye (2011)

N	***T***	***Average Autocorrelation in Explanatory Variable***	
		Reed and Ye (2011)	***Corrected***
5	***5***	0.979	0.602
5	***10***	0.995	0.608
5	***15***	0.970	0.690
5	***20***	0.959	0.798
5	***25***	0.976	0.805
10	***5***	0.980	0.204
10	***10***	0.993	0.581
10	***15***	0.970	0.716
10	***20***	0.964	0.768
10	***25***	0.975	0.751
20	***5***	0.953	0.239
20	***10***	0.971	0.598
20	***15***	0.973	0.645
20	***20***	0.973	0.755
20	***25***	0.979	0.833
50	***5***	0.955	0.261
50	***10***	0.943	0.557
50	***15***	0.963	0.720
50	***20***	0.971	0.777
50	***25***	0.978	0.797
77	***5***	0.926	0.318
77	***10***	0.957	0.615
77	***15***	0.969	0.706
77	***20***	0.974	0.738
77	***25***	0.979	0.789
Minimum		0.926	0.204
Maximum		0.995	0.833
Average		0.969	0.635

Note: Autocorrelations are estimated for the independent variable in Dataset 1 (see Table 1) under the (i) original Reed and Ye (2011) and (ii) corrected experimental designs.

4 Results[10]

Sample characteristics of simulated datasets. Table 6 provides descriptive statistics for the elements of the EVCMs estimated from the simulated datasets. Reported are measures of heteroskedasticity, autocorrelation, and cross-sectional dependence. As before, the top panel reports details about the original RY datasets, while the new datasets are featured in the bottom panel, Within each panel, datasets are divided depending on whether $T \geq N$ or $T < N$. Each dataset produces either ten or eleven observations of $\hat{\beta}_x$, one for each estimator (more on the estimators below). There are fewer observations per dataset when $T < N$, because, as noted above, one of the estimators (the fully specified FGLS with heteroskedasticity, autocorrelation, and cross-sectional dependence; also known as the Parks estimator), cannot be estimated in this case.

Heteroskedasticity is calculated from a given dataset's group-specific variances. We sort the associated standard deviations and take the ratio of the 3rd and 1st quartile values, $\left(\frac{\sqrt{\bar{\hat{\sigma}}_{\varepsilon,3rd\ quartile}}}{\sqrt{\bar{\hat{\sigma}}_{\varepsilon,1st\ quartile}}}\right)$.[11] Larger values indicate greater heteroskedasticity. Autocorrelation is estimated by $\bar{\hat{\rho}}$. These values should range between -1 and 1, with the expectation that most of the AR(1) parameters will be positive. Cross-sectional dependence is measured by the absolute values of the cross-sectional correlations, averaged over all possible pairs of cross-sectional units. These, in turn, are calculated from the respective cross-sectional covariances, $\bar{\hat{\sigma}}_{\varepsilon,ij}$, $i,j = 1,2,\ldots,N$, $i \neq j$. These values should also range between 0 and 1.

Both the original RY datasets and the datasets new for this study demonstrate a wide range of error behaviours. Heteroskedasticity ranges from a low of 1.21 to a high of 40.21.[12] Autocorrelation ranges from –0.06 to 0.79, and cross-sectional correlation from 0.20 to 0.79. The new datasets are generally characterized by greater heteroskedasticity and cross-sectional dependence, but lesser autocorrelation.

Efficiency. This section compares the performance of the respective estimators. All the results follow the procedures discussed above, and incorporate the correction to RY's original experimental design.[13] Table 7 reports average performance measures for efficiency for different subgroups of experiments. The first two columns report average EFFICIENCY values for all experiments according to whether T/N ≤ 1.5 or T/N > 1.5, where EFFICIENCY is calculated using Equation (9). We choose the cut-off of 1.5 to be consistent with RY's first

[10] Data and code to replicate the results in this paper are posted at Dataverse:
https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910%2FDVN%2FYKSATT.

[11] Note that the $\bar{\hat{\sigma}}_{\varepsilon}$ terms are variances, and not standard deviations.

[12] The particularly high heteroskedasticity values come from Datasets 12 and 16, where the dependent variable is zero-one.

[13] See Section 3 for our replication of RY's results without correction to their experimental design.

Table 6: Description of Simulated Datasets Used in the Experiments

		Heteroskedasticity	***Autocorrelation***	***Cross-sectional Dependence***
		REED AND YE (2011) DATASETS		
$N \leq T$ (80 datasets; 880 observations)	***Minimum***	1.21	-0.06	0.20
	Mean	1.68	0.36	0.44
	Maximum	2.35	0.78	0.90
$N > T$ (64 datasets; 640 observations)	***Minimum***	1.34	-0.04	0.22
	Mean	1.76	0.34	0.43
	Maximum	2.25	0.79	0.79
		NEW DATASETS		
$N \leq T$ (72 datasets; 792 observations)	***Minimum***	1.26	0.08	0.22
	Mean	4.47	0.47	0.35
	Maximum	40.21	0.73	0.52
$N > T$ (68 datasets; 680 observations)	***Minimum***	1.47	0.16	0.23
	Mean	6.93	0.46	0.34
	Maximum	34.91	0.73	0.49

Note: For more details on the construction of the simulated datasets, see Section 2 in the text.

Table 7: Comparison of Estimator *EFFICIENCY*

Estimator	***Average EFFICIENCY***		***Percentage of Times the Estimator Is More Efficient Than OLS***	
	T/N > 1.5 ***(1)***	***T/N ≤ 1.5*** ***(2)***	***T/N > 1.5*** ***(3)***	***T/N ≤ 1.5*** ***(4)***
		REED AND YE (2011) DATASETS		
Estimator 5/9/10/11	96.6	84.2	68.8	78.9
Estimator 6	82.8	74.7	75.0	89.1
Estimator 7 (Parks)	45.5	66.1*	100.0	100.0*
Estimator 8 (PCSE)	86.9	89.1	62.5	72.7
		NEW DATASETS		
Estimator 5/9/10/11	70.8	54.5	88.6	97.9
Estimator 6	61.5	48.0	97.7	99.0
Estimator 7 (Parks)	46.9	80.1*	97.7	96.4*
Estimator 8 (PCSE)	85.1	92.1	95.5	80.2

* The results for Estimator 7 are not comparable to the other estimators when *T/N ≤ 1.5* because they are based on a subset of the experiments, since Estimator 7 cannot be estimated when *T/N < 1.0*.

Note: The *EFFICIENCY* measure is defined in Section 2 in the text. Yellow-coloured cells indicate "best" estimator for a given data-ty recommendation, and also based on our own analysis. The next two columns provide a different perspective on efficiency. They report the percent of experiments where a given estimator is more efficient than OLS. The "best" estimators are indicated by yellow-highlighting the respective cells in the table.

The top panel reports the results for the datasets used by RY. According to RY's first recommendation, when researchers are primarily interested in efficiency and *T/N* is greater than *1.5*, they should choose Estimator 7, the Parks estimator. Our findings confirm this recommendation for the Reed and Ye (2011) datasets. When *T/N* > 1.5, the average *EFFICIENCY* of Estimator 7 is 45.5, substantially lower than that of the other estimators. Moreover, Estimator 7 is always better than OLS (100 percent). The other estimators are more efficient than OLS most, but not all, of the time.

The bottom part of the panel reports the results of experiments based on the new datasets. This represents a clean "out of sample" test of RY's recommendation, because none of these datasets were included in RY's analysis. Focussing again on the experiments where *T/N* > 1.5, we see that Estimator 7 (FGLS-Parks) has a much lower average *EFFICIENCY* value than the other estimators. Further, it is more efficient than OLS approximately 98 percent of the time, tied for best most among all estimators.

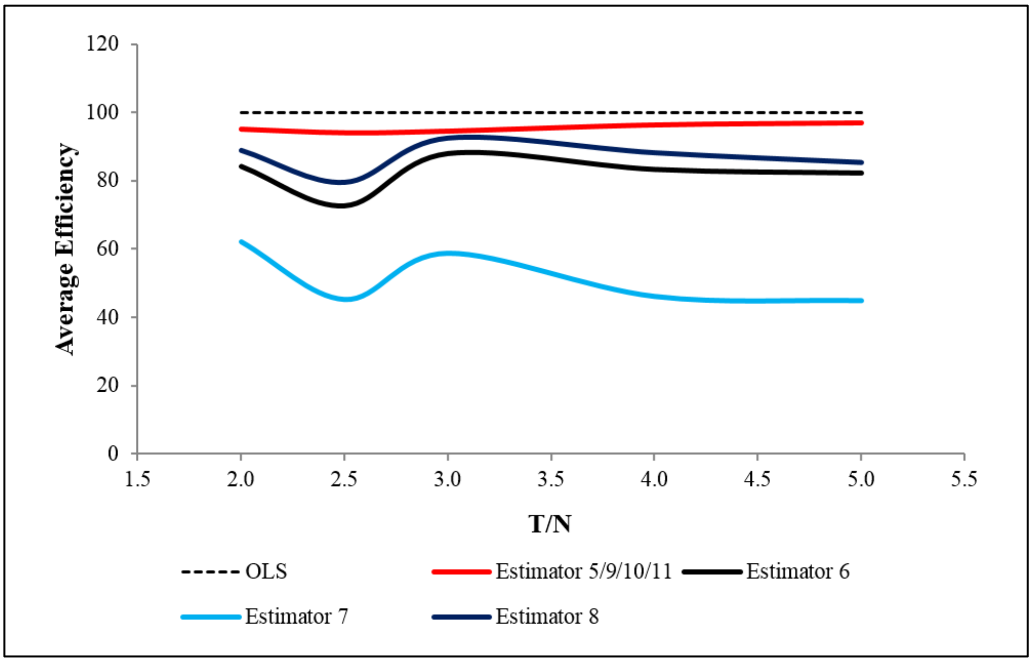
Averages can mask much variation. Accordingly, Figures 1 and 2 plot the average *EFFICIENCY* values for each of the estimators as a function of *T/N* when *T/N > 1.5*. Figure 1 does this for the RY datasets, and Figure 2 does this for the new datasets. The dotted, black line at Average Efficiency = 100 represents the OLS estimator, which serves as a benchmark for the other estimators.

Each line connects a series of points that report average *EFFICIENCY*, where the lines have been smoothed for the sake of readability. There are five points underlying each line in Figure 1 (for *T/N* = 2.0, 2.5, 3.0, 4.0 and 5.0), and seven points in Figure 2 (*T/N* = 1.9, 2.0, 2.5, 3.0, 3.8, 4.0 and 5.0). The reason the lines do not change monotonically with *T/N* is that other characteristics (heteroskedasticity, autocorrelation, cross-sectional dependence) are changing simultaneously with *T/N*. The movement from one *T/N* value to another is, in fact, a movement to a different DGP, with different population EVCM values.

Each of the estimators are color-coded in Figures 1 and 2. It is clear from both figures that the light blue line, corresponding to Estimator 7 (the Parks estimator), strictly dominates the others. For every *T/N* value included in our analysis, the average *EFFICIENCY* value for this estimator lies strictly below that of the other estimators, indicating greater efficiency. This confirms RY's first recommendation.

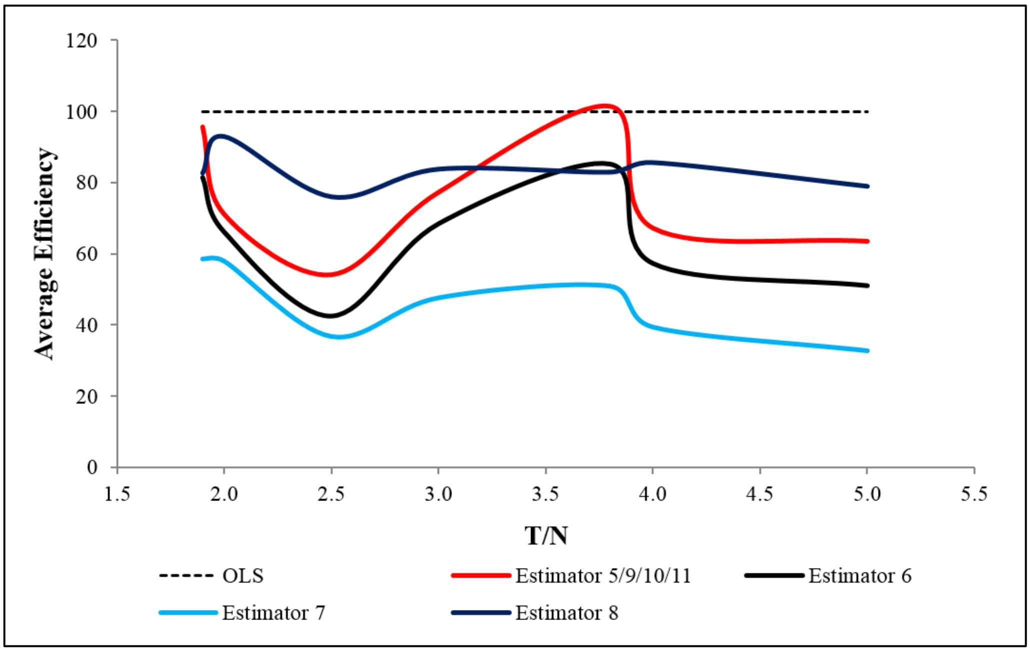
We return to Table 7 and next examine the experiments where *T/N* ≤ 1.5. While Estimator 7 is included in the table, its results are not directly comparable to the other estimators because its results are based on a much smaller number of experiments, since it cannot be estimated when *T/N* < 1.0. Ignoring Estimator 7 for the moment, it is seen that Estimator 6 performs better than the other estimators both in terms of having a lower average *EFFICIENCY* value (74.7 and 48.0 for the RY and new datasets, respectively), and in terms of besting OLS more frequently than the other estimators (89.1 and 99.0 percent, respectively). Estimator 6 is essentially the Parks estimator (Estimator 7), except that it does not accommodate cross-sectional dependence.

Figure 1: Comparison of Estimator *EFFICIENCY*: Reed and Ye (2011) Datasets, T/N > 1.5



Note: The *EFFICIENCY* measure is defined in Section 2 in the text. Estimators are identified in Table 3.

Figure 2: Comparison of Estimator *EFFICIENCY*: New Datasets, T/N > 1.5



Note: The *EFFICIENCY* measure is defined in Section 2 in the text. Estimators are identified in Table 3.

Figures 3 and 4 further highlight the superior performance of Estimator 6 when it comes to efficiency. We first note that the lines in the figures connect a larger number of points than in the preceding figures. There are 16 points underlying each line in Figure 3 (for *T/N* = 0.13, 0.19, 0.20, 0.21, 0.26, 0.30, 0.31, 0.32, 0.40, 0.42, 0.50, 0.52, 0.75, 1.00, 1.25, and 1.50), and 15 points in Figure 4 (*T/N* = 0.13, 0.19, 0.20, 0.25, 0.26, 0.30, 0.32, 0.38, 0.40, 0.50, 0.75, 0.95, 1.00, 1.25, and 1.50). Estimator 6 is represented by the solid black line.

With one exception, Estimator 6 strictly dominates the other estimators over all values of *T/N* reported in Figures 3 and 4. The lone exception involves Estimator 7 in the Reed and Ye (2011) datasets for *T/N* = 1.50. For smaller values of *T/N* (1.00 and 1.25), Estimator 6 lies strictly below Estimator 7 (indicating superior efficiency). When we turn to Figure 4 and the new datasets, we see that Estimator 6 bests Estimator 7 even when *T/N* = 1.50. Thus, our results indicate that *T/N* = 1.50 is a crossing-over point. For values less than that, Estimator 6 is most efficient. For values greater than that, Estimator 7 is most efficient. For values in the immediate vicinity of 1.50, either estimator may be most efficient, depending on other characteristics of the dataset.

It is interesting to note that the superior performance of Estimator 6 for $1.0 \leq T/N < 1.5$ is an example of the "shrinkage principle." This principle "asserts that the imposition of restrictions -- even false restrictions" can improve estimator performance (Diebold, 2007, p. 45). Even though the population EVCM is characterized by cross-sectional dependence, the estimator that "falsely" omits cross-sectional dependence (Estimator 6) outperforms the estimator that correctly includes it (Estimator 7). The reason this "false restriction" is effective in these cases is because there are insufficient observations to obtain reliable estimates of the cross-sectional covariances in $\boldsymbol{\Sigma}$ (cf. Equation 3).

Our findings call for a modification of RY's second recommendation, which states: "When the primary concern is efficiency, *T/N < 1*, and *Heteroskedasticity* > 1.67, use either *Estimator 5* or *Estimator 6*." For one, there is no need to condition the recommendation on heteroskedasticity. Second, Estimator 6 dominates Estimator 5 for all values of $T/N \leq 1.5$, so that Estimator 5 can be omitted as a "best" option. And lastly, the superior performance of Estimator 6 extends for a wider range of *T/N* values than determined by RY.

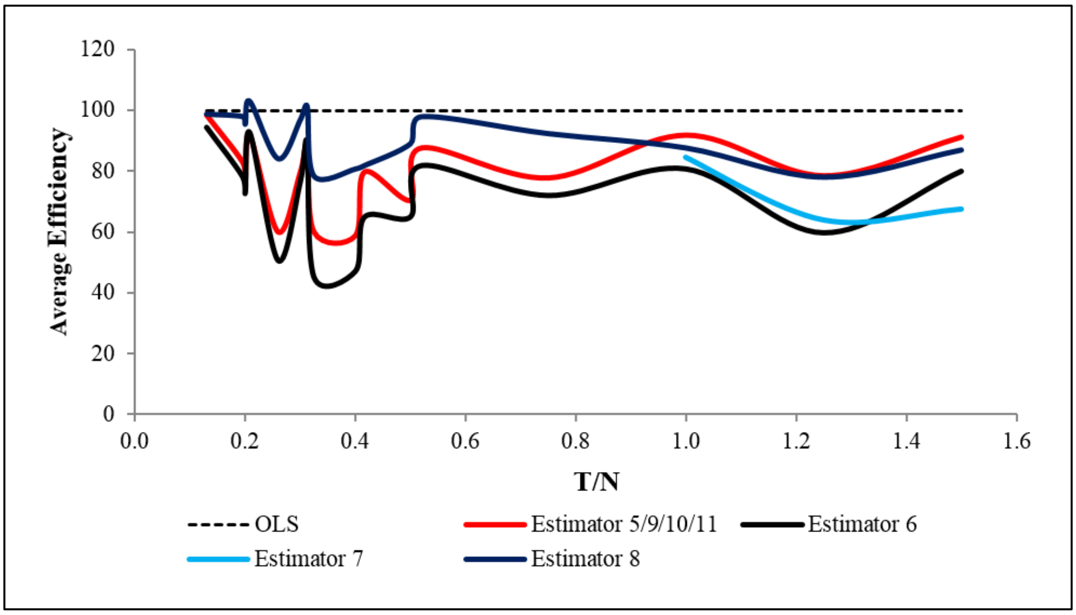
Taken together, the above results sketch a (virtually) complete decision tree for choosing the most efficient panel data estimator, provided the estimators the researcher is choosing from are included in Stata's or Eviews' standard statistical software package. This can be summarized in the following two modified recommendations:

> ***RECOMMENDATION 1*: When the primary concern is efficiency and T/N > 1.50, use Estimator 7 (= Parks estimator).**

> ***RECOMMENDATION 2*: When the primary concern is efficiency and T/N < 1.50, use Estimator 6 (= Parks estimator without cross-sectional dependence).**

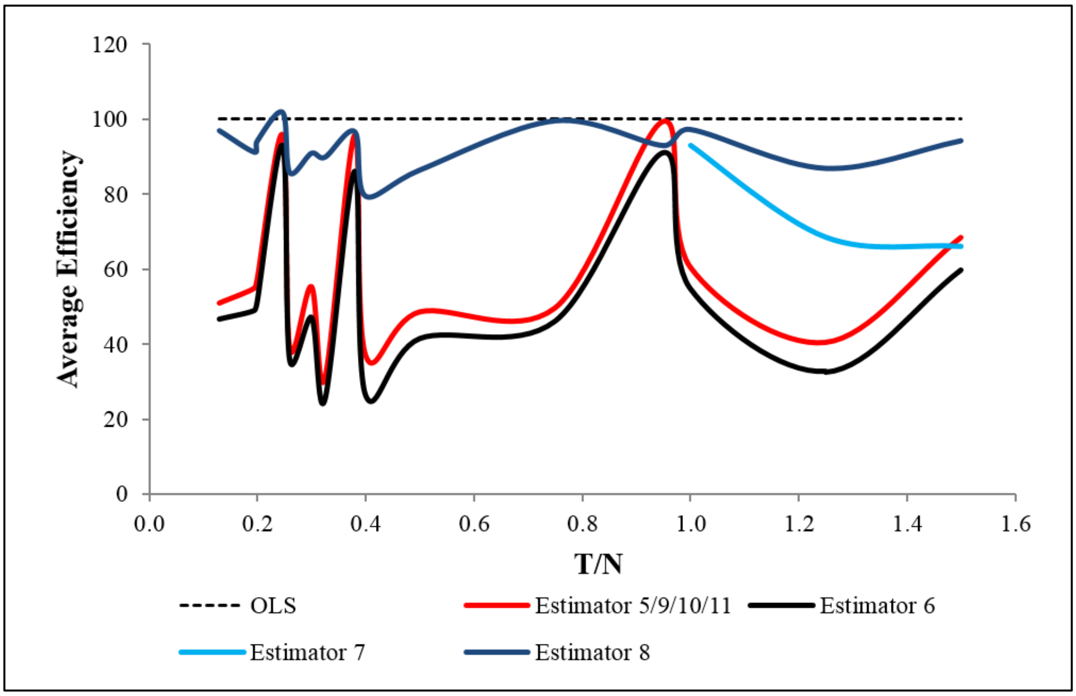
Accuracy in hypothesis testing. Table 8 reports performance results with respect to accuracy in hypothesis testing. The key columns are those that report the average value of the absolute difference between 95 percent and the coverage rate, |95 – *Coverage*|. An estimator should have

Figure 3: Comparison of Estimator *EFFICIENCY*: Reed and Ye (2011) Datasets, T/N ≤ 1.5



Note: The *EFFICIENCY* measure is defined in Section 2 in the text. Estimators are identified in Table 3.

Figure 4: Comparison of Estimator *EFFICIENCY*: New Datasets, T/N ≤ 1.5



Note: The *EFFICIENCY* measure is defined in Section 2 in the text. Estimators are identified in Table 3.

Table 8: Comparison of Estimator Coverage Rates

	Autocorrelation < 0.30		*Autocorrelation ≥ 0.30*	
	Coverage (1)	***|95 – Coverage| (2)***	***Coverage (3)***	***|95 – Coverage| (4)***
	REED AND YE (2011) DATASETS (T/N ≥ 1)			
Estimator 1	65.7	29.3	90.9	6.1
Estimator 2	64.1	30.9	91.0	4.9
Estimator 3	86.5	8.9	88.6	6.5
Estimator 4	60.1	34.9	91.5	3.8
Estimator 5	59.8	35.2	88.6	6.4
Estimator 6	88.0	7.1	90.9	4.4
Estimator 7 (Parks)	42.9	52.1	45.6	49.4
Estimator 8 (PCSE)	89.5	5.5	92.7	2.3
Estimator 9	51.9	43.1	85.6	9.4
Estimator 10	70.5	24.5	77.3	17.7
Estimator 11	58.5	36.5	88.2	6.8
	NEW DATASETS (T/N ≥ 1)			
Estimator 1	87.9	9.3	74.6	21.4
Estimator 2	83.1	11.9	73.2	22.0
Estimator 3	88.4	6.6	90.2	6.7
Estimator 4	83.6	11.4	74.7	20.3
Estimator 5	85.8	9.2	73.4	22.9
Estimator 6	90.9	4.1	90.9	5.7
Estimator 7 (Parks)	38.1	56.9	42.0	53.0
Estimator 8 (PCSE)	91.4	3.6	92.1	3.3
Estimator 9	75.5	19.5	64.7	30.3
Estimator 10	68.9	26.1	72.7	22.4
Estimator 11	80.6	14.4	68.6	26.4
	REED AND YE (2011) DATASETS (T/N < 1)			
Estimator 1	70.4	24.6	93.8	5.8
Estimator 2	67.1	27.9	93.7	4.0
Estimator 3	91.7	5.3	92.4	4.8
Estimator 4	61.9	33.1	94.1	1.3
Estimator 5	60.3	34.7	92.6	3.7
Estimator 6	86.9	8.2	93.2	3.4
Estimator 7 (Parks)	---	---	---	---
Estimator 8 (PCSE)	90.6	4.4	93.3	1.8

	Autocorrelation < 0.30		*Autocorrelation ≥ 0.30*	
	Coverage (1)	*|95 – Coverage|* (2)	*Coverage* (3)	*|95 – Coverage|* (4)
Estimator 9	50.3	44.7	87.2	7.8
Estimator 10	76.6	18.4	86.2	8.8
Estimator 11	59.2	35.8	92.2	3.8
	NEW DATASETS (T/N < 1)			
Estimator 1	84.0	11.2	80.2	15.2
Estimator 2	81.9	13.1	78.2	16.8
Estimator 3	83.5	11.5	93.6	3.3
Estimator 4	88.3	6.7	75.6	19.4
Estimator 5	88.3	8.0	75.8	20.5
Estimator 6	93.6	4.0	93.2	3.9
Estimator 7 (Parks)	---	---	---	---
Estimator 8 (PCSE)	91.5	3.5	92.3	2.7
Estimator 9	71.8	23.2	63.6	31.4
Estimator 10	82.7	12.3	82.9	12.1
Estimator 11	84.4	10.6	71.2	23.8

Note: The performance measures *Coverage* and |95 – *Coverage*| are defined in Section 2 in the text. A yellow-coloured cell indicates that Estimator 8 performs best for a given data-type. A green-coloured cell indicates that this estimator is second best.

a coverage rate close to 95 percent. Coverage rates less than (greater than) than 95 percent will reject the null hypothesis H_0: $\beta_x = its\ true\ value$ too often (not often enough). Both outcomes distort hypothesis testing. Thus the "best" estimator on the dimension of accuracy in hypothesis testing is one for which |95 – *Coverage*| is closest to zero.

The table has four panels. The first two panels report performance results for the experiments where *T/N* ≥ 1 for the Reed and Ye (2011) datasets and the new datasets, respectively. The next two panels report results for *T/N* < 1. *T/N* = 1 is selected as the cut-off because Estimator 7 (the Parks estimator) cannot be estimated when *T/N* is less than this.

The table also has four columns, with the first two columns collecting experiments where the associated datasets are characterized by *Autocorrelation* values less than 0.30, and the next two columns reporting results when *Autocorrelation* ≥ 0.30. This cut-off is motivated by RY's third recommendation. RY reported that Estimator 8 (the PCSE estimator) performed best for hypothesis testing when *Autocorrelation* < 0.30, while no estimator performed acceptably for autocorrelation values larger than this.

In the table, cells where Estimator 8 has the smallest |95 – *Coverage*| values are color-coded yellow. Cells where Estimator 8 has the second smallest |95 – *Coverage*| value are color-coded green. An inspection of the first two columns confirms RY's third recommendation for both the Reed and Ye (2011) datasets and the new datasets, irrespective of the value of *T/N*. Across the four subsets of experiments (Reed and Ye, 2011, $T/N < 1$ and $T/N \geq 1$; and New Datasets, $T/N < 1$ and $T/N \geq 1$), the values of | 95 – *Coverage* | range from a low of 3.5 to a high of 5.5.

However, the results allow one to go even further. When *Autocorrelation* ≥ 0.30, Estimator 8 either has the smallest, or close to the smallest |95 – *Coverage*| value in each of the four subsamples. Further, the corresponding values of |95 – *Coverage*| are quite small, ranging from 1.8 to 3.3.

It is worth noting that both Estimator 7 (Parks) and Estimator 8 (PCSE) estimate the same number of parameters. Yet, as Table 8 makes clear, their performance with respect to inference is markedly different. The main difference between the two is that Estimator 7 inverts $\boldsymbol{\Sigma}$ when calculating the coefficient covariance matrix, while Estimator 8 does not. It is the act of inverting $\boldsymbol{\Sigma}$, especially when T is close to N so that the matrix is barely full rank, that is the source of the Parks estimator's problem in producing reliable standard errors.[14]

Figures 5 to 8 provide more detail about the relative performances of the estimators with respect to hypothesis testing. Unlike the previous figures, there are a great many unique points on the horizontal axis, which causes the lines to be far less regular. For example, each estimator line in Figure 5 connects 80 individual points. Because each estimator has a unique estimate of the estimated coefficient's standard error, there are now either 11 lines (Figures 5 and 7) or 10 (Figures 6 and 8), to include in each figure.

In order to maintain readability, Figures 5 and 6 highlight just three estimators: Estimator 8 (solid black line, PCSE estimator), Estimator 6 (solid red line), and Estimator 7 (solid blue line). The other estimators are represented by identical dotted lines. We focus on Estimator 8 because Table 8 identified this estimator as "best" on the dimension of accuracy in hypothesis testing. We also highlight Estimator 6 because Table 8 indicates that this estimator also does relatively well. And we draw attention to Estimator 7 – the Parks estimator and the estimator chosen as best for efficiency when $T/N > 1.5$ – to show just how poorly this estimator performs when it comes to hypothesis testing. Figures 7 and 8 omit Estimator 7 because it cannot be estimated when $T/N < 1$.

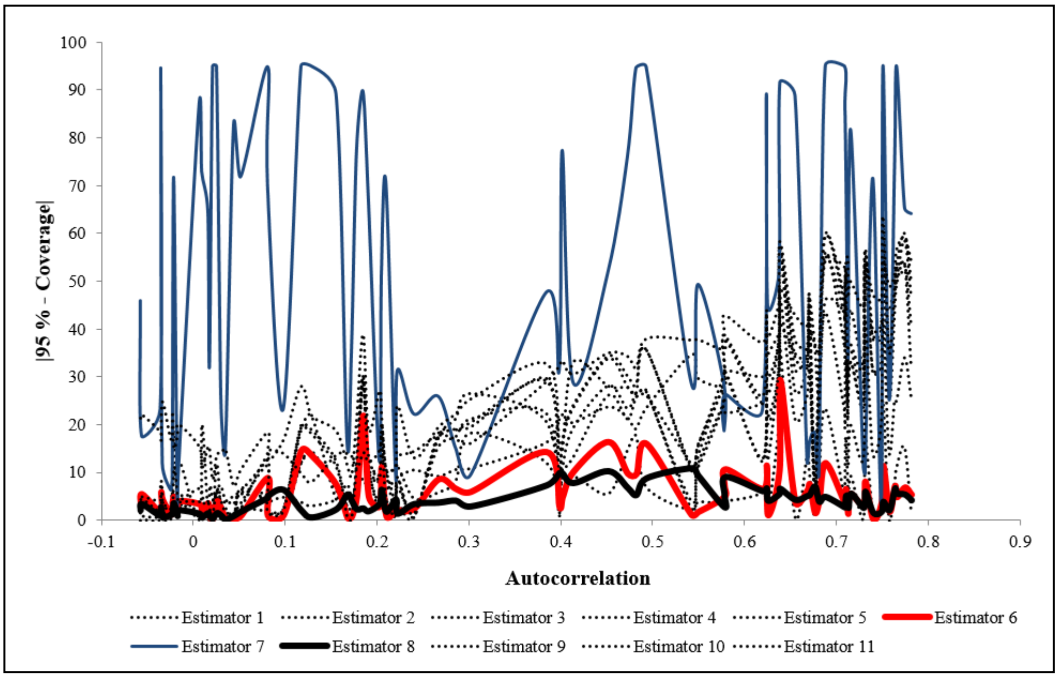
Figures 5 and 6 illustrate the general point that hypothesis testing can be very unreliable when using standard panel data estimators. While the performance of Estimator 7 is uniquely dismal, many of the other estimators also perform unacceptably poorly. Even the "best" estimator, Estimator 8, has instances where its performance is less than stellar.

Looking across all four Figures, it is clear that Estimator 8 (PCSE) generally dominates the other estimators across the diverse collection of experiments represented in Figures 5 through 8. While there are instances where one or more of the other estimators perform better than Estimator 8 in a given experiment, it is difficult to know whether this is anything more than sampling error. Table 8, along with Figures 5 through 8 allow then the following modification to RY's third recommendation:

> ***<u>RECOMMENDATION 3</u>*: When the primary concern is hypothesis testing, use Estimator 8 (PCSE).**

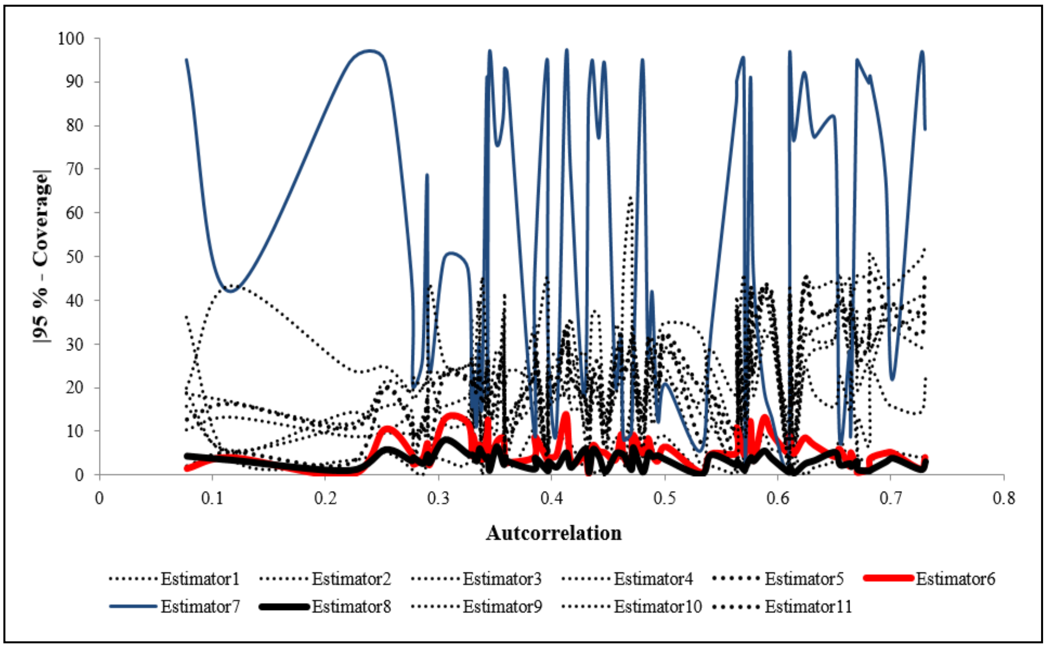
[14] This suggests that an alternative FGLS estimator that restricted the number of cross-sectional elements in $\boldsymbol{\Sigma}$, such as the one employed in O'Connell (1998), could mitigate the problem faced by the Parks estimator in producing reliable standard errors. We thank an anonymous reviewer for suggesting this.

Figure 5: Comparison of | 95 – *Coverage* | Values: Reed and Ye (2011) Datasets, T/N ≥ 1.0



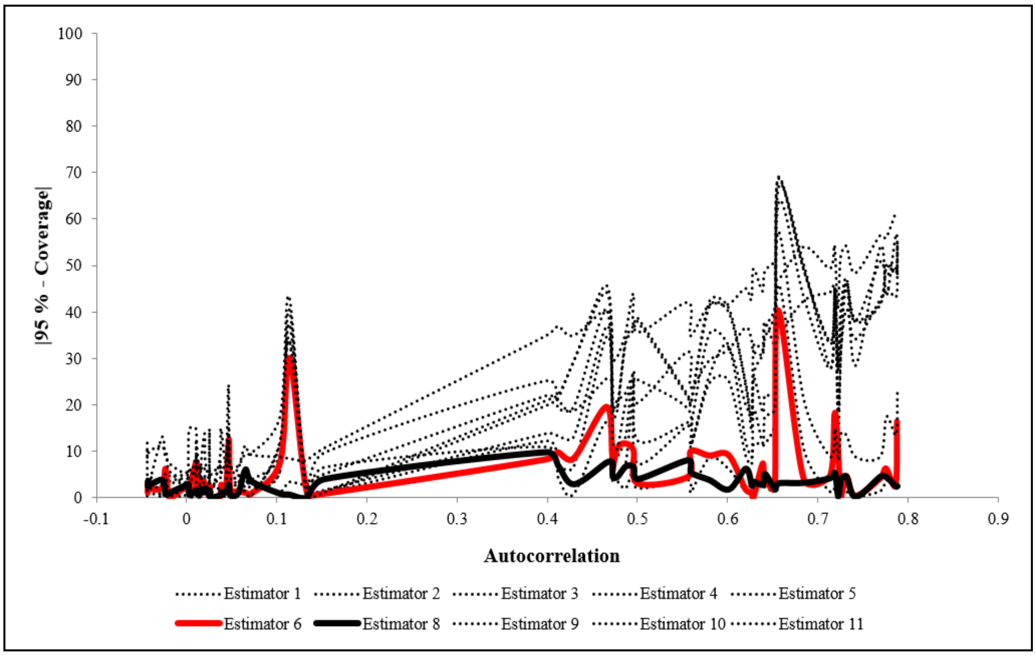
Note: The performance measure | 95 – *Coverage* | is defined in Section 2 in the text. Estimators are identified in Table 3.

Figure 6: Comparison of | 95 – *Coverage* | Values: New Datasets, T/N ≥ 1.0



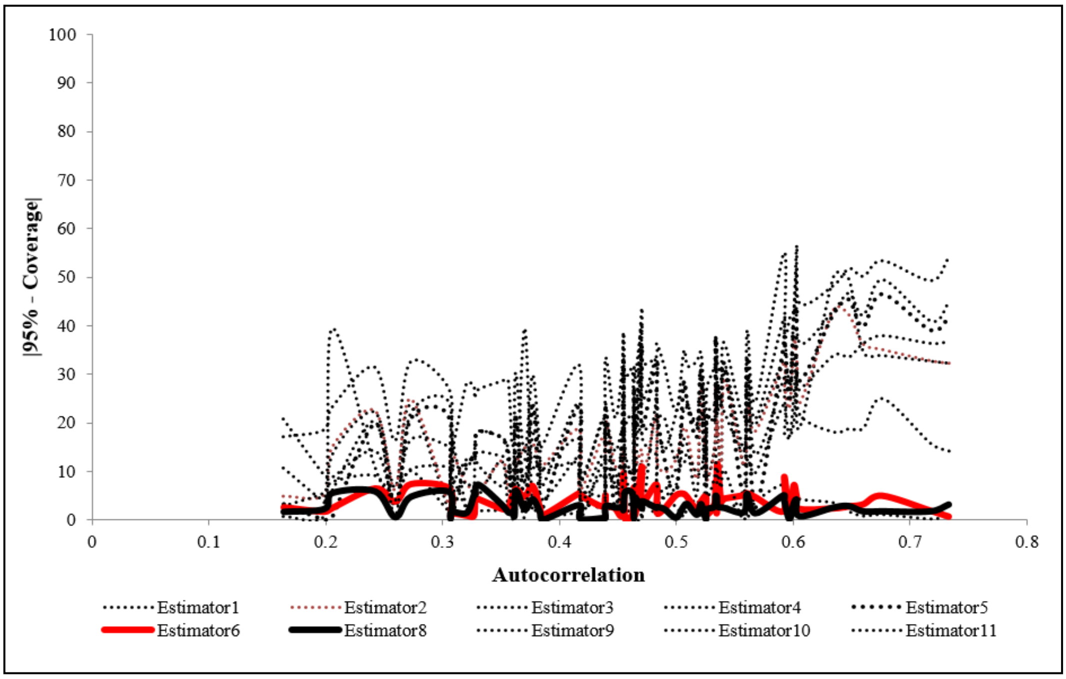
Note: The performance measure | 95 – *Coverage* | is defined in Section 2 in the text. Estimators are identified in Table 3.

Figure 7: Comparison of | 95 – *Coverage* | Values: Reed and Ye (2011) Datasets, T/N < 1.0



Note: The performance measure | 95 – *Coverage* | is defined in Section 2 in the text. Estimators are identified in Table 3.

Figure 8: Comparison of | 95 – *Coverage* | Values: New Datasets, T/N < 1.0



Note: The performance measure | 95 – *Coverage* | is defined in Section 2 in the text. Estimators are identified in Table 3.

Together, Recommendations 1 through 3 allow an applied researcher choosing panel data estimators from Stata or Eviews to easily select the “best” estimator. When it comes to choosing an estimator for efficiency, the researcher only needs to know the size of the panel dataset (N and T). That is sufficient to determine his/her selection. When it comes to choosing an estimator for hypothesis testing, the choice is even simpler: choose Estimator 8, the PCSE estimator.

Bootstrapping. While useful to applied researchers, the recommendations above require one to use different estimators depending on whether the primary interest is coefficient efficiency or accuracy in hypothesis testing. At the very least, this is awkward and difficult to motivate. It would be better if a researcher could use the same estimator for both estimation and inference.

In a recent study, Moundigbaye et al. (2017) develop bootstrap methods for SUR models with autocorrelated errors. In this section, we demonstrate the feasibility of these methods by bootstrapping the Parks estimator. Table 9 compares the accuracy of the PCSE estimator with the parametric bootstrap from Moundigbaye et al. (2017). A full comparison lies beyond the purview of this study. However, the table provides some examples using Dataset 1 (cf. Table 1) for varying N and T values. In every case, the bootstrapping method produces more accurate inference results than the PCSE estimator. For example, when N = 5 and T = 10, 87.3 percent of the 95% confidence intervals calculated from the PCSE estimator contain the true value of β_x. In contrast, 95.2 percent of confidence intervals contain the true value of β_x using the bootstrapped method.

While only an example, this exercise suggests that when $N \leq T$, a single-estimator approach that uses the Parks estimator with bootstrapping can be superior to the two-estimator approach that relies on the Parks estimator for coefficient estimates and the PCSE estimator for hypothesis testing. This is a topic for future research.

Table 9: A Comparison of the PCSE and Bootstrapped Parks Estimators With Respect to Inference: An Example

N	***T***	***PCSE***		***Estimator 7 - Bootstrapped***	
		Coverage (1)	***|95 – Coverage| (2)***	***Coverage (3)***	***|95 – Coverage| (4)***
5	***10***	87.3	7.7	95.2	0.2
5	***15***	89.8	5.2	95.4	0.4
5	***20***	90.0	5.0	95.2	0.2
5	***25***	92.5	2.5	96.4	1.4
10	***10***	88.3	6.7	97.3	2.3
10	***15***	91.0	4.0	98.5	3.5
10	***20***	92.5	2.5	96.4	1.4
10	***25***	93.0	2.0	96.3	1.3

Note: PCSE coverage rates are taken from Monte Carlo experiments using Dataset 1 and the respective N and T values. Bootstrapped coverage rates are calculated using the parametric bootstrap method of Moundigbaye et al. (2016).

5 Conclusion

This study follows up a previous analysis of panel data estimators in Reed and Ye (2011). RY conducted Monte Carlo experiments to study the performance of a wide range of Parks-type panel data estimators. They focused on estimators that are readily available in statistical software packages such as Stata and Eviews, and for which the number of cross-sectional units (*N*) and time periods (*T*) are small to moderate in size. They developed three recommendations for applied researchers seeking guidance about which panel data estimator to use in their research.

We identify a mistake in RY that affects their recommendations. Accordingly, we repeat the Monte Carlo experiments undertaken by RY, correcting their mistake. We also extend their study by including more real-world panel datasets on which to base our simulations. The result is a cleaner and more complete set of recommendations. In particular, we identify two estimators, a FGLS estimator that weights on heteroskedasticity and the Parks estimator, as being most efficient depending on whether T/N is less than or greater than 1.50, respectively. And we identify the PCSE estimator as being best for hypothesis testing in all situations.

A major contribution of our study is that it maps observable characteristics of the data to a specific estimator choice. Our recommendations are based solely on the ratio of *T/N*, which is readily observable. The ability to map observable data characteristics to estimator selection is potentially very valuable for applied researchers.

We note that while OLS with cluster robust standard errors is widely used by applied researchers, our experiments find that it performs relatively poorly on both efficiency and inference grounds for the small to moderately-sized panel datasets studied here. Thus, another contribution of our study is that it alerts researchers that there are better alternatives to OLS when the underlying DGP is assumed to be of the Parks variety.

Our analysis leaves several issues unresolved. One such issue is unbalanced data. All of the experiments above assumed that the panel datasets are balanced. It is not clear how these recommendations need to be modified when this is not the case. Another issue concerns dynamic panel data. All of the experiments above assumed static DGPs. As is well known, dynamic panel data have a number of complications that require special attention. Similarly, our analysis does not include many other panel data estimators, some of which we mention in the introduction above.

While we acknowledge the limitations of our study, it is still the case that the panel data estimators that come packaged in Stata and Eviews are widely used by many researchers. The fact that the best estimators separate out so clearly, across a wide variety of data environments, is striking. While additional work needs to be done, the findings of this study provide a useful start for researchers deciding which panel data estimator they should use.

References

Anselin, L. (2013). *Spatial econometrics: methods and models* (Vol. 4). Springer Science & Business Media.

Bai, J., (2003). Inferential theory for factor models of large dimensions. *Econometrica*, 71: 135–173. http://www.jstor.org/s Table/3082043

Baltagi, B.H., Egger, P., and Pfaffermayr, M. (2013). A generalized spatial panel data model with random effects. *Econometric Reviews*, 32(5–6): 650–685. https://doi.org/10.1080/07474938.2012.742342

Beck, N. and Katz, J.N. (1995). What to do (and not to do) with time series cross-section data. *American Political Science Review*, 89: 634–647. http://www.jstor.org/s Table/2082979

Biagi, B., Brandano, M.G., and Detotto, C. (2012). The effect of tourism on crime in Italy: a dynamic panel approach. *Economics: The Open-Access, Open-Assessment E-Journal*, 6 (2012-25): 1–24. http://dx.doi.org/10.5018/economics-ejournal.ja.2012-25

Bivand, R., and Piras, G. (2015). Comparing implementations of estimation methods for spatial econometrics. *Journal of Statistical Software*, Vol. 63(18). https://www.jstatsoft.org/article/view/v063i18

Casper, G. and Tufis, C. (2003). Correlation versus interchangeability: The limited robustness of empirical findings on democracy using highly correlated data sets. *Political Analysis*, 11(2): 196–203. https://www.researchgate.net/publication/245023372_Correlation_Versus_Interchangeability_The_Limited_Robustness_of_Empirical_Findings_on_Democracy_Using_Highly_Correlated_Data_Sets

Coakley, J., Fuertes, A.-M., and Smith, R.P. (2006). Unobserved heterogeneity in panel time series models. *Computational Statistics & Data Analysis*, 50(9): 2361–2380. https://doi.org/10.1016/j.csda.2004.12.015

Corrado, L., and Fingleton, B. (2012). Where is the economics in spatial econometrics? *Journal of Regional Science*, 52(2): 210–239. http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9787.2011.00726.x/abstract

Diebold, F.X. (2007). *Elements of forecasting*, 4th Edition. Ohio: Thomson, South-Western.

Eberhardt, M., Helmers, C., and Strauss, H. (2013). Do spillovers matter when estimating private returns to R&D? *The Review of Economics and Statistics*, 95(2): 436–448. https://doi.org/10.1162/REST_a_00272

Eberhardt, M., and Teal, F. (2011). Econometrics for grumblers: a new look at the literature on cross-country growth empirics. *Journal of Economic Surveys*, 25(1): 109–155. http://onlinelibrary.wiley.com/doi/10.1111/j.1467-6419.2010.00624.x/abstract

Elhorst, J.P. (2014). Spatial panel data models. In *Spatial Econometrics* (pp. 37–93). Springer: Berlin, Heidelberg. https://www.researchgate.net/publication/226957388_Spatial_Panel_Data_Models

Gujarati, D.N. (2004). *Basic econometrics*, 4th Edition. The McGraw−Hill Companies.

Kapetanios, G., Pesaran, M.H., and Yamagata, T. (2011). Panels with nonstationary multifactor error structures. *Journal of Econometrics*, 160(2): 326–348. https://doi.org/10.1016/j.jeconom.2010.10.001

Kersting, E. and Kilby, C. (2014). Aid and democracy redux. *European Economic Review*, 67: 125–143. https://doi.org/10.1016/j.euroecorev.2014.01.016

Manski, C.F. (1988). *Analog estimation methods in econometrics*. New York: Chapman & Hall.

Moundigbaye, M., Messemer, C., Parks, R.W., and Reed, W.R. (2017) Bootstrap methods for inference in a SUR model with autocorrelated disturbances. Working paper, Department of Economics and Finance, University of Canterbury. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=613081

Nunn, N. and Qian, N. (2014). US food aid and civil conflict. *American Economic Review*, 104(6): 1630–1666. http://dx.doi.org/10.1257/aer.104.6.1630

O'Connell, P.G.J. (1998). The overvaluation of purchasing power parity. *Journal of International Economics*, 44: 1–19. https://doi.org/10.1016/S0022-1996(97)00017-2

Parks, R.W. (1967). Efficient estimation of a system of regression equations when disturbances are both serially and contemporaneously correlated. *Journal of the American Statistical Association*, 62: 500–509. http://www.jstor.org/s Table/2283977

Pesaran, M.H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica*, 74: 967–1012. http://www.jstor.org/s Table/3805914

Pesaran, M.H., and Smith, R. P. (1995). Estimating long-run relationships from dynamic heterogeneous panels. *Journal of Econometrics*, 68: 79–113. https://doi.org/10.1016/0304-4076(94)01644-F

Reed, W.R. (2008). The robust relationship between taxes and U.S. state economic growth. *National Tax Journal*, 61(1): 57–80. http://www.jstor.org/s Table/41790431

Reed, W.R. and Webb, R. (2010). The PCSE estimator is good – just not as good as you think. *Journal of Time Series Econometrics*, 2(1), Article 8. https://www.researchgate.net/publication/46464697_The_PCSE_estimator_is_good_-_just_not_as_good_as_you_think

Reed, W.R. and Ye, H. (2011). Which panel data estimator should I use? *Applied Economics*, 43(8): 985–1000. https://doi.org/10.1080/00036840802600087

Appendix

List and Description of Panel Data Estimators

Estimator	***Package***	***Command***
1	Stata	command = **xtreg**
2	Stata	command = **xtreg** options = **robust**
3	Stata	command = **xtreg** options = **cluster(**name of cross-sectional variable)
4	Stata	command = **xtreg** options = **cluster(**name of time period variable)
5	Stata	command = **xtgls** options = **corr(independent) panels(heteroscedastic)**
6	Stata	command = **xtgls** options = **corr(ar1) panels(heteroscedastic)**
7 (Parks)	Stata	command = **xtgls** options = **corr(ar1) panels(correlated)**
8 (PCSE)	Stata	command = **xtpcse** options = **corr(ar1)**
9	EViews	GLS Weights = **Cross-section weights** Coef covariance method = **White cross-section**
10	EViews	GLS Weights = **Cross-section weights** Coef covariance method = **White period**
11	EViews	GLS Weights = **Cross-section weights** Coef covariance method = **White (diagonal)**

Source: Table 1 in Reed and Ye (2011).