econometrics

MDPI

Article

Missing Values in Panel Data Unit Root Tests

Yiannis Karavias [1,*], Elias Tzavalis [2] and Haotian Zhang [3]

[1] Department of Economics, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK
[2] Department of Economics, Athens University of Economics and Business, 104 34 Athens, Greece; etzavalis@aueb.gr
[3] Department of Economics, The Chinese University of Hong Kong, Shatin, NT, Hong Kong, China; htzhang@link.cuhk.edu.hk
* Correspondence: i.karavias@bham.ac.uk

Abstract: Missing data or missing values are a common phenomenon in applied panel data research and of great interest for panel data unit root testing. The standard approach in the literature is to balance the panel by removing units and/or trimming a common time period for all units. However, this approach can be costly in terms of lost information. Instead, existing panel unit root tests could be extended to the case of unbalanced panels, but this is often difficult because the missing observations affect the bias correction which is usually involved. This paper contributes to the literature in two ways; it extends two popular panel unit root tests to allow for missing values, and secondly, it employs asymptotic local power functions to analytically study the impact of various missing-value methods on power. We find that zeroing-out the missing observations is the method that results in the greater test power, and that this result holds for all deterministic component specifications, such as intercepts, trends and structural breaks.

Keywords: panel unit root tests; local power function; missing values; bias correction; unbalanced panel; structural breaks

JEL Classification: C22; C23

Citation: Karavias, Yiannis, Elias Tzavalis, and Haotian Zhang. 2022. Missing Values in Panel Data Unit Root Tests. *Econometrics* 10: 12. https://doi.org/10.3390/econometrics10010012

Academic Editor: Ryo Okui

Received: 16 December 2021
Accepted: 12 March 2022
Published: 16 March 2022

1. Introduction

It is almost always the case in applied panel data research that some values will be missing, leading to unbalanced panels. Missing observations can happen for various reasons. In macroeconomics, the available data do not start at the same date for all countries, or the frequency of data collection for many variables changed over time; for example, where data were available on an annual basis, now they are also available every quarter, and therefore there are missing values at the quarter dates for the time that only annual observations were available. Firm and bank-level data are plagued by mergers and bankruptcies or the introduction of new banks and firms in the sample. "Windsorising" the data (trimming the outliers), which is standard practice in corporate finance, also creates missing values. In household survey data, many households drop out with time. In financial data, there are missing values on certain days, for example, holidays and weekends.

This paper examines the problem of missing values in panel data unit root testing. Missing observations were first studied in a time series framework with stationary data. Many of the early contributions can be found in Harvey (1989), but the first was by Savin and White (1978), which examined the Durbin and Watson (1950, 1951, 1971) test for serial correlation. Their main result was that ignoring the missing values by closing up the observations leads to a Durbin–Watson statistic that has the same null limiting distribution, and the bounds critical values are still valid. Bhargava (1989) examined the impact of missing values on the power of the Durbin–Watson test using an approximation of the

power function and found that in the presence of an intercept in the model, and without very large gaps in the data, it is still reasonable to use the test. For non-stationary data, the first contributions were those of Shin and Sarkar (1994a, 1994b), which examine the impact of missing values on the instrumental variable unit root tests of Hall (1989) and the Dickey–Fuller test of Dickey and Fuller (1979). Shin and Sarkar (1994b) find that under the null hypothesis of a unit root, the estimator and t-statistics have the same distribution as in the non-missing data case. However, the sampling pattern does affect the power of the tests.

Ryan and Giles (1998) re-examine the two schemes for dealing with missing observations in Dickey–Fuller unit root tests found in Shin and Sarkar (1994b), and propose a third one. The first scheme simply removes the gaps from the series and assumes that the existing observations are continuous. The second scheme replaces the missing values with the last recorded observation before the gap, and the third scheme uses linear interpolation to fill in the missing data by taking the average of two observations, the last one before the gap and the first one after the gap. They first confirm the findings of Shin and Sarkar (1994b) that the first two schemes leave the unit root test null distributions unchanged but show that the third scheme introduces additional terms in the limiting distribution. Furthermore, using extensive Monte Carlo simulations they show that the first scheme delivers the highest power, while linear interpolation provides some empirical size gains but significant power loss. However, the second scheme leads to more power in the augmented Dickey–Fuller test which deals with serial correlation (Dickey and Fuller 1981).

The first contribution of this paper is to extend the popular panel unit root tests of Harris and Tzavalis (1999), Karavias and Tzavalis (2014) to unbalanced panels. This is the first paper that considers how to adjust panel data unit root tests to unbalanced panels. It is not straightforward to adapt the single time series results in panels because most panel unit root tests require bias corrections which will be affected by the pattern and location of the missing values, see, e.g., Levin et al. (2002), Im et al. (2003), among others. The problem is even more serious when the bias correction is estimated by simulations.

The tests of Harris and Tzavalis (1999), Karavias and Tzavalis (2014) are fixed-T tests with a wide range of deterministic component specifications including, individual unit intercepts, linear trends and common structural breaks. This allows us to study the impact of missing values on various settings. We focus on panel unit root tests with a large number of cross-section units N, and a small number of time series observations T because this was the original dynamic panel data framework introduced by Holtz-Eakin et al. (1988) and the framework of the first panel data unit root test, that of Breitung and Meyer (1994). It is also one of the most common in terms of applications, see, e.g., Karavias et al. (2021). The panel data unit root tests of Harris and Tzavalis (1999), Karavias and Tzavalis (2014) are popular in applied research and have been implemented in statistical software.[1] They have several advantages beyond being applicable to short panels; they are invariant to the initial conditions, they allow for flexible and general trend functions, and they allow for cross-section heteroskedasticity.

We adjust the above test statistics for missing values by providing new bias correction formulas and deriving their asymptotic limiting distributions. The adjustment allows for general patterns of missing values that can differ across individuals and leads to excellent test size properties. Under the null hypothesis, the distribution of the adjusted test statistics remains identical to the case without missing values. This result holds for any missing-value correction scheme, unlike the single time series results of Ryan and Giles (1998).

The paper's second contribution is to employ the Harris and Tzavalis (1999), Karavias and Tzavalis (2014) tests and the fixed-T framework to study which method for dealing with missing data results in tests with greater power. The power properties of these tests have been studied previously in Madsen (2010) and Karavias and Tzavalis (2016, 2017). In this paper, we extend the local power functions of Karavias and Tzavalis (2016, 2017) to allow for missing values, and then use these functions to theoretically compare the three missing value schemes from Ryan and Giles (1998). The results show that the zeroing-out

scheme leads to the greatest power for all types of deterministic components. This is also the first paper that examines the nexus between missing values and structural breaks. We find that zeroing-out once more leads to greater power, but the ranking of the other two schemes depends on the relative locations of the missing values and the structural breaks.

The paper is structured as follows. Section 2 presents the tests of Harris and Tzavalis (1999), Karavias and Tzavalis (2014). Section 3 introduces missing values and describes how they can be analysed in the fixed-T framework. Section 4 presents the missing-value-adjusted statistics and their limiting distributions under the null and under local alternatives. Section 5 compares the impact of popular schemes for dealing with missing values, for various deterministic specifications. Section 6 concludes the paper.

2. Panel Unit Root Tests without Missing Values

Assume that there are N cross-section units and T time series observations and consider the following data generating processes:

$$y_{i,t} = a_i + u_{i,t}, \tag{1}$$
$$y_{i,t} = a_i + b_i t + u_{i,t}, \tag{2}$$
$$y_{i,t} = a_{1,i} I(t \leq T_0) + a_{2,i} I(t > T_0) + u_{i,t}, \tag{3}$$
$$y_{i,t} = a_{1,i} I(t \leq T_0) + a_{2,i} I(t > T_0) + b_{1,i} I(t \leq T_0) t + b_{2,i} I(t > T_0) t + u_{i,t}, \tag{4}$$

for $i = 1, \ldots, N$ and $t = 1, \ldots, T$. For notational convenience, we further assume that the initial observation is $y_{i,0}$ and it is observed resulting in a total of $T + 1$ time series observations per unit.

Model (1) is includes individual (or incidental) intercepts and model (2) includes individual intercepts and individual trends. The models in (3) and (4) consider a single structural break in the intercepts and trends of the series, at time T_0. The break is assumed to be common for all units as in Bai (2010). The parameters $a_{1,i}$ and $b_{1,i}$ are the intercept and trend individual effects before the break and $a_{2,i}$ and $b_{2,i}$ are those after the break. The first two models have been considered in Harris and Tzavalis (1999), while models (3) and (4) have been considered in Karavias and Tzavalis (2014).

The error term $u_{i,t}$ is assumed to be an autoregressive process of order one, as follows:

$$u_{i,t} = \rho u_{i,t-1} + \varepsilon_{i,t}, \tag{5}$$

for $i = 1, \ldots, N$ and $t = 1, \ldots, T$. The key parameter of interest is the autoregressive parameter ρ, which determines the stationarity of the panel process.

For models (1) and (2) the null hypothesis of non–stationarity is given by $H_0 : \rho = 1$, while the alternative of stationarity is $H_1 : \rho < 1$. For models (3) and (4) the null hypothesis depends on whether there is a structural break under the null or not. Both choices are considered in Karavias and Tzavalis (2014) and could have been considered here, however, the results in terms of missing values do not change qualitatively and in the following we will only consider the case where a structural break occurs only under the alternative. Explicitly stated, the null hypothesis is $H_0 : \rho = 1$ & $a_{1,i} = a_{2,i}$ for (3), and $H_0 : \rho = 1$ & $a_{1,i} = a_{2,i}$ & $b_{1,i} = b_{2,i}$ for (4). We will further assume that the date of the break is known to the researcher, as the missing values analysis also does not change if the date of the break is unknown.

To remove the individual effects from $y_{i,t}$ we employ the annihilator matrices Q_m, where $m = 1, 2, 3, 4$ corresponds to models (1)–(4). We introduce the following notation: Let I_T be a $T \times T$ identity matrix, e be a $T \times 1$ vector of ones, and $\tau = (1, 2, 3, ..., T)'$. Moreover, let e_1 and τ_1 be $T \times 1$ vectors such that $e_{1,t} = e_t$ and $\tau_{1,t} = \tau_t$ if $t \leq T_0$ and 0 otherwise, and let e_2 and τ_2 be $T \times 1$ vectors such that $e_{2,t} = e_t$ and $\tau_{2,t} = \tau_t$ if $t > T_0$ and 0 otherwise. The vectors e_j and τ_j are effectively "breaking" versions of e and τ. Q_m is an annihilator matrix with the general formula $Q_m = I_T - Z_m (Z_m' Z_m)^{-1} Z_m'$, and where Z_m depends on the model. Define $Z_1 = e$, $Z_2 = \{e, \tau\}$, $Z_3 = \{e_1, e_2\}$ and $Z_4 = \{e_1, e_2, \tau_1, \tau_2\}$.

By premultiplying models (1)–(4) with the corresponding Q_m, the least-squares estimator of the transformed model is given by:

$$\hat{\rho}_m = \left(\sum_{i=1}^{N} y_{i,-1}' Q_m y_{i,-1} \right)^{-1} \left(\sum_{i=1}^{N} y_{i,-1}' Q_m y_i \right), \tag{6}$$

where $y_{i,-1} = (y_{i,0}, y_{i,1}, \ldots, y_{i,T-1})'$ and $y_i = (y_{i,1}, y_{i,2}, \ldots, y_{i,T})'$.

The estimator (6) is inconsistent because it suffers from the well known Hurwicz-Nickell bias.[2] Harris and Tzavalis (1999), Karavias and Tzavalis (2014) derive expressions for this bias when $\rho = 1$ and show that it depends on the deterministic component specification. Furthermore, they show that the bias can be estimated and $\hat{\rho}$ be bias-corrected. The following test statistic, and its asymptotic distribution, is then used for testing the null hypotheses:[3]

$$t_m = \frac{\hat{\rho}_m - B_m - 1}{\sqrt{Var(\hat{\rho}_m)}} \xrightarrow{d} N(0,1), \tag{7}$$

where B_m is the bias correction and it is given by the probability limit of $\hat{\rho}_m - 1$. Harris and Tzavalis (1999), Karavias and Tzavalis (2014) provide explicit formulas for B_m and $Var(\hat{\rho}_m)$, for models (1)–(4). For (3) and (4), the expressions of t_m, B_m and $\hat{\rho}_m$ also depend on the date of the break. This dependence is suppressed in our notation because the date of the break does not have an impact on the theoretical results of the paper.

The first contribution is to examine how the above statistic and its limiting distribution change in the presence of missing values. Missing values are introduced in the next section.

3. Missing Values

So far we have assumed that there are $T+1$ observations of $y_{i,t}$ for every $i \in 1, \ldots, N$. Let the data spawn from $t = \{0, 1, \ldots, T\}$, and let there be a missing value at time t^*, where $1 < t^* < T$. Under the null hypotheses that $H_0 : \rho = 1$ for models (1)–(2) and $H_0 : \rho = 1$ & $a_{1,i} = a_{2,i}$ for (3), and $H_0 : \rho = 1$ & $a_{1,i} = a_{2,i} = a_i$ & $b_{1,i} = b_{2,i} = b_i$ for (4), (1)–(4) imply the following data generating process:

$$y_i = y_{i,-1} + \varepsilon_i \tag{8}$$

$$y_i = y_{i,-1} + b_i e + \varepsilon_i, \tag{9}$$

where $\varepsilon_i = (\varepsilon_{i,1}, ..., \varepsilon_{i,T})'$. In matrix form, the above equations (presenting only the first, to save space) become:

$$\begin{pmatrix} y_{i,1} \\ \vdots \\ y_{i,t^*-2} \\ y_{i,t^*-1} \\ y_{i,t^*} \\ y_{i,t^*+1} \\ y_{i,t^*+2} \\ \vdots \\ y_{i,T} \end{pmatrix} = \begin{pmatrix} y_{i,0} \\ \vdots \\ y_{i,t^*-3} \\ y_{i,t^*-2} \\ y_{i,t^*-1} \\ y_{i,t^*} \\ y_{i,t^*+1} \\ \vdots \\ y_{i,T-1} \end{pmatrix} + \begin{pmatrix} u_{i,1} \\ \vdots \\ u_{i,t^*-2} \\ u_{i,t^*-1} \\ u_{i,t^*} \\ u_{i,t^*+1} \\ u_{i,t^*+2} \\ \vdots \\ u_{i,T} \end{pmatrix}.$$

When there is a missing value y_{i,t^*}, the dynamic nature of the system results in y_{i,t^*} appearing in two equations of the system, the t^* and the t^*+1. This means that one missing value plagues two equations of the system as can also be seen from the above representation.

There is a fundamental difference in the way that missing values are treated in a fixed-T framework compared to single time series analysis. In single time series, it is assumed that the data generating process is based on the index t, as in models (1)–(4), but

only a subset of these observations is available, $t_1, t_2, \ldots, t_T$. A new pseudo series is created based on the index t_i, i.e., $x_k = \rho x_{k-1} + \alpha_k$, where $k = t_1, t_2, \ldots, t_T$ and the estimation of ρ is based on x_k. This approach is reasonable when T is asymptotic because the impact of the index change is asymptotically negligible. When T is fixed as it is here, the idea is to work with a fixed set of equations, see, e.g., Hayashi (2000, sec. 5.3). We will examine methods that keep the asymptotic distribution of the test the same, but we wish to find which method results in maximum power.

Define D_i to be a deterministic matrix that reshuffles the data to deal with missing values in the unit i. Notice that D_i allows for the pattern and number of missing values to differ across units. If there are no missing values, $D_i = I_T$. If there is a missing value for unit i at time t^*, then $[D_i]_{t^*,t^*} = 0$, which is how missing values are introduced into the model. Effectively, D_i multiplies the missing value with 0 and we assume that the outcome of this algebra is 0. Because this is a dynamic model and two equations are affected by a missing value at time t^*, it must also be $[D_i]_{t^*+1,t^*+1} = 0$. The rest of the diagonal elements of D_i are equal to 1, while the off-diagonal elements are the ones performing the missing-value correction scheme.

The different schemes for dealing with missing values, see, e.g., Ryan and Giles (1998), such as closing the gaps, using linear interpolation and using the last available observation, dictate different versions of D_i. In the following section, we will derive the asymptotic distribution of the statistic in (7) and the asymptotic local power without assuming a specific type of D_i. This means that the analysis can be used for comparing other methods for dealing with missing values, beyond the ones in Section 5.

4. Asymptotic Distribution and Local Power Function

The main idea behind the asymptotic analysis comes from Hayashi (2000, p. 338), and to demonstrate it we consider (1), which can be quasi-differenced according to Equation (5) and be written as $y_{i,t} = \rho y_{i,t-1} + (1-\rho)a_i + \varepsilon_{i,t}$. Stacking the model across the time dimension, it becomes:

$$y_i = \rho y_{i,-1} + (1-\rho)a_i e + \varepsilon_i. \tag{10}$$

Then, we premultiply the model by D_i, which is the transformation matrix that deals with the missing values in each unit, and apply a second transformation, the within transformation, to remove the deterministic components. The removal of the individual effects is based on:

$$Q^D_{1,i} = I_T - (D_i e)(e_i' D_i' D_i e)^{-1}(e' D_i). \tag{11}$$

Notice how $Q^D_{1,i}$ now depends on the individual i, since individuals are allowed to differ in terms of the number and location of missing values. The general expression of $Q^D_{m,i}$ is $Q^D_{m,i} = I_T - (D_i Z_m)(Z_m' D_i' D_i Z_m)^{-1}(Z_m' D_i)$. Then, $\hat{\rho}$ is the estimator which minimises the least squares criterion in the transformed model and it equal to:

$$\hat{\rho}^D_m = \left(\sum_{i=1}^{N} y_{i,-1}' D_i' Q^D_{m,i} D_i y_{i,-1} \right)^{-1} \left(\sum_{i=1}^{N} y_{i,-1}' D_i' Q^D_{m,i} D_i y_i \right). \tag{12}$$

The asymptotic analysis is based on the following set of assumptions. These are not the weakest assumptions possible, however, they are useful for allowing us to study the problem of missing values analytically.

Assumption 1 (Errors/No Selectivity Bias). *(i) u_i, for $i = 1, ..., N$, is a sequence of independent random vectors with $E(u_i|D_i) = 0$ and $E(u_i u_i'|D_i) = \sigma^2 I_T$, where $\sigma^2 < \infty$. (ii) u_i follows a multivariate normal distribution.*

Assumption 2 (Variable Independence). *$u_{i,t}$, for $i = 1, ..., N$, and $t = 1, ..., T$, is independent of a_i, b_i, $a_{1,i}$, $a_{2,i}$, $b_{1,i}$, $b_{2,i}$ and $y_{i,0}$, and $Var(y_{i,0}) < \infty$.*

Assumption 3 (Invertibility). *For* $m \in \{1, \ldots, 4\}$, $N^{-1} \sum_{i=1}^{N} y'_{i,-1} D'_i Q_m^D D_i y_{i,-1} > 0$ *with probability 1 for all possible break points,* T *and* N.

Assumption 4 (Missing Values). *As* $N \to \infty$,

$$\frac{1}{N} \sum_{i=1}^{N} E(D'_i Q_i^D D_i) \xrightarrow{p} \Phi_D, \tag{13}$$

where Φ_D *is positive definite.*

Assumption 1 is standard in the literature, see, e.g., Verbeek and Nijman (1992). It implies that the errors are not correlated with the missing values, in other words there is no selectivity bias. Still, the individual effects can be correlated with missing observations. The assumption also implies that conditionally on the missing values, the errors are homoskedastic across both the time series and cross-section dimensions and not serially correlated. Cross-section dependence is not allowed, and finally, the second part imposes normality, which helps in simplify the formulas below. Assumption 2 is also standard and necessary for deriving the asymptotic local power function. Assumption 3 is an invertibility assumption that also restricts the presence of the structural breaks to locations that allow the existence of Q_3 and Q_4. Effectively, it is a high-level assumption that implies the need for trimming the sample at the beginning and the end. For more information, see also Karavias and Tzavalis (2014). Assumption 4, is a regularity condition that allows the existence of the estimator bias in the presence of missing values.

We are now able to present the limiting distribution of $\hat{\rho}_m$ in the presence of missing data:

Proposition 1. *Let Assumptions 1, 3 and 4 hold. Furthermore, for models (3) and (4) let the date of the break be known. Then, under* H_0, *and as* $N \to \infty$:

$$t_m = \frac{\hat{\rho}_m - B_m - 1}{\sqrt{V_m / N}} \xrightarrow{d} N(0, 1). \tag{14}$$

The quantities B_m *and* V_m *are defined as:*

$$B_m = \frac{tr(\Lambda' \Phi_D)}{tr(\Lambda' \Phi_D \Lambda)}, \tag{15}$$

$$V_m = \frac{2tr(\Omega_D)}{[tr(\Lambda' \Phi_D \Lambda)]^2}, \tag{16}$$

where Ω_D *is such that,*

$$\frac{1}{N} \sum_{i=1}^{N} A_{m,i}^2 \xrightarrow{p} \Omega_D, \tag{17}$$

and $A_{m,i} = (1/2)(\Lambda' D'_i Q_{m,i} D_i + D'_i Q_{m,i} D_i \Lambda) - B_m (\Lambda' D'_i Q_{m,i} D_i \Lambda)$. *Finally,* Λ *is a matrix with elements* $[\Lambda]_{i,j} = 1$ *for* $i < j$ *and* $[\Lambda]_{i,j} = 1$ *otherwise, for* $i, j \in \{1, ..., T\}$.

Proposition 1 derives the asymptotic distribution of $\hat{\rho}$ under the null hypothesis and under a general pattern of missing value treatment. The proof of the proposition is based on Harris and Tzavalis (1999), Karavias and Tzavalis (2014) and is omitted. Expressions (15) and (16) make clear how the bias and variance of the estimator should be adjusted in the presence of missing values. The result in (14) differs from those in Harris and Tzavalis (1999), Karavias and Tzavalis (2014) in that it contains the matrix D_i. Proposition 1 demonstrates that the limiting distribution of a statistic adjusted for the missing values remains the same as in the case without missing values. This finding is in line with part of the previous literature, see, e.g., Shin and Sarkar (1994b). Unlike Ryan and Giles (1998)

however, the limiting distribution does not change in the case of linear interpolation, because the distribution does not depend on the type of D_i.

When the date of the break is unknown, Karavias and Tzavalis (2014) suggest computing the minimum of t_m for models $m = 3, 4$, over every permissible break date, as determined by Assumption 2. The limiting distribution in this case will depend on the correlations between the t_m for different break dates, which will be affected by the missing values. Therefore, the critical values of Karavias and Tzavalis (2014) are no longer valid in this case. Instead, one can use the bootstrap proposed in Karavias and Tzavalis (2019) to derive the appropriate critical values. The case of unknown breaks will not be pursued in the analysis below because the impact of missing values is qualitatively the same as in the case of known breaks.

To employ the statistic in (14), it is necessary to a employ consistent estimator of B_m, which is given by:

$$\hat{B}_m = \frac{\sum_{i=1}^{N} \Lambda' D_i' Q_{m,i} D_i}{\sum_{i=1}^{N} \Lambda' D_i' Q_{m,i} D_i \Lambda}, \tag{18}$$

and a consistent estimator for V_m, which is given by:

$$\hat{V}_m = \frac{\frac{1}{N} \sum_{i=1}^{N} 2tr(\hat{A}_{m,i}^2)}{\left[\frac{1}{N} \sum_{i=1}^{N} \Lambda' D_i' Q_{m,i} D_i \Lambda\right]^2}, \tag{19}$$

where $\hat{A}_{m,i} = (1/2)(\Lambda' D_i' Q_{m,i} D_i + D_i' Q_{m,i} D_i \Lambda) - \hat{B}_m (\Lambda' D_i' Q_{m,i} D_i \Lambda)$.

The following proposition examines the behaviour of the t_m statistics under local alternatives, as in Karavias and Tzavalis (2016, 2017). The main advantage of the local power theory is that it allows us to examine analytically the impact of each type of missing value correction, and this is more transparent than doing Monte Carlo simulations, where the results depend also on other parameters in the experimental design. The local power function is an approximation of the power function in a $N^{-1/2}$ neighbourhood of the null hypothesis. The local alternatives are defined as $\rho_N = cN^{-1/2}$, where $c > 0$, because N is the only increasing data dimension.

Proposition 2. *Let Assumptions 1–4 hold. Furthermore, for models (3) and (4) let the date of the break be known. Then, under* $H_1 : \rho_N = cN^{-1/2}$, *and as* $N \to \infty$:

$$t_m = \frac{\hat{\rho}_m - B_m - 1}{\sqrt{V_m / N}} \xrightarrow{d} N(-cK_m, 1). \tag{20}$$

The quantity K_m *is given by:*

$$K_m = \frac{tr(F\Lambda'\Phi_D) + tr(\Lambda'\Phi_D\Lambda) - 2B_m tr(F\Lambda'\Phi_D)}{2tr(\Omega_D)}, \tag{21}$$

where $F = [d\Theta/d\rho]_{\rho=1}$ *and* Θ *is a* $T \times T$ *matrix that has elements:* $[\Theta]_{i,j} = 0$ *if* $i = j$ *or* $i < j$, *and* $[\Theta]_{i,j} = \rho^{(i-j-1)}$.

Proposition 2 states the limiting distribution of t_m under local alternatives. The proof of the proposition is based on Karavias and Tzavalis (2016, 2017) and is omitted. The result is elegant and states that the probability of rejecting the null when it is not true ($c > 0$) is a monotonic function of K_m. It suffices therefore to examine the sign and magnitude of K_m; the larger the K_m, the more powerful the test. $K_m = 0$ means that the test has trivial power, while if $K_m < 0$, the test is biased.

5. Dealing with Missing Values

In this section we evaluate K_m for various types of D_i to see which way of dealing with missing values results in greater test power. We will consider the three schemes which have

previously appeared in the single time series literature (see, e.g., Ryan and Giles (1998)): Closing the gaps, filling in the previous available value, and using linear interpolation.

With respect to interpolation, there are various methods can produce estimates of the missing values, but we follow Ryan and Giles (1998) and only use the average of the two observations before and after the missing values. There are other methods available, which ultimately introduce a bias-variance trade-off problem, but we leave this direction for future research.

The first option that we examine is zeroing-out the missing observations, which is equivalent to "closing the gaps". The zeroing-out happens by creating D_i to be a diagonal matrix with diagonal elements $[D_i]_{t^*,t^*} = [D_i]_{t^*+1,t^*+1} = 0$ if observation y_{i,t^*} is missing, and equal to $[D_i]_{t,t} = 1$ for $t \neq t^*$. The second option in the literature is to substitute the missing value with the previous available value, i.e., $y_{i,t^*} = y_{i,t^*-1}$. This is possible by defining the matrix D_i as:

$$D_i = \begin{bmatrix} 1_{i,1} & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1_{i,t^*-1} & & & & & & \\ & 0_{i,t^*} & 1_{i,t^*} & 0_{i,t^*} & 0_{i,t^*} & & & & \\ & & 1_{i,t^*+1} & 0_{i,t^*+1} & 0_{i,t^*+1} & 0_{i,t^*+1} & 0_{i,t^*+1} & & \\ & & & & & 1_{i,t^*+2} & & & \\ & & & & & & \ddots & & \\ & & & & & & & 1_{i,T} \end{bmatrix}$$

Finally, the third option is to substitute the missing value with the average of y_{i,t^*-1} and y_{i,t^*+2}. This is possible by creating D_i as:

$$D_i = \begin{bmatrix} 1_{i,1} & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1_{i,t^*-1} & & & & & & \\ & 0_{i,t^*} & 1/2_{i,t^*} & 0_{i,t^*} & 0_{i,t^*} & 1/2_{i,t^*} & & & \\ & & 1/2_{i,t^*+1} & 0_{i,t^*+1} & 0_{i,t^*+1} & 1/2_{i,t^*+1} & 0_{i,t^*+1} & & \\ & & & & & 1_{i,t^*+2} & & & \\ & & & & & & \ddots & & \\ & & & & & & & 1_{i,T} \end{bmatrix}$$

In unreported Monte Carlo simulations it is found that all three above transformations have size very close to the nominal.[4] It is therefore of interest which transformation results in higher power.

Table 1 below compares the three methods in terms of local asymptotic power, based on the theoretical results of (21). We consider the case of a single missing value and where $D_i = D$, that is, the missing value appears in all units and at the same place. This is a simplifying assumption, but our interest is not on the power of tests per se, but rather in determining the relative power of the three ways of dealing with missing values. We allow the missing value to appear towards the beginning, the middle, and the end of the sample at times $\lfloor 0.25T \rfloor$, $\lfloor 0.5T \rfloor$ and $\lfloor 0.75T \rfloor$. We further set $T = 20$, and for the Karavias and Tzavalis (2014) we allow the dates of the breaks to take place at the same dates as the missing values. Table 1 below reports K_m from Equation (21), for $m = 1, 3$. A higher value of K_m means higher test power.

Table 1. Asymptotic local power of the Harris and Tzavalis (1999), Karavias and Tzavalis (2014) tests for the model with incidental intercepts and a single missing value.

Harris and Tzavalis (1999)			
Missing Value Location	Zo	Pr	Li
$\lfloor 0.25T \rfloor$	8.545	8	8.26
$\lfloor 0.5T \rfloor$	8.559	8.276	8.423
$\lfloor 0.75T \rfloor$	8.544	8.126	8.26
Karavias and Tzavalis (2014)			
Break fraction: $\lfloor 0.25T \rfloor$			
Missing Value Location	Zo	Pr	Li
$\lfloor 0.25T \rfloor$	6.32	6.304	6.107
$\lfloor 0.5T \rfloor$	6.825	6.384	6.637
$\lfloor 0.75T \rfloor$	6.825	6.506	6.637
Break fraction: $\lfloor 0.5T \rfloor$			
Missing Value Location	Zo	Pr	Li
$\lfloor 0.25T \rfloor$	5.835	5.624	5.798
$\lfloor 0.5T \rfloor$	5.319	5.094	4.692
$\lfloor 0.75T \rfloor$	5.835	5.624	5.798
Break fraction: $\lfloor 0.75T \rfloor$			
Missing Value Location	Zo	Pr	Li
$\lfloor 0.25T \rfloor$	6.825	6.384	6.637
$\lfloor 0.5T \rfloor$	6.825	6.507	6.637
$\lfloor 0.75T \rfloor$	6.32	6.101	6.107

Notes: The above table provides the values of K_m for one missing value in the sample. The missing value appears at the same place for all units at the locations $\lfloor 0.25T \rfloor$, $\lfloor 0.5T \rfloor$ and $\lfloor 0.75T \rfloor$, for $T = 20$. For the Karavias and Tzavalis (2014) test, three break dates are also considered, again at the locations $\lfloor 0.25T \rfloor$, $\lfloor 0.5T \rfloor$ and $\lfloor 0.75T \rfloor$. "Zo" stands for zeroing-out, "Pr" stands for previous observed value and "Li" stands for linear interpolation. A larger value of K_m indicates higher power.

Notes: The above table provides the values of K_m for one missing value in the sample. The missing value appears at the same place for all units at the locations $\lfloor 0.25T \rfloor$, $\lfloor 0.5T \rfloor$ and $\lfloor 0.75T \rfloor$, for $T = 20$. For the Karavias and Tzavalis (2014) test, three break dates are also considered, again at the locations $\lfloor 0.25T \rfloor$, $\lfloor 0.5T \rfloor$ and $\lfloor 0.75T \rfloor$. "Zo" stands for zeroing-out, "Pr" stands for previous observed value and "Li" stands for linear interpolation. A larger value of K_m indicates higher power.

The results of Table 1 demonstrate that the Harris and Tzavalis (1999) test performs best when the missing value is zeroed-out, and the next best method in terms of power is the linear interpolation. The worse method is substituting the missing value with the previous observation. This is the first result in the panel unit root test literature on the impact and treatment of missing values. The fact that zeroing-out produces the highest power agrees with the single time series results of Ryan and Giles (1998). However, they find that linear interpolation performs worst, which is contrary to what is found here, as linear interpolation outperforms the previous value substitution.

In the presence of structural breaks the results are less clear-cut. The zeroing-out method dominates the other two, for all combinations of structural break and missing value dates. However, the ranking between the other two methods depends on the relative location of the missing value and the date of the structural break. If the missing observation, or the next, coincides with the date of the structural break, then the substitution of the last available observation leads to higher power than linear interpolation. For the rest of the cases, linear interpolation leads to greater power.

The presence of linear trends in $m = 2, 4$ results in $K_m = 0$ in (21). This is the known problem of trivial local power in the presence of incidental trends, see, e.g., Moon et al. (2007) and Karavias and Tzavalis (2016, 2017). This result demonstrates that panel unit root tests do not have power in the presence of incidental trends in a neighbourhood of the null hypothesis. For the purposes of the analysis here, this means that the asymptotic local

power functions cannot be used to show which method is the best. However, unreported Monte Carlo simulations show that for alternatives far from the null, the results for the case without trends still hold.

The results of Table 1 allow for a comparison of the relative power of the three missing value methods. The absolute powers can be calculated from $\mathcal{T}(v_\alpha + cK_m)$, where $\mathcal{T}$ is the cumulative distribution function of the standard normal distribution, and v_α is the α-level percentile of said distribution. The absolute power gains of zeroing-out over the other methods are greater when T is smaller, c is closer to 0, and when the number of missing values is larger.

The above conclusions extend to settings with multiple missing values, and cases where the number and locations of missing values differ across units. To save space, we do not present these results but are available upon request. The Harris and Tzavalis (1999), Karavias and Tzavalis (2014) tests can accommodate cross-section dependence in the form of an additive time effect, however, that does not change the above analysis or its conclusions.

6. Conclusions

This paper examined the impact of missing value correction methods in panel data unit root tests. The analysis focused on the fixed-T tests of Harris and Tzavalis (1999) and their extension in the presence of structural breaks by Karavias and Tzavalis (2014).

The first contribution of the paper is the extension of the aforementioned tests to allow for missing observations in the data. The fixed effects estimators in dynamic panel data models are inconsistent and need to be bias-corrected; the present paper shows how this bias correction can be done and provides the appropriate formulas for using the tests in practice.

The second contribution is a study of the power properties of the tests under various methods for dealing with missing data. To carry out this analysis, we derived asymptotic local power functions which can be used to analytically compare different methods. We used the new formulas to compare the methods of zeroing-out (which is equivalent to closing the gaps in the data), replacing the missing value with the last available observation and using linear interpolation in the form of the average of the two adjacent observations. Overall, the results show that the zeroing-out or "closing gaps" methodology dominates the other two and should be the preferred method in practice.

Author Contributions: Conceptualization, Y.K., E.T. and H.Z.; writing—original draft preparation, Y.K., E.T. and H.Z.; writing—review and editing, Y.K., E.T. and H.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank an Associate Editor and three referees for their constructive comments.

Conflicts of Interest: The authors declare no conflict of interest.

Notes

1 In Stata, the test by Harris and Tzavalis (1999) has been implemented in the official "xtunitroot" command, while the test of Karavias and Tzavalis (2014), which allow for structural breaks, have been implemented by the community contributed command of "xtbunitroot" by Chen et al. (2021).

2 In the current context the term "bias" is frequently used to describe inconsistency. This happens because when the time series dimension is assumed fixed, the autoregressive parameter estimator bias (Hurwicz 1950; Nickell 1981) persists asymptotically and creates inconsistency.

[3] The assumptions under which this result holds are presented in Section 4 below.

[4] Ryan and Giles (1998) find size distortions under the null which leads them to consider size-adjusted power. This is not needed here.

Bai, Jushan. 2010. Common breaks in means and variances for panel data. *Journal of Econometrics* 157: 78–92. [CrossRef]

Bhargava, Alok. 1989. Missing observations and the use of Durbin–Watson statistic. *Biometrica* 76: 828–31.

Breitung, Järg, and Wolfgang Meyer. 1994. Testing for unit roots in panel data: Are wages on different bargaining levels cointegrated? *Applied Economics* 26: 353–61. [CrossRef]

Chen, Pengyu, Yiannis Karavias, and Elias Tzavalis. 2021. *Panel Unit Root Tests with Structural Breaks*. Discussion Papers 21-12. Birmingham: Department of Economics, University of Birmingham.

Dickey, David A., and Wayne A. Fuller. 1979. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association* 74: 427–31.

Dickey, David A., and Wayne A. Fuller. 1981. Likelihood ratio statistics for autoregressivev time series with a unit root. *Econometrica* 49: 1087–72. [CrossRef]

Durbin, James, and Geoffrey S. Watson. 1950. Testing for serial correlation in least squares regression I. *Biometrika* 37: 409–28

Durbin, James, and Geoffrey S. Watson. 1951. Testing for serial correlation in least squares regression II. *Biometrika* 38: 159–78. [CrossRef]

Durbin, James, and Geoffrey S. Watson. 1971. Testing for serial correlation in least squares regression III. *Biometrika* 58: 1–19. [CrossRef]

Hall, Alastair. 1989. Testing for a unit root in the presence of moving average errors. *Biometrica* 76: 49–56. [CrossRef]

Harvey, Andrew Charles. 1989. *Forecasting, Structural Time Series and the Kalman Filter*. Cambridge: Cambridge University Press.

Harris, Richard D. F., and Elias Tzavalis. 1999. Inference for unit roots in dynamic panels where the time dimension is fixed. *Journal of Econometrics* 91: 201–26. [CrossRef]

Hayashi, Fumio. 2000. *Econometrics*. Princeton: Princeton University Press.

Holtz-Eakin, Douglas, Whitney Newey, and Harvey S. Rosen. 1988. Estimating Vector Autoregressions with Panel Data. *Econometrica* 56: 1371–95. [CrossRef]

Hurwicz, Leonid. 1950. *Least Squares Bias in Time Series, Statistical Inference in Dynamic Economic Models*. Edited by Tjalling Charles Koopmans. New York: Wiley.

Im, Kyung So, M. Hashem Pesaran, and Yongcheol Shin. 2003. Testing for unit roots in heterogeneous panels. *Journal of Econometrics* 115: 53–74. [CrossRef]

Karavias, Yiannis, and Elias Tzavalis. 2014. Testing for unit roots in short panels allowing for a structural break. *Computational Statistics and Data Analysis* 76: 391–407. [CrossRef]

Karavias, Yiannis, and Elias Tzavalis. 2016. Local power of fixed-T panel unit root tests with serially correlated errors and incidental trends. *Journal of Time Series Analysis* 37: 222–39. [CrossRef]

Karavias, Yiannis, and Elias Tzavalis. 2017. Local power of panel unit root tests allowing for structural breaks. *Econometric Reviews* 36: 1123–56. [CrossRef]

Karavias, Yiannis, and Elias Tzavalis. 2019. Generalized fixed-T panel unit root tests. *Scandinavian Journal of Statistics* 46: 1227–51. [CrossRef]

Karavias, Yiannis, Paresh Narayan, and Joakim Westerlund. 2021. Structural Breaks in Interactive Effects Panels and the Stock Market Reaction to COVID–19. *arXiv*, arXiv:2111.03035v1.

Levin, Andrew, Chien-Fu Lin, and Chia-Shang James Chu. 2002. Unit root tests in panel data: Asymptotic and finite-sample properties. *Journal of Econometrics* 108: 1–24. [CrossRef]

Madsen, Edith. 2010. Unit root inference in panel data models where the time-series dimension is fixed: A comparison of different tests. *Econometrics Journal* 13: 63–94. [CrossRef]

Moon, Hyungsik Roger, Benoit Perron, and Peter C. B. Phillips. 2007. Incidental trends and the power of panel unit root tests. *Journal of Econometrics* 141: 416–59. [CrossRef]

Nickell, Stephen. 1981. Biases in Dynamic Models with Fixed Effects. *Econometrica* 49: 1417–26. [CrossRef]

Ryan, Kevin F., and David E. A. Giles. 1998. Testing for unit roots in economic time-series with missing observations. In *Advances in Econometrics*. Edited by Thomas Fomby and Carter Hill. Greenwich: JAI Press, pp. 203–42.

Savin, N. Eugene, and Kenneth J. White. 1978. Testing for autocorrelation with missing observations. *Econometrica* 46: 59–67. [CrossRef]

Shin, Dong Wan, and Sahadeb Sarkar. 1994a. Unit roots for ARIMA(0,1,q) models with irregularly observed samples. *Statistics and Pobability Letters* 19: 188–94.

Shin, Dong Wan, and Sahadeb Sarkar. 1994b. Likelihood ratio type unit root tests for AR(1) models with nonconsecutive observations. *Comunications in Statistics: Theory and Methods* 23: 1387–97.

Verbeek, Marno, and Theo Nijman. 1992. Testing for Selectivity Bias in Panel Data Models. *International Economic Review* 33: 681–703. [CrossRef]