ORIGINAL ARTICLE

Estimating dynamic spatial panel data models with endogenous regressors using synthetic instruments

Bernard Fingleton[1] ●

Abstract

The paper applies synthetic instruments, initially developed for cross-sectional regression, to estimate dynamic spatial panel data models. These have two main advantages. First, instruments correlated with endogenous variables and yet independent of the errors are difficult to find. Not only are synthetic instruments normally exogenous, but they are usually strongly correlated with endogenous variables, and thus help to avoid the problem of weak instruments. Secondly, they help to reduce instrumental variables proliferation, which is a common result of standard methods of avoiding endogeneity bias. As demonstrated by Monte Carlo simulation, instrument proliferation causes bias in the Sargan–Hansen J test statistic, which is an important indicator of instrument validity and hence estimation consistency. It is also associated with a downward bias in parameter standard error estimates. The paper shows the results of applying synthetic instruments across a variety of different specifications and data generating processes, and it illustrates the method with real data leading to more reliable inference of causal impacts on the level of employment across London districts.

Keywords Dynamic spatial panel data models · Synthetic instruments · Sargan–Hansen J test · Monte Carlo simulation · Inference · Migration · Employment

JEL Classification C31–C33 · R1

1 Introduction

Progress in spatial econometrics has been stimulated by the increasing availability of spatio-temporal data, as shown by Anselin et al. (2008), Elhorst (2014), Pesaran (2015) and Baltagi (2021) among others. Static spatial panel data methods are

✉ Bernard Fingleton
bf100@cam.ac.uk

1 Department of Land Economy, University of Cambridge, 19 Silver Street, CB3 9EP Cambridge, UK

long-established, but more recently dynamic spatial panel data models have come to the fore as highly informative and relevant (see for example Parent and LeSage 2012; Yu and Lee 2008; Elhorst 2014). The predominant approach in the spatial panel model literature has been an assumption that the regressors are exogenous, apart from spatial and/or temporal lag terms relating to the endogenous variable[1] and typically estimation has focused on maximum likelihood and related methods. This presents a challenge in the presence of endogenous regressors. The focus in this paper is on dynamic spatial panel models in which endogeneity extends beyond spatial and temporal lags of the dependent variable. Baltagi et al. (2019) give extensive Monte Carlo simulations demonstrating the consistency of the estimates produced by a spatial version of 'difference generalized method of moments (GMM)', an approach which easily accommodates endogenous regressors.[2]

The main contribution of the paper is via the approach of Le Gallo and Páez (2013), who advocated the use of synthetic instruments to eliminate endogeneity bias in cross-sectional regressions, which is here extended to GMM estimation of dynamic spatial panel models. The intuition behind synthetic instruments is that because they are based on the invariant topology of the geographic space, typically they are exogenous to the spatio-temporal process being modelled. Specifically, because the endogenous variable(s) invariably has a non-random spatial distribution, some inherent dimensions of the topology will often be quite strongly correlated with the endogenous variable. Synthetic instruments deriving from these inherent dimensions typically possess hard-to-find ideal properties of instrumental variables, namely exogeneity and yet correlation with the endogenous variable.

In this paper, synthetic instruments help resolve some pitfalls of GMM estimation arising when there is an overabundance of instrumental variables. This can be exacerbated with spatial data, where the suite of instruments might be enhanced by the use of the spatial lags of variables in addition to variables per se (Kelejian and Prucha 1998, 1999; Pace et al. 2012; Baltagi et al. 2019). Moreover, instruments are often weak, with negligible correlation with endogenous variables. In contrast, synthetic instruments are invariably strong, with typically very high correlations. By replacing weak instruments with fewer synthetic instruments, one is likely to obtain more reliable inference. In particular, reducing the number of instruments eliminates problems relating to the crucial Sargan–Hansen test of overidentifying restrictions. An associated problem is the downwardly biased parameter standard errors associated with two-step GMM estimation, which causes upward bias in t-ratios. To remedy this, two related finite sample corrections are reported: the well-known Windmeijer correction and the 'HKL' double correction (Hwang et al. 2022), which also allows for overidentification bias. Finally as an illustration, synthetic instruments

---

[1] Although there are prominent papers where endogenous regressors are also introduced, including Liu and Lee (2013), Jin and Lee (2013), Drukker, Egger and Prucha (2013), Jeanty et al. (2010), Kelejian and Piras (2014, 2016, 2017, 2018) and Fingleton and Le Gallo (2008).

[2] The estimates obtained in this paper can be replicated using the xtabond2 command available in Stata; the relevant code is available below.

and standard error corrections are applied to real data to provide a stronger inferential basis to published research.

2 A dynamic spatial panel model

Consider first the estimation of the simple dynamic model given by Eq. (1),

$$y_{it} = \gamma y_{it-1} + \rho \sum_{j=1}^{N} w_{ij} y_{it} + \theta \sum_{j=1}^{N} w_{ij} y_{it-1} + \beta_1 x_{it} + \beta_2 \tilde{x}_{it} + \varepsilon_{it}; \quad i = 1, ..., N, t = 1, ..., T \tag{1}$$

In which there are N regions/locations/individuals and T times, x is an exogenous variable, $\tilde{x}$ is an endogenous variable, w_{ij} is the i,j th element of an exogenous time-invariant N by N connectivity matrix $\mathbf{W}_N$, $\gamma, \rho, \theta, \beta_1$ and β_2 are parameters to be estimated. The error term is compound, thus

$$\varepsilon_{it} = \mu_i + v_{it}$$

where μ_i is a set of individual effects, one for each of the N regions, controlling for unobserved time-invariant heterogeneity across regions or locations. The term v_{it} varies both by region and by time and represents other, unpredictable, random effects. The assumption is that each μ_i and v_{it} is a random draw from independent and identically distributed distributions thus $\mu_i \sim iid(0, \sigma_\mu^2)$ and $v_{it} \sim iid(0, \sigma_v^2)$ with μ_i and v_{it} independent of each other and among themselves. Given $\sigma_\mu^2 > 0$, there is interregional heterogeneity with μ_i capturing unmodelled individual effects such as physical geography and also regional variation in unobserved effects.

A more general specification written in matrix terms is

$$\mathbf{y}_t = \mathbf{B}_N^{-1} \mathbf{C}_N \mathbf{y}_{t-1} + \mathbf{B}_N^{-1} \mathbf{x}_t \beta_1 + \mathbf{B}_N^{-1} \tilde{\mathbf{x}}_t \beta_2 + \mathbf{B}_N^{-1} \boldsymbol{\varepsilon}_t \tag{2}$$

in which $\mathbf{B}_N = \left(\mathbf{I}_N - \rho \mathbf{W}_N\right)$, $\mathbf{C}_N = \left(\gamma \mathbf{I}_N + \theta \mathbf{W}_N\right)$, $\mathbf{B}_N, \mathbf{C}_N$ and identity matrix $\mathbf{I}_N$ are matrices of dimension N by N, ρ, γ and θ are scalar coefficients, $\mathbf{y}_t$ is an N by 1 vector, $\mathbf{x}_t$ is an N by k_1 matrix of exogenous regressors, β_1 is a k_1 by 1 vector of coefficients, $\tilde{\mathbf{x}}_t$ is an N by k_2 matrix of endogenous regressors, β_2 is a k_2 vector of coefficients, and vector $\boldsymbol{\varepsilon}_t = \boldsymbol{\mu} + \boldsymbol{v}_t$ is a N by 1 compound error term. Standard assumptions are that $\mathbf{B}_N$ is a non-singular matrix and $\mathbf{W}_N, \mathbf{B}_N^{-1}$ and the regressors are uniformly bounded in absolute value. The model satisfies stationarity conditions only if the maximum absolute characteristic root of $\tilde{\mathbf{A}} = \mathbf{C}_N \mathbf{B}_N^{-1}$ is less than one (Elhorst 2001, 2014), Parent and LeSage (2011, 2012) and Debarsy et al. (2012).

Consistent estimation of Eq. (2) by maximum likelihood is challenged by the presence of the endogenous variables, and assumptions regarding initial conditions and on how T and N tend to infinity. Bond (2002) argues that the distribution of the dependent variable depends in a non-negligible way on what is assumed about the distribution of the initial conditions. For example, the initial condition could be stochastic or non-stochastic, correlated or uncorrelated with the individual effects, or have to satisfy stationarity properties. Different assumptions about the nature of the initial conditions will lead to different likelihood functions, and the resulting

ML estimators can be inconsistent when the assumptions on the initial conditions are misspecified. Hsiao (2003, pp. 80–135) has more details. GMM makes weaker assumptions about the initial conditions, indeed according to Baltagi (2013), p. 158, the 'GMM estimator requires no knowledge concerning the initial conditions', and naturally accommodates the presence of endogenous regressors. The paper therefore focusses on GMM which is very well documented in the literature so only a brief summary is provided here.

Following Arellano and Bond (1991), estimation of linear GMM panel data regressions involves first differences[3] to avoid dynamic panel bias (Nickell 1981), eliminating the individual effects μ_i which would otherwise be correlated with the spatial lag and time lag of the dependent variable. First differencing Eq. (2) gives

$$\Delta \mathbf{y}_t = \gamma \Delta \mathbf{y}_{t-1} + \rho \mathbf{W}_N \Delta \mathbf{y}_t + \theta \mathbf{W}_N \Delta \mathbf{y}_{t-1} + \Delta \mathbf{x}_t \beta_1 + \Delta \tilde{\mathbf{x}}_t \beta_2 + \Delta \boldsymbol{\varepsilon}_t \qquad (3)$$

Because of the presence of endogenous variables, instrumental variables are required for consistent estimation. Typically, instruments that are correlated with endogenous variables and yet independent of the errors are difficult to find. One solution is to use lags of regressors already present in the model, but typically, for IV estimation generally, more lags means less data. However, the usual instrument set for difference GMM, namely HENR instruments (after Holtz-Eakin, Newey and Rosin 1988), avoid this by zeroing out missing observations while including separate instruments for each time period. So with HENR one has one instrument per variable, time period and lag distance. This amounts to $(T-2)(T-1)/2$ instruments for each endogenous variable, since endogenous variables are contemporaneously correlated with the errors, provided the v_{it} are not serially autocorrelated of order one, regressors lagged by two periods are appropriate in order to satisfy the orthogonality conditions relating to instruments and differenced errors. Arellano and Bond (1991) provide a test for serial correlation, the m_2 test statistic, which tests for second order serial correlation in the first differenced residuals, and which is assumed to be asymptotically normal under the null of zero correlation. Additionally, following the approach adopted by Baltagi et al. (2019), given data with identifiable spatial locations, spatially weighted earlier time-lagged levels of the dependent and explanatory variables are also potentially viable instruments. Accordingly, Baltagi et al. (2019) set out moments equations thus

$$E\left(y_{il} \Delta v_{it}\right) = 0 \text{ hence } \sum_i y_{il} \Delta v_{it} = 0 \quad , \forall i,\, l = 0, 1, ..., t-2, t = 2, 3, ..., T \qquad (4)$$

$$E\left(\mathbf{w}_i \mathbf{y}_l \Delta v_{it}\right) = 0 \text{ hence } \sum_i \sum_{i \neq j} w_{ij} y_{il} \Delta v_{it} = 0 \quad \forall i,\, l = 0, 1, ..., t-2, t = 2, 3, ..., T \qquad (5)$$

where $\mathbf{w}_i = \left(w_{i1}, ..., w_{iN}\right)$ is a 1 by N vector which corresponds to the $i'th$ row of $\mathbf{W}_N$. Similar expressions giving additional moments equations involve the lagged endogenous regressors, for regressor j, $\tilde{x}_{j,il}$ and $\tilde{\mathbf{x}}_{j,l}$ replaces y_{il} and $\mathbf{y}_l$ in Eqs. (4)

---

[3] System GMM estimation is not considered in this paper.

and (5). With regard to strictly exogenous regressors, there is no feedback from the dependent variable. In this case, the moments conditions are

$$E\left(x_{j,im}\Delta v_{it}\right) = 0, \, \forall i,j, \, m = 1, ..., T; \quad t = 2, ..., T \tag{6}$$

$$E\left(\mathbf{w}_i \mathbf{x}_{j,m} \Delta v_{it}\right) = 0, \, \forall i,j, \, m = 1, ..., T; \quad t = 2, ..., T \tag{7}$$

Baltagi et al. (2019) also introduce a spatial moving average (SMA) error dependence process, but for simplicity the Monte Carlo simulations assume that the errors are spatially independent.[4]

HENR instruments lead to quadratic growth in the number of instruments with respect to T, so there is the possibility of an overabundance of instruments. One can limit the number of lags applying to endogenous variables in the moments equations. A second solution is to collapse the instrument matrix so that there is one instrument for each variable and lag distance, rather than one for each time period, variable, and lag distance, which amounts to adding together columns of the instrument matrix. Also, both solutions can be combined. Collapsing means that one replaces the set of instruments, one for each period, into one column. Therefore, the set of moments equations given by Eqs. (4) and (5) are replaced by

$$\sum_{i,t} y_{it-2} \Delta v_{it} = 0 \tag{8}$$

$$\sum_{i,t} \sum_{i \neq j} w_{ij} y_{it-2} \Delta v_{it} = 0 \tag{9}$$

with similar expressions for the other endogenous regressors.

However, these approaches have limitations, limiting lags alone may not solve the instrument proliferation problem, depending on the context. Collapsing, by omitting time variation, will tend to give less precise estimates. Alternatively, strictly exogenous variables can be introduced as a single column in the matrix of instruments for each exogenous variable, thus producing far fewer instruments than by using the moments conditions of Eqs. (6) and (7). These are referred to these as IV-style instruments rather than HENR-type instruments.

3 Consequences of instruments proliferation

3.1 Parameter standard errors

With numerous instruments, the estimated asymptotic standard errors of the efficient, two-step, GMM estimator are downward biased in small samples. Windmeijer (2005) corrects for the bias which results from estimating the optimal weight matrix

---

[4] SMA error dependence is included in the example using real data.

used in the second step of linear two-step GMM. The optimal weight matrix is the inverse of the covariance of the sample moments leading to the smallest covariance matrix for the GMM estimator. The bias results from the weight matrix being evaluated at estimated, rather than the true values of parameters. However, additional bias may also occur because of overidentification bias, which affects the finite sample bias in the GMM estimator itself. Hwang et al. (2022)[5] also correct for overidentification bias.

3.2 The Sargan–Hansen J test statistic

Theoretically, the moments conditions require that instruments should be orthogonal to the error term. However, with more instruments than variables to be instrumented, one has overidentification and a problem of not being able to exactly satisfy all the moments equations simultaneously. The solution is to attempt to satisfy the moments as closely as possible, and the success of this is given by the outcome of Sargan–Hansen's J test (Sargan 1958; Hansen 1982), as defined by Eq. (10), which tests the null hypothesis of joint validity of the moments conditions under overidentification. Though it is robust to non-sphericity of the errors, it can be greatly weakened by instrument proliferation (Anderson and Sørenson 1996; Bowsher 2002; Roodman 2009a,b).

The J test statistic is given by

$$J = \mathbf{S}_1 \mathbf{A} \mathbf{S}'_2 / N \tag{10}$$

$$\mathbf{S}_1 = \sum_{i=1}^{N} \Delta v'_{i2} \mathbf{Z}_i$$

$$\mathbf{S}_2 = \sum_{i=1}^{N} \mathbf{Z}'_i \Delta v_{i2}$$

$$\mathbf{A} = \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{Z}_i \Delta v_{i1} \Delta v'_{i1} \mathbf{Z}_i \right)^{-1}$$

In $\mathbf{S}_1$ and $\mathbf{S}_2$, Δv_{i2} are differenced second-step errors, $\mathbf{Z}$ is the matrix of instruments, comprised of N $\mathbf{Z}_i$ s, each of dimension $(T-2, p)$ where p is the number of instruments. Under the null hypothesis that the moments conditions are valid, J is distributed as χ^2_{p-k}, where k is the number of estimated parameters and $p > k$, so if J exceeds the relevant critical value of χ^2_{p-k}, some or all of the moments conditions are not supported by the data. The preliminary (one-step) consistent estimator giving the differenced first-step errors is based on

---

[5] Thanks to Jay Lee who provided software support.

$$\mathbf{A_1} = \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{Z}'_i \mathbf{H} \mathbf{Z}_i \right)^{-1} \tag{11}$$

In the above, $\mathbf{H}$ is a $(T - 2, T - 2)$ matrix of 2 s on the main diagonal and -1 s on the adjacent upper and lower diagonals.

4 Synthetic instruments

Normally, it is difficult to find legitimate external exogenous instruments that correlate with the endogenous variables and which are yet unrelated to the error term. Typically, instruments that correlate closely with endogenous variables tend to also correlate with the errors, but trying to avoid correlation with the errors commonly leads to instruments that are weak and irrelevant with respect to the endogenous variable. A solution to the problem is based on the spatial filtering literature (notably Griffith 1988, 1996, 2000, 2003; Getis and Griffith 2002; Boots and Tiefelsdorf 2000; Patuelli et al. 2006) which is the basis for the construction of synthetic instruments, as advocated by Le Gallo and Páez (2013) for cross-sectional regression. Synthetic instruments have the properties of ideal instruments, because normally they are well correlated with the endogenous variables and yet independent of the errors. In fact, the synthetic instruments are the fitted values resulting from regressing the endogenous regressors and their spatial lags on weighted linear combinations of subsets of orthogonal eigenvectors deriving from a symmetrical N by N contiguity matrix[6] $\mathbf{M}_N$, in which the $m_{ij}, i = 1, ..., N, j = 1, ...N$, take the values

$$m_{ij} = 1 \text{ if } i \text{ and } j \text{ are neighbours}$$
$$m_{ij} = 0 \text{ otherwise}$$

$\mathbf{M}_N$ simply reflects the spatial connectivity of N regions and this is normally unaffected by the data under analysis. Likewise, the eigenvectors are exogenous, in other words not determined by $\mathbf{y}_t$, and so are an appropriate basis for synthetic instruments.

The effectiveness of the eigenvectors as instruments derives from the fact that each one represents a different orthogonal latent map pattern and so it is likely that one or more will correlate strongly with a non-randomly spatially distributed endogenous variable. Following Griffith (2000) and much related literature, we first consider the Moran Coefficient spatial autocorrelation index MC_t which measures the spatial autocorrelation in $\mathbf{y}$ at time t as given by

---

[6] However, the starting point could be other connectivity matrices, as illustrated by Patuelli et al. (2006). With the contiguity matrix, regions sharing a common boundary will be classed as neighbours, but other criteria could be applied such as being within a certain distance of each other, or having some other form of common attribute.

$$MC_t = \frac{N}{\mathbf{1}'_N \mathbf{M}_N \mathbf{1}_N} \frac{\mathbf{y}'_t \mathbf{P}_N \mathbf{y}_t}{\mathbf{y}'_t (\mathbf{I}_N - \mathbf{1}_N \mathbf{1}'_N / N) \mathbf{y}_t} \tag{12}$$

where

$$\mathbf{P}_N = (\mathbf{I}_N - \mathbf{1}_N \mathbf{1}'_N / N) \mathbf{M}_N (\mathbf{I}_N - \mathbf{1}_N \mathbf{1}'_N / N)$$

in which $\mathbf{I}_N$ is an N by N identity matrix, and $\mathbf{1}_N$ denotes an N by 1 vector of ones. The N by N matrix $\mathbf{P}_N$ yields N 'orthogonal' eigenvectors[7] $\mathbf{E}_i, i = 1, ..., N$, each of which is of dimension N by 1. Replacing $\mathbf{y}_t$ in Eq. (12) by $\mathbf{E}_i$ measures the spatial autocorrelation for eigenvector $\mathbf{E}_i$. So, each of the eigenvectors of $\mathbf{P}_N$ can be understood as a distinctive map pattern, with a separate MC, ranging from strongly positive to strongly negative autocorrelation, given by the different $\mathbf{E}_i$ values distributed across the regions implied by the connectivity matrix $\mathbf{M}_N$.

The synthetic instruments actually employed are weighted linear combinations of subsets of the $\mathbf{E}_i$ which are referred to in the literature as spatial filters (see Griffith 2003; Boots and Tiefelsdorf 2000). In the empirical analysis below, we use iterative regression to identify which subset of the $\mathbf{E}_i$ are appropriate for a given endogenous regressor, with the regression coefficient estimates giving the weights to apply in the weighted linear combination.

Applying such a spatial filter to even a completely spatially random variable will tend to find a significant relationship and a moderately strong correlation between the random variable and the synthetic instrument because it is the outcome of a search through many candidate $\mathbf{E}_i$ s, but with a spatially organised variable, for example, a quadratic trend surface defined by its geographic coordinates, the outcome will typically be a much more significant correlation.[8] Because spatio-temporal panel data are unlikely to be randomly distributed and almost invariably spatially organised in some way, the spatial filter can be used to obtain a synthetic instrument which is highly correlated with an endogenous variable, and thus, one has a way to generate a relevant instrumental variables that are unrelated to the error term in the data, and yet which are highly correlated with endogenous variables which are related to the error term. This is very helpful, because relevant and exogenous instrumental variables are difficult to find. As noted by Le Gallo and Páez (2013), working in the context of cross-sectional data, 'Synthetic variables, being artificial map patterns derived from the spatial configuration of the system, provide a near ideal solution—as long as spatial partitioning is not codetermined with other variables, which is typically the case'.

The aim is to obtain spatial filters for endogenous regressors $\tilde{x}_{itk}; i = 1, ..., N, t = 1, ..., T, k = 1, ..., k_2$. The approach adopted involves iteratively

---

[7] One of which is constant and correlated with the others.

[8] A simple experiment supports this. Generating 1000 different uniformly distributed random variables on a 20 by 20 lattice produces correlations between random variable and synthetic instruments in the range 0 to 0.5, with a mean correlation equal to 0.2883. Generating 1000 different quadratic surfaces, with parameters chosen on each replication from N(0,1) distributions, gives correlations ranging from 0.875 to 0.95 with a mean equal to 0.9282.

fitting regressions in which the dependent variable is the $k'th$ endogenous variable $\tilde{\mathbf{x}}_{tk}$ and the independent variables are the eigenvectors $\mathbf{E}_i$. The outcome is the isolation of the relevant subset of $\mathbf{E}_i$ and their relative weights as given by estimated regression coefficients. So then for each endogenous variable we can form a weighted linear combination of the subset so as to give an appropriate synthetic instrument. This can be summarised thus

(1) Set up empty vectors $\mathbf{z}_k$; $k = 1, ..., k_2$ to ultimately contain synthetic instrument for variable k

(2) Use data for panel at time $t = 1$

(3) $k = 1$

(4) Set N by 1 vector $\mathbf{V} = 1$

(5) $j = 1$

(6) For variable k and eigenvector j, regress $\tilde{\mathbf{x}}_{tk}$ on $\mathbf{E}_j$

(7) If the regression coefficient β is significantly different from $0, \mathbf{V} = [\mathbf{V} + \beta \mathbf{E}_j]$

(8) $j = j + 1$, if $j \leq N$ go to 6)

(9) Regress $\tilde{\mathbf{x}}_{tk}$ on $\mathbf{V}$ to obtain $\hat{\tilde{\mathbf{x}}}_{tk}$

(10) Append $\hat{\tilde{\mathbf{x}}}_{tk}$ to the column matrix so that $\mathbf{z}_k = \left[\mathbf{z}_k ; \hat{\tilde{\mathbf{x}}}_{tk} \right]$

(11) $k = k + 1$, if $k \leq k_2$ go to 4)

(12) $t = t + 1$, if $t \leq T$ go to 3)

The NT by 1 vectors $\mathbf{z}_k, k = 1, ..., k_2$ are then used as external synthetic instruments in GMM estimation.

5 The Monte Carlo simulations

5.1 The basic DGP

Our simulations assume that there are four regressors, though these are subsequently extended to give a spatial Durbin type of specification. At its simplest, our data generating process (DGP) is based on a version of Eqs. (2) and (3) thus

$$
\begin{aligned}
y_{it} &= \gamma y_{it-1} + \rho \sum_{j=1}^{N} w_{ij} y_{it} + \theta \sum_{j=1}^{N} w_{ij} y_{it-1} + \beta_1 x_{1it} + \beta_2 x_{2it} + ... \\
&\quad \beta_3 \tilde{x}_{3it} + \beta_4 \tilde{x}_{4it} + \varepsilon_{it}; \quad i = 1, ..., N, t = 1, ..., T \\
\varepsilon_{it} &= \mu_i + v_{it}
\end{aligned}
\tag{13}
$$

The aim is to devise a design that captures all sources of endogeneity, the ultimate outcome of which is $\tilde{x}_{3it}$ and $\tilde{x}_{4it}$ being correlated with ε_{it}. The approach adopted is similar to Liu and Saraiva (2015), but in the context of compound errors, so that endogeneity occurs because of correlation between the regressors and v and hence ε. In this simple initial case, the DGP draws from the Gaussian multivariate distribution,

$$
\begin{pmatrix} v \\ \tilde{v} \\ \mathbf{x}_1 \\ \mathbf{x}_2 \\ \tilde{\mathbf{x}}_3 \\ \tilde{\mathbf{x}}_4 \end{pmatrix} \sim N \left(\left(\begin{matrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{matrix} \right) \left[\begin{matrix} \sigma_v^2 & p_1 & 0 & 0 & 0 & 0 \\ p_1 & 1 & 0 & 0 & p_2 & p_3 \\ 0 & 0 & 1 & p_4 & 0 & 0 \\ 0 & 0 & p_4 & 1 & 0 & 0 \\ 0 & p_2 & 0 & 0 & 1 & 0 \\ 0 & p_3 & 0 & 0 & 0 & 1 \end{matrix} \right] \right) \tag{14}
$$

in which the leading diagonal of the covariance matrix contains the variances and the off-diagonals, p_1, p_2, p_3 and p_4, are the covariances between the random variables. The set-up in Eq. (14) indicates that the exogenous variables are unrelated to other variables, except that they are correlated with each other via p_4. The endogenous regressors $\tilde{\mathbf{x}}_3$ and $\tilde{\mathbf{x}}_4$ are correlated with $\tilde{v}$ which is a separate equation system, but $\tilde{v}$ is correlated with the remainder error component v (with $\tilde{v}$ separate from v, there is the option of different $\tilde{v}$ s for different endogenous regressors, as applied subsequently). The outcome is a set of NT by 1 random vectors. The individual error component μ_i is generated via a univariate normal distribution with zero mean and variance σ_μ^2. Combining the error components μ_i and v_{it} gives ε_{it}, so that $\tilde{x}_{3it}$ and $\tilde{x}_{4it}$ are correlated with ε_{it}. Note that we apply a similar approach in the context of non-spatial data and the spatial Durbin specification subsequently.

Given true values of the various parameters, and drawing in each replication from the multivariate normal distribution provides numerous realisations of $y_{it}, i = 1, ..., N; t = 1, ..., T$. Note that the draws leading to a maximum absolute characteristic root of $\tilde{\mathbf{A}}$ equal to or greater than 1 are rejected, so the simulation data sets are all dynamically stable and stationary. These data are the basis of estimates of the model parameters. The aim is to compare the resulting estimates with the true values of the parameters of Eq. (13).

In practice, for the purposes of simulation, various alternative true parameter values have been considered, but the results presented subsequently for the DGP are based on $\sigma_\mu^2 = 0.2, 0.8$ and $\sigma_v^2 = 0.8, 0.2$. The simulations thus encompass low and high individual heterogeneity, and low and high levels of remainder variance. Also, $p_1 = 0.5, p_2 = 0.75$, $p_3 = 0.25$ and $p_4 = 0.3$, so $\tilde{x}_{3it}$ is strongly endogenous, and $\tilde{x}_{4it}$ is weakly endogenous. Also, it is assumed that $\gamma = 0.75, \rho = 0.3, \theta = -0.2$ and $\beta_1 = 4, \beta_2 = 3, \beta_3 = 2, \beta_4 = 1$.

The reported outcomes for this simple specification are based on a 'r ahead and r behind' connectivity matrix of Kelejian and Prucha (1999), which is subsequently row-standardised. Assuming $r = 5$ means that each row of spatial matrix $\mathbf{W}_N$ (i.e. w_{ij}, with $i = 1, ..., N, j = 1, ..., N$) has up to 10 connections (five ahead and five behind each with equal weights), with zeros elsewhere and on the main diagonal. Additionally, we subsequently consider results based on a dense $\mathbf{W}_N$ matrix in the context of the spatial Durbin specification.

Results are reported for 100 replications, which nullifies aberrant outcomes and is sufficient to highlight the main traits in the simulation. In each replication, initial 51 simulation outcomes of $x_{1it}, x_{2it}, \tilde{x}_{3it}$ and $\tilde{x}_{4it}$ are ignored in order to minimise any potential effect of initial values at $t = -50$ of zero (i.e. simulation outcomes for $t = -50, -49, ..., 0$ are discarded). Also, $T = 10$ and there are $N = 100$ regions.

	Number of instruments			
Table 1 Mean parameter estimates: $\sigma_\mu^2 = 0.8\,\sigma_\nu^2 = 0.2$	True value	292	152	40	19
	$\gamma = 0.75$	0.7496	0.7497	0.7506	0.7507
	$\rho = 0.3$	0.3051	0.3039	0.301	0.3011
	$\beta_1 = 4$	3.9976	3.9965	3.9958	4.0006
	$\beta_2 = 3$	2.9922	2.9945	2.9972	2.9963
	$\beta_3 = 2$	2.0408	2.0528	2.051	2.055
	$\beta_4 = 1$	1.0119	1.0117	1.0097	1.0084
	$\theta = -0.6$	-0.6036	-0.6033	-0.6025	-0.6067

5.2 Results for simple spatial DGP

The results obtained depend of various set-ups regarding instruments. The idea is that we wish to show the impact of having fewer instruments on estimation as well as the effect of different error distribution assumptions. The largest number of instruments is given by applying the standard solution to the existence of endogenous variables, namely HENR instrumentation. Temporally lagged values of the endogenous variables y_{it} and Wy_{it} together with temporally lagged values of $Wy_{it-1}, \tilde{x}_{3it}, \tilde{x}_{4it}$ and their spatial lags $W^2 y_{it-1}, W\tilde{x}_{3it}, W\tilde{x}_{4it}$, with full HENR instrumentation amount to $8(T-1)(T-2)/2 = 288$ instruments. In addition, there are four exogenous IV-style instruments equal to the exogenous variables and their spatial lags, namely $x_{1it}, x_{2it}, Wx_{1it}, Wx_{2it}$. The result is $288 + 4 = 292$ instruments overall.

One side-issue relating to the existence of many instruments is the possibility that some are almost collinear. One could drop them from the instrument set, but this would change the number of degrees of freedom available for the Sargan–Hansen J statistics. Moreover, the definition of collinearity is somewhat subjective[9] and eliminating collinear instruments has minimal impact on outcomes.

Synthetic instruments are usually strongly correlated. For example, on the basis of 100 replications the mean correlation between $\tilde{x}_{4it}$ and its spatial lag $W\tilde{x}_{4it}$ and their respective synthetic instruments is 0.6687 and 0.7205. Applying them to the endogenous regressors and their spatial lags, $\tilde{x}_{3it}, W\tilde{x}_{3it}$ and $\tilde{x}_{4it}, W\tilde{x}_{4it}$, together with the exogenous regressors and their spatial lags gives eight IV-style instruments. Full HENR instrumentation for $y_{it}, Wy_{it}, Wy_{it-1}, W^2 y_{it-1}$, adds $4(T-1)(T-2)/2 = 144$ instruments, giving a total of 152 instruments overall. There is considerable reduction in the number of instruments resulting from both collapsing and using synthetic instruments. Collapsing the standard HENR instrumentation for $y_{it}, Wy_{it}, Wy_{it-1}$

---

[9] Belsley collinearity diagnostics (Belsley et al. 1980) suggest a condition index with a default tolerance greater than 30, combined with variance decomposition proportions exceeding a default tolerance of 0.5, but other (arbitrary) tolerances could be applied. Applying the default tolerances to the instrument matrix resulting from a single Monte Carlo simulation identifies a handful of collinear instruments (note that these do not have a correlation coefficient equal to 1.0, for example, a correlation of 0.8555 is deemed sufficient for collinearity).

Table 2 Mean parameter estimates: $\sigma_\mu^2 = 0.2\,\sigma_\nu^2 = 0.8$

	Number of instruments			
True value	292	152	40	19
$\gamma = 0.75$	0.747	0.7485	0.7494	0.7486
$\rho = 0.3$	0.312	0.3108	0.3038	0.3015
$\beta_1 = 4$	3.9994	4.002	4.0063	4.0072
$\beta_2 = 3$	2.9941	2.995	2.9963	3.0004
$\beta_3 = 2$	1.9895	1.9907	1.9894	1.9976
$\beta_4 = 1$	1.0011	1.0109	1.0143	1.022
$\theta = -0.6$	-0.6043	-0.6057	-0.6051	-0.6083

Table 3 Mean parameter bias and root mean squared error (RMSE) for selected data generated processes (DGP)

	DGP	Number of instruments			
		292	152	40	19
Mean abs param. bias	$\sigma_\mu^2 = 0.8\,\sigma_\nu^2 = 0.2$	0.0103	0.0116	0.0103	0.0109
	$\sigma_\mu^2 = 0.2\,\sigma_\nu^2 = 0.8$	0.0053	0.0065	0.0063	0.0062
Mean RMSE	$\sigma_\mu^2 = 0.8\,\sigma_\nu^2 = 0.2$	0.0285	0.0271	0.0302	0.0307
	$\sigma_\mu^2 = 0.2\,\sigma_\nu^2 = 0.8$	0.0438	0.0391	0.0429	0.0442

and $W^2 y_{it-1}$ gives $4(T-2)=32$ instruments. Adding the aforementioned eight IV-style instruments gives an overall total of $4(T-2)+8=40$ instruments. Collapsing standard HENR instrumentation for y_{it} alone gives $(T-2)=8$ instruments. Using synthetic instruments for Wy_{it}, Wy_{it-1} and $W^2 y_{it-1}$, together with $\tilde{x}_{3it}, W\tilde{x}_{3it}, \tilde{x}_{4it}, W\tilde{x}_{4it}$ creates seven IV-style instruments, and there are four more from the exogenous variables and their spatial lags $x_{1it}, x_{2it}, Wx_{1it}, Wx_{2it}$. Combined, there are 19 instruments in total.

The results obtained show remarkably little negative consequence as a result of reducing the number of instruments, plus the significant benefit of an unbiased Sargan–Hansen J statistic and less biased standard errors. Tables 1 and 2 show the mean parameter estimates resulting from 100 Monte Carlo simulations according to the number of instruments, true parameter values, and error distribution assumptions.[10] It is clear that the estimates closely approximate the true values. Table 3 summarises the outcomes, showing the mean absolute parameter bias, obtained by averaging across the mean absolute bias for each parameter, and the average of the mean RMSE's, again averaging across the mean RMSE of each parameter. Bias tends to be smaller with larger σ_ν^2, but the opposite is true for RMSE. There is little variation in either as the number of instruments

---

[10] To save space, we do not consider direct, indirect and total short- and long-run effects that take account of spillovers.

Table 4 Mean standard errors: $\sigma_{\mu}^2 = 0.8\ \sigma_{\nu}^2 = 0.2$

Param.	Number of instruments			
	292	152	40	19
γ	0.0047	0.0042	0.0057	0.0061
ρ	0.0115	0.0086	0.0104	0.0111
β_1	0.0299	0.0243	0.0277	0.0286
β_2	0.0300	0.0238	0.0271	0.0284
β_3	0.0563	0.0322	0.0391	0.0414
β_4	0.0560	0.0324	0.0392	0.0416
θ	0.0106	0.0087	0.0115	0.0157
Classic two-step	0.0170	0.0082	0.0188	0.0230
Windmeijer	0.0362	0.0249	0.0261	0.0261
HKL	0.0321	0.0244	0.0239	0.0250

Table 5 Mean standard errors: $\sigma_{\mu}^2 = 0.2\ \sigma_{\nu}^2 = 0.8$

Param.	Number of instruments			
	292	152	40	19
γ	0.0074	0.0065	0.0089	0.0094
ρ	0.0187	0.0141	0.0167	0.0185
β_1	0.0495	0.0395	0.0457	0.0475
β_2	0.0499	0.0400	0.0457	0.0473
β_3	0.0915	0.0523	0.0637	0.0683
β_4	0.0937	0.0537	0.0653	0.0694
θ	0.0175	0.0140	0.0182	0.0258
Classic two-step	0.0278	0.0135	0.0309	0.0378
Windmeijer	0.0596	0.0408	0.0431	0.0435
HKL	0.0532	0.0400	0.0393	0.0413

varies. In Tables 4 and 5, for each table cell relating to each parameter, we give the average of the mean simulation outcomes averaging across classic two-step, Windmeijer and HKL standard errors. Also, the mean classic two-step, Windmeijer and HKL standard errors are obtained by taking the average of the mean standard errors, averaging across parameter standard errors. The tables indicate that standard errors tend to rise as the number of instruments falls, though this is confounded by the fact that with 152 and 292 instruments the weight matrix is not symmetric positive definite and a generalised inverse is applied, so the standard errors cannot be guaranteed to be accurate. Compared with the uncorrected classic two-step standard errors, the corrections due to Windmeijer and the HKL correction are associated with larger standard errors. Tables 4 and 5

Table 6 Mean value of diagnostics:$\sigma_\mu^2 = 0.8\sigma_v^2 = 0.2$

	Statistic	Number of instruments			
		292	152	40	19
First order	m_1 (z ratio)	−7.2713	−7.2372	−7.2344	−7.2327
Second order	m_2 (z ratio)	−0.0845	−0.1206	−0.1629	−0.1501
Dynamic stability, stationarity	Max abs e'value of $\tilde{A}$	0.8665	0.9151	0.8682	0.8696
Instrument validity	Sargan–Hansen J statistic	94.3761	93.3202	31.8091	11.4474
	p-value	1	0.9996	0.5352	0.5330
	d. of freedom	285	145	33	12

Table 7 Mean value of diagnostics:$\sigma_\mu^2 = 0.2\sigma_v^2 = 0.8$

	Statistic	Number of instruments			
		292	152	40	19
First order	m_1(z ratio)	−7.2012	−7.1616	−7.1501	−7.1285
Second order	m_2(z ratio)	−0.0150	−0.0088	−0.0327	−0.0297
Dynamic stability, stationarity	Max abs e'value of $\tilde{A}$	0.8626	0.8646	0.8672	0.8681
Instrument validity	Sargan–Hansen J statistic	94.0651	93.4647	32.0978	12.3537
	p-value	1	0.9996	0.5245	0.4618
	d. of freedom	285	145	33	12

also indicate that standard errors are higher when remainder variance is high and individual variance is low.

Tables 6 and 7 summarise diagnostic indicators in terms of means taken across the 100 Monte Carlo simulations, again broken down by the number of instruments and error distribution assumptions. A prominent feature is the downward bias in the Sargan–Hansen J statistic for overidentifying restrictions with a large number of instruments. The J test statistics reflect the problem of low power associated with too many moments conditions. This leads to an implausibly high mean p-value of 1.000 (Anderson and Sørenson 1996, Bowsher 2002, Roodman 2009b). Reducing from 292 to 152 instruments is insufficient to reduce the mean J test statistic much below 1.0. While this is an unreliable indicator of instrument validity, the z-ratios for error serial correlation point to significant negative first order correlation (m_1) and effectively zero second-order serial correlation (m_2) in the first differenced residuals (see Arellano and Bond 1991) when referred to the N(0,1) distribution. This points clearly to consistent estimates. The tables of diagnostics also show that the estimates obtained are dynamically stable, as measured by the maximum absolute eigenvalue of $\tilde{A}$.

Table 8 Mean parameter estimates: $\sigma_\mu^2 = 0.8\sigma_\nu^2 = 0.2$

True value	Number of instruments			
	110	40	26	12
$\gamma = 0.2$	0.1967	0.1981	0.2	0.1999
$\beta_1 = 1$	0.8703	0.8669	0.8766	0.8683
$\beta_2 = 0.5$	0.5449	0.5452	0.5419	0.5439
$\beta_3 = 0.75$	0.848	0.8499	0.8322	0.8529
$\beta_4 = 1$	1.0891	1.0866	1.0915	1.0898

Table 9 Mean parameter estimates: $\sigma_\mu^2 = 0.2\sigma_\nu^2 = 0.8$

True value	Number of instruments			
	110	40	26	12
$\gamma = 0.2$	0.1962	0.2001	0.2009	0.2025
$\beta_1 = 1$	0.9585	0.9462	0.9656	0.9616
$\beta_2 = 0.5$	0.5069	0.5118	0.5058	0.5071
$\beta_3 = 0.75$	0.7854	0.7827	0.7693	0.7753
$\beta_4 = 1$	1.0338	1.0357	1.0404	1.0322

Table 10 Mean parameter bias and root mean squared error (RMSE) for selected data generated processes (DGP)

	DGP	number of instruments			
		110	40	26	12
Mean abs param. bias	$\sigma_\mu^2 = 0.8\ \sigma_\nu^2 = 0.2$	0.0593	0.0733	0.0551	0.0737
	$\sigma_\mu^2 = 0.2\ \sigma_\nu^2 = 0.8$	0.0103	0.0268	0.0202	0.0211
Mean RMSE	$\sigma_\mu^2 = 0.8\ \sigma_\nu^2 = 0.2$	0.0873	0.0856	0.1229	0.0859
	$\sigma_\mu^2 = 0.2\ \sigma_\nu^2 = 0.8$	0.0285	0.0728	0.1672	0.0721

5.3 Results with non-spatial data

Thus far, we have reduced instrument proliferation using collapsing combined with synthetic instruments derived from the N by N contiguity matrix $\mathbf{M}_N$. In this section, we explore the efficacy of this approach in the absence of spatial effects. Consider therefore the DGP as in Eq. (13) but with $\rho = 0, \theta = 0$. There are four estimators under consideration. First, full HENR instruments based on endogenous variables y_{it}, $\tilde{x}_{3it}$ and $\tilde{x}_{4it}$, with IV-style instruments for the exogenous x_{1it} and x_{2it} gives $3(T-2)(T-1)/2+2=110$ instruments. Secondly, full HENR instruments based on y_{it}, with synthetic instruments for $\tilde{x}_{3it}$ and $\tilde{x}_{4it}$ which become IV-style instruments alongside x_{1it} and x_{2it} gives $(T-2)(T-1)/2+4=40$ instruments. Thirdly, collapsing the HENR instruments based on y_{it} plus the 4 above-mentioned IV-style instruments

Table 11 Mean standard errors: $\sigma_\mu^2 = 0.8\sigma_\nu^2 = 0.2$

Parm.	Number of instruments			
	110	40	26	12
γ	0.0086	0.0102	0.0122	0.0113
β_1	0.0470	0.0439	0.1237	0.0480
β_2	0.0251	0.0278	0.0467	0.0296
β_3	0.0531	0.0442	0.1490	0.0481
β_4	0.0585	0.0489	0.1705	0.0541
Classic two-step	0.0127	0.0277	0.0750	0.0364
Windmeijer	0.0462	0.0400	0.0972	0.0395
HKL	0.0565	0.0372	0.1290	0.0387

Table 12 Mean standard errors: $\sigma_\mu^2 = 0.2 \ \sigma_\nu^2 = 0.8$

Parm.	Number of instruments			
	110	40	26	12
γ	0.0047	0.0157	0.0192	0.0176
β_1	0.0115	0.0758	0.2059	0.0839
β_2	0.0299	0.0478	0.0777	0.0507
β_3	0.0300	0.0745	0.2595	0.0831
β_4	0.0563	0.0817	0.2774	0.0922
Classic two-step	0.0170	0.0470	0.1295	0.0626
Windmeijer	0.0362	0.0674	0.1607	0.0677
HKL	0.0321	0.0628	0.2136	0.0662

Table 13 Mean value of diagnostics: $\sigma_\mu^2 = 0.8\sigma_\nu^2 = 0.2$

	Statistic	Number of instruments			
		110	40	26	12
First order	m_1 (z ratio)	-7.1597	-7.1555	-6.6186	-7.1350
Second order	m_2 (z ratio)	-0.0872	-0.1406	-0.0521	-0.1160
Instrument validity	Sargan–Hansen J statistic	95.2712	36.0049	20.0380	6.7002
	p-value	0.7372	0.4456	0.5337	0.5024
	d. of freedom	105	35	21	7

gives $(T-2)+4=12$ instruments. Fourthly, collapsing the HENR instruments based on $y_{it} \ \tilde{x}_{3it}$ and $\tilde{x}_{4it}$ gives $3(T-2)+2=26$ instruments.

Data are generated by again randomly sampling from a multivariate normal distribution as in Eq. (15). In this case correlation between $\mathbf{x}_1$ and

Table 14 Mean value of diagnostics:$\sigma_\mu^2 = 0.2\sigma_v^2 = 0.8$

	Statistic	Number of instruments			
		110	40	26	12
First order	m_1 (z ratio)	-7.1827	-7.1956	-6.5890	-7.1738
Second order	m_2 (z ratio)	-0.0720	-0.0700	-0.0339	-0.0558
Instrument validity	Sargan–Hansen J statistic	95.6816	34.8609	19.8630	6.8066
	p-value	0.7280	0.4923	0.5440	0.5071
	d. of freedom	105	35	21	7

endogenous regressors $\tilde{\mathbf{x}}_3$ and $\tilde{\mathbf{x}}_4$ is introduced via $p_5 = 0.4$ and $p_6 = 0.5$. Also $p_1 = 0.5, p_2 = 0.75, p_3 = 0.25$ and $p_4 = 0.3$.

$$
\begin{pmatrix} v \\ \tilde{v} \\ \mathbf{x}_1 \\ \mathbf{x}_2 \\ \tilde{\mathbf{x}}_3 \\ \tilde{\mathbf{x}}_4 \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \begin{bmatrix} \sigma_v^2 & p_1 & 0 & 0 & 0 & 0 \\ p_1 & 1 & 0 & 0 & p_2 & p_3 \\ 0 & 0 & 1 & p_4 & p_5 & p_6 \\ 0 & 0 & p_4 & 1 & 0 & 0 \\ 0 & p_2 & p_5 & 0 & 1 & 0 \\ 0 & p_3 & p_6 & 0 & 0 & 1 \end{bmatrix} \right)
\tag{15}
$$

Tables 8, 9, 10, 11, 12, 13 and 14 summarise the outcomes over a range of simulation set-ups. Broadly stated, there is minimal impact resulting from reducing the number of instruments, apart from the usual effects of eliminating bias in the Sargan Hansen J test static and in the parameter standard errors. However, the fact that estimates of β_1 are persistently downwardly biased, and β_2, β_3 and β_4 upwardly biased is worthy of mention. Also, Table 10 indicates that bias tends to be larger with greater individual heterogeneity, but no obvious change in bias and RMSE occurs with varying numbers of instruments. In contrast, as shown in Tables 11 and 12, there is a clear increase in parameter standard errors as the number of instruments diminishes. Also, comparing Tables 11 and 12, standard errors tend to be larger as remainder error variance increases and individual heterogeneity reduces. Again, the corrected standard errors are invariably larger than the classic two-step standard errors. Tables 13 and 14 highlight the usual downward bias in the Sargan–Hansen J statistic, reflecting the rule-of-thumb that the J statistic should be more reliable when the number of instruments is less than the number of individuals ($N = 100$).

5.4 Results for Spatial Durbin specification with dense connectivity matrix

The data generating process for spatial data implemented a row-normalised 'five ahead and five behind' matrix $\mathbf{W}_N$ for the spatial lag of the dependent variable and for instruments. In this section, we replace this with a Lehmer matrix, which is a symmetric and positive definite matrix $\mathbf{W}_N^*(i.e.\ w_{ij}^*$, with $i = 1, ..., N, j = 1, ..., N)$ in which

Fig. 1 Based on Lehmer matrix
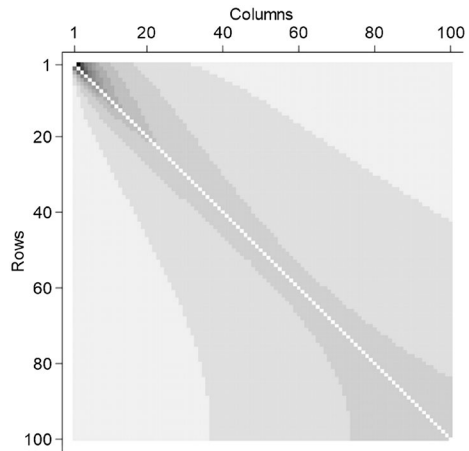
$$w_{ij}^* = i/j, j \geq i$$
$$w_{ij}^* = j/i, j < i$$
(16)

$\mathbf{W}_N^*$ is a dense matrix with cell values diminishing away from the main diagonal. $\mathbf{W}_N$ is $\mathbf{W}_N^*$ with a main diagonal by zeros, then row normalised, as illustrated in Fig. 1. As with the other $\mathbf{W}_N$ the row and column sums are uniformly bounded.

To create added diversity, an attempt is made to approximate the reality of many analytical situations in which all variables, denoted by $\tilde{x}_{1it}, \tilde{x}_{2it}, \tilde{x}_{3it}$ and $\tilde{x}_{4it}$, are endogenous. Also, we include spatial lags of explanatory variables to give a spatial Durbin specification (see Halleck Vega and Elhorst 2015, for discussion). Inclusion of spatial lags $W\tilde{x}_{1it}, W\tilde{x}_{2it}, W\tilde{x}_{3it}, W\tilde{x}_{4it}$ with corresponding parameters $\beta_5, \beta_6, \beta_7$ and β_8, gives

$$y_{it} = \gamma y_{it-1} + \rho \sum_{j=1}^{N} w_{ij} y_{it} + \theta \sum_{j=1}^{N} w_{ij} y_{it-1} + \beta_1 \tilde{x}_{1it} + \beta_2 \tilde{x}_{2it} + \beta_3 \tilde{x}_{3it} + \beta_4 \tilde{x}_{4it} ...$$

$$+\beta_5 \sum_{j=1}^{N} w_{ij} \tilde{x}_{1it} + \beta_6 \sum_{j=1}^{N} w_{ij} \tilde{x}_{2it} + \beta_7 \sum_{j=1}^{N} w_{ij} \tilde{x}_{3it} + \beta_8 \sum_{j=1}^{N} w_{ij} \tilde{x}_{4it} + \varepsilon_{it}; i = 1, ..., N, t = 1, ..., T$$

$$\varepsilon_{it} = \mu_i + v_{it}$$
(17)

Again the variables are generated from a multivariate Gaussian distribution, in this case allowing each endogenous variable to depend on a separate error process, as might occur in reality with diverse sources of endogeneity. Thus, $\tilde{\mathbf{x}}_1$ correlates with $\tilde{v}_1$ via p_7 in Eq. (18), p_8 is the $\tilde{\mathbf{x}}_2, \tilde{v}_2$ correlation, p_2 is the $\tilde{\mathbf{x}}_3, \tilde{v}_3$ correlation and p_3 is the correlation between $\tilde{\mathbf{x}}_4$ and $\tilde{v}_4$. The correlation between the $\tilde{v}$ s and v is p_1, and the correlations between the regressors are p_4, p_5 and p_6.

Table 15 Mean parameter estimates: spatial Durbin and simple specifications

True value	$\sigma_\mu^2 = 0.8, \sigma_\nu^2 = 0.2, p_1 = 0.5$ etc			$\sigma_\mu^2 = 0.8, \sigma_\nu^2 = 0.2, p_1 = 0.75$ etc $\sigma_\nu^2 = 0.2, p_1 = 0.5$ etc		
	Number of instruments			Number of instruments		
	152(148)	82 (78)	26(22)	152(148)	82(78) (78) (78) (78)	26(22)
$\gamma = 0.3$	0.2999	0.3004	0.3016	0.2987	0.2992	0.2994
	0.29900	0.2990	0.2993	0.2996	0.2998	0.3011
$\rho = 0.2$	0.2518	0.2649	0.2584	0.2859	0.2917	0.266
	0.2073	0.2078	0.2064	0.2258	0.2223	0.2257
$\beta_1 = 4$	4.064	4.0643	4.0602	4.3972	4.3816	4.3991
	4.059	4.0668	4.048	4.3896	4.3931	4.3994
$\beta_2 = 3$	2.9766	2.971	2.9662	3.2347	3.2279	3.241
	2.9789	2.9767	2.9856	3.2246	3.2212	3.2337
$\beta_3 = 2$	2.0186	2.0211	2.0167	2.1543	2.1505	2.157
	2.035	2.035	2.0409	2.1637	2.1639	2.1688
$\beta_4 = 1$	0.9682	0.9729	0.9754	0.7222	0.7361	0.7161
	0.9746	0.9689	0.9829	0.7309	0.727	0.7352
$\beta_5 = 1$	0.6264	0.5917	0.6783	0.44	0.3782	0.6678
$\beta_6 = 2$	1.749	1.6404	1.6552	1.463	1.3879	1.6064
$\beta_7 = 3$	2.7628	2.6768	2.7227	2.5413	2.4926	2.6283
$\beta_8 = 4$	3.7552	3.6327	3.6477	3.6171	3.5805	3.611
$\theta = -0.4$	-0.3992	-0.3982	-0.3974	-0.3871	-0.387	-0.3898
	-0.4018	-0.4033	-0.4043	-0.4000	-0.4006	-0.3959

n.b. italics denote estimates from simple specification

$$
\begin{pmatrix} \nu \\ \tilde{\nu}_1 \\ \tilde{\nu}_2 \\ \tilde{\nu}_3 \\ \tilde{\nu}_4 \\ \tilde{\mathbf{x}}_1 \\ \tilde{\mathbf{x}}_2 \\ \tilde{\mathbf{x}}_3 \\ \tilde{\mathbf{x}}_4 \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \sigma_\nu^2 & p_1 & p_1 & p_1 & p_1 & 0 & 0 & 0 & 0 \\ p_1 & 1 & 0 & 0 & 0 & p_7 & 0 & 0 & 0 \\ p_1 & 0 & 1 & 0 & 0 & 0 & p_8 & 0 & 0 \\ p_1 & 0 & 0 & 1 & 0 & 0 & 0 & p_2 & 0 \\ p_1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & p_3 \\ 0 & p_7 & 0 & 0 & 0 & 1 & p_4 & p_5 & p_6 \\ 0 & 0 & p_8 & 0 & 0 & p_4 & 1 & 0 & 0 \\ 0 & 0 & 0 & p_2 & 0 & p_5 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & p_3 & p_6 & 0 & 0 & 1 \end{bmatrix} \right)
\tag{18}
$$

Variables entering Eq. (17) are generated on the basis of $p_1 = 0.5, p_2 = 0.75, p_3 = 0.25$, $p_4 = 0.3$, $p_5 = 0.4$, $p_6 = 0.5, p_7 = 0.8$ and $p_8 = 0.2$. Additionally, data are generated on the basis of $p_1 = 0.75, p_2 = 0.75, p_3 = 0.25$, $p_4 = -0.5$, $p_5 = -0.2$, $p_6 = 0.75$, $p_7 = 0.8$ and $p_8 = 0.3$. In this case, the closer correlation between ν and the $\tilde{\nu}$ s suggests increased endogeneity.

The true parameter values utilised in Eq. (17) are $\gamma = 0.3$, $\rho = 0.2$, $\theta = -0.4$ and $\beta_1 = 4, \beta_2 = 3, \beta_3 = 2, \beta_4 = 1, \beta_5 = 1, \beta_6 = 2, \beta_7 = 3$ and $\beta_8 = 4$. Consider first an estimator which applies synthetic instruments, based on $\mathbf{M}_N$, to the endogenous variables $\tilde{x}_{1it}, \tilde{x}_{2it}, \tilde{x}_{3it}, \tilde{x}_{4it}$ their spatial lags $W\tilde{x}_{1it}, W\tilde{x}_{2it}, W\tilde{x}_{3it}, W\tilde{x}_{4it}$. Combining

Table 16 Mean parameter estimates: spatial Durbin and simple specifications

True value	$\sigma_\mu^2 = 0.2, \sigma_v^2 = 0.8, p_1 = 0.5$ etc			$\sigma_\mu^2 = 0.2, \sigma_v^2 = 0.8, p_1 = 0.75$ etc $\sigma_v^2 = 0.2, p_1 = 0.5$ etc		
	Number of instruments			Number of instruments		
	152(148)	82 (78)	26(22)	152(148)	82 (78)	26(22)
$\gamma = 0.3$	0.2999	0.3011	0.3019	0.2977	0.298	0.2986
	0.2987	*0.2988*	*0.2995*	*0.2989*	*0.2993*	*0.3004*
$\rho = 0.2$	0.2875	0.2959	0.2839	0.3337	0.3533	0.3209
	0.2094	*0.2113*	*0.2096*	*0.234*	*0.2349*	*0.2351*
$\beta_1 = 4$	4.0391	4.0464	4.0506	4.3781	4.3828	4.3874
	4.0566	*4.0636*	*4.0629*	*4.3664*	*4.3845*	*4.3898*
$\beta_2 = 3$	2.9811	2.9867	2.983	3.2161	3.219	3.2151
	2.9735	*2.9647*	*2.9657*	*3.2036*	*3.2147*	*3.2115*
$\beta_3 = 2$	1.9893	1.9955	1.993	2.1294	2.1344	2.137
	2.0006	*1.9974*	*2.0034*	*2.1282*	*2.133*	*2.14*
$\beta_4 = 1$	0.9736	0.9682	0.9666	0.713	0.7152	0.7015
	0.9646	*0.9636*	*0.9577*	*0.7341*	*0.7225*	*0.7173*
$\beta_5 = 1$	0.4515	0.3572	0.4135	0.2404	0.104	0.3
$\beta_6 = 2$	1.5375	1.5065	1.5926	1.2748	1.1367	1.2407
$\beta_7 = 3$	2.5685	2.5061	2.5962	2.2725	2.1484	2.3076
$\beta_8 = 4$	3.6557	3.489	3.4948	3.3835	3.2878	3.4922
$\theta = -0.4$	-0.3914	-0.3958	-0.3962	-0.3844	-0.3819	-0.3878
	-0.3965	*-0.3973*	*-0.397*	*-0.4032*	*-0.401*	*-0.3984*

n.b. italics denote estimates from simple specification

these eight IV-style instruments with HENR instruments from $y_{it}, Wy_{it}, Wy_{it-1}$ and $W^2 y_{it-1}$, gives s $4(T-1)(T-2)/2 + 8 = 152$ instrumental variables. Secondly, applying synthetic instruments to all variables except y_{it} and its spatial lag Wy_{it}, results in $(T-1)(T-2) + 10 = 82$ instruments. Collapsing these latter two HENR sets of instruments gives $2(T-2) + 10 = 26$ instruments. A simple comparator assumes $\beta_5 = 0, \beta_6 = 0, \beta_7 = 0$ and $\beta_8 = 0$, so that the spatial lags and their respective instruments are eliminated, leaving 148, 78 and 22 instruments, respectively.

Tables 15, 16, 17, 18, 19, 20, 21, 22 and 23 give the mean outcomes, broken down by number of instruments, the values assigned to σ_μ^2 and σ_v^2 and the values of $p_1, ..., p_8$ denoted either by $p_1 = 0.5$ etc or $p_1 = 0.75$ etc. The italicised numbers are the outcomes for a simple comparator. The salient aspects are the usual bias with many instruments in the Sargan–Hansen J statistic and parameter standard errors. For the spatial Durbin specification, some parameter estimation bias is apparent in Tables 15 and 16; notably, the ρ and θ estimates are persistently upwardly biased regardless of the number of instruments or error assumptions, and likewise the spatial lag parameters $\beta_5, \beta_6, \beta_7$ and β_8 are negatively biased.[11]

---

[11] Again, we do not consider direct, indirect and total short- and long-run effects.

Table 17 Mean parameter bias and root mean squared error (RMSE) for selected data generated processes (DGP): spatial Durbin and simple specifications

	DGP $p_1 = 0.5$ etc	Number of instruments		
		152 (148)	82 (78)	26 (22)
Mean abs param. bias	$\sigma_\mu^2 = 0.8 \sigma_v^2 = 0.2$	0.1179	0.1515	0.1358
		0.0215	*0.0240*	*0.0188*
	$\sigma_\mu^2 = 0.2 \sigma_v^2 = 0.8$	0.1798	0.2126	0.1910
		0.0190	*0.0219*	*0.0223*
Mean RMSE	$\sigma_\mu^2 = 0.8 \sigma_v^2 = 0.2$	0.2726	0.3170	0.3553
		0.0491	*0.05573*	*0.0620*
	$\sigma_\mu^2 = 0.2\ \sigma_v^2 = 0.8$	0.3404	0.4152	0.4691
		0.0577	*0.0635*	*0.0744*
	DGP $p_1 = 0.75$ etc	152 (148)	82 (78)	26 (22)
Mean abs param. bias	$\sigma_\mu^2 = 0.8 \sigma_v^2 = 0.2$	0.2821	0.2991	0.2404
		0.1533	*0.1535*	*0.1568*
	$\sigma_\mu^2 = 0.2 \sigma_v^2 = 0.8$	0.3628	0.4107	0.3484
		0.1432	*0.1495*	*0.1516*
Mean RMSE	$\sigma_\mu^2 = 0.8 \sigma_v^2 = 0.2$	0.4312	0.4608	0.5154
		0.1721	*0.1755*	*0.1834*
	$\sigma_\mu^2 = 0.2\ \sigma_v^2 = 0.8$	0.5291	0.5745	0.6179
		0.1694	*0.1814*	*0.1905*
		0.1694		

n.b. italics denote estimates from simple specification

Table 18 Mean standard errors: spatial Durbin specification

True value	$\sigma_\mu^2 = 0.8, \sigma_v^2 = 0.2, p_1 = 0.5$ etc			$\sigma_\mu^2 = 0.8, \sigma_v^2 = 0.2, p_1 = 0.75$ etc $\sigma_\mu^2 = 0.2, p_1 = 0.5$ etc $\sigma_v^2 = 0.2, p_1 = 0.5$ etc		
	Number of instruments			Number of instruments		
	152	82	26	152	82	26
γ	0.0036	0.0038	0.0049	0.0057	0.0062	0.0079
ρ	0.0568	0.0658	0.0967	0.0745	0.0851	0.1322
β_1	0.0577	0.0706	0.0976	0.0825	0.1004	0.1473
β_2	0.0424	0.0549	0.0724	0.0586	0.0699	0.0981
β_3	0.0453	0.0521	0.0699	0.0462	0.0540	0.0712
β_4	0.0505	0.0584	0.0794	0.0743	0.0897	0.1278
β_5	0.4349	0.5095	0.7755	0.6180	0.7042	1.0777
β_6	0.3883	0.4597	0.6835	0.5288	0.6138	0.9384
β_7	0.3946	0.4513	0.6777	0.4845	0.5565	0.8609
β_8	0.3865	0.4559	0.6915	0.5161	0.5871	0.9099
θ	0.0170	0.0184	0.0251	0.0234	0.0260	0.0350
Classic two-step	0.0463	0.0915	0.2644	0.0603	0.1193	0.3535
Windmeijer	0.2322	0.2592	0.3227	0.3090	0.3393	0.4360
HKL	0.2336	0.2494	0.3058	0.3160	0.3304	0.4122

Table 19 Mean standard errors: spatial Durbin specification

True value	$\sigma_\mu^2 = 0.2, \sigma_v^2 = 0.8,\ p_1 = 0.5$ etc			$\sigma_\mu^2 = 0.2, \sigma_v^2 = 0.8, p_1 = 0.75$ etc $\sigma_v^2 = 0.2, p_1 = 0.5$ etc $\sigma_v^2 = 0.2, p_1 = 0.5$ etc		
	Number of instruments			Number of instruments		
	152	82	26	152	82	26
γ	0.0045	0.0049	0.0062	0.0068	0.0075	0.0094
ρ	0.0737	0.0828	0.1273	0.0878	0.1033	0.1556
β_1	0.0737	0.0923	0.1247	0.1019	0.1301	0.1806
β_2	0.0555	0.0708	0.0910	0.0719	0.0882	0.1208
β_3	0.0587	0.0706	0.0915	0.0563	0.0665	0.0838
β_4	0.0654	0.0750	0.0992	0.0910	0.1150	0.1585
β_5	0.5729	0.6397	0.9592	0.7442	0.8803	1.3031
β_6	0.4935	0.5926	0.9088	0.6301	0.7446	1.0907
β_7	0.5116	0.5787	0.8672	0.5728	0.6733	1.0469
β_8	0.5039	0.5607	0.8676	0.6055	0.7115	1.0681
θ	0.0212	0.0241	0.0328	0.0292	0.0331	0.0445
Classic two-step	0.0585	0.1145	0.3354	0.0720	0.1441	0.4236
Windmeijer	0.3011	0.3286	0.4137	0.3704	0.4176	0.5180
HKL	0.3044	0.3184	0.3896	0.3751	0.4074	0.4935

In contrast, the bias for the comparator is negligible, and since everything else is the same, it appears that the primary cause of the larger bias is the presence of spatial lags. For the spatial Durbin specification, Table 17 shows that on the whole bias tends to increase as the number of instruments decreases. It is also larger as endogeneity increases, with p_1 going from 0.5 to 0.75. Additionally, it increases as error variance goes from $\sigma_v^2 = 0.2$ to $\sigma_v^2 = 0.8$. Mean RMSE follows a similar pattern. For the simple comparator, there are similar traits, but the bias

Table 20 Spatial Durbin specification mean value of diagnostics $\sigma_\mu^2 = 0.8 \sigma_v^2 = 0.2, p_1 = 0.5$ etc

	Statistic	Number of instruments		
		152	82	26
Error serial correlation	m_1(z ratio)	-7.1784	-7.2143	-7.1769
cor	m_2(z ratio)	-0.0692	-0.0903	-0.0677
Dynamic stability, stationarity	Max abs e'value of $\tilde{A}$	0.3290	0.3290	0.3303
Instrument validity	Sargan–Hansen J statistic	88.5696	69.1083	15.0208
	p-value	0.9997	0.5368	0.4904
	d. of freedom	141	71	15

Table 21 Spatial Durbin specification mean value of diagnostics $\sigma_\mu^2 = 0.2 \sigma_v^2 = 0.8, p_1 = 0.5$ etc

	Statistic	Number of instruments		
		152	82	26
Error serial correlation	m_1 (z ratio)	-7.2127	-7.1755	-7.1484
cor.	m_2 (z ratio)	-0.0072	-0.2095	-0.2270
Dynamic stability, stationarity	Max abs e'value of $\tilde{\mathbf{A}}$	0.3273	0.3286	0.3298
Instrument validity	Sargan–Hansen J statistic	88.0099	71.3155	15.8747
	p-value	0.9997	0.4739	0.4430
	d. of freedom	141	71	15

Table 22 Spatial Durbin specification mean value of diagnostics $\sigma_\mu^2 = 0.8 \sigma_v^2 = 0.2, p_1 = 0.75$ etc

	Statistic	Number of instruments		
		152	82	26
Error serial correlation	m_1 (z ratio)	-7.2585	-7.2755	-7.2315
cor.	m_2 (z ratio)	-0.0953	-0.1064	-0.1200
Dynamic stability, stationarity	Max abs e'value of $\tilde{\mathbf{A}}$	0.3258	0.3261	0.3273
Instrument validity	Sargan–Hansen J statistic	88.7034	69.5378	15.4282
	p-value	0.9997	0.5271	0.4563
	d. of freedom	141	71	15

Table 23 Spatial Durbin specification mean value of diagnostics $\sigma_\mu^2 = 0.2 \sigma_v^2 = 0.8, p_1 = 0.75$ etc

	Statistic	Number of instruments		
		152	82	26
Error serial correlation	m_1 (z ratio)	-7.1955	-7.1984	-7.1533
cor.	m_2 (z ratio)	-0.2064	-0.2103	-0.2385
Dynamic stability, stationarity	Max abs e'value of $\tilde{\mathbf{A}}$	0.3232	0.3235	0.3279
Instrument validity	Sargan–Hansen J statistic	88.9886	71.8841	14.2590
	p-value	0.9996	0.4554	0.5326
	d. of freedom	141	71	15

and RMSE are always much smaller. Tables 18 and 19 show that for the spatial Durbin specification we again see that more instruments are associated with extra downward bias in standard errors and therefore upward bias in t-ratios. Downward bias is also evident from comparing the classic two-step standard errors and the corrected standard errors due to Windmeijer and HKL. The spatial Durbin diagnostics in Tables 20, 21, 22, 23 all point to the viability of the instruments,

Table 24 Model diagnostics

	statistic	
Error serial correlation cor.	m_1 (z ratio)	-2.8477
	m_2 (z ratio)	*******
Dynamic stability, stationarity	Max abs e'value of $\tilde{\mathbf{A}}$	0.6148
Instrument validity	Sargan–Hansen J statistic	9.5946
	p-value	0.4768
	d. of freedom	10

according to the evidence provided by the m_2 statistic, although again the J test statistic is severely undersized when there are a large number of instruments.

6 Example with real data

Fingleton et al. (2020) provides a good example where the application of synthetic instruments produces valuable information regarding the reality of causal effects. The data analysed are taken from successive UK censuses carried out in 1971, 1981, 1991, 2001 and 2011 and relate to employment across each of 760 small areas of Greater London known as wards. For each ward and each census, data are available on the level of employment, and the number of people born in Ireland, India, Pakistan, mainland Europe, the UK, and the rest of the world (i.e. elsewhere), so-called country-of-birth cohorts. Thus, for each ward and each census year there are six country-of-birth cohorts. Additionally, data are available on the location quotient[12] of the unemployment rate. Therefore, there are seven possibly endogenous right hand side explanatory variables. These data, in logarithmic form, are analysed via a dynamic spatial panel data model in the spirit of Eqs. (1) and (2), in which log employment is the dependent variable and the other variables are explanatory variables, and employment depends on the level of employment in the previous census, the level of employment in nearby (contiguous wards), and the level of employment in contiguous wards in the previous census. Additionally, unlike Fingleton et al. (2020), year dummies are introduced, since they are statistically significant. The data only allow two, because of the effect of lagging and differencing which leaves only three years, and one year dummy has to be dropped to avoid perfect collinearity. The choice of dummy years has no effect on the estimates obtained, apart from the dummy variable parameter estimates. Also, time-invariant heterogeneity across wards is eliminated by differencing. The analysis employs a row-standardised contiguity matrix $\mathbf{W}_N$ in order to capture spatial spillover effects. Additionally, following Baltagi et al. (2019), spatial error dependence is represented by a spatial moving average (SMA) error process. As they explain, this should control for omitted spatial

---

[12] The location quotient is defined as the share of the economically active that are unemployed in ward i at time t divided by the share at time t in Greater London as a whole.

Table 25 Parameter estimates using synthetic instruments

Variable	Param	Estimate	se[a]	t-ratio	se[b]	t- ratio	se[c]	t- ratio
y(−1)	γ	0.42	0.235	1.78	0.293	1.43	0.382	1.1
w*y	ρ	0.413	0.173	2.39	0.199	2.07	0.263	1.57
Irish	β_1	−0.354	0.094	−3.79	0.101	−3.52	0.136	−2.62
Indian	β_2	0.158	0.072	2.19	0.080	1.97	0.113	1.4
Pakistani	β_3	−0.085	0.030	−2.86	0.033	−2.56	0.040	−2.1
European	β_4	−0.027	0.063	−0.43	0.063	−0.43	0.079	−0.34
RoW	β_5	0.22	0.052	4.22	0.055	3.96	0.070	3.12
UK_born	β_6	0.588	0.206	2.86	0.237	2.48	0.324	1.81
ln_unemp_LQ	β_7	−0.204	0.091	−2.24	0.099	−2.07	0.136	−1.51
Year dummy	β_8	0.176	0.0444	3.96	0.0443	3.96	0.0511	3.44
Year dummy	β_9	0.049	0.0236	2.1	0.0224	2.21	0.0252	1.96
w*y(−1)	θ	−0.059	0.2982	−0.2	0.3578	−0.16	0.4607	−0.13
SMA error	λ	−0.3266	0.0480	−6.791				
	σ_μ^2	0.0317						
	σ_ν^2	0.0060						

SMA error parameter λ null reference distribution obtained from 100 Bootstrap reps.

se[a] = classic two-step; se[b] = Windmeijer corrected; se[c] = doubly corrected (HKL)

lags of regressors typical of the spatial Durbin specification. With SMA errors, the negative coefficient for the spatial error parameter λ indicates positive spatial error dependence.

Fingleton et al. (2020) consider two scenarios: one is that the explanatory variables are exogenous, and the other assumes that the regressors are endogenous, with instruments controlling for endogeneity inducing effects, for example, due to reverse causation where an increase in the level of employment, the dependent variable, causes country-of-birth numbers to increase, maybe resulting from inward migration attracted by employment opportunities. Eliminating this kind of feedback effect is likely to provide evidence of causal impacts, isolating the effect of an increase in a country-of-birth cohort on the level of employment. However, their conclusions are cautionary because of the short time period considered. Given that the data allow only $T = 4$ effective census years, with 1971 providing the lagged data, this means that one is unable to calculate the Arellano and Bond (1991) m_2 test statistic for zero second order serial correlation in the first differenced residuals, which is only defined for $T >= 5$.

Applying synthetic instruments involves just 22 instruments and leads to a Sargan–Hansen J test that does support an assumption of consistent estimation. The instrument set includes 14 IV-style synthetic instruments derived from the seven presumably endogenous explanatory variables and their spatial lags, plus eight collapsed HENR instruments based on lagged values of $y_{it}, Wy_{it}, Wy_{it-1}$ and $W^2 y_{it-1}$, giving 22 instruments in total. The synthetic instruments are based on the eigenvectors $\mathbf{E}_i, i = 1, ..., N$ deriving from the matrix $\mathbf{P}_N$ and the symmetric contiguity matrix $\mathbf{M}_N$. Given the large number of eigenvectors ($N = 760$), a quite stringent rule

is applied to rule out spurious correlation between eigenvectors and each endogenous variable, with a t-ratio of 2.58 applied, which is equivalent to a two-sided p value = 1%. The estimates are given in Tables 24 and 25.

The diagnostics in Table 24 indicate acceptable model estimates. The maximum characteristic root of $\tilde{\mathbf{A}}$ points to a dynamically stable and stationary model. Moreover, there is evidence from Sargan–Hansen's J that the 10 overidentifying restrictions are valid, indicating that endogeneity bias has been purged, estimates are consistent, and the effects are interpreted as causal.

The interpretation of Table 25 reflects to some extent the analysis of Fingleton et al. (2020). The signs of the parameter estimates are not dissimilar, although the presence of the year dummies in the current specification has an impact; for example, the estimate of the spatial lag parameter ρ is smaller though still significantly different from zero (see Lee and Yu 2010). Some estimates are larger and although we have larger classic standard errors, t-ratios are larger and pick out some significant effects. On the basis of the uncorrected standard errors, the significant variables are Irish, Indian, UK, Pakistani and rest of the world countries-of-birth, plus a significant negative effect due to the unemployment location quotient and significant positive spatial spillovers from log employment (ρ) and residuals (λ) in contiguous districts. A 1% increase in Indian-born residents evidently causes an approximately 0.16% increase in employment. For UK-born residents, the estimate is a 0.6% increase, and for migrants from the rest of the world, we estimate a 0.22% increase. In contrast, a 1% increase in the Irish-born population evidently causes a 0.35% reduction in employment, and for Pakistani-born residents there is a 0.085% reduction. While we might infer causality due to the avoidance of endogeneity-bias in our estimation, but caution is required, partly because these conclusions are based on uncorrected standard errors. As shown in Table 25, correction due to Windmeijer and HKL does increase standard errors, but the significant effects persist.

An additional consideration is the fact that the parameter point estimates are not the true elasticities as a result of the significant spatial spillover effect (ρ) which magnifies the initial parameter point estimates. Accordingly, the true elasticity for Indian-born residents (variable x_2) is derived from

$$\left[\frac{dy}{dx_{12}} \cdots \frac{dy}{dx_{N2}} \right]_t = \hat{\beta}_2 \mathbf{B}_N^{-1} \tag{19}$$

which is an N by N matrix of partial derivatives. From this, a simplified average measure of the total effect of a 1% increase in the Indian-born population at time t is given by the mean column sum of $\hat{\beta}_2 \mathbf{B}_N^{-1}$, which is equal to 0.269%. For migrants born in the rest of the world, the elasticity is 0.374%. By comparison, the true elasticity for the UK-born cohort is 1.01%. On the negative side, the elasticity for Pakistani-born residents is -0.144% and for the Irish-born it is -0.603%.

Extending beyond time t, assuming dynamic stability and stationarity, elasticities converge in the long-run to steady-state levels given by

$$\left[\frac{dy}{dx_{1k}} \cdots \frac{dy}{dx_{Nk}}\right] = \left[-\mathbf{C}_N + \mathbf{B}_N\right]^{-1} \left(\hat{\beta}_k \mathbf{I}_N\right) \qquad (20)$$

The total long-run effect of a 1% increase in the Indian-born population is given by the mean column sum of Eq. (20) with $k = 2$, which is equal to 0.699%. For migrants born in the rest of the world, the elasticity is 0.971%. By comparison, the true elasticity for the UK-born cohort is 2.599%. On the negative side, the elasticity for Pakistani-born residents is -0.374% and for the Irish-born it is -1.567%.

7 Conclusion

One of the main problems in analysing dynamic (spatial) panel data estimated by GMM with endogenous explanatory variables is the difficulty in finding appropriate instruments. A standard approach is to use instruments that are internal to the model, lagging the regressors appropriately in order to satisfy the moments conditions which require instruments to be orthogonal to differenced errors. The default approach using HENR instrumentation typically generates a large number of internal instruments, as could using multiple spatial lags as suggested in the spatial econometrics literature. But instrument proliferation makes the Hansen–Sargan J statistic severely undersized and possessing low power and therefore unreliable as an indicator of estimation consistency. There are several options available to reduce the problem, and this paper proposes a new one, which is to extend the use of synthetic instruments as advocated by Le Gallo and Páez (2013) for cross-sectional data to dynamic spatial panel data modelling. The key to the approach is a set of exogenous instruments derived from a connectivity matrix which is not causally related to the data. Nevertheless, the synthetic instruments thus created tend to be strongly correlated with endogenous regressors and thus can remedy the problem of weak instruments. Using Monte Carlo simulation, the paper provides evidence that collapsing and also applying synthetic instruments, and as a result significantly reducing the number of instruments, reduces problems associated with instrument overabundance while producing, on the whole, plausible parameter estimates, although there is evidence of some parameter estimation bias. An important additional problem is the downward bias in parameter standard error estimates, which can result in serious upward bias in *t*-ratios, although collapsing and the use of synthetic instruments causes this problem to greatly diminish. The results of Monte Carlo simulation are of course conditional on assumptions made, so one cannot be too dogmatic regarding the generality of the conclusions reached. However, application to real data given in Fingleton et al. (2020) produces plausible outcomes and new insights which put the analysis provided in that paper on a stronger inferential footing.

Appendix

Computing

(i) Data

The full data set for each simulation comprises 100 regions and 10 time periods. Table 26 gives a sample of data for 10 periods.

In Table 26, yr denotes time period and regee denotes region. The dependent variable is y_it, with regressors x_1it, x_2it, x_3it, x_4it and their spatial lags are W_x_1,W_x_2,W_x_3 and W_x_4. The temporal lag of y is lag_y, the spatial lag of y is Wy, the spatial lag of the temporal lag is Wy_ and the spatial lag of the spatial lag of the temporal lag is WWy_.

zz_1 is the synthetic instrument for endogenous variable x_it1, and W_zz_1 is the spatial lag of zz_1. The synthetic instruments for x_it2, x_it3 and x_it4 and their spatial lags zz_2, zz_3, zz_4, W_zz_2, W_zz_3,and W_zz_4 are omitted to save space. zz_ev1, zz_ev2 and zz_ev3 are the synthetic instruments for Wy, Wy_ and WWy_.

(ii) Stata code

Table 26 Sample of data used in Monte Carlo simulations

Year	regee	y_it	x_1it	x_2it	x_3it	x_4it	W_x_1	W_x_2	W_x_3	W_x_4
1	1	−0.75	2.01	−0.03	−2.60	1.54	0.14	0.17	−0.38	0.08
2	1	−1.36	1.58	−0.12	−2.45	1.31	0.27	−0.09	−0.06	0.02
3	1	−0.41	1.29	−0.82	−1.17	1.06	0.16	−0.25	0.12	−0.03
4	1	−0.29	3.01	−2.01	−0.45	0.48	0.02	0.14	0.09	−0.21
5	1	0.75	3.49	−1.23	−1.04	0.94	0.01	−0.05	−0.14	0.12
6	1	0.02	1.61	−0.19	−0.88	0.62	−0.02	−0.35	−0.20	0.39
7	1	1.04	1.86	−1.26	0.50	0.28	−0.07	−0.34	−0.02	0.22
8	1	0.81	1.62	−2.10	0.75	0.27	−0.04	−0.17	0.08	0.01
9	1	1.14	2.01	−1.92	0.96	−0.21	−0.15	−0.20	0.43	−0.28
10	1	−0.04	0.47	−0.25	0.05	−0.24	−0.32	0.08	0.23	−0.24
1	2	−1.11	−0.39	0.78	−1.46	0.64	0.32	0.08	−0.38	0.10

Year	regee	Wy_	Wy	WWy_	zz_1	W_zz_1	zz_ev1	zz_ev2	zz_ev3	lag_y
1	1	0.86	0.01	0.99	0.19	0.13	0.26	0.95	0.98	0.47
2	1	0.01	0.26	0.35	0.24	0.13	0.37	0.26	0.42	−0.75
3	1	0.26	0.33	0.26	0.10	0.12	0.37	0.37	0.24	−1.36
4	1	0.33	0.08	0.27	0.00	0.06	0.07	0.37	0.24	−0.41
5	1	0.08	−0.33	0.08	−0.02	0.04	−0.24	0.07	0.13	−0.29
6	1	−0.33	−0.53	−0.19	−0.01	0.06	−0.45	−0.24	−0.14	0.75
7	1	−0.53	−0.33	−0.38	−0.02	0.03	−0.23	−0.45	−0.32	0.02
8	1	−0.33	−0.07	−0.19	−0.06	−0.10	−0.02	−0.23	−0.19	1.04
9	1	−0.07	0.24	−0.11	−0.07	0.01	0.24	−0.02	−0.15	0.81
10	1	0.24	0.21	0.17	−0.19	−0.09	0.19	0.24	0.14	1.14
1	2	1.01	0.11	0.98	0.19	0.12	0.24	0.95	0.98	−0.55

The code runs on Stata/BE 17.0 for Windows (64 bitx86-64).

Using data such as in Table 26, the following code applies the estimator xtabond2, developed by David Roodman.

Roodman D (2009) 'How to Do xtabond2: An Introduction to "Difference" and "System" GMM in Stata', The Stata Journal (2009) Vol. 9, Number 1, pp. 86–136.

xtabond2,version 03.07.00.

Spatial specification

* 292 instruments.

xtabond2 y_it l.y_it Wy x_1it x_2it x_3it x_4it Wy_, gmmstyle(y_it Wy Wy_ WWy_ x_3it W_x_3 x_4it W_x_4, laglimits(2.)) iv(x_1it x_2it W_x_1 W_x_2, passthru) noconstant noleveleq twostep robust nomata.

* 152 instruments.

xtabond2 y_it l.y_it Wy x_1it x_2it x_3it x_4it Wy_, gmmstyle(y_it Wy Wy_ WWy_, laglimits(2.)) iv(x_1it x_2it zz_3 zz_4 W_x_1 W_x_2 W_zz_3 W_zz_4, passthru) noconstant noleveleq twostep robust nomata.

* 40 instruments.

xtabond2 y_it l.y_it Wy x_1it x_2it x_3it x_4it Wy_, gmmstyle(y_it Wy Wy_ WWy_, collapse laglimits(2.)) iv(x_1it x_2it zz_3 zz_4 W_x_1 W_x_2 W_zz_3 W_zz_4, passthru) noconstant noleveleq twostep robust nomata.

* 19 instruments.

xtabond2 y_it l.y_it Wy x_1it x_2it x_3it x_4it Wy_, gmmstyle(y_it, collapse laglimits(2.)) iv(x_1it x_2it zz_3 zz_4 W_x_1 W_x_2 W_zz_3 W_zz_4 zz_ev1 zz_ev2 zz_ev3, passthru) noconstant noleveleq twostep robust nomata.

Non-spatial specification

* 110 instruments.

xtabond2 y_it l.y_it x_1it x_2it x_3it x_4it, gmmstyle(y_it x_3it x_4it, laglimits(2.)) iv(x_1it x_2it, passthru) twostep robust artests(2) noconstant noleveleq nomata.

* 40 instruments.

xtabond2 y_it l.y_it x_1it x_2it x_3it x_4it, gmmstyle(y_it, laglimits(2.)) iv(x_1it x_2it zz_3 zz_4, passthru) twostep robust artests(2) noconstant noleveleq nomata.

*26 instruments.

xtabond2 y_it l.y_it x_1it x_2it x_3it x_4it, gmmstyle(y_it x_3it x_4it, collapse laglimits(2.)) iv(x_1it x_2it, passthru) twostep robust artests(2) noconstant noleveleq nomata.

*12 instruments.

xtabond2 y_it l.y_it x_1it x_2it x_3it x_4it, gmmstyle(y_it, collapse laglimits(2.)) iv(x_1it x_2it zz_3 zz_4, passthru) twostep robust artests(2) noconstant noleveleq nomata.

Spatial Durbin

* 152 instruments.

xtabond2 y_it l.y_it Wy x_1it x_2it x_3it x_4it W_x_1 W_x_2 W_x_3 W_x_4 Wy_, gmmstyle(y_it Wy Wy_ WWy_, laglimits(2.)) iv(zz_1 zz_2 zz_3 zz_4 W_zz_1 W_zz_2 W_zz_3 W_zz_4, passthru) noconstant noleveleq twostep robust nomata.

* 82 instruments.

xtabond2 y_it l.y_it Wy x_1it x_2it x_3it x_4it W_x_1 W_x_2 W_x_3 W_x_4 Wy_, gmmstyle(y_it Wy, laglimits(2.)) iv(zz_1 zz_2 zz_3 zz_4 W_zz_1 W_zz_2 W_zz_3 W_zz_4 zz_ev2 zz_ev3, passthru) noconstant noleveleq twostep robust nomata.

* 26 instruments.

xtabond2 y_it l.y_it Wy x_1it x_2it x_3it x_4it W_x_1 W_x_2 W_x_3 W_x_4 Wy_, gmmstyle(y_it Wy, collapse laglimits(2.)) iv(zz_1 zz_2 zz_3 zz_4 W_zz_1 W_zz_2 W_zz_3 W_zz_4 zz_ev2 zz_ev3, passthru) noconstant noleveleq twostep robust nomata.

Declarations

References

Andersen TG, Sørensen BE (1996) GMM estimation of a stochastic volatility model: a Monte Carlo study. Journal of Business and Economic Statistics 14:328–352

Anselin L, LeGallo J, Jayet H (2008) Spatial panel econometrics. In: Mátyás L, Sevestre P (eds) The econometrics of panel data, fundamentals and recent developments in theory and practice, 3rd edn. Kluwer, Dordrecht, pp 627–662

Arellano M, Bond S (1991) Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. Rev Econ Stud 58:277–297

Baltagi B (2013) Econometrics of panel data, 5th edn. Springer, New York

Baltagi BH (2021) Econometric analysis of panel data, 6th edn. Springer, New York

Baltagi BH, Fingleton B, Pirotte A (2019) A time-space dynamic panel data model with spatial moving average errors. Reg Sci Urban Econ 76:13–31

Belsley DA, Kuh E, Welsh RE (1980) Regression diagnostics. Wiley, New York

Bond SR (2002) Dynamic panel data models: a guide to micro data methods and practice. Portuguese Econ J 1(2):141–162. https://doi.org/10.1007/s10258-002-0009-9

Boots B, Tiefelsdorf M (2000) Global and local spatial autocorrelation in bounded regular Tessellations. J Geogr Syst 2:319–348

Bowsher CG (2002) On testing overidentifying restrictions in dynamic panel data models. Econ Lett 77:211–220

Debarsy N, Ertur C, LeSage JP (2012) Interpreting dynamic space–time panel data models. Statist Methodol 9(1–2):158–171

Drukker DM, Egger P, Prucha IR (2013) On two-step estimation of a spatial autoregressive model with autoregressive disturbances and endogenous regressors. Economet Rev 32:686–733

Elhorst JP (2001) Dynamic models in space and time. Geogr Anal 33(2):119–140

Elhorst JP (2014) Spatial econometrics: from cross sectional data to spatial panels. Springer, Heidelberg

Fingleton B, Le Gallo J (2008) Estimating spatial models with endogenous variables, a spatial lag and spatially dependent disturbances: finite sample properties. Pap Reg Sci 87:319–339

Fingleton B, Olner D, Pryce G (2020) Estimating the local employment impacts of immigration: a dynamic spatial panel model. Urban Stud 57:2646–2662

Getis A, Griffith DA (2002) Comparative spatial filtering in regression analysis. Geogr Anal 34(2):130–140

Griffith DA (1988) Advanced spatial statistics. Kluwer Academic, Dordrecht

Griffith DA (1996) Spatial autocorrelation and eigenfunctions of the geographic weights matrix accompanying geo-referenced data. Can Geogr 40:351–367

Griffith DA (2000) A linear regression solution to the spatial autocorrelation problem. J Geogr Syst 2:141–156

Griffith DA (2003) Spatial autocorrelation and spatial filtering: gaining understanding through theory and scientific visualization. Springer, New York

Halleck Vega S, Elhorst JP (2015) The SLX model. J Reg Sci 55:339–363

Hansen L (1982) Large sample properties of generalized method of moments estimators. Econometrica 50:1029–1054

Holtz-Eakin D, Newey W, Rosen HS (1988) Estimating vector autoregressions with panel data. Econometrica 56:1371–1395

Hsiao C (2003) Analysis of panel data, 2nd edn. Cambridge University Press, Cambridge

Hwang J, Kang B, Lee S (2022) 'A doubly corrected robust variance estimator for Linear GMM' forthcoming. J Econometr. https://doi.org/10.1016/j.jeconom.2020.09.010

Jeanty P, Partridge M, Irwin E (2010) Estimation of a spatial simultaneous equation model of population migration and housing price dynamics. Reg Sci Urban Econ 40:343–352

Jin F, Lee LF (2013) Generalized spatial two stage least squares estimation of spatial autoregressive models with autoregressive disturbances in the presence of endogenous regressors and many instruments. Econometrics 1:71–114

Kelejian HH, Piras G (2014) Estimation of spatial models with endogenous weighting matrices, and an application to a demand model for cigarettes. Reg Sci Urban Econ 46:140–149

Kelejian HH, Piras G (2016) An extension of the J -test to a spatial panel data framework. J Appl Econ 31:387–402

Kelejian HH, Piras G (2017) Spatial econometrics. Elsevier, Amsterdam

Kelejian HH, Piras G (2018) Important overlooked IVs in spatial models. Empirical Econ 55:69–83

Kelejian HH, Prucha IR (1998) A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. J Estate Finan Econ 17:99–121

Kelejian HH, Prucha IR (1999) A generalized moments estimator for the autoregressive parameter in a spatial model. Int Econ Rev 40:509–533

Le Gallo J, Páez A (2013) Using synthetic variables in instrumental variable estimation of spatial series models. Environ Plan A 45:2227–2242

Lee LF, Yu J (2010) Some recent developments in spatial panel data models. Reg Sci Urban Econ 40:255–271

Liu X, Lee LF (2013) Two stage least squares estimation of spatial autoregressive models with endogenous regressors and many instruments. Economet Rev 32:734–753

Liu X, Saraiva P (2015) GMM estimation of SAR models with endogenous regressors. Reg Sci Urban Econ 55(5–6):68–79

Nickell S (1981) Biases in dynamic models with fixed effects. Econometrica 49:1417–1426

Pace RK, LeSage JP, Zhu S (2012) Spatial dependence in regressors. In: Fomby TB, Carter Hill R, Jeliazkov I, Escanciano JC, Hillebrand E (eds) Advances in econometrics. Emerald Group Publishing Limited, Somerville, pp 257–295

Parent O, LeSage JP (2011) A space–time filter for panel data models containing random effects. Comput Stat Data Anal 55(1):475–490

Parent O, LeSage JP (2012) Spatial dynamic panel data models with random effects. Reg Sci Urban Econ 42(4):727–738

Patuelli R, Griffith DA, Tiefelsdorf M, Nijkamp P (2006) The use of spatial filtering techniques, Tinbergen Institute Discussion Paper 2006-049/3

Pesaran MH (2015) Time series and panel data econometrics. Oxford University Press, Oxford

Roodman D (2009a) A note on the theme of too many instruments. Oxford Bull Econ Stat 71:135–158

Roodman D (2009b) How to do xtabond2: an introduction to difference and system GMM in Stata. Stand Genom Sci 9(1):86–136

Sargan JD (1958) The estimation of economic relationships using instrumental variables. Econometrica 26:393–415

Windmeijer F (2005) A finite sample correction for the variance of linear efficient two-step GMM estimators. J Econometr 126:25–51

Yu J, de Jong R, Lee LF (2008) Quasi-maximum likelihood estimators for spatial dynamic panel data with fixed effects when both N and T are large. J Econometr 146:118–134

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.