Fall 9-2023

The Mundlak Spatial Estimator

Badi H. Baltagi
Syracuse University, bbaltagi@maxwell.syr.edu

CENTER FOR POLICY RESEARCH
THE MAXWELL SCHOOL

WORKING PAPER SERIES

The Mundlak Spatial Estimator

Badi Baltagi

Paper No. 258
September 2023

ISSN: 1525-3066
426 Eggers Hall
Syracuse University
Syracuse, NY 13244-1020
T 315.443.3114 **E** ctrpol@syr.edu
https://surface.syr.edu/cpr_workingpapers/

Syracuse University
Maxwell School of
Citizenship & Public Affairs
Center for Policy Research

CENTER FOR POLICY RESEARCH - Fall 2023

Shannon Monnat, Director
Professor of Sociology, Lerner Chair in Public Health Promotion & Population Health

Associate Director
Margaret Austin

SENIOR RESEARCH ASSOCIATES

Badi Baltagi, ECN
Robert Bifulco, PAIA
Carmen Carrión-Flores, ECN
Monica Deza, ECN
Sean Drake, SOC
Amy Fairchild, MAX
Alfonso Flores-Lagunes, ECN
Sarah Hamersma, PAIA
Madonna Harrington Meyer, SOC
Colleen Heflin, PAIA
Yilin Hou, PAIA

Hugo Jales, ECN
Gabriela Kirk-Werner, SOC
Jeffrey Kubik, ECN
Yoonseok Lee, ECN
Leonard M. Lopoo, PAIA
Amy Lutz, SOC
Yingyi Ma, SOC
Shannon Monnat, SOC
Tomas Olivier, PAIA
Jan Ondrich, ECN
David Popp, PAIA

Stuart Rosenthal, ECN
Michah Rothbart, PAIA
Alexander Rothenberg, ECN
Rebecca Schewe, SOC
Ying Shi, PAIA
Saba Siddiki, PAIA
Perry Singleton, ECN
Michiko Ueda-Ballmer, PAIA
Yulong Wang, ECN
Maria Zhu, ECN

RESEARCH AFFILIATES

Len Burman, PAIA*
Christopher Faricy, POL. SCI.
Shana Gadarian, POL. SCI
William Horrace, ECN
Jun Li, PAIA
Wei Li, MATH

Jianxuan Liu, MATH
Andrew London, SOC
Jerry Miner, ECN*
Jennifer Karas Montez, SOC.
Merril Silverstein, SOC
Rick Welsh, FST

Emily Wiemers, PAIA
Peter Wilcoxen, PAIA
Janet Wilmoth, SOC
Douglas Wolf, PAIA*
John Yinger, PAIA

GRADUATE ASSOCIATES

Graham Ambrose, PAIA
Roseanna Benser, SOC
Mayra Cervantes, PAIA
Brandon Charles, PAIA
Kyuhan Choi, ECN
Adam Cucchiara, PAIA
Ehsan Dowlatabadi, ECN
William Clay Fannin, PAIA
Myriam Gregoire-Zawilski, PAIA
Joshua Grove, SOC

Lyuan Han, PAIA
Ashraf Haque, PAIA
Harneet Kaur, SOC
Shanel Khaliq, SOC
Hyojeong Kim, PAIA
Dong Lee, PAIA
Mitchell McFarlane, PAIA
Nicholas Oesterling, PAIA
Raquel Prado, PAIA
Michael Quinn, ECN

Shaneya Simmelkjaer, SOC
Andrew Smith, ECN
Sarah Souders, PAIA
Yue Sun, SOC
Juan Uribe-Quintero, PAIA
Nastassia Vaitsiakhovich, SOC
Francisco Villarroel, PAIA
Shuping Wang, PAIA
Jingni Zhang, PAIA
Yuwei Zhang, PAIA
Bo Zheng, PAIA

POSTDOCTORAL SCHOLARS

Michael Dunaway, Postdoctoral Research Scholar
Xue Zhang, Postdoctoral Research Scholar

RESEARCH STAFF

Jack Baldwin, Senior Associate, Maxwell X Lab
Emily Graham, Assistant Director, Lerner Center
Hannah Patnaik, Managing Director, Maxwell X Lab
Alex Punch, Director, Lerner Center

CPR STAFF

Katrina Fiacchi, Assistant Director
Zia Jackson, Office Coordinator
Alyssa Kirk, Office Coordinator
Davor Mondom, CPDG Center Coordinator
Candi Patterson, Computer Consultant

*Emeritus

Abstract

The spatial Mundlak model first considered by Debarsy (2012) is an alternative to fixed effects and random effects estimation for spatial panel data models. Mundlak modelled the correlated random individual effects as a linear combination of the averaged regressors over time plus a random time-invariant error. This paper shows that if spatial correlation is present whether spatial lag or spatial error or both, the standard Mundlak result in panel data does not hold and random effects does not reduce to its fixed effects counterpart. However, using maximum likelihood one can still estimate these spatial Mundlak models and test the correlated random effects specification of Mundlak using Likelihood ratio tests as demonstrated by Debarsy for the Mundlak spatial Durbin model.

JEL No.: C33

Keywords: Mundlak Regression, Panel Data, Fixed and Random Effects, Spatial error model, Spatial Durbin model

Authors: Badi H. Baltagi, Distinguished Professor and Senior Research Associate, Department of Economics and Center for Policy Research, The Maxwell School of Citizenship and Public Affairs, Syracuse University, 426 Eggers Hall, Syracuse, NY 13244, bbaltagi@syr.edu

1 Introduction

Debarsy (2012) is the first to extend the Mundlak (1978) approach to the spatial Durbin panel data model (SDM). [1]This adds an auxiliary regression for the correlated random individual effects that is a function of the explanatory variables along with their spatial weighted averages both averaged over time. A likelihood ratio (LR) test that assesses the significance of the correlation between regressors and individual effects is proposed, and its properties are investigated using Monte Carlo simulations as well as an empirical example to explain housing price variations across 588 municipalities in Belgium over the period 2004 to 2007. This paper revisits the Mundlak (1978) model but in the context of Anselin's (1988) spatial error model (SEM) with a spatial autoregressive remainder term. It shows that what Mundlak showed for panel data does not extend to the spatial panel SEM, i.e., random effects does not reduce to fixed effects once the average regressors over time are included in the random effects SEM regression. This is because of the presence of spatial correlation in the remainder disturbances. Of course, OLS on this Mundlak SEM model, ignoring spatial correlation in the remainder disturbances, still yields the fixed effects estimator as in the panel data case of Mundlak (1978). However, GLS on this Mundlak SEM model accounting for spatial correlation in the remainder error does not yield the fixed effects estimator. So random effects does not reduce to fixed effects for the SEM Mundlak model. This is different from the panel data case with no spatial correlation where Mundlak's correlated random effects model reduces to the fixed effects estimator and OLS is equivalent to GLS, see Baltagi (2006). All is not lost, however, using maximum likelihood estimation (MLE) under normality, Mundlak's (1978) idea of modelling the random effects as correlated with the regressors can be tested using Likelihood ratio (LR) tests. This is exactly the recommendation of Debarsy (2012) for the Mundlak spatial Durbin model (with spatial lag on the dependent variable and the regressors but not on the error). The presence of spatial lag introduces additional endogeneity besides the random correlated individual effects and Debarsy (2012) recommended MLE and LR testing and illustrated it using an empirical example to explain housing price variations in Belgium. We illustrate these results with xsmle in Stata applied to Belotti, Hughes, and Piano Mortari

[1]Note that Debarsy's (2012) SDM does not have a spatial error model in the remainder disturbance and hence does not nest the SEM considered in this paper.

(2017) data set on residential demand for electricity covering the 48 states in the continental United States plus the district of Columbia for the period 1990-2010. Mundlak's adding regressors averages over time instead of fixed effects in a correlated random effects model is a useful tool in panel data. Perhaps it will be in Spatial panel models. This paper warns that although it can be used as Debarsy (2012) did, it does not yield exactly the same results as fixed effects estimation. For both the SEM and SDM we show that the spatial Mundlak correlated random effects estimator does not reduce to its fixed effects MLE counterpart. The LR test shows that these Mundlak averages are jointly significant in the SEM and SDM for residential demand for electricity in the United States.

2 The Mundlak Spatial Error Component Regression Model

Consider the spatial error component regression model:

$$y_{ti} = X'_{ti}\beta + u_{ti}, \; i = 1, .., N; t = 1, \cdots, T, \tag{1}$$

where y_{ti} is the observation on the ith country for the tth time period, X_{ti} denotes the $k \times 1$ vector of observations on the non-stochastic regressors and u_{ti} is the regression disturbance. In vector form, the disturbance is assumed to have random country effects as well as a spatially remainder error term (SEM), see Anselin (1988):

$$u_t = \mu + \eta_t \tag{2}$$

with η_t following an autoregressive specification[2]

$$\eta_t = \lambda W_N \eta_t + \nu_t \tag{3}$$

where $\mu' = (\mu_1, \cdots, \mu_N)$ denote the vector of random country effects which are assumed to be $\text{IIN}(0, \sigma^2_\mu)$. λ is the scalar spatial autoregressive coefficient with $\mid \lambda \mid < 1$. W_N is a known $N \times N$ spatial weight matrix whose diagonal

[2]The same result, i.e., OLS is not equivalent to GLS can be similarly shown for the correlated random effects spatial Mundlak model when the remainder disturbance in SEM is a spatial moving average. This is not shown here to save space.

elements are zero. W_N also satisfies the condition that $(I_N - \lambda W_N)$ is non-singular. $\nu'_t = (\nu_{t1}, \cdots, \nu_{tN})$, where ν_{ti} is assumed to be $\text{IIN}(0, \sigma^2_\nu)$ and also independent of μ_i. In vector form, one can rewrite this as

$$\eta_t = (I_N - \lambda W_N)^{-1}\nu_t = B^{-1}\nu_t \tag{4}$$

where $B = I_N - \lambda W_N$ and I_N is an identity matrix of dimension N. The regression model can be rewritten in matrix notation as

$$y = X\beta + u \tag{5}$$

where y is now of dimension $NT \times 1$, X is $NT \times k$, β is $k \times 1$ and u is $NT \times 1$. The observations are sorted such that the slow index is t and the fast index is i. X is assumed to be of full column rank and its elements are assumed to be bounded in absolute value. The disturbance term can be written in vector form as:

$$u = Z_\mu \mu + (I_T \otimes B^{-1})\nu \tag{6}$$

where $\nu' = (\nu'_1, \cdots, \nu'_T)$. Let $Z_\mu = (\iota_T \otimes I_N)$, where ι_T is a vector of ones of dimension T. I_N is an identity matrix of dimension N. In this case, $J_T = \iota_T \iota'_T$ is a matrix of ones of dimension T, so that $\bar{J}_T = J_T/T$ is the averaging matrix over T. Also, let $E_T = I_T - \bar{J}_T$, be the deviations from average matrix. Define P to be the projection matrix on $Z_\mu = (\iota_T \otimes I_N)$, i.e., $P = \bar{J}_T \otimes I_N$, and $Q = (I_{NT} - P) = E_T \otimes I_N$. The fixed effects for the Anselin (1988) model premultiplies the regression model by Q to get the within model

$$Qy = QX\beta + Qu \tag{7}$$

where

$$Qu = (E_T \otimes B^{-1})\nu \tag{8}$$

and the individual effects are wiped out, since $E_T \ \iota_T = 0$. The remainder disturbance has variance covariance matrix $\Phi = \sigma^2_\nu(E_T \otimes (B'B)^{-1})$ and is not $\sigma^2_\nu Q$ unless $\lambda = 0$. Note that, the within panel regression (7), ignoring spatial correlation in the remainder disturbances, i.e., OLS on (7) yields the standard fixed effects panel data estimator or within estimator, see Baltagi (2021)

$$\widehat{\beta}_{OLS} = (X'QX)^{-1}X'Qy = \widetilde{\beta}_w \tag{9}$$

This is different from performing GLS on this spatial within model which yields

$$\begin{aligned}\widehat{\beta}_{GLS} &= (X'Q(E_T \otimes (B'B))QX)^{-1}X'Q(E_T \otimes (B'B))Qy \qquad (10)\\ &= (X'(E_T \otimes (B'B))X)^{-1}X'(E_T \otimes (B'B))y\end{aligned}$$

using the generalized inverse $\Phi^- = (E_T \otimes (B'B)/\sigma_\nu^2$ and $Q\Phi^-Q = E_T \otimes (B'B)/\sigma_\nu^2$. This reduces to the within estimator in (9) when $\lambda = 0$. In this case, the standard fixed effects estimator based on OLS yields consistent but not efficient (Best Linear Unbiased) estimates of β, while $\widehat{\beta}_{GLS}$ yields (Best Linear Unbiased) estimates of β. Also, the standard errors of the fixed effects estimates ignoring the spatial error correlation will yield misleading inference. Therefore for the Anselin (1988) SEM, the within estimator has to be performed with GLS or (MLE under normality) in order to get the efficient estimator and proper inference.

Mundlak (1978) argued that if the individual effects are random and correlated with "all" the explanatory variables through their averages over time, i.e.

$$\mu_i = \bar{X}'_{.i}\pi_\mu + \epsilon_i \qquad i = 1, 2, .., N \qquad (11)$$

where $\epsilon_i \sim \text{IIN}(0, \sigma_\epsilon^2)$ and $\bar{X}'_{.i}$ is $1 \times K$ vector of observations on the explanatory variables averaged over time, random effects estimation reduces to fixed effects estimation for β. These individual effects are uncorrelated with the explanatory variables if and only if $\pi_\mu = 0$, otherwise this is a correlated random effects regression. Mundlak (1978) assumed, without loss of generality, that the X's are in deviations from their sample mean. In vector form (11) can be written as follows:

$$\mu = Z'_\mu X\pi_\mu/T + \epsilon \qquad (12)$$

where $\mu' = (\mu_1, \ldots, \mu_N)$, and $\epsilon' = (\epsilon_1, \ldots, \epsilon_N)$. Substituting this auxiliary regressions for μ, defined in (12) into (6), one gets

$$u = PX\pi_\mu + Z_\mu\epsilon + (\iota_T \otimes B^{-1})\nu \qquad (13)$$

where $P = \bar{J}_T \otimes I_N$ is the averaging matrix defined earlier. The Mundlak SEM regression becomes

$$y = X\beta + PX\pi_\mu + Z_\mu\epsilon + (\iota_T \otimes B^{-1})\nu \qquad (14)$$

The variance-covariance matrix for the disturbances of this SEM Mundlak regression is given by

$$\Omega = \sigma_\epsilon^2(J_T \otimes I_N) + \sigma_\nu^2(I_T \otimes (B'B)^{-1}) \qquad (15)$$

This matrix can be rewritten as:

$$\Omega = \sigma_\nu^2 \Big[\bar{J}_T \otimes (T\phi I_N + (B'B)^{-1}) + E_T \otimes (B'B)^{-1} \Big] = \sigma_\upsilon^2 \Sigma \tag{16}$$

where $\phi = \sigma_\epsilon^2 / \sigma_\nu^2$. Using results in Wansbeek and Kapteyn (1982), Σ^{-1} is given by

$$\Sigma^{-1} = \bar{J}_T \otimes (T\phi I_N + (B'B)^{-1})^{-1} + E_T \otimes B'B. \tag{17}$$

GLS on (14) using Σ^{-1} yields the SEM Mundlak random effects estimator from the augmented regression with averaged regressors over time. Obviously, this is different from the fixed effects SEM estimates for β given in (10) . The Mundlak standard panel data result where correlated random effects yields the fixed effects estimator does not extend to the SEM Mundlak model as we showed above.

Note that $|\Sigma| = |T\phi I_N + (B'B)^{-1}| \cdot |(B'B)^{-1}|^{T-1}$. Under the assumption of normality, the log-likelihood function for this model is given by

$$\begin{aligned} L &= -\frac{NT}{2} \ln 2\pi\sigma_\nu^2 - \frac{1}{2} \ln |\Sigma| - \frac{1}{2\sigma_\nu^2} d'\Sigma^{-1} d \\ &= -\frac{NT}{2} \ln 2\pi\sigma_\nu^2 - \frac{1}{2} \ln[|T\phi I_N + (B'B)^{-1}|] + \frac{(T-1)}{2} \ln |B'B| \\ &\quad -\frac{1}{2\sigma_\nu^2} d'\Sigma^{-1} d \end{aligned} \tag{18}$$

with $d = y - X\beta - PX\pi_\mu$. When $\pi_\mu = 0$, the model reverts to the SEM random effects panel data model and the first-order conditions of MLE as well as the LM test for $\lambda = 0$ for this model, are given by Anselin (1988).[3]When $\pi_\mu \neq 0$, this a a correlated random effects SEM model and performing GLS on this augmented Mundlak spatial regression does not yield the fixed effects SEM estimates as in Mundlak's (1978) panel data case without spatial correlation. However performing MLE yields asymptotically efficient estimates for the regression coefficients and the Likelihood ratio (LR) test for $\pi_\mu = 0$ yields a valid test for whether the random effects are correlated with the regressors through their averages as Mundlak suggested.

[3]As an extension to this work, Baltagi, Song and Koh (2003) derived the joint LM test for spatial error correlation as well as random country effects. Additionally, they derived conditional LM tests, which test for random country effects given the presence of spatial error correlation. Also, spatial error correlation given the presence of random country effects.

3 Empirical Illustration: Residential demand for electricity

We illustrate the results with the xsmle command in Stata applied to Belotti, Hughes, and Piano Mortari (2017) data set on residential demand for electricity covering the 48 states in the continental United States plus the district of Columbia for the period 1990-2010. The dependent variable is the log of residential electricity sales and it is modelled as a function of log real per-capita income, log of real average residential price of electricity, log of housing units per capita, log of cooling degree and heating degree days. The Stata data set is available as **state_spatial_dbf.dta**. Using a rook W matrix for spatial contiguity of the 48 states in the continental United States plus the district of Columbia, **xsmle** allows us to estimate several spatial panel models including SEM for Anselin (1988) and SDM for Debarsy (2012). Column 1 of Table 1 shows the fixed effects results using xtreg ignoring spatial correlation, this matches Table 5 of Belotti, Hughes, and Piano Mortari (2017). These fixed effects are significant yielding an F(48, 1024) = 358.4 which rejects the null that the state effects are zero. This is equivalent to running Mundlak's (1978) augmented regression with time averages of all the regressors given in column 2 of Table 1. The time averages of heating units and degree cooling days are insignificant, but the remainder time averages show significance implying correlation between some of the explanatory variables and the random individual state effects. The joint F-test for the significance of these Mundlak averages is F(5, 1067) = 32.28 which rejects the null that these Mundlak averages are all zero.

Column 3 of Table 1 gives the SEM fixed effects maximum likelihood results which match the results in Table 5 of Belotti, Hughes, and Piano Mortari (2017) that they label (SEM). The spatial effects in the remainder term (lambda) is significant implying that inference from fixed effects ignoring this spatial correlation may be misleading. Next, we add the Mundlak (1978) time averages of the regressors and run the SEM model with random effects. The results are given in Column 4 of Table 1. Although these results are close, they are not exactly the same as the fixed effects SEM given in column 3 of Table 1. So, Mundlak's SEM with random effects does not yield Fixed effects SEM. Note also that the time averages of real income per capita, heating units and degree cooling days are insignificant. The LR statistic for the joint significance of the Mundlak average regressors is 53.6

which is asymptotically distributed as χ^2_5 and is significant.

Column 5 of Table 1 gives the SDM with fixed effects as in Debarsy (2012) but now for the residential electricity demand. The spatial lag (rho) is significant. Note that these results differ from the results in Table 5 of Belotti, Hughes, and Piano Mortari (2017) labelled SDM because here all the regressors where spatially lagged whereas Belotti, et al. selected only one regressor to be spatially lagged. Note that the spatially weighted degree cooling as well as the spatially weighted degree heating days are insignificant. Next, we add the Mundlak time averages and run random effects SDM. These are given in column 6 of Table 1. These results are slightly different from column 5 of Table 1. So, the Mundlak SDM with random effects does not yield the fixed effects SDM estimator. Interestingly, column 6 of Table 1 indicates that the time averages of two regressors are insignificant, but all the time averages of the *spatially weighted* regressors are insignificant. The LR statistic for the joint significance of the Mundlak average regressors is 56.1 which is asymptotically distributed as χ^2_5 and is significant.[4]

4 Conclusion

This paper revisits the Mundlak (1978) spatial model considered by Debarsy (2012) in the context of a spatial panel Durbin (SDM) model. It starts by showing that for the random effects spatial error model (SEM) considered by Anselin (1988), the spatial Mundlak random effects estimator does not reduce to the fixed effects SEM estimator. This is different from the standard Mundlak panel result where random effects reduces to fixed effects once the random individual effects are modeled as a linear function of *all* the averaged regressors over time. This non equivalence result between fixed effect and random effect also holds for other spatial Mundlak models including the Mundlak spatial Durbin model (SDM) considered by Debarsy (2012). The SDM model considered by Debarsy does not have a spatial error model but does have a spatial lag and spatially weighted regressors so it does not nest the SEM considered in this paper. As demonstrated by Debarsy (2012), one can use maximum likelihood estimation to estimate these spatial Mundlak models and test the

[4] It is important to note that direct and indirect effects can be computed as described in LeSage and Pace (2009), Elhorst (2014), Debarsy (2012) and computed using Stata's command xsmle by Belotti, et al. (2017). These results were not reported here as the purpose of this application is to demonstrate the difference between fixed effects spatial and Mundlak random effects spatial models.

Table 1 Mundlak Spatial Estimates for Residential Electricity Demand

ln_sales_rpop	**FE**	**Mundlak**	**FE-SEM**	**SEM-Mundlak**	**FE-SDM**	**Mundlak-SDM**
x						
ln_rinc_cap	.391**	.391**	.375**	.374**	.139**	.140**
ln_gprice_res	-.235**	-.235**	-.271**	-.271**	-.294**	-.294**
ln_hunit_pop	1.019**	1.019**	.818**	.818**	.566**	.568**
ln_degday_cool	.075**	.075**	.071**	.071**	.058**	.058**
ln_degday_heat	.188**	.188**	.156**	.156**	.130**	.130**
xbar						
mrinc		-.192**		-.176		.003
mgprice		-.615**		-.579**		-.594**
mhunit		.261		.462		.800**
mcool		.018		.022		.055
mheat		-.336**		-.304**		-.245**
Spatial						
lambda			.390**	.390**		
Wx						
ln_rinc_cap					.107**	.109**
ln_gprice_res					.183**	.181**
ln_hunit_pop					.224**	.231**
ln_degday_cool					-.003	-.002
ln_degday_heat					.010	.012
Wxbar						
mrinc						-.139
mgprice						.182
mhunit						-.689
mcool						-.126
mheat						-.110
Spatial						
rho					.359**	.351**
LR statistic						
Mundlak terms				53.6**		56.1**

** indicates significance at the .05 level. x indicates the regressors. xbar indicates the time averages. Wx indicates the spatially weighted regressors for Durbin and Wxbar is their time averages.

Mundlak correlated random effects specification with likelihood ratio tests. We demonstrate this non-equivalence results using xsmle in Stata applied to Belotti, Hughes, and Piano Mortari (2017) data set on residential demand for electricity covering the 48 states in the continental United States plus the district of Columbia for the period 1990-2010.

5 Funding and/or Conflicts of interests/Competing interests

No funding. Declarations of interest: none. No conflicts of interest.

6 References

Anselin, L., 1988, *Spatial Econometrics: Methods and Models* (Kluwer Academic Publishers, Dordrecht).

Baltagi, B.H., 2021, *Econometric Analysis of Panel Data,* 6th edition (Springer, Switzerland).

Baltagi, B.H., 2006, An alternative derivation of Mundlak's fixed effects results using system estimation, *Econometric Theory* 22, 1191-1194.

Baltagi, B.H., S.H. Song and W. Koh, 2003, Testing panel data regression models with spatial error correlation, *Journal of Econometrics* 117, 123-150.

Belotti, F., G. Hughes, and A. Piano Mortari, 2017, Spatial panel data models using Stata, *The Stata Journal* 17(1), 139-180.

Debarsy, N., 2012, The Mundlak approach in the spatial Durbin panel data model. *Spatial Economic Analysis* 7, 109-131.

Elhorst, J. P., 2014, *Spatial Econometrics: From Cross-Sectional Data to Spatial Panels* (Springer, Heidelberg).

LeSage, J. P., and R. K. Pace, 2009, *Introduction to Spatial Econometrics* (Taylor & Francis, Boca Raton).

Mundlak, Y., 1978, On the pooling of time series and cross-section data, *Econometrica* 46, 69–85.

Wansbeek, T.J. and A. Kapteyn, 1982, A simple way to obtain the spectral decomposition of variance components models for balanced data, *Communications in Statistics* A11, 2105–2112.