Robust Inference in Short Linear Panels with Fixed Effects with Endogenous Covariates in a Spatial Setting

Steven Wu-Chaves

Abstract

I propose a simple way to obtain robust standard errors in linear panels in a spatial context with endogenous covariates where the number of time periods is small relative to the cross sectional dimension. The method is based on applying a Spatial HAC to an average of moment conditions across time to obtain a covariance estimator that is robust to both spatial and serial correlation (HACSC). I also present a control function approach (CF) alternative to estimate the parameters and extend the HACSC estimator to this case, where the standard errors require an adjustment to account for the sampling variability induced by the first stage estimation. In addition, I derive the Fixed Effects-Random Effects equivalence under a Correlated Random Effects framework in the presence of a spatial lag of the dependent variable to obtain a fully-robust Hausman-type test using the HACSC estimator. I run a Monte Carlo experiment and show that the HACSC estimator is robust to strong patterns of serial and spatial correlation. Furthermore, I also find that whenever the CF assumptions hold, the CF approach is more efficient than Two-Stage Least Squares. Finally, I estimate the effect of school district spending on the performance of fourth-grade students in Michigan, allowing for spillovers across districts. I find that the expenditure from neighboring districts has a positive and non-negligible impact on test passing rates.

1 Introduction

The assumption of independent data is widespread in empirical economics since it simplifies many of the estimation methods. However, in many fields such as international trade, urban economics, public policy or even network analysis, this assumption might not hold since the outcome variable of an individual might be affected by other observations' actions, which leads to (spatially) dependent data. Furthermore, many of the tools used to develop the asymptotic theory behind popular econometric methods such

as the Central Limit Theorem and Law of Large Numbers often rely on independent and identically distributed (i.i.d.) data. This facilitates both the estimation and inference, but if this assumption is violated, then the latter becomes more difficult even if the parameters are estimated consistently.

Additionally, the increasing availability of data sets over time has increased the popularity of panel methods in recent years as they allow to incorporate time effects and to estimate richer models. Nevertheless, they also introduce complications because the presence of unobserved heterogeneity could generate inconsistency problems both in the parameters and standard errors if it is not properly handled. When combining both spatially dependent observations and panel data, inference becomes more challenging since the error term can be both serially and spatially correlated.

To address the spatial correlation, the literature in the field has usually resorted to assume and model a particular structure of the error term, as it was common to do with time series data. However, since the seminal work of White (1980), the common practice nowadays in the latter is to use standard errors that are robust to general forms of heteroskedasticity and autocorrelation (HAC). This procedure has been extended to the spatial framework (SHAC) by Conley (1999) and Kelejian and Prucha (2007) in a cross sectional setting. However, to the best of my knowledge and surprisingly enough, this has not been extended to the panel case where the time dimension is fixed and the number of units of observation goes to infinity, even in the linear case.[1] Admittedly, there are many cases in which the time dimension is also large, however there are also instances where the number of observations across time is considerably smaller compared to the cross sectional dimension.

This generates issues because ignoring the serial correlation could still generate biased standard errors, even if the associated covariance matrix is robust to spatial correlation. In fact, one of the few routines implemented in Stata for panel data in a spatial context corrects the standard errors for spatial correlation, but assumes serial independence of the error terms. The main purpose of this paper is to propose a simple way to obtain robust standard errors in a linear panel that are robust to heteroskedasticity and both to spatial and serial correlation (HACSC), without imposing any structure on the time dimension and using a Fixed Effects framework with endogenous covariates. I also extend this procedure to the case of the control function approach, where the computation of standard errors is more difficult due to the presence of a generated regressor.

HAC estimators have been extensively used in the time series literature since they avoid having to model the error term structurally, which can lead to inconsistency

[1]Perhaps one of the reasons is that econometricians assume that it is obvious what to do, but many methods make strong assumptions in the time dimension like serial independence.

issues if that process is misspecified. Newey and West (1987) were the first to extend White's estimator to allow for general forms of heteroskedasticity and autocorrelation. In the panel case, Arellano (1987) introduced the panel clustered standard errors, which are robust to heteroskedasticity and autocorrelation but require that the observations between clusters to be uncorrelated.

In spatial panels, multiple authors have made important contributions to the field, extending many of the methods developed in the time series literature. For example, Driscoll and Kraay (1998) presented how to deal with spatially dependent panel data in a GMM context by averaging the moment conditions in the cross section dimension index, N. Their approach relies on holding fixed N and letting time dimension $T \to \infty$. Vogelsang (2012) develops asymptotic theory for linear spatial panels with fixed effects in a fixed-b framework by averaging HAC estimators and by computing the HAC for averages as in Driscoll and Kraay (1998). In this case, the asymptotics rely again in $T \to \infty$ and allowing N to remain fixed or to grow. In a similar context Kim and Sun (2013) proposed a bivariate kernel HACSC estimator, which requires that both the cross section and time dimensions to go to infinity. Bester et al. (2011) suggested a cluster covariance matrix that is applicable when the data is dependent in the context of time series, spatial and panel data. More recently, Müller and Watson (2022a) introduced a new methodology to construct confidence intervals based on population principal components with the property that the resulting interval will have a coverage probability of 95% for a set of spatial patterns in a cross sectional setting. Müller and Watson (2022b) extended this framework to spatial panels to cover estimation techniques like difference-in-difference setups.

At the cross sectional level, Conley (1999) was the first to develop a Spatial HAC (SHAC) estimator in a GMM context. His approach is based on the assumption that the data generating process is spatially stationary. When working with dependent data and allowing $N \to \infty$, it is common to assume some sort of weak dependence mechanism, analogous to the time series literature, so that the influence of one observation on other units diminishes as the distance between them increases. In this case, Conley assumes that the data is spatially $\alpha-$mixing. Bester et al. (2016) provide a fixed-b analysis of Conley's SHAC estimator. Kelejian and Prucha (2007) relax the spatial stationarity assumption and model the spatial dependence in terms of a weighting matrix, arguing that having a different number of neighbors, as it is common in empirical work, violates the assumption. It is important to note that their SHAC estimator is based on consistent estimates of the error terms, but they do not provide any parameter estimation framework.

Kim and Sun (2011) generalize this estimator to allow general linear and non linear models using moment conditions. Conley and Molinari (2007) performed a Monte Carlo

study in which they compared the performance of multiple covariance estimators with dependent data in the context of locations measured with error and they concluded that non parametric estimators work better compared to parametric ones such as GMM and maximum likelihood estimators. In this paper, I follow Driskoll and Kraay's approach, but instead of averaging the moment conditions over the cross sectional dimension, I average the moment conditions over time and construct a GMM estimator and then apply Kelejian and Prucha's SHAC over the corresponding residuals. By doing this, I avoid imposing any assumptions over the serial correlation and hence, construct a covariance estimator that is robust to both serial and spatial correlation.

Beyond testing the statistical significance of the effect of a covariate on the response variable, robust inference is also important when trying to choose the correct specification of a model. More specifically, the correlated random effects (CRE) approach has been very popular in recent years because it is a simple way to test between Random Effects (RE) and Fixed Effects (FE) specifications and it allows to include time constant variables as noted by Joshi and Wooldridge (2019). Furthermore, we can obtain the FE coefficients of the time varying variables by including the time average of these on the right hand side of the equation in a Pooled OLS or RE regression, a result attributed to Mundlak (1978). Debarsy (2012) was the first to extend the Mundlak approach to the spatial setting. More recently, Li and Yang (2020) showed that when the model includes a structurally modeled error term (which involves maximum likelihood estimation), the equivalence holds conditional on the parameter associated with the error term, however, the equivalence breaks unconditionally, i.e., when this parameter has to be estimated jointly with the rest of parameters. In this paper, I show that the result holds in a specific setting; namely, if the model does not include a structurally modeled error term.

One of the additional advantages of not imposing a particular spatial structure on the error term is that some estimation methods become readily available such as Two Stage Least Squares (2SLS) or a Control Function (CF) approach (Blundell and Powell 2003) whenever the researcher suspects an endogenous variable is in the model. In fact, adding a spatial lag of the response variable as a covariate yields the spatial autoregressive model (SAR), a very popular model in this literature. However, Kelejian and Prucha (1998) showed that this term induces an endogeneity problem, which is why the researcher has to resort to an Instrumental Variable (IV) procedure. In terms of the estimation of parameters, both 2SLS and the CF approach require the availability of instruments, however one important difference is that the latter imposes additional assumptions and is therefore less robust than 2SLS. On the other hand, if the assumptions hold, the CF allows to deal with the endogeneity in a more parsimonious way

if multiple functions[2] of the endogenous variable appear on the right hand side of the equation and is probably more efficient (Wooldridge 2010). Note that this parsimony is relevant in the spatial case since it is common to include spillover effects in the models and therefore, the likelihood of having multiple functions of a variable increases in this context.

In a spatial setup, Basile (2009) and Basile et al. (2014) extended the CF to additive non-parametric models. In terms of inference, Basile et al. (2014) recommends to use bootstrap to obtain confidence intervals, a practice that is common even in the i.i.d. case. However, as pointed out by Kunsch (1989), the independence assumption plays a critical role on the validity of the bootstrap, so besides the computational cost, in a spatial context this is not a trivial procedure due to the dependence between observations. Intuitively, if we just randomly sample the data in a time series setting at each bootstrap repetition, the serial correlation structure would be lost and a similar issue occurs in the spatial case. This is why different bootstrap methods have been proposed in the time series literature (see Politis and White [2004] for a brief overview), nevertheless their extension to the spatial case is not straightforward due to the absence of a natural ordering of the observations. Given this, it might be desirable to obtain a closed-form formula for the covariance matrix when the empirical researcher is working with parametric linear models with panel data in a spatial context. This paper tries to fill this hole in the literature by adjusting the HACSC estimator to the CF setting. This adjustment is necessary because in addition to deal with the spatial and serial correlation, it is necessary to take into account the sampling error induced by the first stage estimation.

The rest of the paper is organized a follows. Section 2 discusses the model and the assumptions used to obtain the estimator of the covariance matrix. Section 3 presents the HACSC estimator and its asymptotic properties. Section 4 derives the FE and RE equivalence using the correlated random effects approach in a spatial context. Section 5 presents an additional application of the HACSC estimator under a Feasible GLS context. Section 6 presents the control function approach and a discussion of the additional assumptions imposed in this context. Section 7 shows an empirical application of the HACSC estimator using data from the Michigan education system. Section 9 concludes.

[2]A well known result in the literature is that 2SLS and the CF give the same numerical coefficients if only one function of the endogenous variable is in the model. This carries over to the spatial case under the settings outlined at the beginning of the paragraph.

2 The Model

2.1 Estimation of the parameters

Consider the following model[3]:

$$\begin{aligned} y_{it} &= x_{1it}\beta_1 + x_{2it}\beta_2 + W_i X_{1t}\gamma_1 + W_i X_{2t}\gamma_2 + \lambda W_i y_t + c_i + u_{it} \\ &= x_{it}\beta + W_i X_t \gamma + \lambda W_i y_t + c_i + u_{it}, \;\; i = 1 \dots N, \; t = 1 \dots T \end{aligned} \tag{2.1}$$

where y_{it} is the dependent variable, x_{1it} is a $1 \times (k_1+1)$, vector of explanatory exogenous variables (including an intercept), x_{2it} is a $(1 \times k_2)$ vector of endogenous variables. The sense in which x_{1it} is exogenous will be clarified below. W_i is the i-th row of the $N \times N$ time invariant weighting matrix W, whose diagonal elements are zero, X_{1t} and X_{2t} are the $N \times k_1$ and $N \times k_2$ matrices of exogenous and endogenous covariates, respectively, for all observations at time t, y_t is the vector of dependent variables at time t, c_i is the individual heterogeneity and u_{it} is the idiosyncratic error. Hence β, γ and λ are the parameters of interest and they are of dimension $(k+1) \times 1$, $k \times 1$ and 1×1 respectively. Throughout the rest of the paper, I assume that $N \to \infty$ while T remains fixed.

We assume that there exist a set of instruments z_{2it} for x_{2it} of dimension $l \geq k_2$ (so that $W_i Z_{2t}$ are the instruments for $W_i X_{2t}$). As previously showed by Kelejian and Prucha (1998), the inclusion of a spatial lag of the dependent variable on the right hand side also induces an endogeneity issue for which we also need instruments. Kelejian et al. (2004) and Lee (2003) determined that the optimal set of instruments for this variable is a sequence of the form $W^j X_t$, for $j = 1...s, s \in \mathbb{N}$ (in this case, we would only include higher power spatial lags of X_{1t}). If we let $w^j_{rit} \equiv W^j_i X_{rt}$, $r = 1, 2$, and $\mathfrak{Z}_{2it} \equiv W_i Z_{2t}$, $A_{it} \equiv (x_{1it} \quad x_{2it} \quad w_{1it} \quad w_{2it} \quad W_i y_t)$ and $\theta \equiv (\beta_1' \quad \beta_2' \quad \gamma_1' \quad \gamma_2' \quad \lambda)'$, then the model can be written more compactly as:

$$y_{it} = A_{it}\theta + c_i + u_{it} \tag{2.2}$$

Since we are not assuming a particular structure for the error term, we can estimate the parameters of (2.2) with the Fixed Effects 2SLS estimator. To do so, we can apply the within transformation to all the variables, so let $\ddot{y}_{it} = y_{it} - \bar{y}_i$, where $\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}$ and similarly for the independent variables and the instruments. Then we can use

[3]The model includes a spatial lag of the dependent variable on the right hand side for the sake of generality and because this is a widely spread practice in the spatial literature. Nevertheless, it is important to emphasize that its inclusion precludes the interpretation of (2.1) as a conditional mean function and also complicates the interpretation of the coefficients. As such, in some sections of the paper this variable will be omitted.

Pooled 2SLS to the transformed model

$$\ddot{y}_{it} = \ddot{A}_{it}\theta + \ddot{u}_{it} \tag{2.3}$$

using the instruments $\ddot{Z}_{it} = (\ddot{x}_{1it} \quad \ddot{w}_{1it} \quad \ddot{z}_{2it} \quad \ddot{\mathfrak{Z}}_{2it} \quad \ddot{w}^2_{1it} \quad \ddot{w}^3_{1it} \dots \ddot{w}^s_{1it})$. Note that all the individual unobserved effects have been removed. To obtain consistent parameters, we need the following orthogonality condition:

$$\mathbb{E}(\ddot{Z}'_{it}\ddot{u}_{it}) = \mathbb{E}[g_{it}(Z_{it}, \theta)] = 0, \;\; t = 1 \dots T \tag{2.4}$$

which is implied by the stronger strict exogeneity condition:

$$\mathbb{E}(u_{it}|Z) = \mathbb{E}(u_{it}|Z, W) = 0$$

where Z is the $NT \times [(s+1)k_1 + 2l + 1]$ matrix of exogenous variables for all cross sectional units and all time periods. We note that in this spatial setting, this condition is stronger than in the non-spatial case because here we are conditioning the expected value of u_{it} with respect to all other units and not only i's independent variables [see Wooldridge (2010), pp. 301 for more details].

The $g_{it}(Z_{it}, \theta)$ function is of dimension $(s+1)k_1 + 2l + 1 = r$, hence for each i, there are $T \times r$ moment conditions. Under this framework, we could use many more moment conditions because our strict exogeneity assumption implies orthogonality conditions for each pair of time periods and cross sectional units [i.e. $\mathbb{E}(\ddot{Z}_{it}\ddot{u}_{js})$, $i, j = 1 \dots N$ and $t, s = 1 \dots T$], however we will only use the conditions implied by the FE estimator. Using a similar idea as Driscoll and Kraay (1998), for each observation i we can average these moment conditions over time[4], so let:

$$g_i(Z_i, \theta) = \frac{1}{T} \sum_{t=1}^{T} g_{it}(Z_{it}, \theta) \tag{2.5}$$

From this, one can construct a GMM estimator, which will be defined as follows:

$$\hat{\theta} = \min_{\theta \in \Theta} \left[\frac{1}{N} \sum_{i=1}^{N} g_i(Z_i, \theta) \right]' \hat{\Omega} \left[\frac{1}{N} \sum_{i=1}^{N} g_i(Z_i, \theta) \right] \tag{2.6}$$

where $\hat{\Omega}$ is a $r \times r$ positive definite, symmetric, weighting matrix. Admittedly, as noted above we could estimate θ by running Pooled 2SLS on (2.3), however, the GMM framework allows for more generality. For instance, averaging the moment conditions over time for each observation can be done in other setups different than fixed effects. Furthermore, this averaging might not be the most efficient approach, but obtaining

[4]Note however that Driskoll and Kraay's case is based on having N fixed ant $T \to \infty$ and they average across i for all t.

the optimal GMM in a two-step procedure might provide some efficiency gains with respect to Pooled 2SLS.

2.2 Assumptions

The consistency and normality of this estimator can be obtained from a Uniform Law of Large Numbers (ULLN) and Central Limit Theorem (CLT) derived by Nazgul and Prucha (2009) for non-stationary random fields in a possibly uneven lattice. Before stating their assumptions, we need some definitions. Let $D \subset \mathbb{R}^d$, $d \geq 1$ be an uneven lattice and let $\rho(i,j) = \max_{1\leq k\leq d} |j_k - i_k|$ and $|i| = \max_{1\leq k\leq d} |i_k|$, where i_k denotes the k-th component of i, be a metric and norm, respectively, of $\mathbb{R}^d$. The minimum distance between two subsets E, F of D is defined as $\rho(E,F) = \inf[\rho(i,j) : i \in E \text{ and } j \in F]$ and let $|E|$ denote the cardinality of a subset $E \in D$. Other definitions used throughout this section can be found in the Appendix.

We now state the assumptions required to obtain the consistency and asymptotic normality of $\hat{\theta}$. We note that the N subscript in the random fields and scalars of the assumptions are to explicitly indicate that the ULLN and CLT can accommodate for triangular arrays, which are common in the spatial literature and particularly in Cliff-Ord type models. However, for notation simplicity, it will be suppressed in many sections for the remainder of the paper.

Assumption 1
The lattice $D \subset \mathbb{R}^d, d \geq 1$ is infinite countable and there exists a distance ρ_0 such that $\rho(i,j) \geq \rho_0 \; \forall i,j \in D$. Without loss of generality, suppose that $\rho_0 > 1$.

Assumption 1 provides the necessary structure to the lattice. Note that the existence of the distance is essential in order to obtain non parametric estimators of the covariance matrix and it is analogous to the time difference between observations in the time series literature. Furthermore, it is possible that the distance *observed* by the researcher, between two observations i and j, $\rho^*(i,j)$, is measured with error. Note that the existence and availability of this distance measure is not trivial, even in the leading case of a geographical region. As shown in Figure 1, there are instances in which using the linear distance between many pairs of points in that territory would not represent the *real* burden to arrive from one location to another (e.g. driving), while there are other cases in which this measure would be appropriate (e.g. pollution).

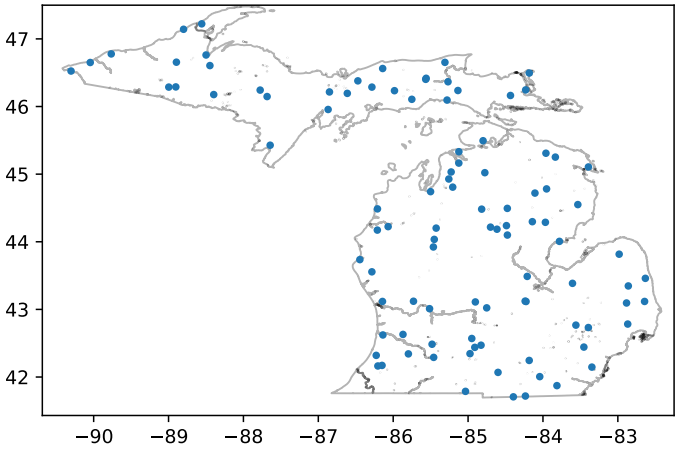

Figure 1: Points in an irregular geographic region.

Now we state conditions related to the the $g_i(\cdot)$ functions and $Z_{i,N}$, where $Z_{i,N}$ represents an α-mixing random field such that $i \in D$. At this point, it is important to note that since we are working with panel data and time averages for estimation purposes, the random field considered in the assumptions is the one constructed with the time averages for each observation.

Assumption 2 (Uniform L_2 integrability)
There is an array of positive real constants $\{c_{i,N}\}$ such that

$$\lim_{k\to\infty} \sup_N \sup_{i\in D_N} \mathbb{E}\left[|Z_{i,N}/c_{i,N}|^2 \mathbb{1}\left(|Z_{i,N}/c_{i,N}| > k\right)\right] = 0$$

Where $\mathbb{1}(\cdot)$ denotes an indicator function. Note that Assumption 2 allows for the possibility of asymptotic unbounded second moments, however for the remainder of the paper we will focus on the case of bounded moments, in which case we can set $c_{i,N} = 1 \quad \forall i$. The next assumption put some restrictions on the α coefficients of the random field.

Assumption 3 (α-mixing)
Let $\bar{Q}^k_{i,N} \coloneqq Q_{|X_{i,N}/c_{i,N}|\mathbb{1}(|Z_{i,N}/c_{i,N}|>k)}$ denote the upper tail quantile function of $|Z_{i,N}/c_{i,N}|\mathbb{1}(|Z_{i,N}/c_{i,N}| > k)$ and recall that $\alpha_{\text{inv}}(u)$ is the inverse function of $\bar{\alpha}_{1,1}(m)$ as in the definition specified in the Appendix. The α-mixing coefficients satisfy:

1. $\lim_{k\to\infty} \sup_N \sup_{i\in D_N} \int_0^1 \alpha^d_{\text{inv}}(u) \left[\bar{Q}^{(k)}_{i,N}(u)\right]^2 du = 0.$

2. $\sum_{m=1}^{\infty} m^{d-1} \bar{\alpha}_{k,h}(m) < \infty$ for $k + h \leq 4$.

3. $\bar{\alpha}_{1,\infty}(m) = \mathcal{O}_p(m^{-p-\varepsilon})$ for some $\varepsilon > 0$.

Under Assumptions 2 and 3.2 with $k = h = 1$ and letting $\{D_N\}$ be a sequence of finite subsets of D that satisfies Assumption 1 such that $|D_N| \to \infty$ as $N \to \infty$, a direct application of Theorem 3 in Nazgul and Prucha (2009) leads to the conclusion that

$$\frac{1}{|D_N|} \sum_{i \in D} Z_{i,N} - \mathbb{E}(Z_{i,N}) \xrightarrow{p} 0$$

Note that one could relax Assumption 2 to L_1 uniform integrability for the theorem to hold, nevertheless, the below CLT requires L_2 uniform integrability. In order to apply this pointwise WLLN to the $g_i(\cdot, \theta)$ functions, we assume that these satisfy the regularity conditions specified in Assumption A.1 presented in the Appendix. Given the fact that any measurable function of an α-mixing process is α-mixing, the $g_i(Z_{i,N}, \theta)$ also satisfy a pointwise WLLN, i.e.

$$\frac{1}{|D_N|} \sum_{i \in D} g_i(Z_{i,N}, \theta) - \mathbb{E}[g_i(Z_{i,N}, \theta)] \xrightarrow{p} 0 \tag{2.7}$$

With this Weak Law of Large Numbers, in order for the above GMM estimator to be consistent, we need an Uniform LLN for which we need the additional regularity conditions on the $g_i(\cdot, \cdot)$ functions stated in Assumption A.2. Under these assumptions, we have the following proposition, which is a special case of Theorem 2 in Nazgul and Prucha (2009).

Proposition 1. *Let $\{D_N\}$ be a sequence of finite subsets of D that satisfies Assumption 1 such that $|D_N| \to \infty$ as $N \to \infty$ and let $Q_N(\theta) = \frac{1}{|D_N|} \sum_{i \in D_N} g_i(Z_{i,N}, \theta)$. Suppose (Θ, ν) is a compact metric space and consider a sequence of real valued functions $\{g_i(Z_{i,N}, \theta) : i \in D_N, N \in \mathbb{N}\}$ satisfying Assumption A.2 and that for all θ in Θ, these functions satisfy the WLLN in* (2.7)*. Then*

$$\sup_{\theta \in \Theta} |Q_N(\theta) - \mathbb{E}[Q_N(\theta)]| \xrightarrow{p} 0$$

With these tools at hand, define the following functions:

$$Q_N(\theta) \equiv \left[\frac{1}{N} \sum_{i=1}^{N} g_i(Z_i, \theta) \right]' \hat{\Omega} \left[\frac{1}{N} \sum_{i=1}^{N} g_i(Z_i, \theta) \right]$$

$$Q(\theta_0) \equiv \mathbb{E}[g_i(Z_{i,N}, \theta_0)]' \, \Omega_0 \, \mathbb{E}[g_i(Z_{i,N}, \theta_0)]$$

And suppose that $\hat{\Omega} \xrightarrow{p} \Omega_0$, where Ω_0 is a positive definite matrix. Recalling that $\mathbb{E}[g_i(Z_i, \theta)] = 0$ only when $\theta = \theta_0$, the true population value, the following proposition summarize the conditions under which the GMM estimator will be consistent:

Proposition 2. *Suppose that all the conditions of Proposition 1 hold. Additionally, assume that (i) $g_i(Z_i, \cdot)$ are continuous for all $\theta \in \Theta$, (ii) $\hat{\Omega} \xrightarrow{p} \Omega_0$, an $r \times r$ positive definite matrix and (iii) θ_0 is the only vector for which the moment condition in* (2.4) *holds. Then $Q_N(\hat{\theta})$ converges uniformly to $Q(\theta_0)$ and $\hat{\theta} \xrightarrow{p} \theta_0$, the unique minimizer of $Q(\theta)$.*

Note that since $\frac{1}{N}\sum_{i=1}^{N} g_i(Z_i, \theta)$ satisfies the ULLN of Proposition 1 and $\hat{\Omega} \xrightarrow{p} \Omega_0$, the proof of this proposition follows from Theorem 4.1.1 in Amemiya (1985). To obtain the asymptotic distribution of $\hat{\theta}$, we assume the following condition, which guaranties that the sum is not dominated by any term.

Assumption 4
If we define $\tilde{\sigma}_n^2 = \mathrm{Var}(S_n)$ and $S_n = \sum_{i \in D_N} Z_{i,N}$. Then the following condition is satisfied:

$$\liminf_{n \to \infty} |D_N|^{-1} \tilde{\sigma}_n^2 > 0$$

Under this assumption, Theorem 1 in Nazgul and Prucha (2009) ensures the asymptotic normality of the random variables Z_i.

Proposition 3. *Let $\{D_N\}$ be a sequence of finite subsets of D that satisfies Assumption 1 such that $|D_N| \to \infty$ as $N \to \infty$ and let $\{Z_i : i \in D_N, n \in \mathbb{N}\}$ be a sequence of zero mean real-valued random variables that satisfy Assumption 2. Furthermore, assume that the random field is α-mixing satisfying Assumption 3. Then,*

$$\tilde{\sigma}_n^{-1} S_n \xrightarrow{d} N(0, 1)$$

Once again, the previous proposition applies directly to the underlying random fields, however, we need a result to for the $g_i(Z_{i,N}, \theta)$ functions. Assuming that the latter satisfy the standard regularity conditions of Assumption A.3 , the first order conditions for the GMM estimator are

$$\left[\frac{1}{N}\sum_{i=1}^{N} \nabla_\theta g_i(Z_i, \hat{\theta})\right]' \hat{\Omega} \left[\frac{1}{N}\sum_{i=1}^{N} g_i(Z_i, \hat{\theta})\right] = 0 \tag{2.8}$$

Taking a mean value expansion of the last term around θ_0 yields the following expression:

$$g_i(\hat{\theta}) = g_i(\theta_0) + \nabla_\theta g_i(\tilde{\theta})(\hat{\theta} - \theta_0) + \text{ remainder} \tag{2.9}$$

for $\tilde{\theta}$ between $\hat{\theta}$ and θ_0 element-wise and where I suppressed the dependence of g_i on

Z_i for notation simplicity. Replacing (2.9) in (2.8) yields:

$$\sqrt{N}(\hat{\theta}-\theta_0) = -\left\{\left[\frac{1}{N}\sum_{i=1}^{N}\nabla_\theta g_i(\hat{\theta})\right]'\hat{\Omega}\left[\frac{1}{N}\sum_{i=1}^{N}\nabla_\theta g_i(,\tilde{\theta})\right]\right\}^{-1} \left[\frac{1}{N}\sum_{i=1}^{N}\nabla_\theta g_i(\hat{\theta})\right]'\hat{\Omega}\left[\frac{1}{\sqrt{N}}\sum_{i=1}^{N} g_i(\theta_0)\right] + \text{remainder} \tag{2.10}$$

Noting again that the $\nabla_\theta g_i(\theta)$ preserve the mixing conditions, then $\frac{1}{N}\sum_{i=1}^{N}\nabla_\theta g_i(\hat{\theta}) \xrightarrow{p} \mathbb{E}\left[\nabla_\theta g_i(\theta_0)\right]$ by the WLLN above. Since $g_i(\theta)$ is continuously differentiable, by Slutzky's Theorem, the first term of (2.10) converges in probability to

$$\left\{\left[\mathbb{E}(\nabla_\theta g_i(\theta_0)\right]'\Omega_0\left[\mathbb{E}(\nabla_\theta g_i(\theta_0)\right]\right\}^{-1}$$

Furthermore, by the CLT,

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} g_i(\theta_0) = \mathcal{O}_p(1)$$

Therefore, taking the probability limit of (2.10), we obtain

$$\begin{aligned}\sqrt{N}(\hat{\theta}-\theta_0) = &-\left\{\mathbb{E}\left[\nabla_\theta g_i(\theta_0)\right]'\Omega_0\mathbb{E}\left[\nabla_\theta g_i(\theta_0)\right]\right\}^{-1} \\ &\mathbb{E}\left[\nabla_\theta g_i(\theta_0)\right]'\Omega_0\left[\frac{1}{\sqrt{N}}\sum_{i=1}^{N} g_i(\theta_0)\right] + o_p(1) \\ &\xrightarrow{d} N(0, C^{-1}\Sigma C^{-1})\end{aligned} \tag{2.11}$$

where

$$C = \left\{\mathbb{E}\left[\nabla_\theta g_i(\theta_0)\right]'\Omega_0\mathbb{E}\left[\nabla_\theta g_i(\theta_0)\right]\right\}^{-1}\mathbb{E}\left[\nabla_\theta g_i(\theta_0)\right]'\Omega_0$$

and

$$\Sigma = \mathbb{E}[g_i(\theta_0)g_i(\theta_0)'] = \text{Var}[g_i(\theta_0)] \tag{2.12}$$

Note that for the cases considered in this paper, C is just a matrix of data, so we do not need to estimate it. On the other hand, we need an estimator of the variance of the moment conditions, which we present in the next section. From an empirical implementation point of view, it is important to note that this GMM framework includes the simple estimators mentioned at the beginning of the section as special cases. For example, in the case of A_{it} containing only exogenous variables, then the GMM reduces to the same solution as estimating (2.3) with Pooled OLS. If A_{it} has some endogenous variables like in the model (2.1), and assuming that we have a set of instruments Z_{it}, then the Fixed Effects 2SLS can be obtained from the GMM estimator by setting $\hat{\Omega} = \ddot{Z}'\ddot{Z}$, where Z is the stacked $NT \times r$ matrix of instruments. Furthermore, we

would need that the well known matrices of these estimators are of full column rank and to converge in probability to non-singular finite matrices.

Another empirical consideration is the specification of the weighting matrix W since in the model, the dependence of the outcome variable on other observations is generated by this matrix. In practice, there exist different ways to specify W. For example, one could assign weights as the inverse of the distance between two observations and set to zero the weights after a threshold or use a k-neighbors scheme. When dealing with geographic units, one could assign an equal weight for all the units j that share a border with unit i (rook type) or if they share an edge or a vertex (queen type) like in Figure XX, or even assign an equal weight to all other units in the sample [see LeSage and Pace (2009) for a discussion on weighting matrices].

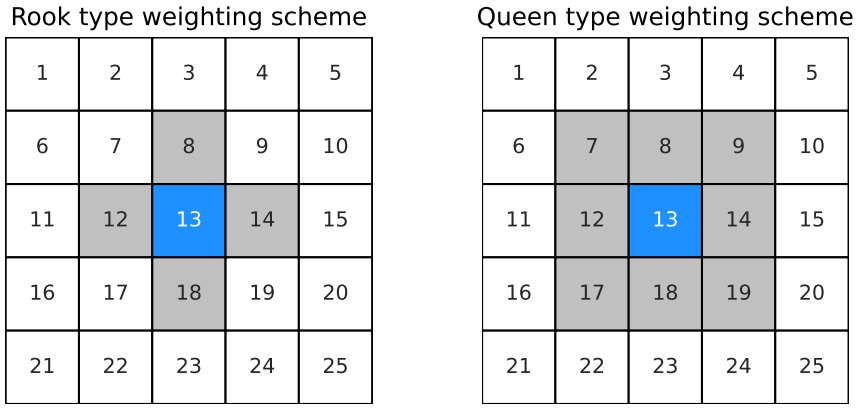

Figure 2: Rook and queen type weighting schemes. On the rook type scheme, if W is row normalized, only units 8, 12, 14 and 18 will receive a weight of $\frac{1}{4}$ in row 13. Analogously, if a queen type scheme is used, units 7, 8, 9, 12, 14, 17, 18 and 19 will have a weight of $\frac{1}{8}$ in row 13 of W.

Nonetheless, some of these specifications might violate the assumptions stated in this section. In particular, recall that we are working with an α-mixing random field, which implies that the dependence between the observations decays as they are farther apart. In this respect, it is clear that assigning an equal weight to all other observations violates this assumption. In a similar fashion, a k-neighbors pattern might not satisfy the α-mixing condition in cases where there are isolated units (e.g. a unit located alone in an island). Note that these restrictions to W also apply in cases where the distance measure is of economic nature or derived from a network perspective (e.g. degree of centrality).

3 The HACSC Estimator

To obtain robust standard errors, recall that because for each observation we took the time average of their corresponding moment conditions, essentially we are working with a cross sectional problem. The idea is therefore to apply Kelejian and Prucha's (2007) estimator of the covariance matrix in this context, for which we need consistent estimates of the error terms. Analogous to the time series literature, their estimator requires a kernel function $K(\cdot)$, which will provide weights to the covariance terms entering the sums. In principle, only the covariance of observations that are close relative to some distance measure will receive a positive weight, while observations that are far away will receive a weight of zero. In other words, this function will operationalize the weak dependence assumption between observations to the error terms. Note however that this kernel will provide weights at the cross sectional dimension and not across time. To fix ideas, the researcher will need to choose a distance ρ_b such that $\rho_b \to \infty$ as $N \to \infty$ that will play the role of the truncation lag in a time series context. The next assumption imposes additional restrictions on the kernel function.

Assumption 5
The kernel $K : \mathbb{R} \to [-1, 1]$, satisfies the following conditions:

1. $K(0) = 1$
2. $K(x) = K(-x)$
3. $K(x) = 0$ for $x > 1$
4. $|K(x) - 1| \leq c_K |x|^{\alpha_K}$, $|x| \leq 1$ for some $\alpha_K \geq 1$ and $0 < c_K < \infty$.

As pointed out by Kelejian and Prucha (2007), Assumption 5 is satisfied by many kernels such as the rectangular kernel, Bartlett, the triangular kernel, among others. The next assumption imposes some structure for the error terms.

Assumption 6
The $N \times 1$ vector of errors is generated as follows:

$$u = R\varepsilon \tag{3.1}$$

where the ε is a $N \times 1$ vector of i.i.d. random variables with mean 0, variance of 1 and $\mathbb{E}[|\varepsilon|^q] < \infty$ for $q \geq 4$ and the R is a $N \times N$ non-singular unknown matrix whose row and column sums are uniformly bounded.

In light of Assumption 6, recall that although theoretically we are working with a cross sectional problem because we took the time average of the moment conditions, the underlying structure of the data is a panel. In this sense, (3.1) can also be seen as an average, so for each i, we have:

$$u_i = \begin{pmatrix} u_{i1} \\ u_{i2} \\ \vdots \\ u_{iT} \end{pmatrix} \tag{3.2}$$

where it t-th row of u_i is:

$$u_{i,t} = \sum_{s=1}^{t} R_{i,s}\varepsilon_s$$

and $R_{i,s}$ is the i-th row of R_s, a matrix with similar properties than R defined above, at time s. This implies that in each time period, the disturbances will depend on other unit's disturbances, past own values of disturbances, and past values of other unit's disturbances. In other words, this structure allows for *both* spatial correlation and serial correlation, "spatial serial" correlation and heteroskedasticity. Nevertheless, the uniform boundedness condition for R guaranties that the correlation between units is restricted at the cross sectional dimension, analogous to the time series case. Given the distance ρ_b, we can denote with v_i the number of pseudo-neighbors for i:

$$v_i = \sum_{j=1}^{N} \mathbb{1}[\rho^*(i,j) \leq \rho_b]$$

and let $v = \max_i v_i$. In words, v_i denotes the number of units j that are at a distance less than ρ_b from unit i. The following assumption is related to v.

Assumption 7

The random variable v satisfies the following conditions:

1. $\mathbb{E}(v^2) = o_p(N^{2\tau})$, where $\tau \leq \left(\frac{1}{2}\right)\frac{q-2}{q-1}$ and q is defined in Assumption 6.
2. $\sum_{j=1}^{N} |\sigma_{ij}|\rho(i,j)^{\alpha_S} \leq c_S$, for $\alpha_S \geq 1$ and $0 < c_S < \infty$ and σ_{ij} is the (i,j)-th element of Σ (defined below).

Assumption 7 plays a role in terms of limiting the degree of correlation between units, as well as ensuring that the estimator of the covariance matrix is consistent given the fact that we are using residuals instead of errors to estimate it. Assumptions 8 and 9 provide an identification condition and bound the measurement error of the distance,

respectively.

Assumption 8

The matrix of exogenous variables, $\ddot{Z}$, has full column rank and its elements are uniformly bounded in absolute value by the finite constant $0 < c_Z < 0$. For a fixed and finite T, the matrices:

1. $\lim_{N\to\infty} (NT)^{-1} \ddot{Z}'\ddot{Z} = Q_{ZZ}$.
2. $\lim_{N\to\infty} (NT)^{-1} \ddot{Z}'RR'\ddot{Z} = Q_{ZRRZ}$.
3. $\operatorname{plim}_{N\to\infty} (NT)^{-1} \ddot{Z}'\ddot{Z} = Q_{ZZ}$.

are finite and non-singular. Furthermore, the matrix $\operatorname{plim}_{N\to\infty} (NT)^{-1} \ddot{Z}'\ddot{A} = Q_{ZA}$ has full column rank $2k$. Similarly, the diagonal elements of W are zero and all of its elements are uniformly bounded by a finite constant $0 < c_W < \infty$.

Assumption 9

The distance measure used by the empirical researcher $\rho^*(\cdot,\cdot)$ is potentially measured with error, i.e.

$$\rho^*(i,j) = \rho(i,j) + e_{ij} \geq 0$$

where $e_{ij} = e_{ji}$ denotes the measurement errors which are bounded in absolute value by the finite constant $0 < c_e < \infty$. Furthermore, $\{e_{ij}\}$ is independent of $\{\varepsilon_i\}$.

We need an additional assumption to account for the fact that we are using residuals instead of the actual error terms. This condition is provided in Assumption A.4 and should be satisfied by most $N^{\frac{1}{2}}$-consistent estimators. An extensive discussion of this and the previous assumptions is provided by Kelejian and Prucha (2007).

Note that given equations (2.4) and (2.5) and the matrix Σ specified in (2.12), we have the following:

$$\mathbb{E}[g_i(\theta_0)g_i(\theta_0)'] = \mathbb{E}\left[\ddot{Z}_i'\ddot{u}_i\ddot{u}_i'\ddot{Z}_i\right] \tag{3.3}$$

Because all the analysis is conditional on Z and W and by applying the Law of Iterated Expectations, from (3.3) and Assumption 6 we get that $\mathbb{E}(uu') = RR' = \Sigma$, where u is the $N \times 1$ vector of stacked error terms. In practical terms and recalling that $g_i(\cdot,\cdot)$ was defined as an average over time, we can estimate (3.3) by replacing the error terms by their residual counterparts and the expected value by an average applying the WLLN. Therefore, for the proposed estimator $\hat{\Sigma}$, its (r,s)-th element can be obtained as follows:

$$\hat{\Sigma}_{rs} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{l=1}^{T}\ddot{Z}_{it,r}\ddot{Z}_{jl,s}\hat{\ddot{u}}_{it}\hat{\ddot{u}}_{jl}K\left[\frac{\rho^*(i,j)}{\rho_b}\right] \tag{3.4}$$

where $\ddot{Z}_{it,r}$ is the value of the covariate r for observation i at time t, while its population counterpart is given by the following expression:

$$\Sigma_{rs} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{l=1}^{T}\ddot{Z}_{it,r}\ddot{Z}_{jl,l}\sigma_{it,jl} \tag{3.5}$$

The following proposition establishes the consistency of $\hat{\Sigma}$.

Proposition 4. *Consider the model in* (2.1) *and Assumptions 5-9 and A.4. Suppose that the* (r,s)*-th elements of* Σ *and* $\hat{\Sigma}$ *are given by* (3.5) *and* (3.4) *respectively. Then* $\hat{\Sigma} \xrightarrow{p} \Sigma$.

Given the fact that we have assumed that T is fixed from the beginning, the proof of this proposition is virtually the same as in Kelejian and Prucha (2007). Note that we can re-write (3.4) as follows:

$$\begin{aligned}\hat{\Sigma}_{rs} = \frac{1}{NT}\Bigg\{&\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{l=1}^{T}\ddot{Z}_{it,r}\ddot{Z}_{il,s}\hat{\ddot{u}}_{it}\hat{\ddot{u}}_{is}\cdot K\left[0\right] \\ &+\sum_{i=1}^{N}\sum_{i\neq j}^{N}\sum_{t=1}^{T}\sum_{l=1}^{T}\ddot{Z}_{it,r}\ddot{Z}_{il,s}\hat{\ddot{u}}_{it}\hat{\ddot{u}}_{js}K\left[\frac{\rho^*(i,j)}{\rho_b}\right]\Bigg\}\end{aligned} \tag{3.6}$$

The first term of (3.6) makes it clear that there are no restrictions imposed on the serial correlation for a particular observation, as the terms are not being down-weighted.

4 Correlated Random Effects

A direct application of the HACSC proposed in the previous section is related to the Correlated Random Effects (CRE) context. One of the most popular method applied in a panel setting is the fixed effects estimator since it allows the unobserved heterogeneity c_i to be arbitrarily correlated with the explanatory variables in the model. On the other side of the spectrum, the random effects approach imposes no correlation between c_i and the independent variables. A typical task that the empirical researcher must face is to choose between these two specifications, for which the literature has suggested multiple approaches. One of these is the CRE framework, which imposes restrictions on the distribution of the individual heterogeneity conditional on the regressors (Wooldridge, 2010).

One option is to follow Mundlak (1978) suggestion, which assumes that c_i can be

modeled as a linear function of the averages of the time varying independent variables. More specifically, consider the following model:

$$y_{it} = x_{it}\beta + c_i + u_{it}, \;\; i = 1 \dots N, \; t = 1 \dots T \tag{4.1}$$

Assuming that the x_i's are time varying, Mundlak considered the following specification:

$$c_i = \eta + \bar{x}_i\delta + e_i \tag{4.2}$$

where e_i is uncorrelated with $\bar{x}_i$ by assumption. Replacing (4.2) in (4.1) yields:

$$y_{it} = x_{it}\beta + \bar{x}_i\delta + e_i + u_{it}, \;\; i = 1 \dots N, \; t = 1 \dots T \tag{4.3}$$

Mundlak (1978) showed that estimating β in (4.3) by pooled OLS (POLS) or random effects yields the same β than estimating (4.1) by fixed effects. In addition, we can perform a Hausman-type test using this equation by testing $\delta = 0$ to determine the suitability of one estimator versus the other one. It turns out that this FE-RE equivalence carries over the spatial setting under a particular setting, namely a model such as in equation (2.1), i.e. no autoregressive process of the error term u (Li and Yang (2020) showed that the equivalence breaks if we try to model structurally). Furthermore, this result carries over to the case of endogenous variables, which is a common issue in empirical work.

More concretely, consider the model in (2.1) and using the same notation, the Fixed Effects Two Stage Least Squares (FE2SLS) coefficients can be obtained by running Pooled 2SLS on the following equation:

$$\ddot{y}_{it} = \ddot{x}_{1it}\beta_1 + \ddot{x}_{2it}\beta_2 + \ddot{w}_{1it}\gamma_1 + \ddot{w}_{2it}\gamma_2 + \rho W_i\ddot{y}_t \tag{4.4}$$

using the instrumental variables $(\ddot{z}_{2it} \;\; \ddot{\mathfrak{Z}}_{2it} \;\; \ddot{w}^2_{1it} \;\; \ddot{w}^3_{1it} \dots \ddot{w}^s_{1it})$, $s \in \mathbb{N}$. Then, it can be shown that running Pooled 2SLS on:

$$\begin{aligned} y_{it} - \eta\bar{y}_i = {} & (x_{1it} - \eta\bar{x}_{1i})\beta_1 + (x_{2it} - \eta\bar{x}_{2i})\beta_2 + (w_{1it} - \eta\bar{w}_{1i})\gamma_1 + (w_{2it} - \eta\bar{w}_{2i})\gamma_2 \\ & + \rho W_i(y_t - \eta\bar{y}) + (1-\eta)\bar{x}_{1i}\delta_1 + (1-\eta)\bar{z}_{2i}\delta_2 + (1-\eta)\bar{w}_{1i}\lambda_1 \\ & + (1-\eta)\bar{\mathfrak{Z}}_{2i}\lambda_2 + (1-\eta)\sum_{j=2}^{s}\bar{w}^j_{1i}\zeta_j \end{aligned} \tag{4.5}$$

using IV's: $\begin{aligned}[t] & [(z_{2it} - \eta\bar{z}_{2i}) \;\; (\mathfrak{Z}_{2it} - \eta\bar{\mathfrak{Z}}_{2i}) \;\; (w^2_{1it} - \eta\bar{w}^2_{1i}) \dots \; (w^s_{1it} - \eta\bar{w}^s_{1i}) \\ & (1-\eta)\bar{\mathfrak{Z}}_{2i} \;\; (1-\eta)W_i\bar{Z}_2 \;\; (1-\eta)\bar{w}^2_{1i} \dots (1-\eta)w^s_{1i}] \end{aligned}$

yields the same $(\beta_1 \;\; \beta_2 \;\; \gamma_1 \;\; \gamma_2 \;\; \rho)$ as in (4.4) and where $\eta = 1 - \left[\sigma_u^2/(\sigma_u^2 + T\sigma_c^2)\right]^{1/2}$ is assumed to be known. The following proposition summarizes this result.

Proposition 5. *Suppose* $\tilde{\Gamma} = (\tilde{\beta}_2 \quad \tilde{\beta}_2 \quad \tilde{\gamma}_1 \quad \tilde{\gamma}_2 \quad \tilde{\rho})$ *is the coefficient vector obtained by running Pooled 2SLS to equation* (4.5)*. Then* $\tilde{\Gamma} = \hat{\Gamma}_{FE2SLS}$*, the coefficient vector obtained by running Pooled 2SLS to equation* (4.4)*.*

The proof of this proposition can be found in the Appendix. Note that we have included the time averages of the instruments in (4.5), but this might introduce some distortions in the sense that the dimension of the z's might be larger than the original dimension of the x_2's. In practice, this will impact the degrees of freedom employed to perform the hypothesis testing to choose between FE and RE. Although when the cross sectional dimension is large this might not matter, in small samples this could have a significant impact in the statistical significance of the coefficients.

It is important to note that this FE-RE equivalence is an algebraic result, and as it turns out, one can obtain the FE coefficients of $(\beta \quad \gamma \quad \rho)$ in (4.5) by replacing the average of the instruments by the time averages of the predicted values of a regression of the endogenous variables on all of the exogenous variables, i.e.

$$\begin{aligned} y_{it} - \eta\bar{y}_i = {} & (x_{1it} - \eta\bar{x}_{1i})\beta_1 + (\hat{x}_{2it} - \eta\hat{\bar{x}}_{2i})\beta_2 + (w_{1it} - \eta\bar{w}_{1i})\gamma_1 + (\hat{w}_{2it} - \eta\hat{\bar{w}}_{2i})\gamma_2 \\ & + \rho W_i(\hat{y}_t - \eta\hat{\bar{y}}) + (1-\eta)\bar{x}_{1i}\delta_1 + (1-\eta)\hat{\bar{x}}_{2i}\delta_2 + (1-\eta)\bar{w}_{1i}\lambda_1 \\ & + (1-\eta)\hat{\bar{w}}_{2i}\lambda_2 + (1-\eta)W_i\hat{\bar{y}}\zeta_1 \end{aligned} \tag{4.6}$$

This will "correct" the degrees of freedom issue mentioned above, at the expense of making the asymptotic theory harder since we have to take into account that we are using the predicted values instead of the original instrument averages. Proposition 6 summarizes this result and is proved in the Appendix.

Proposition 6. *Suppose* $\check{\Gamma} = (\check{\beta}_1 \quad \check{\beta}_2 \quad \check{\gamma}_1 \quad \check{\gamma}_2 \quad \check{\rho})$ *is the coefficient vector obtained by running Pooled OLS to equation* (4.6)*, where the* $\hat{}$ *represent the linear projections of the endogenous variables on the exogenous covariates. Then* $\check{\Gamma} = \hat{\Gamma}_{FE2SLS}$*, the coefficient vector obtained by running Pooled 2SLS to equation* (4.4)*.*

Once the researcher estimates the coefficients of (4.5) or (4.6), the next natural step is to test the hypothesis $\Xi = (\delta \quad \lambda \quad \zeta) = 0$ [here ζ denotes either $(\zeta_2 \dots \zeta_s)$ in (4.5) or ζ_1 in (4.6)] to decide between FE and RE specifications. Even if model (2.1) does not have an explicit functional form for the error term, the u_{it} could still be serially or spatially correlated, therefore, we can use the HACSC estimator proposed in section 3 to conduct a fully robust Hausman-type test in a simple way. Specifically, one would need to get the Wald statistic as $\mathcal{W} = (\mathbf{R}\Xi)'(\mathbf{R}\hat{\Sigma}\mathbf{R}')^{-1}(\mathbf{R}\Xi)$, where $\mathbf{R}$ includes the set of restrictions on the coefficients, Ξ is the full set of coefficients estimated and $\hat{\Sigma}$ is the estimated HACSC robust covariance matrix.

5 Feasible GLS

As previously stated and analogous to the time series literature, it is common practice in empirical work to assume a particular structure of the error term in a spatial context. In particular, consider the following model:

$$\begin{aligned} y_t &= X_t\beta + v_t \\ v_t &= \rho W v_t + \varepsilon_t \\ \varepsilon_t &= c + u_t \end{aligned} \tag{5.1}$$

where y_t is a $N \times 1$ vector, X_t is a $N \times k$ matrix of covariates, c denotes the vector of individual heterogeneity and u_t is a vector of idiosyncratic errors at time t. In this model, X_t may contain spatial lags of the independent variables. In what follows, the conditioning on both X_t and W of all the analysis is implicit. By stacking the equations by time period, the model can be rewritten as follows:

$$\begin{aligned} y &= X\beta + v \\ v &= (\mathbf{I}_T \otimes \rho W)v + \varepsilon \\ \varepsilon &= (e_t \otimes \mathbf{I}_N)c + u \end{aligned} \tag{5.2}$$

where e_t represents a $T \times 1$ vector of ones. At this point, the researcher needs to make an assumption about the orthogonality condition between the independent variables and the composite error term and more specifically, the vector c. A typical choice is to assume that all the explanatory variables X are exogenous with respect to both vectors c and u, with each element of these being i.i.d. with zero mean and finite variances σ_c^2 and σ_u^2 respectively, and both vectors being independent from each other. Note that this working assumption is stronger than the one required to obtain the consistency of the fixed effects estimator described in previous sections (as in the rest of the paper, I assume that T is fixed and $N \to \infty$).

Given these assumptions, from (5.1) we can write $\mathbb{E}(v_t v_t')$ as follows:

$$\mathbb{E}(v_t v_t') = (\sigma_c^2 + \sigma_u^2)(\mathbf{I}_N - \rho W)^{-1}(\mathbf{I}_N - \rho W)^{-1} \tag{5.3}$$

Or using the stacked version of (5.2) instead, then we can write $\mathbb{E}(\varepsilon\varepsilon') = \Omega_\varepsilon$ in the following way:

$$\Omega_\varepsilon = \sigma_c^2(J_T \otimes \mathbf{I}_N) + \sigma_u^2 \mathbf{I}_{NT} \tag{5.4}$$

where $J_T = e_t e_t'$. Therefore it follows that,

$$\mathbb{E}(vv') = \left[\mathbf{I}_T \otimes (\mathbf{I}_N - \rho W)^{-1}\right] \left[\sigma_c^2(J_T \otimes \mathbf{I}_N) + \sigma_u^2 \mathbf{I}_{NT}\right] \left[\mathbf{I}_T \otimes (\mathbf{I}_N - \rho W)^{-1}\right] \tag{5.5}$$

Note that the middle of this matrix has a classic random effects structure. In order to compute this covariance matrix, it is assumed that the matrix $(\mathbf{I}_N - \rho W)$ is invertible and that $|\rho| < 1$ just as in the previous sections. Following the time series case and to facilitate the computation of the middle of (5.5), note that

$$\Omega_\varepsilon = \sigma_u^2 Q_0 + \sigma_1^2 Q_1 \tag{5.6}$$

where $Q_0 = \left(\mathbf{I}_T - \frac{J_T}{T}\right) \otimes \mathbf{I}_N$, $Q_1 = \frac{J_T}{T} \otimes \mathbf{I}_N$ and $\sigma_1^2 = \sigma_u^2 + T\sigma_c^2$. Noting that Q_0 and Q_1 are idempotent, symmetric, $Q_0 + Q_1 = \mathbf{I}_{NT}$ and that $Q_0 Q_1 = \mathbf{0}_{NT}$, it follows that $\Omega_\varepsilon^{-1} = \sigma_u^{-2} Q_0 + \sigma_1^{-2} Q_1$ and $\Omega_\varepsilon^{-\frac{1}{2}} = \sigma_u^{-1} Q_0 + \sigma_1^{-1} Q_1$. In short, if the researcher is willing to impose that the covariates are orthogonal to the individual heterogeneity vector c and the error term in (5.2) follows a spatial AR(1) process, then the matrix $\mathbb{E}(vv')$ will have a particular form that depends only on three parameters.

Knowing this, one can obtain an estimator that is potentially more efficient than the FE estimator. More specifically, the researcher can exploit the structure of the error term in (5.2) to remove the spatial correlation by performing a spatial Cochrane-Orcutt type transformation. Let

$$\begin{aligned} y^* &= y - (\mathbf{I}_T \otimes \rho W) y \\ X^* &= X - (\mathbf{I}_T \otimes \rho W) X \\ v^* &= v - (\mathbf{I}_T \otimes \rho W) v \end{aligned}$$

Therefore, the transformed model is

$$y^* = X^* \beta + v^* \tag{5.7}$$

Note that $v^* = \varepsilon$ so that (5.7) contains a classical composite error term. Given the structure of ε, we can perform a second transformation by multiplying (5.7) by $\Omega_\varepsilon^{-\frac{1}{2}}$ to obtain

$$\check{y} = \check{X}\beta + \check{\varepsilon} \tag{5.8}$$

where $\check{y} = \Omega_\varepsilon^{-\frac{1}{2}} y^*$ and similarly for the rest of the terms. Note that

$$\begin{aligned} \mathbb{E}\left(\check{\varepsilon}\check{\varepsilon}'\right) &= \Omega_\varepsilon^{-\frac{1}{2}} \mathbb{E}(\varepsilon\varepsilon') \Omega_\varepsilon^{-\frac{1}{2}} \\ &= (\sigma_u^{-1} Q_0 + \sigma_1^{-1} Q_1)(\sigma_u^2 Q_0 + \sigma_1^2 Q_1)(\sigma_u^{-1} Q_0 + \sigma_1^{-1} Q_1) \\ &= Q_0 + Q_1 \\ &= \mathbf{I}_{NT} \end{aligned} \tag{5.9}$$

Hence (5.8) can be estimated by Pooled OLS to obtain a GLS-type estimator to obtain efficiency gains, denoted by β_{GLS}. If all the relevant matrices are well behaved as $N \to \infty$ and non-singular, Kapoor et al. (2007) showed that

$$(NT)^{\frac{1}{2}}\left(\hat{\beta}_{GLS}-\beta\right)\xrightarrow{d}N(0,\Psi)\text{ as }N\to\infty \tag{5.10}$$

where $\Psi=\left(\sigma_u^2M_{XX}^0+\sigma_1^2M_{XX}^1\right)^{-1}$ and $M_{XX}^j=\lim_{N\to\infty}\frac{1}{NT}X^{*\prime}Q_jX^*$ for $j=0,1$. The previous analysis requires knowledge of σ_c^2,σ_u^2 and ρ and therefore it is not feasible. Kapoor et al. (2007) proposed generalized moments estimators of these parameters and they showed that if $\hat{\beta}_{FGLS}$ is the Pooled OLS estimator of (5.8) using *any* consistent estimators $\hat{\sigma}_c^2,\hat{\sigma}_u^2$ and $\hat{\rho}$ instead of σ_c^2,σ_u^2 and ρ, then

$$(NT)^{\frac{1}{2}}\left(\hat{\beta}_{GLS}-\hat{\beta}_{FGLS}\right)\xrightarrow{p}0\text{ and }\hat{\Psi}-\Psi\xrightarrow{p}0 \tag{5.11}$$

where $\hat{\Psi}=\left(\frac{1}{NT}\hat{X}^{*\prime}\hat{\Omega}_\varepsilon^{-1}\hat{X}^*\right)^{-1}$, provided that the working assumptions used to derive (5.10) hold. Note that the hats over the components of $\hat{\Psi}$ denote the dependence of the terms on $\hat{\sigma}_c^2,\hat{\sigma}_u^2$ and $\hat{\rho}$.

The validity of the previous covariance matrix Ψ rests on the working assumptions that the error term v follows a spatial AR(1) and the conditions imposed on each element of c and u hold. However, from an empirical perspective it is always possible that the structure of Ω_ε does not have the RE form due to the presence of heteroskedasticity or serial correlation on u_i for example. It is important to stress out that even if Ω_ε does not have the same structure as in (5.4), $\hat{\beta}_{FGLS}$ remains consistent, provided that the strict exogeneity condition (more formally this would mean that $\mathbb{E}[X\otimes c]=0$ and $\mathbb{E}[X\otimes u]=0$) and the corresponding rank condition continue to hold.

Nevertheless, if the researcher is unsure about the assumptions related to the vectors of individual heterogeneity c or the idiosyncratic errors u made in this section, it is wise to make robust inference. In these instances, the HACSC estimator presented in this paper can be useful to achieve this purpose. More specifically, consider the residuals

$$\ddot{\check{\varepsilon}}_t=\check{y}_t-\check{X}_t\hat{\beta}_{FGLS},\quad t=1\dots T.$$

where $\hat{\beta}_{FGLS}$ is obtained by estimating (5.8). In this context, the (r,s)-th element of the middle of the robust covariance matrix is

$$\hat{\Sigma}_{rs}=\frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{l=1}^{T}\check{X}_{it,r}\check{X}_{jl,s}\ddot{\check{\varepsilon}}_{it}\ddot{\check{\varepsilon}}_{jl}K\left[\frac{\rho^*(i,j)}{\rho_b}\right] \tag{5.12}$$

And the fully robust covariance matrix is:

$$\check{\Psi}=\left(\check{X}'\check{X}\right)^{-1}\left\{\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{l=1}^{T}\ddot{\check{\varepsilon}}_{it}\ddot{\check{\varepsilon}}_{jl}\check{X}_{it}'\check{X}_{jl}K\left[\frac{\rho^*(i,j)}{\rho_b}\right]\right\}\left(\check{X}'\check{X}\right)^{-1} \tag{5.13}$$

where $\check{X}_{it}$ is the $1\times k$ vector of covariates at time t for observation i. Note that

the computation of $\check{\Psi}$ requires the use of the transformed variables and not the original ones, which is consistent with the *estimating* equation (5.8). As in the previous sections, the kernel function $K(\cdot)$ will provide weights so that the (possible) spatial correlation decreases for observations that are far apart according to the distance measure $\rho(\cdot,\cdot)$. Naturally, $\check{\Psi}$ will be valid whether the RE structure of Ω_ε holds or not and will be robust to arbitrary serial and spatial correlation, as well as heteroskedasticity.

Throughout this section we have assumed that all the elements of the explanatory variables are uncorrelated with the error term u. If some elements of X are endogenous (i.e. $\mathbb{E}[x_{it}' u_{it}] \neq 0$) and the researcher has available instruments Z, then the extension to an IV procedure is straightforward as discussed in Mutl and Pfaffermayr (2010) and Baltagi and Liu (2011). The estimation approach would be to apply Pooled 2SLS to the estimating equation 5.8 using instruments $\check{Z}$, where the $\check{}$ denotes the same transformations made earlier in the section. In this instance, the computation of the covariance matrix using the HACSC estimator would look like (3.4), but the researcher would need to use the transformed variables as in this section instead.

6 Alternative estimation: a Control Function Approach

It is well known that Instrumental Variables estimation procedures such as 2SLS deliver consistent estimates of the parameters at the expense of losing precision when compared to OLS as pointed out by Cameron and Trivedi (2005). In such instances, if the researcher is willing to impose additional assumptions, she can resort to the control function approach (Blundell and Powell 2003), which can deliver estimates that are (potentially) more efficient as it will be shown in simulations. Consider the model shown in (6.1), which is very similar to (2.1) but without the spatial lag of the dependent variable on the right hand side[5], which will allow us to interpret it as a conditional mean function and for simplicity we will assume that there's only one element in x_{2it}:

$$\begin{aligned} y_{it} &= x_{1it}\beta_1 + x_{2it}\beta_2 + W_i X_{1t}\gamma_1 + W_i X_{2t}\gamma_2 + c_i + u_{it} \\ &= x_{it}\beta + W_i X_t \gamma + c_i + u_{it}, \;\; i = 1 \dots N, \; t = 1 \dots T \end{aligned} \tag{6.1}$$

where the definitions are the same as in Section 1. By applying the within transformation, we obtain the estimating equation:

$$\ddot{y}_{it} = \ddot{x}_{it}\beta + W_i \ddot{X}_t \gamma + \ddot{u}_{it} \tag{6.2}$$

[5]It is certainly possible to use the control function approach with the spatial lag of the dependent variable as a covariate

As with the 2SLS case and using obvious notation, this approach also requires the availability of a set of instruments $\ddot{Z}_{it} = (\ddot{x}_{1it} \;\; \ddot{w}_{1it} \;\; \ddot{z}_{2it} \;\; \ddot{\mathfrak{Z}}_{2it})$. The first two assumptions of the Control Function (CF) approach are the same as with 2SLS, namely: $\mathbb{E}(\ddot{Z}_{it}\ddot{u}_{jt}) = 0$ for $i, j = 1 \ldots N$ and $t = 1 \ldots T$ and the identification condition $\text{rank}[\mathbb{E}(\ddot{Z}'\ddot{A})] = 2k-1$. The first stage of the estimation involves the reduced form of the endogenous variable on the instruments and obtaining the disturbances $\ddot{v}_{2it}$, i.e.

$$\ddot{v}_{2it} = \ddot{x}_{2it} - \ddot{Z}_{it}\psi \tag{6.3}$$

where $\mathbb{E}(\ddot{Z}'_{it}\ddot{v}_{it}) = 0$. Given that $\mathbb{E}(\ddot{Z}'_{it}\ddot{u}_{it}) = 0$, note that $\ddot{x}_{2it}$ and $\ddot{w}_{2it}$ are endogenous if and only if $\ddot{u}_{it}$ is correlated with $\ddot{v}_{2it}$ and $W_i\ddot{v}_{2t}$. At this point we state the additional assumption required by the CF approach:

$$\mathbb{E}(\ddot{u}_{it}|Z, X_2, W) = \mathbb{E}(\ddot{u}_{it}|Z, \ddot{v}_2, W) = \mathbb{E}(\ddot{u}_{it}|\ddot{v}_2, W) = \mu_1\ddot{v}_{2it} + \mu_2 W_i\ddot{v}_{2t} \tag{6.4}$$

This equation has two strong implicit restrictions. First, the second equality would hold under independence of Z and $(\ddot{u}, \ddot{v}_2, W)$ and second, we are assuming a linear conditional expectation of $\ddot{u}_{it}$ on the parameters. Given this, we can write

$$\ddot{u}_{it} = \mu_1\ddot{v}_{2it} + \mu_2 W_i\ddot{v}_{2t} + \ddot{e}_{it} \tag{6.5}$$

Replacing (6.5) in (6.2) yields:

$$\ddot{y}_{it} = \ddot{x}_{it}\beta + W_i\ddot{X}_t\gamma + \mu_1\ddot{v}_{2it} + \mu_2 W_i\ddot{v}_{2t} + \ddot{e}_{it} \tag{6.6}$$

Stacking again all the explanatory variables into a matrix A and the coefficients into a vector θ yields:

$$\ddot{y}_{it} = \ddot{a}_{it}\theta + \ddot{e}_{it} \tag{6.7}$$

The error term in (6.7) is uncorrelated with the rest of variables in the equation (including $\ddot{x}_{2it}$ and $\ddot{w}_{2it}$), so the parameters can be consistently estimated using Pooled OLS by replacing the disturbances with the computed residuals from the first stage. Therefore, the estimating equation for the main model becomes:

$$\ddot{y}_{it} = \hat{\ddot{a}}_{it}\theta + \ddot{e}_{it} \tag{6.8}$$

where the $\hat{}$ denotes that we are using generated regressors. Two important observations from equation (6.6) is that by including both $\ddot{v}_{2it}$ and $W_i\ddot{v}_{2t}$, the parameters obtained from this estimation will be numerically the same as 2SLS.[6] Second, *if* $\mu_2 = 0$, then it would be enough to include only $\ddot{v}_{2it}$ in the estimating equation to get consistent estimates of θ and in this scenario, they would be different than 2SLS'. Furthermore, it

[6] In this sense, we do not get any efficiency gains compared to 2SLS by including both terms.

is precisely by excluding $W_i\ddot{v}_{2it}$ from the estimation that the CF would probably more efficient than 2SLS in this case, as it would be using additional information from this restriction.

From this point onward, one has to decide how to deal with the error term. One option is to impose some structure to it and apply a Feasible GLS procedure in order to obtain further efficiency gains. Note that this is possible because in (6.4) we have conditioned on the whole set of exogenous variables and the weighting matrix. However, this would not be possible if we slightly modify the model. So far we have assumed that the model also contains spatial spillovers of the endogenous variable $\ddot{x}_{2it}$, but suppose that for some theoretical reason, the model does not include $W_i\ddot{x}_{2it}$. In this case we could relax (6.4) to

$$\mathbb{E}(\ddot{u}_{it}|Z_{it}, x_{2it}) = \mathbb{E}(\ddot{u}_{it}|Z_{it}, \ddot{v}_{2it}) = \mathbb{E}(\ddot{u}_{it}|\ddot{v}_{2it}) = \mu_1\ddot{v}_{2it} \tag{6.9}$$

Note that we are now conditioning only on the own control function. In this instance one could still estimate the transformed model by Pooled OLS, however it would preclude to apply a Feasible GLS procedure because the strict spatial exogeneity assumption would be violated since it will involve the weighting matrix W and the error terms of other observations.

Alternatively, the researcher can treat the error term non-parametrically and apply the HACSC estimator proposed in this paper to obtain robust standard errors. Nevertheless, in this case there's an additional layer of complication on top of the spatio-temporal correlation and the heteroskedasticity: by including $\hat{\ddot{v}}_{2it}$ in the estimating equation, we now have a generated regressor and therefore, the covariance matrix of the parameters needs to be adjusted to take into account the sampling error induced by the first stage estimation (i.e. we are getting *estimates* of ψ). Although Basile et al. (2014) recommends to perform a bootstrap to obtain the standard errors in a CF setup, sampling with spatially dependent data is not a trivial matter so having a formula is useful in practice.

In this setup, the fully robust covariance matrix is

$$B^{-1}MB^{-1} \tag{6.10}$$

where
$B = \mathbb{E}\left(\sum_i^N \sum_t^T \ddot{a}'_{it}\ddot{a}_{it}\right)$.
$M = \text{Var}\left[\sum_i^N \sum_t^T (\ddot{z}_{it}\psi)'(\ddot{e}_{it} + \ddot{v}_{it}\theta) - G \cdot r_{it}(\psi)\theta\right] = \text{Var}\left[\sum_i^N \sum_t^T m_{it}\right]$.
$G = \mathbb{E}\left[\sum_i^N \sum_t^T (\ddot{z}_{it}\psi)'\ddot{z}_{it}\right]$
$r_{it}(\delta) = \left(\frac{1}{NT}\sum_i^N \sum_t^T \ddot{z}'_{it}\ddot{z}_{it}\right)^{-1}\left[(NT)^{-\frac{1}{2}}\sum_i^N \sum_t^T \ddot{z}'_{it}\ddot{v}_{it}\right]$.

The derivation of (6.10) can be found in the Appendix. To estimate it, we can replace the population quantities by their sample analogues so that
$\hat{B} = \frac{1}{NT}\sum_i^N \sum_t^T \hat{\ddot{a}}'_{it}\hat{\ddot{a}}_{it}$.
$\hat{m}_{it} = (\ddot{z}_{it}\hat{\psi})'(\hat{\ddot{e}}_{it} + \hat{\ddot{v}}_{it}\hat{\theta}) - \hat{G}\cdot\hat{r}_{it}(\hat{\psi})\hat{\theta}$.
With these quantities calculated, the (r,s)-th element of M can be estimated as

$$\hat{M}_{rs} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{l=1}^{T}\hat{\ddot{m}}_{it,r}\hat{\ddot{m}}_{jl,s}K\left[\frac{\rho^*(i,j)}{\rho_b}\right]$$

Note that (6.10) also has a sandwich type form, very similar to the HACSC estimator presented earlier. Similarly, the kernel function is also used to operationalize the weak spatial dependence assumption, however in this case the terms it multiplies (m_{it} instead of $\ddot{z}'_{it}\ddot{u}_{it}$) have a different structure to take into account the first stage sampling error.

7 Simulations

7.1 Design

To test the performance of the HACSC estimator and the CF version of it, I performed a Monte Carlo study. In this experiment, the units of observation live in a squared regular grid of 20×20 and the distance between two adjacent individuals is normalized to one. To evaluate the performance of the estimator, consider the following data generating process:

$$\begin{aligned}
y_{it} &= \beta_0 + x_{1it}\beta_1 + x_{2it}\beta_2 + x_{1it}x_{2it}\beta_3 + c_i + u_{it}\\
x_{1it} &= \delta_0 + \delta_1 z_{1it} + v_{it}\\
c_i &= (I - \rho W)_i^{-1}C\\
u_{it} &= \alpha v_{it} + e_{it}\\
e_t &= (I - \rho W)^{-1}a_t\\
a_{it} &= \psi a_{i,t-1} + \varepsilon_{it}\\
&\mathbb{E}(x_{1it}u_{it}) \neq 0, \mathbb{E}(x_{2it}c_i) \neq 0, \mathbb{E}(z_{1it}c_i) \neq 0
\end{aligned}$$

where $[\beta_0 \;\; \beta_1 \;\; \beta_2 \;\; \beta_3]' = [2 \;\; 0.7 \;\; 0.6 \;\; 0.3]'$ and ε_{it}, and C are independent and identically distributed random variables following normal distributions and are independent from each other. z_{1it} is an instrument for x_{1it} and x_{2it} is exogenous with respect to the error term u_{it} and they follow a normal and gamma distributions respectively. Note that there is an interaction term between the endogenous and exogenous variable, for which we have a readily available instrument, $z_{1it}x_{1it}$.

In this setup, the error term u_{it} satisfy the CF assumption given that it depends

linearly on the error term from the reduced-form equation, v_{it}. The error terms e and a follow a spatial and temporal AR(1) process respectively. The strength of the spatial correlation is governed by the parameter ρ, while the persistence of the serial correlation is moderated by ψ. Note also that the individual heterogeneity also follows a Spatial AR(1) model, however, since I am going to apply the within transformation for the estimation, its DGP does will not affect the results.

For the weighting matrix W, I used a rook-type weighting scheme so that each observation will have between two and four pseudo-neighbors and each of those will have an equal weight. W is row-normalized to ensure that $(I - \rho W)$ is invertible. I estimated the model using both FE 2SLS and the CF approach with $N = 400$ and $T = 5$ using 1,000 replications. I am interested in comparing the estimates of the coefficients by the two methods to see if there are some efficiency gains by using the CF approach. Furthermore, I also want to evaluate the performance of four different estimators of the covariance matrix: the HACSC proposed in this paper, a SHAC assuming no serial correlation, the cluster robust and the "regular" ones without any adjustment. In the case of the CF approach, I will compare the standard errors presented in Section 6 that account for the first stage and a HACSC that ignores the two-step procedure.

I conducted a simulation for every combination of $\rho = [0,\ 0.3,\ 0.7]$ and $\psi = [0,\ 0.3,\ 0.7]$. I used the Bartlett Kernel to perform the analysis, contrary to Kelejian and Prucha (2007), who used the Parzen Kernel. An important parameter in this experiment is the threshold distance ρ_b at which the Kernel will assign a zero weight for units that are apart by more than ρ_p. Following the recommendation of the authors mentioned above, I set $\rho_b = N^{\frac{1}{4}}$, i.e. the integer part of $N^{\frac{1}{4}}$. At each iteration, I draw a new set of covariates and keep it fixed across the iterations of the ρ and ψ parameters.

7.2 Results

This section describes the results of the simulations using two metrics for the estimated coefficients: the mean and the corresponding standard deviation across the 1,000 replications for different values of ρ and ψ. Table 7.1 presents the outcomes of this experiment and it shows that both estimators provided unbiased estimates of the parameters in the sense that the average of the estimated coefficients is centered around the true values for any combination of ρ and ψ. This is expected since in this exercise the CF assumption is true.

However, when analyzing the standard deviations, the CF consistently shows a lower value than 2SLS (e.g. 0.049 against 0.084 for β_3 when $\rho = \psi = 0.3$). Figure 3 exemplifies this finding: note that the distribution of the estimated parameters is tighter around the true value for the CF estimates compared to 2SLS'. Therefore, whenever the CF assumption holds, this estimator seems to be more efficient, which can be explained

by the fact that we are using additional information when performing the estimation. Interestingly, these efficiency gains are more evident for β_1 and β_3, the coefficients associated with the endogenous variables, whereas for the coefficient of the exogenous covariate β_2, the differences between the standard deviations of both estimators are more modest across all pairs of ρ and ψ.

Table 7.1: Average estimated coefficients and standard deviation across the 1000 replications using a rook type weighting matrix, N=400 and T=5.

		β_1		β_2		β_3	
		CF	2SLS	CF	2SLS	CF	2SLS
ρ	ψ						
0.0	0.0	0.704	0.698	0.606	0.604	0.298	0.300
		(0.196)	(0.294)	(0.269)	(0.283)	(0.049)	(0.089)
	0.3	0.696	0.692	0.595	0.593	0.300	0.301
		(0.188)	(0.269)	(0.254)	(0.272)	(0.047)	(0.080)
	0.7	0.703	0.705	0.600	0.601	0.300	0.299
		(0.18)	(0.266)	(0.250)	(0.265)	(0.045)	(0.079)
0.3	0.0	0.690	0.683	0.589	0.584	0.303	0.305
		(0.207)	(0.301)	(0.275)	(0.299)	(0.052)	(0.090)
	0.3	0.706	0.718	0.603	0.608	0.299	0.294
		(0.191)	(0.281)	(0.276)	(0.294)	(0.049)	(0.084)
	0.7	0.704	0.697	0.599	0.595	0.299	0.301
		(0.189)	(0.288)	(0.254)	(0.282)	(0.048)	(0.085)
0.7	0.0	0.695	0.698	0.573	0.575	0.302	0.301
		(0.236)	(0.330)	(0.352)	(0.371)	(0.057)	(0.095)
	0.3	0.691	0.693	0.580	0.581	0.303	0.301
		(0.231)	(0.334)	(0.335)	(0.360)	(0.054)	(0.095)
	0.7	0.704	0.693	0.603	0.600	0.300	0.303
		(0.221)	(0.309)	(0.317)	(0.336)	(0.054)	(0.089)

To analyze the performance of the HACSC estimator, I use two metrics: first the average of the variance[7] estimated for each coefficient for each pair of ρ and ψ across the 1,000 replications and I compare it with the "true value", which is computed as the variance of the set of estimated coefficients for each pair of ρ and ψ across the 1,000 replications. Tables A1-A3 present this comparison and the first thing to note in

[7]I used the estimated variances instead of the standard errors because the non linearity of the square root function could affect the results.

the case of the CF is that both estimated variances, with and without the first stage correction, are very close to the true value so at first glance, using this metric the correction does not seem to make an impact.

For the 2SLS estimator, the differences are more substantial. The HACSC estimator is consistently closer to the true value across all pairs of *rho* and *psi* compared to the SHAC that imposes no serial correlation and the non-robust one. In general, the variance estimated with the HACSC is on average larger compared to the one computed with these two alternatives. Admittedly, in this case the cluster-robust variances are also very close to the true value. Overall these results suggest making the standard errors robust to spatial correlation at the expense of imposing no serial correlation can result in unreliable inference. Furthermore, as shown in Figure A1, using the HACSC estimator will provided standard errors that are, on average, properly centered around the true value.

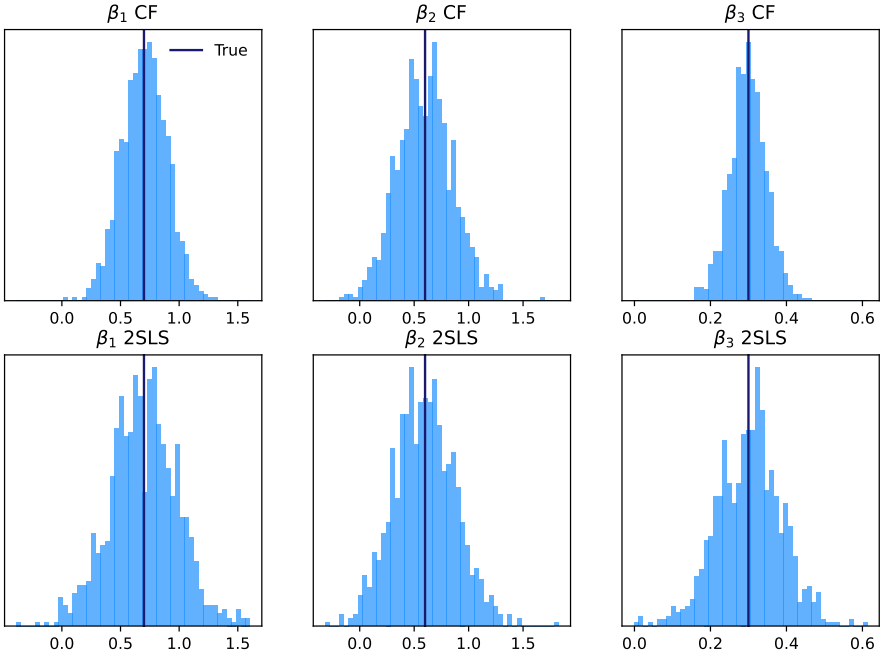

Figure 3: Distribution of coefficients estimated by 2SLS and the Control Function approach for $\rho = 0.3$ and $\psi = 0.7$ using a rook type weighting matrix.

As a second method to analyze the HACSC in this setup, I tested the null hypothesis $H_0 : \beta_3 = 0.3$ at a 5% of significance using a t-test over the 1,000 replications using the standard errors computed with the different estimators and I obtained the rejection probabilities. Using this metric, an estimator is performs better if the rejection probability is closer to 5%. Table 7.2 presents the results of this exercise.[8]

[8]Tables A4 and A5 show the results for $H_0 : \beta_1 = 0.7$ and $H_0 : \beta_2 = 0.6$ respectively.

For the case of the CF approach, the rejection probabilities using the adjustment are slightly closer to 5% compared to the estimator that ignores the first stage so in this sense, the adjustment seems important to obtain more reliable inference if the researcher uses the CF approach. On the other hand, if we use 2SLS to estimate the coefficients, the HACSC estimator rejection probabilities are closer to the 5% compared to the SHAC and non-robust standard errors, which are over rejecting the null hypothesis. Using this metric, the cluster-robust standard errors seem to perform just as well as the HACSC estimator.

Table 7.2: Rejection probabilities for the null hypothesis $H_0 : \beta_3 = 0.3$ at a 5% of significance using a t-test over the 1,000 replications with a rook type weighting matrix, N = 400, T=5.

ρ	ψ	CF	CF_no1	HACSC	SHAC	Cluster	Non-Robust
	0.0	0.050	0.060	0.067	0.088	0.058	0.082
0.0	0.3	0.046	0.060	0.054	0.072	0.046	0.068
	0.7	0.045	0.061	0.050	0.075	0.043	0.072
	0.0	0.045	0.061	0.068	0.096	0.058	0.091
0.3	0.3	0.050	0.064	0.047	0.074	0.040	0.067
	0.7	0.051	0.062	0.068	0.085	0.058	0.080
	0.0	0.050	0.072	0.057	0.077	0.048	0.066
0.7	0.3	0.041	0.057	0.066	0.095	0.065	0.090
	0.7	0.044	0.060	0.056	0.076	0.041	0.073

CF is the HACSC estimator using the first stage correction and CF_no1 refers to the HACSC estimator ignoring the first stage estimation using a CF approach.

Overall, the results suggest that the HACSC estimator, both in the case of 2SLS and the CF approach with the correction, provide more reliable inference compared to the existing SHAC.

8 Empirical Application

To test the performance of the HACSC estimator with real world data, I revisit the problem of analyzing the effect of spending on the educational outcome of fourth graders in Michigan studied by Papke and Wooldridge (2008) using district level data from 1993 to 2001[9]. In short, Michigan changed the way schools were funded in 1994, going from a property-tax based system to a statewide system, which was possible trough an increase in the sales tax and lottery profits .

To measure the effect of spending on the academic achievement of students, the authors used as the dependent variable the fraction of fourth-graders that passed the math test (math4_{it}) of the Michigan Education Assessment Program (MEAP) given that the definition of this subject and the way it is evaluated has remained relatively constant over time. On the other hand, in addition to the current level of spending on a student, the authors also allow for the possibility that the level spending on the previous three years to play a role in the test scores. This is indeed a sensible choice given that one could argue that the previous years of education lay the foundations in the learning process of students.

The model also includes the proportion of students eligible for the free and reduced-price lunch program (lunch_{it}), the district enrollment (enroll_{it}) and time dummies. More details about the full model can be found in Papke (2005). Borrowing their notation, the estimated model is:

$$\text{math4}_{it} = \theta_t + \beta_1 \log(\text{avgrexp}_{it}) + \beta_2 \text{lunch}_{it} + \beta_3 \log(\text{enroll}_{it}) + c_{i1} + u_{it} \tag{8.1}$$

where avgrexp_{it} denotes the simple average of real spending from the current and previous three years. It is important to note that in addition to the linear probability model (LPM), Papke and Wooldridge (2008) also estimate the model with other non linear estimators but because they find that the LPM is a good approximation to the non linear estimates and since this paper focuses on linear models, we will compare the results only with their LPM results.

In order to replicate their results and use the HACSC estimator, we need a distance measure between the school districts. As mentioned in previous sections, this is not a trivial matter when we are working with geographical units but in this case, we will work with the geographic distance between the centroids of each district.[10] However, there have been changes in the school districts since 2001, which is why I could only use 98.6% of the original sample used by Papke and Wooldridge (2008). The main reason for this is that some districts have merged with others and in these cases, I used the data point of the district that absorbed the one disappearing. Table 8.1 compares the summary statistics from the original and new data sets and the t-tests show that there are no statistically significant differences between them.

[9] I want to thank Dr. Papke and Dr. Wooldridge for kindly sharing their data set.

[10] Roughly speaking, a centroid can be interpreted as the center of mass of a geometry.

Table 8.1: Sample means (standard deviations) of the original and new data sets and corresponding t-tests (p-values).

	1995			2001		
	Original	New	t-test	Original	New	t-test
Pass rate on fourth-grade math test	0.62 (0.13)	0.62 (0.13)	-0.30 (0.76)	0.76 (0.13)	0.76 (0.12)	-0.43 (0.67)
Real expenditure per pupil (2001$)	6329 (986)	6317 (978)	0.20 (0.85)	7161 (933)	7147 (916)	0.25 (0.80)
Real foundation grant (2001$)	5962 (1031)	5959 (1035)	0.05 (0.96)	6348 (689)	6347 (692)	0.03 (0.98)
Fraction of eligible for free and reduced lunch	0.28 (0.15)	0.28 (0.15)	0.27 (0.79)	0.31 (0.17)	0.30 (0.17)	0.34 (0.73)
Enrollment	3076 (8156)	3099 (8210)	-0.04 (0.97)	3078 (7293)	3103 (7341)	-0.05 (0.96)
Number of observations	501	494	-	501	494	-

As a first step, I estimate model (8.1) using the Fixed Effects estimator assuming that all the independent variables are exogenous with respect to the error term u_{it}. Table 8.2 shows the estimates using the new data set and the ones reported by Papke and Wooldridge (2008). The coefficient associated with the average real expenditure is virtually the same, whereas the ones of lunch and the enrollment are negative with the new estimates. Nevertheless, the magnitudes of the latter are small and none of them are statistically significant in the original estimation either.

Table 8.2: Estimates assuming that all the explanatory variables are exogenous.

	Original results	New results	Standard errors for new results with different bandwidth values						
	Coefficient	Coefficient	ρ_b=1	ρ_b=100	ρ_b=200	ρ_b=300	ρ_b=400	ρ_b=500	ρ_b=600
log(avgrexp)	0.377 (0.071)	0.372 (0.071)	0.070	0.072	0.067	0.066	0.066	0.063	0.058
lunch	-0.042 (0.073)	0.029 (0.064)	0.064	0.077	0.079	0.072	0.061	0.060	0.061
log(enroll)	0.002 (0.049)	-0.02 (0.048)	0.048	0.045	0.033	0.028	0.026	0.023	0.022
Number of districts	501	493	-	-	-	-	-	-	-

Table 8.2 also shows the standard errors computed with the HACSC estimator using different bandwidth values. As expected and because the minimum distance between any two school districts in the data set is 1.05 kilometers, when the bandwidth is 1 kilometer the HACSC estimator is effectively treating the observations as if they have no effect on their neighbors (i.e. no spatial correlation) and consequently the standard errors are very similar to the ones computed

using an estimator that is robust to heteroskedasticity and serial correlation. Interestingly, as the bandwidth increases, the standard error for each coefficients behaves differently: for the average spending, it first increases and then decreases, for enrollment it decreases monotonically whereas for lunch, there is not an evident pattern. Note that this exercise shows that even if the covariance matrix is robust to heteroskedasticity, spatial and serial correlation, this does *not* mean that the standard errors will be necessarily larger.

One of the issues with the estimates previously discussed is that the spending from a school district might be endogenous, mainly due to the fact that a school district might adjust its current spending if they suspect that the (bad) performance of a cohort throughout the year will be reflected on the pass rates of the MEAP test (Papke and Wooldridge 2008). Fortunately, the change in the way that school districts brought with it a natural instrument: in the 1993/1994 school year, each district started to receive a per-student "foundation grant" based on the initial funding in 1994 that sought to increase the spending per student to a baseline level and had the effect of reducing the differences in spending between the districts across the state of Michigan by the year of 2001 (see Figure 4). The details of why this is a suitable instrument are discussed in Papke and Wooldridge (2008), but in broad terms, the identification assumption is that the idiosyncratic error term has a smooth relationship with both the dependent variable and the initial funding. On the other hand, the foundation grant depended on the initial funding in a non-smooth way [see Table 1 in Papke and Wooldridge (2008) to see this].

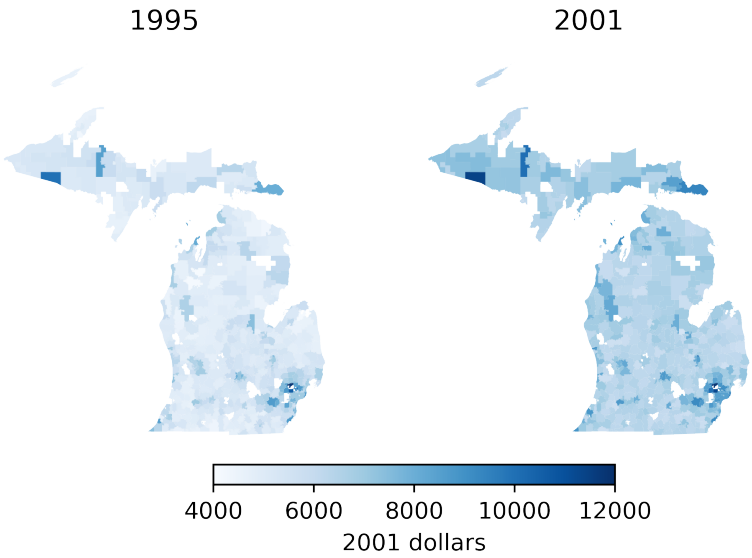

Figure 4: Average real expenditure per student across the Michigan school districts in 1995 and 2001.

As a result of this concern, Papke and Wooldridge (2008) augmented the model by also including the real spending from 1994 with interactions with the time dummies, along with the time averages of lunch and enrollment, using as instruments the foundation grant interacted

with the year binary variables. The new estimated model using instrumental variables is then

$$\begin{aligned} \text{math4}_{it} = \theta_t + \beta_1 \log(\text{avgrexp}_{it}) + \beta_2 \text{lunch}_{it} + \beta_3 \log(\text{enroll}_{it}) \\ + \beta_{4t} \log(\text{rexppp}_{i,1994}) + \xi_1 \overline{\text{lunch}}_i + \xi_2 \overline{\log(\text{enroll}_i)} + v_{it1}. \end{aligned} \tag{8.2}$$

Note that because we have a single endogenous variable, in this case using Two Stage Least Squares (2SLS) would be numerically the same as estimating the model with the control function approach, and because of this, I used the latter. Table 8.3 shows the estimates from this model and once again, the coefficients obtained using the new sample are very similar to the ones computed using the original data set. In particular, the coefficient of the spending is considerable larger than the OLS estimate, which can be explained in the context of the local average treatment effect literature or by the fact that district authorities can decide to increase spending whenever they think the cohort might underperform Papke and Wooldridge (2008).

Table 8.3: Estimates assuming that the spending variable is endogenous.

	Original results	New results	Standard errors for new results with different bandwidth values						
	Coefficient	Coefficient	ρ_b=1	ρ_b=100	ρ_b=200	ρ_b=300	ρ_b=400	ρ_b=500	ρ_b=600
log(avgrexp)	0.555	0.546	0.221	0.265	0.292	0.253	0.221	0.202	0.187
	(0.208)	(0.211)							
lunch	-0.062	0.008	0.066	0.077	0.083	0.079	0.07	0.068	0.067
	(0.075)	(0.067)							
log(enroll)	0.046	0.023	0.069	0.075	0.079	0.071	0.065	0.058	0.054
	(0.067)	(0.066)							
v	-0.421	-0.476	0.250	0.349	0.411	0.383	0.365	0.357	0.353
	(0.232)	(0.236)							
Number of districts	501	493	-	-	-	-	-	-	-

Contrary to the case where all the independent variables were treated as exogenous, the standard errors computed using the HACSC estimator when the bandwidth parameter is set to 1 kilometer are somewhat different to the ones computed using an estimator that is only robust to serial correlation and heteroskedasticity, which is expected because the latter does not take into account the first stage estimation. Once again this results show that the standard errors can be larger or smaller depending on the value selected for the bandwidth.

So far I have assumed that there is only spatial correlation in the error term. However in this scenario there could be spatial spillovers from neighboring units that could be affecting the student performance on the math test. Figure 4 not only shows that the average real expenditure per student increased between 1995 and 2001 in all the school districts, but it also shows the spatial distribution of it. Note that there are districts where the surrounding neighbors have a very similar level of spending, for example, in 1995 the Detroit region shows multiple school districts with higher levels of expenditure compared to the rest of the state. Similarly, in Figure 5 the Upper Peninsula shows several neighboring school districts with higher passing rates than the rest of the region.

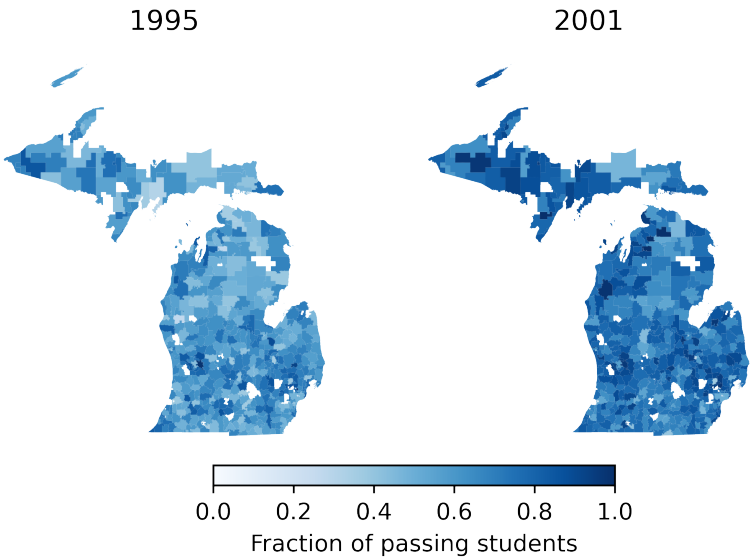

Figure 5: Average real expenditure per student across the Michigan school districts in 1995 and 2001.

Multiple reasons could be behind this pattern. For instance, it could be the case that parents with students that are underperforming identify school districts that are increasing spending and throughout the year, move to one of these districts in order to increase help their children to improve their grades. From the labor side, school districts might need to increase the expenditure in teachers' salaries to avoid losing them to other school districts within a reasonable commuting distance. All in all, it seems important to control for spillover effects of expenditure from neighbors, so I augment the models previously estimated with this additional variable[11] and Table 8.4 shows the estimates of this regression assuming that all the independent variables are exogenous with respect to the error term.

Table 8.4: OLS with extension

	Coefficient	Standard errors with different bandwidth values						
	(st. error)	ρ_b=1	ρ_b=100	ρ_b=200	ρ_b=300	ρ_b=400	ρ_b=500	ρ_b=600
log(avgrexp)	0.281 (0.076)	0.076	0.077	0.071	0.067	0.065	0.061	0.056
lunch	0.030 (0.063)	0.063	0.077	0.082	0.076	0.066	0.064	0.064
log(enroll)	-0.008 (0.047)	0.047	0.044	0.035	0.03	0.028	0.025	0.024
W· log(avgrexp)	0.324 (0.090)	0.088	0.076	0.071	0.057	0.049	0.047	0.047
Number of districts	493	-	-	-	-	-	-	-

[11]For this estimation, I used a rook type weighting matrix

Note that the coefficient on the average expenditure has decreased significantly so that an increase of approximately 10% in spending will now lead to an increase in the pass rate of about 2.8%. On the other hand, if neighboring school districts of unit i increase their expenditure around 10%, the pass rate in i is expected to improve around 3.2%, a larger effect than the own spending. To address the endogeneity issue, I also augmented the model 8.2 with the spending spillover variable using the control function approach[12] and the results are shown in Table 8.5. Once again, in this case the effect of the own expenditure is larger than in the exogenous case, but it is smaller compared to the original estimate. The spillover effect is significantly reduced to a marginal increase of around 0.7% in the pass rates due to an increase in the spending in surrounding school districts and moreover, the coefficient is not statistically significant.

Overall, the difference in the magnitude of the coefficients obtained for the spending in neighboring units make it difficult to interpret the effect of this variable. However in both cases it was positive, which supports the hypothesis that parents may move to school districts where the spending per student is higher. Of course, one cannot rule out the possibility that larger spending by neighboring school districts can attract better teachers to the area that are willing to commute, however, more detailed data may be needed to separate these effects.

Table 8.5: IV extension

	Coefficient	Standard errors for new results with different bandwidth values						
	(st. error)	ρ_b=1	ρ_b=100	ρ_b=200	ρ_b=300	ρ_b=400	ρ_b=500	ρ_b=600
log(avgrexp)	0.408 (0.231)	0.234	0.317	0.361	0.310	0.262	0.234	0.219
lunch	0.016 (0.067)	0.066	0.078	0.087	0.082	0.074	0.072	0.071
log(enroll)	-0.001 (0.067)	0.068	0.08	0.088	0.079	0.069	0.062	0.058
W $\cdot$ log(avgrexp)	0.071 (0.057)	0.056	0.076	0.083	0.077	0.07	0.067	0.065
v	-0.249 (0.254)	0.260	0.379	0.435	0.385	0.346	0.328	0.318
Number of districts	493	-	-	-	-	-	-	-

[12] I used $W \cdot \log(\text{found})$ to instrument for $W \cdot \log(\text{avgrexp})$

9 Conclusion

In this paper, I present a simple way to obtain standard errors that are robust to heteroskedasticity and both serial and spatial correlation in short panels with fixed effects and endogenous covariates. This is important because to the best of my knowledge, the current SHAC estimators do not explicitly allow for serial correlation in this context (admittedly the literature does not ignore this issue when $T \to \infty$). The estimator relies on averaging the moment conditions for a single individual across time, which allows to treat the estimation like a cross sectional problem without imposing any restrictions on the serial correlation of the residuals. This will help empirical researchers to obtain more reliable standard errors in different fields such as urban economics or international trade.

The proposed HACSC estimator can be directly applied in a Correlated Random Effects framework to obtain a fully robust Hausman-type test, which can help empirical researchers to choose between Fixed Effects and Random Effects specifications. In this paper I also showed that the Mundlak equivalence also holds in a particular spatial setting, which will allows to obtain the Fixed Effects coefficients of the time varying covariates in a Random Effects context. Similarly, the HACSC estimator can be used in a RE estimation procedure, whenever the researcher suspects that the structure imposed of the spatial error term might be misspecified.

I also presented a control function approach and the required assumptions to estimate the parameters of the model. Although even in the i.i.d. case it is a standard practice to use bootstrap to obtain the standard errors with this approach, in a spatial setting this is not a trivial procedure given the dependence between observations. For this reason, I also extended the HACSC estimator to this setup, which requires an adjustment of the covariance matrix to take into account the sampling error of the first stage estimation.

The Monte-Carlo experiment performed showed that the HACSC estimator works well in the presence of strong or moderate serial and spatial correlation compared to other methods used by the literature in terms of obtaining unbiased standard errors. As expected, the estimator also shows higher variance than such estimators, especially in settings with low spatial and/or serial correlation. The simulations also showed that if the CF assumptions hold, we can obtain efficiency gains compared to 2SLS.

An avenue for future research is to extend the Monte Carlo experiments in different directions. First, it would be interesting to use different weighting schemes for the weighting matrix W based on distance or a k-neighbor scheme in an irregular lattice, as well as different kernel functions. Analogous to the time series literature, the threshold for the distance bandwidth most certainly plays an important role on the finite sample behavior of the estimator, so implementing a data driven procedure to choose it is also a possibility to explore, particularly when the spatial correlation is strong.

Appendix

Assumptions

Assumption A.1

The functions $g_i(\cdot, \theta)$ satisfy these conditions:

1. $g_i(\cdot, \theta)$ are Borel measurable on $\mathcal{Z}$, the σ-algebra generated by Z, for all $\theta \in \Theta$.
2. $\sup_N \sup_{i \in D_N} \mathbb{E}[|g_i(Z_{i,N}, \theta)|^{2+\eta}] < \infty \ \ \forall \theta \in \Theta$ for some $\eta > 0$.

Assumption A.2

The $g(\cdot, \cdot)$ satisfy the following conditions:

1. For some $p \geq 1$:

$$\limsup_{n \to \infty} \frac{1}{|D_N|} \sum_{i \in D_N} \mathbb{E}\left[d_{i,N}^p \mathbb{1}(d_{i,N}^p > k)\right] \to 0 \text{ as } k \to \infty$$

   where $d_{i,N} = \sup_{\theta \in \Theta} |g_{i,N}(Z_{i,N}, \theta)|$.

2. $g_i(Z_{i,N}, \theta)$ are L_0 stochastically equicontinuous.

Assumption A.3

The true parameter θ_0 and the $g_i(\cdot, \cdot)$ satisfy these conditions:

1. $\theta_0 \in \text{int}(\Theta)$.
2. $g_i(Z_i, \cdot)$ is continuously differentiable on the interior of Θ.
3. $|\nabla_\theta g_i(Z_i, \theta)| < \infty$, where ∇_θ denotes the gradient of $g_i(Z_i, \theta)$ with respect to the parameter vector θ.
4. $\nabla_\theta g_i(Z_i, \theta)$ is Borel measurable, $\mathbb{E}[\nabla_\theta g_i(Z_i, \theta)]$ exists and $\text{rank}\{\mathbb{E}[\nabla_\theta g_i(A_i, \theta)]\} = P$, where $P = \dim(\theta_0)$.
5. $\mathbb{E}[|g_i(Z_i, \theta_0)|^{2+\epsilon}] < \infty$ for some $\epsilon > 0$.

Assumption A.4

There exist finite dimensional vectors m_i and Δ such that $\hat{u_i} - u_i = m_i \Delta$ and

$$\frac{1}{N} \sum_{i=1}^{N} ||z_i||^2 = \mathcal{O}_p(1) \quad \text{and} \quad N^{\frac{1}{2}} ||\Delta|| = \mathcal{O}_p(1)$$

Definitions

α-mixing for random fields

Let D_N be a subset of D. For $U \subseteq D_N$ and $V \subseteq D_N$, let $\sigma_n(U) = \sigma(X_{i,N} : i \in U)$, $\alpha_N(U,V) = \alpha(\sigma_n(U), \sigma_n(V))$. Then the α-mixing coefficients for the random field $\{X_{i,N} : i \in D_N, N \in \mathbb{N}\}$ is defined as follows:

$$\alpha_{k,l,N}(r) = \sup(\alpha_n(U,V), |U| \leq k, |V| \leq l, \rho(U,V) \geq r)$$

for $k, l, r, n \in \mathbb{N}$. Define also

$$\bar{\alpha}_{k,l}(r) = \sup_N \alpha_{k,l,N}(r)$$

Upper tail quantile function

Let X be a random variable. Then the upper quantile function $Q_X : (0,1) \to [0,\infty)$ is defined as:

$$Q_X(u) = \inf\{t : P(X > t) \leq u\}$$

"Inverse" function of mixing coefficients

For the non-increasing sequence of the mixing coefficients $\{\bar{\alpha}_{1,1}\}_{m=1}^{\infty}$, set $\bar{\alpha}_{1,1}(0) = 1$ and define its "inverse" function $\alpha_{\text{inv}}(u) : (0,1) \to \mathbb{N} \cup \{0\}$ as:

$$\alpha_{\text{inv}}(u) = \max\{m \geq 0 : \bar{\alpha}_{1,1}(m) > u\}$$

Stochastic equicontinuity

The array of random functions $\{f_{i,N}(Z_{i,N}, \theta) : i \in D_N, n \geq 1\}$ is:

1. L_0 stochastically equicontinuous on Θ iff for every $\varepsilon > 0$,

$$\limsup_{N\to\infty} \frac{1}{|D_N|} \sum_{i \in D_N} P\left[\sup_{\theta' \in \Theta} \sup_{\theta \in B(\theta', \delta)} |f_{i,N}(Z_{i,N}, \theta) - f_{i,N}(Z_{i,N}, \theta')| > \varepsilon\right] \to 0 \text{ as } \delta \to 0.$$

2. L_p stochastically equicontinuous, $p > 0$, on Θ iff

$$\limsup_{N\to\infty} \frac{1}{|D_N|} \sum_{i \in D_N} \mathbb{E}\left[\sup_{\theta' \in \Theta} \sup_{\theta \in B(\theta', \delta)} |f_{i,N}(Z_{i,N}, \theta) - f_{i,N}(Z_{i,N}, \theta')|^p\right] \to 0 \text{ as } \delta \to 0.$$

3. a.s. stochastically equicontinuous on Θ iff

$$\limsup_{N\to\infty} \frac{1}{|D_N|} \sum_{i \in D_N} \sup_{\theta' \in \Theta} \sup_{\theta \in B(\theta', \delta)} |f_{i,N}(Z_{i,N}, \theta) - f_{i,N}(Z_{i,N}, \theta')| \to 0 \text{ a.s. as } \delta \to 0.$$

Proof of Proposition 5

For notation simplicity, we will assume that $W_i y_t$ is included in x_{2it}, $x_{it} = [x_{1it} \;\; x_{2it}]$, where the x_2 are $k_2 + 1$ endogenous variables and $z_{it} = [x_{1it} \;\; z_{2it} \;\; w^2_{1it} \dots w^s_{1it}]$, where z_2 is a vector of L_2 instruments for x_2, with $L_2 \geq k_2$, and similarly for the spatial variables (note however that $W_i y_t$ is not in $W_i X_t$). Therefore, the problem is to apply Pooled 2SLS to the following equation:

$$\begin{aligned} y_{it} - \eta_i \bar{y}_i &= (x_{it} - \eta_i \bar{x}_i)\beta + W_i(X_t - \eta_i \bar{X})\gamma + (1-\eta_i)\bar{z}_i\delta + (1-\eta_i)W_i\bar{Z}\lambda \\ &= (x_{it} - \eta_i \bar{x}_i)\beta + (w_{it} - \eta_i \bar{w}_i)\gamma + (1-\eta_i)\bar{z}_i\delta + (1-\eta_i)\bar{\mathfrak{Z}}_i\lambda \end{aligned}$$

using IV's: $[(z_{it} - \eta_i \bar{z}_i) \;\; (\mathfrak{Z}_{it} - \eta_i \bar{\mathfrak{Z}}_i) \;\; (1-\eta_i)\bar{z}_{2i} \;\; (1-\eta_i)\bar{\mathfrak{Z}}_i]$.

We first orthogonalize the IV's, i.e., we run $z_{it} - \eta_i \bar{z}_i = (1-\eta_i)\bar{z}_i\epsilon_1 + (1-\theta_i)\bar{\mathfrak{Z}}_i\epsilon_2$ and obtain the residuals r_{it} and $\mathfrak{Z}_{it} - \eta_i \bar{\mathfrak{Z}}_i = (1-\eta_i)\bar{z}_i\epsilon_3 + (1-\theta_i)\bar{\mathfrak{Z}}_i\epsilon_4$ and get the residuals s_{it}. To do so, we use the Frish-Waugh-Lovell theorem sequentially.

1.a) $z_{it} - \eta_i \bar{z}_i$ on $(1-\eta_i)\bar{z}_i$. The coefficient will be:

$$\begin{aligned} \tilde{\epsilon}_1 &= \left[\sum_{i=1}^{N}\sum_{t=1}^{T}(1-\eta_i)^2\bar{z}'_{1i}\bar{z}_{1i}\right]^{-1}\left[\sum_{i=1}^{N}\sum_{t=1}^{T}(1-\eta_i)^2\bar{z}'_{1i}\bar{z}_i\right] \\ &= \left[\sum_{i=1}^{N}\sum_{t=1}^{T}(1-\eta_i)^2\bar{z}'_{1i}\bar{z}_{1i}\right]^{-1}\left[\sum_i(1-\eta_i)\bar{z}'_i\sum_t z_{it} - \sum_i T(1-\eta_i)\eta_i\bar{z}'_i\bar{z}_i\right] \\ &= \left[\sum_{i=1}^{N}T(1-\eta_i)^2\bar{z}'_i\bar{z}_i\right]^{-1}\left[\sum_{i=1}^{N}T(1-\eta_i)^2\bar{z}'_{1i}\bar{z}_{1i}\right] \\ &= \left[\sum_{i=1}^{N}(1-\eta_i)^2\bar{z}'_{1i}\bar{z}_{1i}\right]^{-1}\left[\sum_{i=1}^{N}(1-\eta_i)^2\bar{z}'_{1i}\bar{z}_{1i}\right] = \mathbf{I}_L \end{aligned}$$

Therefor the residuals will be $v_{it} = z_{it} - \bar{z}_i$.

1.b) Run $(1-\eta_i)\bar{\mathfrak{Z}}_i$ on $(1-\eta)\bar{z}$. In this case the coefficient and the residuals will depend only on the i index, call the latter f_i.

1.c) Run v_{it} on f_i to get ϵ_2. The coefficient will be:

$$\begin{aligned} \epsilon_2 &= \left[\sum_i\sum_t f'_i f_i\right]^{-1}\left[\sum_i\sum_t f'_i v_{it}\right] \\ &= \left[\sum_i\sum_t f'_i f_i\right]^{-1}\left[\sum_i f'_i\sum_t v_{it}\right] \\ &= \left[\sum_i\sum_t f'_i f_i\right]^{-1}\left[\sum_i f'_i\sum_t (z_{it} - \bar{z}_i)\right] = \mathbf{0}_L \end{aligned}$$

where we used the fact that the sum of deviations from the mean add up to zero for all i in the second term. This implies that $\epsilon_1 = \mathbf{I}_L$ and therefore, $r_{it} = z_{it} - \bar{z}_i$.

Using very similar steps, it can be shown that if we run $\mathfrak{Z}_{it} - \eta_i \bar{\mathfrak{Z}}_i = (1-\eta_i)\bar{z}_i\epsilon_3 + (1-\theta_i)\bar{\mathfrak{Z}}_i\epsilon_4$,

then $\epsilon_3 = \mathbf{0}_L$ and $\epsilon_4 = \mathbf{I}_L$, and therefore the residuals of this regression will be $s_{it} = \mathfrak{Z}_{it} - \bar{\mathfrak{Z}}_i$. Since we have orthogonalized the instrumental variables with respect to $(1-\eta)\bar{z}_i$ and $(1-\eta)\bar{\mathfrak{Z}}_i$, we now have to apply Pooled 2SLS to the following equation:

$$y_{it} - \eta_i \bar{y}_i = (x_{it} - \eta_i \bar{x}_i)\beta + (w_{it} - \eta_i \bar{w})\gamma$$

using IV's $[(z_{it} - \bar{z}_i) \quad (\mathfrak{Z}_{it} - \bar{\mathfrak{Z}}_i)]$. We now define the following notation: $\ddot{z}_{it} = z_{it} - \bar{z}_i$, $\ddot{\mathfrak{Z}}_{it} = \mathfrak{Z}_{it} - \bar{\mathfrak{Z}}_i$, $\hat{z}_{it} = [\ddot{z}_{it} \quad \ddot{\mathfrak{Z}}_{it}]$, $\tilde{y}_{it} = y_{it} - \eta_i \bar{y}_i$, $\tilde{x}_{it} = [(x_{it} - \eta_i \bar{x}_i) \quad (w_{it} - \eta_i \bar{w}_i)]$, $\hat{y}_{it} = y_{it} - \bar{y}_i$ and $\hat{x}_{it} = [(x_{it} - \bar{x}_i) \quad (w_{it} - \bar{w}_i)]$. Then the $\Gamma = (\beta \ \ \gamma)$ from the previous problem can be obtained as:

$$\begin{aligned}\hat{\Gamma}_{2SLS} = &\left[\left(\sum_i \sum_{t=1} \tilde{x}'_{it}\hat{z}_{it}\right)\left(\sum_i \sum_{t=1} \hat{z}'_{it}\hat{z}_{it}\right)^{-1}\left(\sum_i \sum_{t=1} \hat{z}'_{it}\tilde{x}_{it}\right)\right]^{-1} \cdot \\ &\left(\sum_i \sum_{t=1} \tilde{x}'_{it}\hat{z}_{it}\right)\left(\sum_i \sum_{t=1} \hat{z}'_{it}\hat{z}_{it}\right)^{-1}\left(\sum_i \sum_t \hat{z}'_{it}\tilde{y}_{it}\right) \end{aligned} \tag{A.1}$$

The first term of the square bracket term can be rewritten as follows (the third term of that inverse matrix can also be written in a similar way):

$$\begin{aligned}\sum_i \sum_t \tilde{x}'_{it}\hat{z}_{it} &= \sum_i \sum_t \begin{bmatrix}(x_{it} - \eta_i \bar{x}_i)' \\ (w_{it} - \eta_i \bar{w}_i)'\end{bmatrix}\begin{bmatrix}\ddot{z}_{it} & \ddot{\mathfrak{Z}}_{it}\end{bmatrix} \\ &= \sum_i \sum_t \begin{bmatrix}(x_{it} - \eta_i \bar{x}_i)'\ddot{z}_{it} & (x_{it} - \eta_i \bar{x}_i)'\ddot{\mathfrak{Z}}_{it} \\ (w_{it} - \eta_i \bar{w}_i)'\ddot{z}_{it} & (w_{it} - \eta_i \bar{w}_i)'\ddot{\mathfrak{Z}}'_{it}\end{bmatrix}\end{aligned} \tag{A.2}$$

We focus on the (1,1) term, but the following algebraic manipulation holds for the rest of the terms in the matrix and for the second term in (A.1):

$$\begin{aligned}\sum_i \sum_t (x_{it} - \eta_i \bar{x}_i)'\ddot{z}_{it} &= \sum_i \sum_t x'_{it}\ddot{z}_{it} - \sum_i \eta_i \bar{x}'_i \sum_t \ddot{z}_{it} \\ &= \sum_i \sum_t x'_{it}\ddot{z}_{it} - \sum_i \eta_i \bar{x}'_i \sum_t (z_{it} - \bar{z}_i) \\ &= \sum_i \sum_t x'_{it}\ddot{z}_{it} \\ &= \sum_i \sum_t x'_{it}\ddot{z}_{it} - \sum_i \bar{x}'_i \sum_t (z_{it} - \bar{z}_i)' \\ &= \sum_i \sum_t (x_{it} - \bar{x}_i)'\ddot{z}_{it}\end{aligned}$$

where in the second and fourth lines we used the fact that the sum of deviations from the mean over t add up to zero for all observations. Therefore, (A.2) can be rewritten as:

$$\begin{aligned}\sum_i \sum_t \begin{bmatrix}(x_{it} - \eta_i \bar{x}_i)'\ddot{z}_{it} & (x_{it} - \eta_i \bar{x}_i)'\ddot{\mathfrak{Z}}_{it} \\ (w_{it} - \eta_i \bar{w}_i)'\ddot{z}_{it} & (w_{it} - \eta_i \bar{w}_i)'\ddot{\mathfrak{Z}}'_{it}\end{bmatrix} &= \sum_i \sum_t \begin{bmatrix}(x_{it} - \bar{x}_i)'\ddot{z}_{it} & (x_{it} - \bar{x}_i)'\ddot{\mathfrak{Z}}_{it} \\ (w_{it} - \bar{w}_i)'\ddot{z}_{it} & (w_{it} - \bar{w}_i)'\ddot{\mathfrak{Z}}'_{it}\end{bmatrix} \\ &= \sum_i \sum_t \hat{x}'_{it}\hat{z}_{it}\end{aligned}$$

Similarly,

$$\sum_i \sum_t \hat{z}'_{it} \tilde{y}_{it} = \sum_i \sum_t \hat{z}'_{it} \hat{y}_{it}$$

Therefore,

$$\begin{aligned}
\hat{\Gamma}_{2SLS} = & \left[\left(\sum_i \sum_{t=1} \tilde{x}'_{it} \hat{z}_{it} \right) \left(\sum_i \sum_{t=1} \hat{z}'_{it} \hat{z}_{it} \right)^{-1} \left(\sum_i \sum_{t=1} \hat{z}'_{it} \tilde{x}_{it} \right) \right]^{-1} \cdot \\
& \left(\sum_i \sum_{t=1} \tilde{x}'_{it} \hat{z}_{it} \right) \left(\sum_i \sum_{t=1} \hat{z}'_{it} \hat{z}_{it} \right)^{-1} \left(\sum_i \sum_t \hat{z}'_{it} \tilde{y}_{it} \right) \\
= & \left[\left(\sum_i \sum_{t=1} \hat{x}'_{it} \hat{z}_{it} \right) \left(\sum_i \sum_{t=1} \hat{z}'_{it} \hat{z}_{it} \right)^{-1} \left(\sum_i \sum_{t=1} \hat{z}'_{it} \hat{x}_{it} \right) \right]^{-1} \cdot \\
& \left(\sum_i \sum_{t=1} \hat{x}'_{it} \hat{z}_{it} \right) \left(\sum_i \sum_{t=1} \hat{z}'_{it} \hat{z}_{it} \right)^{-1} \left(\sum_i \sum_t \hat{z}'_{it} \hat{y}_{it} \right) \\
= & \hat{\Gamma}_{FE2SLS}
\end{aligned}$$

Proof of Proposition 6

For notation simplicity and without loss of generality, I will omit $W_i y_t$ in the proof. This term can be treated as an additional endogenous variable included in x_{2it} with its respective instruments $[w_{1it}^2 \dots w_{1it}^s]$. Let $x_{it} = (x_{1it} \quad x_{2it})$, where x_{1it} is a $1 \times \mathrm{k}_1$ vector of exogenous variables and x_{2it} is a $1 \times \mathrm{k}_2$ vector of endogenous covariates.

Similarly, $X_t = (X_{1t} \quad X_{2t})$, $z_{it} = (x_{1it} \quad z_{2it})$, $\bar{z}_i = (\bar{x}_{1i} \quad \bar{z}_{2i})$, $Z_t = (X_{1t} \quad Z_{2t})$ and $\bar{Z}_t = (\bar{X}_1 \quad \bar{Z}_2)$, $\mathfrak{Z}_{2it} = W_i Z_{2it}$, $\bar{\mathfrak{Z}}_{2i} = W_i \bar{Z}_2$.

Finally denote $\hat{x}_{it} = (x_{1it} \quad \hat{x}_{2it})$, $\hat{\bar{x}}_i = (\bar{x}_{1i} \quad \hat{\bar{x}}_{2i})$, $\hat{\bar{X}} = (\bar{X}_1 \quad \hat{\bar{X}}_2)$, where the hats denote the linear projections of x_2 on $(x_1 \quad z_2)$ and their spatial lags.

In a spatial setting, $(\beta \quad \gamma)_{FE2SLS}$ can be obtained by applying Pooled 2SLS to

$$y_{it} - \bar{y}_i = (x_{1it} - \bar{x}_{1i})\beta_1 + (x_{2it} - \bar{x}_{2i})\beta_2 + W_i(X_{1t} - \bar{X}_1)\gamma_1 + W_i(X_{2t} - \bar{X}_2)\gamma_2 + (u_{it} - u_i)$$

using IV's: $[(z_{2it} - \bar{z}_{2i}) \quad W_i(Z_{2t} - \bar{Z}_2)]$

We want to show that applying Pooled 2SLS to:

$$\begin{aligned} y_{it} - \theta_i \bar{y}_i = {} & (x_{1it} - \theta_i \bar{x}_{1i})\beta_1 + (x_{2it} - \theta_i \bar{x}_{2i})\beta_2 + W_i(X_{1t} - \theta_i \bar{X}_1)\gamma_1 + W_i(X_{2t} - \theta_i \bar{X}_2)\gamma_2 \\ & + (1-\theta_i)\bar{x}_{1i}\delta_1 + (1-\theta_i)\bar{x}_{2i}\delta_2 + (1-\theta_i)\bar{W}_i\bar{X}_1\lambda_1 + (1-\theta_i)\bar{W}_i\bar{X}_2\lambda_2 + u_{it} \end{aligned}$$

using IV's: $[(z_{2it} - \theta_i \bar{z}_{2i}) \quad W_i(Z_{2t} - \theta_i \bar{Z}_2) \quad (1-\theta_i)\bar{z}_{2i} \quad (1-\theta_i)W_i\bar{Z}_2]$ yields the same $(\beta \quad \gamma)$.

In order to proof the result, I will follow these steps:

1. Orthogonalize with respect to $[(1-\theta_i)\bar{x}_{1i} \quad (1-\theta_i)\bar{w}_{1i}]$ the instrumental variables and $[(x_{1it} - \theta_i \bar{x}_{1i}) \quad (w_{1it} - \theta_i \bar{w}_{1i})]$
2. Orthogonalize with respect to $[(1-\theta_i)\bar{z}_{2i} \quad (1-\theta_i)\bar{\mathfrak{Z}}_{2i}]$ in the first stage equation.
3. Show that we get the same predicted values using the orthogonalized variables and the original ones.
4. Use the Frisch-Waugh-Lovell (WFL) theorem to show the equivalence.

So the model is:

$$\begin{aligned} y_{it} - \theta_i \bar{y}_i = {} & (x_{1it} - \theta_i \bar{x}_{1i})\beta_1 + (x_{2it} - \theta_i \bar{x}_{2i})\beta_2 + (w_{1it} - \theta_i \bar{w}_{i1})\gamma_1 + (w_{2it} - \theta_i \bar{w}_{i2})\gamma_2 \\ & + (1-\theta_i)\bar{x}_{1i}\delta_1 + (1-\theta_i)\bar{x}_{2i}\delta_2 + (1-\theta_i)\bar{w}_{i1}\lambda_1 + (1-\theta_i)\bar{w}_{i2}\lambda_2 + u_{it} \end{aligned}$$

using IV's: $[(z_{2it} - \theta_i \bar{z}_{2i}) \quad (\mathfrak{Z}_{2it} - \theta_i \bar{\mathfrak{Z}}_{2i}) \quad (1-\theta_i)\bar{z}_{2i} \quad (1-\theta_i)\bar{\mathfrak{Z}}_{2i}]$.

Step 1

a. $z_{2it} - \theta_i \bar{z}_{2i}$ on $(1-\theta_i)\bar{x}_{1i}$, $(1-\theta_i)\bar{w}_{1i}$

The residuals will be: $z_{2it} - \theta_i \bar{z}_{2i} - (1-\theta_i)\bar{x}_{1i}\hat{\eta}_1 - (1-\theta_i)\bar{w}_{1i}\hat{\eta}_2 = l_{it}$

Applying the FWL theorem: for $(1-\theta_i)\bar{x}_{1i}$ on $(1-\theta_i)\bar{w}_{1i}$, the coefficient will be:

$$\begin{aligned}
\hat{\mu}_1 &= \left[\sum_{i=1}^{N}\sum_{t=1}^{T}(1-\theta_i)^2\bar{w}_{1i}'\bar{w}_{1i}\right]^{-1}\left[\sum_{i=1}^{N}\sum_{t=1}^{T}(1-\theta_i)^2\bar{w}_{1i}'\bar{x}_{1i}\right] \\
&= \left[\sum_{i=1}^{N}T(1-\theta_i)^2\bar{w}_{1i}'\bar{w}_{1i}\right]^{-1}\left[\sum_{i=1}^{N}T(1-\theta_i)^2\bar{w}_{1i}'\bar{x}_{1i}\right] \\
&= \left[\sum_{i=1}^{N}(1-\theta_i)^2\bar{w}_{1i}'\bar{w}_{1i}\right]^{-1}\left[\sum_{i=1}^{N}(1-\theta_i)^2\bar{w}_{1i}'\bar{x}_{1i}\right]
\end{aligned}$$

The residuals will be $(1-\theta_i)\bar{x}_{1i} - (1-\theta_i)\bar{w}_{1i}\hat{\mu}_1 = s_i$.

Now we regress $z_{2it} - \theta_i\bar{z}_{2i}$ on $(1-\theta_i)\bar{w}_{1i}$. The coefficient will be:

$$\begin{aligned}
\hat{\mu}_2 &= \left[\sum_{i=1}^{N}\sum_{t=1}^{T}(1-\theta_i)^2\bar{w}_{1i}'\bar{w}_{1i}\right]^{-1}\left[\sum_{i=1}^{N}\sum_{t=1}^{T}(1-\theta_i)^2\bar{w}_{1i}'(z_{2it}-\theta_i\bar{z}_{2i})\right] \\
&= \left[\sum_{i=1}^{N}T(1-\theta_i)^2\bar{w}_{1i}'\bar{w}_{1i}\right]^{-1}\left[\sum_{i=1}^{N}(1-\theta_i)^2\bar{w}_{1i}'\sum_{t=1}^{T}(z_{2it}-\theta_i\bar{z}_{2i})\right] \\
&= \left[\sum_{i=1}^{N}T(1-\theta_i)^2\bar{w}_{1i}'\bar{w}_{1i}\right]^{-1}\left[\sum_{i=1}^{N}(1-\theta_i)^2\bar{w}_{1i}'\{T\times(\bar{z}_{2i}-\theta_i\bar{z}_{2i})\}\right] \\
&= \left[\sum_{i=1}^{N}(1-\theta_i)^2\bar{w}_{1i}'\bar{w}_{1i}\right]^{-1}\left[\sum_{i=1}^{N}(1-\theta_i)^2\bar{w}_{1i}'\bar{z}_{2i}\right]
\end{aligned}$$

The residuals will be $z_{2it} - \theta_i\bar{z}_{2i} - (1-\theta_i)\bar{w}_{1i}\hat{\mu}_2 = g_{it}$.

Finally, we run g_{it} on s_i. The coefficient will be:

$$\begin{aligned}
\hat{\eta}_1 &= \left[\sum_{i=1}^{N}\sum_{t=1}^{T}s_i's_i\right]^{-1}\left[\sum_{i=1}^{N}\sum_{t=1}^{T}s_i'g_{it}\right] \\
&= \left[\sum_{i=1}^{N}T\times s_i's_i\right]^{-1}\left[\sum_{i=1}^{N}s_i'\sum_{t=1}^{T}g_{it}\right] \\
&= \left[\sum_{i=1}^{N}T\times s_i's_i\right]^{-1}\left[\sum_{i=1}^{N}T\times s_i'\bar{g}_i\right] \\
&= \left[\sum_{i=1}^{N}s_i's_i\right]^{-1}\left[\sum_{i=1}^{N}(1-\theta_i)s_i'(\bar{z}_{2i}-\bar{w}_{1i}\hat{\mu}_2)\right]
\end{aligned}$$

Using similar steps, $\hat{\eta}_2$ will be:

$$\hat{\eta}_2 = \left[\sum_{i=1}^{N}s_i^{*\prime}s_i^*\right]^{-1}\left[\sum_{i=1}^{N}(1-\theta_i)s_i^{*\prime}(\bar{z}_{2i}-\bar{x}_{1i}\hat{\mu}_2^*)\right]$$

where $\hat{\mu}_2^*$ is the coefficient of regressing $z_{2it} - \theta_i\bar{z}_{2i}$ on $(1-\theta_i)\bar{x}_{1i}$ and s_i^* are the residuals of regressing $(1-\theta_i)\bar{w}_{1i}$ on $(1-\theta_i)\bar{x}_{1i}$

b. $(\mathfrak{Z}_{2it} - \theta_i\bar{\mathfrak{Z}}_{2i})$ on $(1-\theta_i)\bar{x}_{1i}$, $(1-\theta_i)\bar{w}_{1i}$.

The residuals will be $(\mathfrak{Z}_{2it} - \theta_i \bar{\mathfrak{Z}}_{2i}) - (1-\theta_i)\bar{x}_{1i}\hat{\eta}_3 - (1-\theta_i)\bar{w}_{1i}\hat{\eta}_4 = m_{it}$.

c. $(1-\theta_i)\bar{z}_{2i}$ on $(1-\theta_i)\bar{x}_{1i}$, $(1-\theta_i)\bar{w}_i$.

The residuals are: $(1-\theta_i)\bar{z}_i - (1-\theta_i)\hat{\eta}_5 - (1-\theta)\bar{w}_{1i}\hat{\eta}_6 = v_i$, which only depend on the i subscript.

Applying the FWL theorem, regressing $(1-\theta)\bar{x}_{1i}$ on $(1-\theta_i)\bar{w}_{1i}$ yields $\hat{\mu}_1$, the same as in step 1a. The residuals will be only a function of the i subscript, say f_i.

Finally, run f_i on s_i and the coefficient will be:

$$\left[\sum_{i=1} T s_i' s_i\right]^{-1}\left[\sum_{i=1} T s_i' f_i\right] = \left[\sum_{i=1} T s_i' s_i\right]^{-1}\left[\sum_{i=1} s_i'(1-\theta_i)(\bar{z}_{2i} - \bar{w}_{1i}\hat{\mu}_2)\right] = \hat{\eta}_5 = \hat{\eta}_1$$

The same coefficient as above. Following similar steps, it can be shown that

$$\hat{\eta}_6 = \hat{\eta}_2 = \left[\sum_{i=1} T s_i^{*\prime} s_i^*\right]\left[\sum_{i=1} s_i^{*\prime}(1-\theta_i)(\bar{z}_{2i} - \bar{x}_{1i}\hat{\mu}_2^*)\right]$$

where $\hat{\mu}_2^*$ is defined in step 1a.

$$\therefore v_i = (1-\theta_i)\bar{z}_{2i} - (1-\theta_i)\bar{x}_{1i}\hat{\eta}_5 - (1-\theta_i)\bar{w}_{1i}\hat{\eta}_6 = (1-\theta_i)\bar{z}_{2i} - (1-\theta_i)\bar{x}_{1i}\hat{\eta}_1 - (1-\theta_i)\bar{w}_{1i}\hat{\eta}_2$$

d. $(1-\theta_i)\bar{z}_{2i}$ on $(1-\theta_i)\bar{x}_{1i}, (1-\theta_i)\bar{w}_{1i}$

The coefficients will only depend in i, denote them by r_i. If $(1-\theta_i)\bar{z}_{2i} = (1-\theta_i)\bar{x}_{1i}\hat{\eta}_7 + (1-\theta_i)\bar{w}_{1i}\hat{\eta}_8$, it can be shown using similar arguments than in the previous step that $\hat{\eta}_7 = \hat{\eta}_3$ and $\hat{\eta}_8 = \hat{\eta}_4$.

e. $x_{1it} - \theta_i\bar{x}_{1i}$ on $(1-\theta_i)\bar{x}_{1i}, (1-\theta_i)\bar{w}_{1i}$

We can apply the FWL theorem to get the coefficients:

i. First if we regress $x_{1it} - \theta_i\bar{x}_{1i}$ on $(1-\theta_i)\bar{x}_{1i}$. The coefficient is:

$$\begin{aligned}
&\left[\sum_i\sum_t (1-\theta_i)^2\bar{x}_{1i}'\bar{x}_{1i}\right]^{-1}\left[\sum_i\sum_t (1-\theta_i)\bar{x}_{1i}'(x_{1it} - \theta_i\bar{x}_{1i})\right] \\
=&\left[\sum_i T(1-\theta_i)^2\bar{x}_{1i}'\bar{x}_{1i}\right]^{-1}\left[\sum_i (1-\theta_i)\bar{x}_{1i}'\sum_t (x_{1it} - \theta_i\bar{x}_{1i})\right] \\
=&\left[\sum_i T(1-\theta_i)^2\bar{x}_{1i}'\bar{x}_{1i}\right]^{-1}\left[\sum_i (1-\theta_i)\bar{x}_{1i}' T(\bar{x}_{it} - \theta_i\bar{x}_{1i})\right] \\
=&\left[\sum_i T(1-\theta_i)^2\bar{x}_{1i}'\bar{x}_{1i}\right]^{-1}\left[\sum_i T(1-\theta_i)^2\bar{x}_{1i}'\bar{x}_{1i}\right] \\
=&\mathbf{I}_{k_1}
\end{aligned}$$

where $\mathbf{I}_{k_1}$ denotes an identity matrix of size k_1. Therefore, the residuals will be $x_{1it} - \bar{x}_{1i}$.

ii. Now regress $(1-\theta_i)\bar{w}_{1i}$ on $(1-\theta_i)\bar{x}_{1i}$.
The coefficients and residuals will only depend on i. Denote the later by d_i.

iii. Finally regress $x_{1it} - \bar{x}_{1i}$ on d_i. The coefficient will be:

$$\left[\sum_i \sum_t d_i' d_i\right]^{-1} \left[\sum_i \sum_t d_i'(x_{1it} - \bar{x}_{1i})\right]$$
$$= \left[\sum_i \sum_t d_i' d_i\right]^{-1} \left[\sum_i d_i' \sum_t (x_{1it} - \bar{x}_{1i})\right] = \mathbf{0}_{k_1}$$

where we used the fact that $\sum_t (x_{1it} - \bar{x}_{1i}) = 0$. Therefore $x_{1it} - \theta_i \bar{x}_{1i} = (1-\theta_i)\bar{x}_{1i}\mathbf{I}_{k_1} + (1-\theta_i)\bar{w}_{1i}\mathbf{0}_{k_1}$ and the residuals will be $x_{1it} - \bar{x}_{1i}$.

f. $w_{1it} - \theta_i \bar{w}_{1i}$ on $(1-\theta_i)\bar{x}_{1i}$ and $(1-\theta_i)\bar{w}_{1i}$.

Applying the FWL theorem in a similar way than the previous step, we get the following relationship:

$w_{1it} - \theta_i \bar{w}_{1i} = (1-\theta_i)\bar{x}_{1i}\mathbf{0}_{k_1} + (1-\theta_i)\bar{w}_{1i}\mathbf{I}_{k_1}$ and the residuals will be $w_{1it} - \bar{w}_{1i}$.

Therefore, after orthogonalizing, we can apply Pooled 2SLS to:

$$y_{it} - \theta_i \bar{y}_i = (x_{1it} - \bar{x}_{1i})\beta_1 + (x_{2it} - \theta_i \bar{x}_{2i})\beta_2 + (w_{1it} - \bar{w}_{i1})\gamma_1 + (w_{2it} - \theta_i \bar{w}_{i2})\gamma_2$$
$$+ (1-\theta_i)\bar{x}_{2i}\delta_2 + (1-\theta_i)\bar{w}_{i2}\lambda_2 + u_{it}$$

using IV's: $[l_{it} \quad m_{it} \quad v_i \quad r_i]$.

Step 2

In this step we orthogonalize with respect to v_i and r_i in the first stage equation. Note that these are the residuals from the previous step associated with $(1-\theta_i)\bar{z}_{2i}$ and $(1-\theta_i)\bar{\mathfrak{Z}}_{2i}$ respectively, the instrumental variables.

a. $l_{it} = v_i\zeta_1 + r_i\zeta_2 + \varepsilon_1$.

i. l_{it} on v_i. The coefficient will be:

$$\tilde{\eta}_1 = \left[\sum_i \sum_t v_i' v_i\right]^{-1} \left[\sum_i \sum_t v_i' l_{it}\right]$$
$$= \left[\sum_i T v_i' v_i\right]^{-1} \left[\sum_i v_i' \sum_t l_{it}\right]$$
$$= \left[\sum_i v_i' v_i\right]^{-1} \left[\sum_i v_i' \bar{l}_i\right]$$

Note that $l_{it} = z_{2it} - \theta_i \bar{z}_{2i} - (1-\theta_i)\bar{x}_{1i}\hat{\eta}_1 - (1-\theta_i)\bar{w}_{1i}\hat{\eta}_2$, therefore

$$\bar{l}_i = \frac{1}{T}\sum_t [z_{2it} - \theta_i \bar{z}_{2i} - (1-\theta_i)\bar{x}_{1i}\hat{\eta}_1 - (1-\theta_i)\bar{w}_{1i}\hat{\eta}_2]$$
$$= (1-\theta_i)\bar{z}_{2i} - (1-\theta_i)\bar{x}_{1i}\hat{\eta}_1 - (1-\theta_i)\bar{w}_{1i}\hat{\eta}_2$$
$$= (1-\theta_i)(\bar{z}_{2i} - \bar{x}_{1i}\hat{\eta}_1 - \bar{w}_{1i}\hat{\eta}_2) = v_i$$

Therefore, $\tilde{\eta}_1 = \mathbf{I}_l$ since $\hat{\eta}_1 = \hat{\eta}_5$ and $\hat{\eta}_2 = \hat{\eta}_6$. The residuals are $z_{2it} - \bar{z}_{2i}$.

ii. r_i on v_i. In this case, both the coefficient and the residuals are going to depend only on i, call them h_i.

iii. Regress $z_{2it} - \bar{z}_{2i}$ on h_i. The coefficient is:

$$\begin{aligned}\hat{\eta}_3 &= \left[\sum_i \sum_t h_i' h_i\right]^{-1} \left[\sum_i \sum_t h_i'(z_{2it} - \bar{z}_{2i})\right] \\ &= \left[\sum_i \sum_t h_i' h_i\right]^{-1} \left[\sum_i h_i' \sum_t (z_{2it} - \bar{z}_{2i})\right] = \mathbf{0}_l\end{aligned}$$

Because the sum of deviations from the mean add up to zero. Therefore $l_{it} = v_i \mathbf{I}_l + r_i \mathbf{0}_l + \varepsilon$ and the residuals will be $z_{2it} - \bar{z}_{2i}$.

b. $m_{it} = v_i \pi_1 + r_i \pi_2 + \varepsilon_2$

i. m_{it} on r_i
The coefficient will be, after some algebra, $\tilde{\pi}_2 = \left[\sum_i r_i' r_i\right]^{-1} \left[\sum_i r_i' \bar{m}_i\right]$. Noting that

$$\begin{aligned}\bar{m}_i &= \frac{1}{T} \sum_t \left[(\mathfrak{Z}_{2it} - \theta_i \bar{\mathfrak{Z}}_{2i}) - (1-\theta_i)\bar{x}_{1i}\hat{\eta}_3 - (1-\theta_i)\bar{w}_{1i}\hat{\eta}_4\right] \\ &= (1-\theta_i)(\bar{\mathfrak{Z}}_{2i} - \bar{x}_{1i}\hat{\eta}_3 - \bar{w}_{1i}\hat{\eta}_4) \\ &= (1-\theta_i)(\bar{\mathfrak{Z}}_{2i} - \bar{x}_{1i}\hat{\eta}_7 - \bar{w}_{1i}\hat{\eta}_8) = r_i\end{aligned}$$

We conclude that $\tilde{\pi}_2 = \mathbf{I}_l$ and the residuals are $\mathfrak{Z}_{2it} - \bar{\mathfrak{Z}}_{2i}$.

ii. v_i on r_i. The coefficient will be denoted by $\tilde{\pi}_1 = \left[\sum_i r_i' r_i\right]^{-1} \left[\sum_i r_i' v_i\right]$, and the residuals will depend on i, call them $\tilde{h}_i$.

iii. $\mathfrak{Z}_{2it} - \bar{\mathfrak{Z}}_{2i}$ on $\tilde{h}_i$.
Using again the fact that $\sum_t \mathfrak{Z}_{2it} - \bar{\mathfrak{Z}}_{2i} = 0$, we conclude that $\pi_1 = \mathbf{0}_l$, which implies that $\tilde{\pi}_2 = \pi_2 = \mathbf{I}_l$ and therefore, the residuals will be $\mathfrak{Z}_{2it} - \bar{\mathfrak{Z}}_{2i}$.

In the original first stage we have:

$$\begin{aligned}x_{2it} - \theta_i \bar{x}_{2it} = (x_{1it} - \theta_i \bar{x}_{1i})\phi_1 + (w_{1it} - \theta_i \bar{w}_{1i})\phi_2 + (\mathfrak{Z}_{2it} - \theta_i \bar{\mathfrak{Z}}_{2i})\phi_3 + (w_{2it} - \theta_i \bar{w}_{i2})\phi_4 \\ + (1-\theta_i)\bar{x}_{1i}\rho_1 + (1-\theta_i)\bar{w}_{1i}\rho_2 + (1-\theta_i)\bar{z}_{2i}\rho_3 + (1-\theta_i)\bar{\mathfrak{Z}}_{2i}\rho_4 + \varepsilon_{FS}\end{aligned}$$

After orthogonalizing with respect to $[(1-\theta_i)\bar{x}_{1i} \;\; (1-\theta_i)\bar{w}_{1i} \;\; (1-\theta_i)\bar{z}_{2i} \;\; (1-\theta_i)\bar{\mathfrak{Z}}_{2i}]$, to get $\Phi = (\phi_1 \; \phi_2 \; \phi_3 \; \phi_4)$, we have to regress
$x_{2it} - \theta_i \bar{x}_{2it}$ on $[(x_{2it} - \bar{x}_{2it}) \;\; (x_{1it} - \bar{x}_{1i}) \;\; (w_{1it} - \bar{w}_{1i}) \;\; (\mathfrak{Z}_{2it} - \bar{\mathfrak{Z}}_{2i})]$.

We note that if $z_{it} = [x_{1it} \;\; w_{1it} \;\; z_{2it} \;\; \mathfrak{Z}_{2it}]$, then the coefficient of $x_{2it} - \theta_i \bar{x}_{2it}$ on $z_{it} - \bar{z}_i$ is

$$\begin{aligned}
\check{\Phi} &= \left[\sum_i \sum_t (z_{it} - \bar{z}_i)'(z_{it} - \bar{z}_i)\right]^{-1} \left[\sum_i \sum_t (z_{it} - \bar{z}_i)'(x_{2it} - \theta_i \bar{x}_{2i})\right] \\
&= \left[\sum_i \sum_t (z_{it} - \bar{z}_i)'(z_{it} - \bar{z}_i)\right]^{-1} \left[\sum_i \sum_t (z_{it} - \bar{z}_i)' x_{2it} - \sum_i \left\{\sum_t (z_{it} - \bar{z}_i)'\right\} \theta_i \bar{x}_{2i}\right] \\
&= \left[\sum_i \sum_t (z_{it} - \bar{z}_i)'(z_{it} - \bar{z}_i)\right]^{-1} \left[\sum_i \sum_t (z_{it} - \bar{z}_i)' x_{2it} - \sum_i \left\{\sum_t (z_{it} - \bar{z}_i)'\right\} \bar{x}_{2i}\right] \\
&= \left[\sum_i \sum_t (z_{it} - \bar{z}_i)'(z_{it} - \bar{z}_i)\right]^{-1} \left[\sum_i \sum_t (z_{it} - \bar{z}_i)'(x_{2it} - \bar{x}_{2i})\right]
\end{aligned}$$

Where we used the fact that the terms in curly brackets are zero. Therefore, Φ can also be obtained by regressing $(x_{2it} - \bar{x}_{2it})$ on $[(x_{2it} - \bar{x}_{2it}) \;\; (x_{1it} - \bar{x}_{1i}) \;\; (w_{1it} - \bar{w}_{1i}) \;\; (\mathfrak{Z}_{2it} - \bar{\mathfrak{Z}}_{2i})]$.

Step 3

In this step we show that $\widehat{x_{2it} - \theta_i} \bar{x}_{2i} = \widetilde{x_{2it} - \theta_i} \bar{x}_{2i}$, where

$$\begin{aligned}
\widehat{x_{2it} - \theta_i} \bar{x}_{2i} &= (x_{1it} - \theta_i \bar{x}_{1i})\hat{\phi}_1 + (w_{1it} - \theta_i \bar{w}_{1i})\hat{\phi}_2 + (z_{2it} - \theta_i \bar{z}_{2i})\hat{\phi}_3 + (\mathfrak{Z}_{2it} - \theta_i \bar{\mathfrak{Z}}_{2i})\hat{\phi}_4 \\
&\quad + (1 - \theta_i)\bar{x}_{1i}\hat{\rho}_1 + (1 - \theta_i)\bar{w}_{1i}\hat{\rho}_2 + (1 - \theta_i)\bar{z}_{2i}\hat{\rho}_3 + (1 - \theta_i)\bar{\mathfrak{Z}}_{2i}\hat{\rho}_4 \\
\widetilde{x_{2it} - \theta_i} \bar{x}_{2i} &= (x_{1it} - \bar{x}_{1i})\tilde{\phi}_1 + (w_{1it} - \bar{w}_{1i})\tilde{\phi}_2 + (z_{2it} - \bar{z}_{2i})\tilde{\phi}_3 + (\mathfrak{Z}_{2it} - \bar{\mathfrak{Z}}_{2i})\tilde{\phi}_4 \\
&\quad + (1 - \theta_i)\bar{x}_{1i}\tilde{\rho}_1 + (1 - \theta_i)\bar{w}_{1i}\tilde{\rho}_2 + (1 - \theta_i)\bar{z}_{2i}\tilde{\rho}_3 + (1 - \theta_i)\bar{\mathfrak{Z}}_{2i}\tilde{\rho}_4
\end{aligned}$$

First we note that $\hat{\phi}_j = \tilde{\phi}_j$ for $j = 1, 2, 3, 4$ because in the second equation the respective explanatory variables are orthogonalized with respect to the terms related to the time averages of the independent variables. Given this fact and after some algebra, we have that $\widehat{x_{2it} - \theta_i} \bar{x}_{2i} = \widetilde{x_{2it} - \theta_i} \bar{x}_{2i}$ if $\hat{\phi}_j + \hat{\rho}_j = \tilde{\rho}_j$ for $j = 1, 2, 3, 4$.

To show that the previous equality holds, we start with $\widetilde{x_{2it} - \theta_i} \bar{x}_{2i}$. Because $z_{it} = [x_{1it} \;\; w_{1it} \;\; z_{2it} \;\; \mathfrak{Z}_{2it}]$ as above, we have
$z_{it} - \bar{z}_i = \left[(x_{1it} - \bar{x}_{1i}) \;\; (w_{1it} - \bar{w}_{1i}) \;\; (z_{2it} - \bar{z}_{2i}) \;\; (\mathfrak{Z}_{2it} - \bar{\mathfrak{Z}}_{2i})\right]$, $\tilde{\rho} = (\tilde{\rho}_1' \quad \tilde{\rho}_2' \quad \tilde{\rho}_3' \quad \tilde{\rho}_4')'$ and $\tilde{\phi} = (\tilde{\phi}_1' \quad \tilde{\phi}_2' \quad \tilde{\phi}_3' \quad \tilde{\phi}_4')'$, therefore, $\widetilde{x_{2it} - \theta_i} \bar{x}_{2i} = (z_{it} - \bar{z}_i)\hat{\phi} + (1 - \theta_i)\bar{z}_i\tilde{\rho}$. Greene (2007) shows that given $\hat{\phi}$, one can get $\tilde{\rho}$ as:

$$\begin{aligned}
\tilde{\rho} &= \left[\sum_i \sum_t (1 - \theta_i)^2 \bar{z}_i' \bar{z}_i\right]^{-1} \left[\sum_i \sum_t (1 - \theta_i)\bar{z}_i' \left\{x_{2it} - \theta_i \bar{x}_{2i} - (z_{it} - \bar{z}_i)\hat{\phi}\right\}\right] \\
&= \left[\sum_i T(1 - \theta_i)^2 \bar{z}_i' \bar{z}_i\right]^{-1} \left[\sum_i \left((1 - \theta_i)\bar{z}_i' \sum_t (x_{2it} - \theta_i \bar{x}_{2i}) - (1 - \theta_i)\bar{z}_i' \left\{\sum_t (z_{it} - \bar{z}_i)\right\}\hat{\phi}\right)\right] \\
&= \left[\sum_i (1 - \theta_i)^2 \bar{z}_i' \bar{z}_i\right]^{-1} \left[\sum_i (1 - \theta_i)^2 \bar{z}_i' \bar{x}_{2i}\right]
\end{aligned}$$

where we used the fact that $\sum_t (z_{it} - \bar{z}_i) = 0$ on the second line.

We turn now to $\widehat{x_{2it} - \theta_i \bar{x}_{2i}}$. With similar definitions as above, given $\hat{\phi}$, we get $\hat{\rho}$ as:

$$\begin{aligned}
\hat{\rho} &= \left[\sum_i \sum_t (1-\theta_i)^2 \bar{z}_i' \bar{z}_i\right]^{-1} \left[\sum_i \sum_t (1-\theta_i)\bar{z}_i' \left\{(x_{2it} - \theta_i \bar{x}_{2i}) - (z_{it} - \theta_i \bar{z}_i)\hat{\phi}\right\}\right] \\
&= \left[\sum_i \sum_t (1-\theta_i)^2 \bar{z}_i' \bar{z}_i\right]^{-1} \left[\sum_i (1-\theta_i)\bar{z}_i' \left\{\sum_t (x_{2it} - \theta_i \bar{x}_{2i}) - \sum_t (z_{it} - \theta_i \bar{z}_i)\hat{\phi}\right\}\right] \\
&= \left[\sum_i T(1-\theta_i)^2 \bar{z}_i' \bar{z}_i\right]^{-1} \left[\sum_i T(1-\theta_i)^2 \bar{z}_i' \bar{x}_{2i}\right] \\
&- \left[\sum_i T(1-\theta_i)^2 \bar{z}_i' \bar{z}_i\right]^{-1} \left[\sum_i (1-\theta_i)\bar{z}_i (T\bar{z}_i - T\theta_i \bar{z}_i)\hat{\phi}\right] \\
&= \tilde{\rho} - \left[\sum_i T(1-\theta_i)^2 \bar{z}_i' \bar{z}_i\right]^{-1} \left[\sum_i T(1-\theta_i)^2 \bar{z}_i' \bar{z}_i\right] \hat{\phi} \\
&= \tilde{\rho} - \mathbf{I}_{2(k_1+l)+1}\hat{\phi} = \tilde{\rho} - \hat{\phi}
\end{aligned}$$

Therefore, $\tilde{\rho} = \hat{\rho} + \hat{\phi}$ and hence $\widehat{x_{2it} - \theta_i \bar{x}_{2i}} = \widetilde{x_{2it} - \theta_i \bar{x}_{2i}}$.
In a similar way and using obvious notation, it can be shown that
$\widehat{w_{2it} - \theta_i \bar{w}_{2i}} = \widetilde{w_{2it} - \theta_i \bar{w}_{2i}}$.

Step 4

Given the previous step, the problem becomes:

$$\begin{aligned}
y_{it} - \theta_i \bar{y}_i = &(x_{1it} - \bar{x}_{1i})\beta_1 + (x_{2it} - \theta_i \bar{x}_{2i})\beta_2 + (w_{1it} - \bar{w}_{i1})\gamma_1 + (w_{2it} - \theta_i \bar{w}_{i2})\gamma_2 \\
&+ (1-\theta_i)\bar{x}_{2i}\delta_2 + (1-\theta_i)\bar{w}_{i2}\lambda_2 + u_{it}
\end{aligned}$$

using IV's: $[(z_{2it} - \bar{z}_{2i}) \quad (\mathfrak{Z}_{2it} - \bar{\mathfrak{Z}}_{2i}) \quad (1-\theta_i)\bar{z}_{2i} \quad (1-\theta_i)\bar{\mathfrak{Z}}_{2i}]$. At this point however, it is important to note that although we have orthogonalized with respect to $(1-\theta_i)[\bar{x}_{1i} \quad \bar{w}_{1i}]$, we still have to include in the first stage equation to obtain the predicted values of the endogenous variables. Given this, the second stage equation is:

$$\begin{aligned}
y_{it} - \theta_i \bar{y}_i = &(x_{1it} - \bar{x}_{1i})\beta_1 + (\hat{x}_{2it} - \theta_i \hat{\bar{x}}_{2i})\beta_2 + (w_{1it} - \bar{w}_{i1})\gamma_1 + (\hat{w}_{2it} - \theta_i \hat{\bar{w}}_{i2})\gamma_2 \\
&+ (1-\theta_i)\hat{\bar{x}}_{2i}\delta_2 + (1-\theta_i)\hat{\bar{w}}_{i2}\lambda_2
\end{aligned}$$

where the $\hat{}$ denote the first stage projections on the instrumental variables. To obtain $(\beta \quad \gamma)$, we orthogonalize with respect to $(1-\theta_i)\hat{\bar{x}}_{2i}$ and $(1-\theta_i)\hat{\bar{w}}_{i2}$.

a. $(x_{1it} - \bar{x}_{1i})$ on $(1-\theta_i)\hat{\bar{x}}_{2i}$ and $(1-\theta_i)\hat{\bar{w}}_{i2}$.

   i. $(x_{1it} - \bar{x}_{1i})$ on $(1-\theta_i)\hat{\bar{x}}_{2i}$. The coefficient will be:

$$\left[\sum_i \sum_t (1-\theta_i)^2 \hat{\bar{x}}_{2i}' \hat{\bar{x}}_{2i}\right]^{-1} \left[\sum_i (1-\theta_i)\hat{\bar{x}}_{2i}' \sum_t (x_{1it} - \bar{x}_{1i})\right] = \mathbf{0}_{k_2}$$

   where we used that the sums of deviations from the mean are zero for all i and the residuals will be $x_{1it} - \bar{x}_{1i}$.

ii. $(1-\theta_i)\hat{\bar{w}}_{i2}$ on $(1-\theta_i)\hat{\bar{x}}_{2i}$. In this case the coefficients and the residuals will depend only on i, call them $\tilde{u}_i$.

iii. $x_{1it} - \bar{x}_{1i}$ on $\tilde{u}_i$. By a similar argument to point i just above, the coefficient is $\mathbf{0}_{k_2}$ and so, $x_{1it} - \bar{x}_{1i}$ is orthogonal to both variables.

b . $(w_{1it} - \bar{w}_{1i})$ on $(1-\theta_i)\hat{\bar{x}}_{2i}$ and $(1-\theta_i)\hat{\bar{w}}_{i2}$. Using a similar argument as in a) above, $w_{1it} - \bar{w}_{1i}$ is orthogonal to both variables.

c. $(\hat{x}_{2it} - \theta_i\hat{\bar{x}}_{2i})$ on $(1-\theta_i)\hat{\bar{x}}_{2i}$ and $(1-\theta_i)\hat{\bar{w}}_{i2}$.

i. $(1-\theta_i)\hat{\bar{w}}_{2i}$ on $(1-\theta_i)\hat{\bar{x}}_{2i}$. The coefficient and residuals depend only on i, call them $\check{u}_i$.

ii. $(\hat{x}_{2it} - \theta_i\hat{\bar{x}}_{2i})$ on $(1-\theta_i)\hat{\bar{x}}_{2i}$. By arguments very similar to previous steps, one can show that the coefficient is $\mathbf{I}_{k_2}$ and the residuals will be $(\hat{x}_{2it} - \hat{\bar{x}}_{2i})$.

iii. $(\hat{x}_{2it} - \hat{\bar{x}}_{2i})$ on $\check{u}_i$. By analogous arguments as above, the coefficient of this regression will be $\mathbf{0}_{k_2}$.

Therefore, the residuals of this regression will be $(\hat{x}_{2it} - \hat{\bar{x}}_{2i})$

d. $\hat{w}_{2it} - \theta_i\hat{\bar{w}}_{i2}$ on $(1-\theta_i)\hat{\bar{x}}_{2i}$ and $(1-\theta_i)\hat{\bar{w}}_{i2}$. Using similar ideas as in c) above, the residuals of this regression are $\hat{w}_{2it} - \hat{\bar{w}}_{i2}$.

Therefore, to find $(\beta_1 \quad \beta_2 \quad \gamma_1 \quad \gamma_2)$, we run

$$y_{it} - \theta_i\bar{y}_i = (x_{1it} - \bar{x}_{1i})\beta_1 + (\hat{x}_{2it} - \hat{\bar{x}}_{2i})\beta_2 + (w_{1it} - \bar{w}_{i1})\gamma_1 + (\hat{w}_{2it} - \hat{\bar{w}}_{i2})\gamma_2$$

If we collect all the covariates of this regression into a vector $\hat{x}_{it} - \hat{\bar{x}}_i$ (where the x_{1it} and w_{1it} are their own projections), then:

$$\begin{aligned}
(\beta \quad \gamma) &= \left[\sum_i\sum_t (\hat{x}_{it} - \hat{\bar{x}}_i)'(\hat{x}_{it} - \hat{\bar{x}}_i)\right]^{-1}\left[\sum_i\sum_t (\hat{x}_{it} - \hat{\bar{x}}_i)'(y_{it} - \theta_i\bar{y}_i)\right] \\
&= \left[\sum_i\sum_t (\hat{x}_{it} - \hat{\bar{x}}_i)'(\hat{x}_{it} - \hat{\bar{x}}_i)\right]^{-1}\left[\sum_i\sum_t (\hat{x}_{it} - \hat{\bar{x}}_i)'y_{it} - \sum_i\left\{\sum_t (\hat{x}_{it} - \hat{\bar{x}}_i)\right\}\theta_i\bar{y}_i\right] \\
&= \left[\sum_i\sum_t (\hat{x}_{it} - \hat{\bar{x}}_i)'(\hat{x}_{it} - \hat{\bar{x}}_i)\right]^{-1}\left[\sum_i\sum_t (\hat{x}_{it} - \hat{\bar{x}}_i)'(y_{it} - \bar{y}_i)\right]
\end{aligned}$$

where we use again the fact that the term in curly brackets in the second line is zero. Therefore, $(\beta \quad \gamma)$ can be obtained by regressing
$y_{it} - \bar{y}_i$ on $\left[(x_{1it} - \bar{x}_{1i}) \quad \widehat{(x_{2it} - \bar{x}_{2i})} \quad (w_{1it} - \bar{w}_{1i}) \quad \widehat{(w_{2it} - \bar{w}_{2i})}\right]$,
which is exactly the same problem that the Fixed Effects 2SLS estimator solves.

Derivation of the covariance matrix under a control function approach

Consider the estimating equation in (6.8):

$$\ddot{y}_{it} = \hat{\ddot{a}}_{it}\theta + \ddot{e}_{it}$$

where we can write $\ddot{a}_{it} = \ddot{z}_{it}\psi + \ddot{v}_{it}$. Because every element in $\ddot{a}_{it}$ is exogenous with respect to the error term $\ddot{e}_{it}$, we can write:

$$\begin{aligned}
\hat{\theta} &= \left[\frac{1}{NT}\sum_i^N\sum_t^T \hat{\ddot{a}}'_{it}\hat{\ddot{a}}_{it}\right]^{-1}\left[\frac{1}{NT}\sum_i^N\sum_t^T \hat{\ddot{a}}'_{it}\ddot{y}_{it}\right] \\
&= \left[\frac{1}{NT}\sum_i^N\sum_t^T \hat{\ddot{a}}'_{it}\hat{\ddot{a}}_{it}\right]^{-1}\left[\frac{1}{NT}\sum_i^N\sum_t^T \hat{\ddot{a}}'_{it}(\ddot{a}_{it}\theta + e_{it})\right] \\
&= \left[\frac{1}{NT}\sum_i^N\sum_t^T \hat{\ddot{a}}'_{it}\hat{\ddot{a}}_{it}\right]^{-1}\left[\frac{1}{NT}\sum_i^N\sum_t^T \hat{\ddot{a}}'_{it}(\ddot{a}_{it}\theta + \hat{\ddot{a}}_{it}\theta - \hat{\ddot{a}}_{it}\theta + \ddot{e}_{it})\right] \\
\implies \sqrt{NT}(\hat{\theta}-\theta) &= \left[\frac{1}{NT}\sum_i^N\sum_t^T \hat{\ddot{a}}'_{it}\hat{\ddot{a}}_{it}\right]^{-1}\left\{(NT)^{-\frac{1}{2}}\sum_i^N\sum_t^T \hat{\ddot{a}}'_{it}\left[\underbrace{(\ddot{a}_{it}-\hat{\ddot{a}}_{it})\theta}_{\text{Part 2}} + \underbrace{\ddot{e}_{it}}_{\text{Part 1}}\right]\right\}
\end{aligned}$$

Note that because $\hat{\psi} \xrightarrow{p} \psi$, the first matrix on the right hand side will converge in probability to $\mathbb{E}\left(\sum_i^N\sum_t^T \ddot{a}'_{it}\ddot{a}_{it}\right) = B$. Consider now Part 1:

$$\begin{aligned}
(NT)^{-\frac{1}{2}}\sum_i^N\sum_t^T \hat{\ddot{a}}'_{it}\ddot{e}_{it} &= (NT)^{-\frac{1}{2}}\sum_i^N\sum_t^T(\ddot{z}_{it}\hat{\psi})'\ddot{e}_{it} \\
&= (NT)^{-\frac{1}{2}}\sum_i^N\sum_t^T(\ddot{z}_{it}\hat{\psi} + \ddot{z}_{it}\psi - \ddot{z}_{it}\psi)'\ddot{e}_{it} \\
&= (NT)^{-\frac{1}{2}}\sum_i^N\sum_t^T(\ddot{z}_{it}\psi)'\ddot{e}_{it} + \frac{1}{NT}\sum_i^N\sum_t^T\left[\ddot{z}_{it}\sqrt{NT}(\hat{\psi}-\psi)\right]'\ddot{e}_{it} \\
&= (NT)^{-\frac{1}{2}}\sum_i^N\sum_t^T(\ddot{z}_{it}\psi)'\ddot{e}_{it} + \underbrace{\sqrt{NT}(\hat{\psi}-\psi)'}_{\mathcal{O}_p(1)}\cdot\underbrace{\frac{1}{NT}\sum_i^N\sum_t^T\ddot{z}'_{it}\ddot{e}_{it}}_{o_p(1)}
\end{aligned}$$

Because $\mathcal{O}_p(1)\cdot o_p(1) = o_p(1)$, Part 1 converges to $\left[(NT)^{-\frac{1}{2}}\sum_i^N\sum_t^T(\ddot{z}_{it}\psi)'\ddot{e}_{it}\right]$. Now consider Part 2:

$$\begin{aligned}
(NT)^{-\frac{1}{2}}\sum_i^N\sum_t^T \hat{\ddot{a}}'_{it}(\ddot{a}_{it}-\hat{\ddot{a}}_{it})\theta &= (NT)^{-\frac{1}{2}}\sum_i^N\sum_t^T(\ddot{z}_{it}\hat{\psi})'[\ddot{a}_{it}-\ddot{z}_{it}\hat{\psi}]\theta \\
&= (NT)^{-\frac{1}{2}}\sum_i^N\sum_t^T(\ddot{z}_{it}\hat{\psi})'[\ddot{a}_{it}-\ddot{z}_{it}\psi+\ddot{z}_{it}\psi-\ddot{z}_{it}\hat{\psi}]\theta \\
&= (NT)^{-\frac{1}{2}}\sum_i^N\sum_t^T(\underbrace{\ddot{z}_{it}\hat{\psi})'v_{it}\theta}_{\text{Part 2.1}} - \underbrace{(\ddot{z}_{it}\hat{\psi})'\ddot{z}_{it}(\hat{\psi}-\psi)\theta}_{\text{Part 2.2}}
\end{aligned}$$

Starting with Part 2.1:

$$
\begin{aligned}
(NT)^{-\frac{1}{2}} \sum_i^N \sum_t^T (\ddot{z}_{it}\hat{\psi})' v_{it}\theta &= (NT)^{-\frac{1}{2}} \sum_i^N \sum_t^T (\ddot{z}_{it}\psi + \ddot{z}_{it}\hat{\psi} - \ddot{z}_{it}\psi)' v_{it}\theta \\
&= (NT)^{-\frac{1}{2}} \sum_i^N \sum_t^T (\ddot{z}_{it}\psi)' v_{it}\theta + \left[\ddot{z}_{it}(\hat{\psi}-\psi)\right]' v_{it}\theta \\
&= (NT)^{-\frac{1}{2}} \left[\sum_i^N \sum_t^T (\ddot{z}_{it}\psi)' v_{it}\theta\right] + \underbrace{\sqrt{NT}(\hat{\psi}-\psi)'}_{\mathcal{O}_p(1)} \underbrace{\frac{1}{NT}\sum_i^N \sum_t^T \ddot{z}_{it}'\ddot{v}_{it}\theta}_{\xrightarrow{p}\mathbb{E}(\ddot{z}_{it}'\ddot{v}_{it})=0}
\end{aligned}
$$

So in the last line we have $\mathcal{O}_p(1) \cdot o_p(1) = o_p(1)$ and therefore the last term will vanish as $N \to \infty$ and only $(NT)^{-\frac{1}{2}} \sum_i^N \sum_t^T (\ddot{z}_{it}\psi)' v_{it}\theta$ will remain. Using similar algebra, it can be shown that part 2.2 will converge to

$$
-\frac{1}{NT} \sum_i^N \sum_t^T (\ddot{z}_{it}\psi)' \ddot{z}_{it} \sqrt{NT}(\hat{\psi}-\psi)\theta
$$

Note that

$$
\hat{\psi} - \psi = \left(\sum_i^N \sum_t^T \ddot{z}_{it}'\ddot{z}_{it}\right)^{-1} \left(\sum_i^N \sum_t^T \ddot{z}_{it}'\ddot{v}_{it}\right)
$$

$$
\implies \sqrt{NT}(\hat{\psi}-\psi) = \left(\frac{1}{NT}\sum_i^N \sum_t^T \ddot{z}_{it}'\ddot{z}_{it}\right)^{-1} \left[(NT)^{-\frac{1}{2}} \sum_i^N \sum_t^T \ddot{z}_{it}'\ddot{v}_{it}\right]
$$

Putting everything together we have

$$
\begin{aligned}
&\left[\frac{1}{NT}\sum_i^N \sum_t^T \hat{\ddot{a}}_{it}'\hat{\ddot{a}}_{it}\right]^{-1} \left\{(NT)^{-\frac{1}{2}} \sum_i^N \sum_t^T \hat{\ddot{a}}_{it}' \left[(\ddot{a}_{it} - \hat{\ddot{a}}_{it})\theta + \ddot{e}_{it}\right]\right\} \\
=&B^{-1} \left\{(NT)^{-\frac{1}{2}} \sum_i^N \sum_t^T (\ddot{z}_{it}\psi)' \left[\ddot{e}_{it} + \ddot{v}_{it}\theta - \ddot{z}_{it}(\hat{\psi}-\psi)\theta\right]\right\} + o_p(1) \\
=&B^{-1} \left\{(NT)^{-\frac{1}{2}} \left[\sum_i^N \sum_t^T \{(\ddot{z}_{it}\psi)'(\ddot{e}_{it} + \ddot{v}_{it}\theta)\}\right] - \frac{1}{NT}\sum_i^N \sum_t^T \{(\ddot{z}_{it}\psi)'\ddot{z}_{it}\} \sqrt{NT}(\hat{\psi}-\psi)\theta\right\} + o_p(1)
\end{aligned}
$$

where $\sqrt{NT}(\hat{\psi}-\psi) = \left(\frac{1}{NT}\sum_i^N \sum_t^T \ddot{z}_{it}'\ddot{z}_{it}\right)^{-1} \left[(NT)^{-\frac{1}{2}} \sum_i^N \sum_t^T \ddot{z}_{it}'\ddot{v}_{it}\right]$.
Let

$$
G = \mathbb{E}\left[\sum_i^N \sum_t^T (\ddot{z}_{it}\psi)'\ddot{z}_{it}\right]
$$

and

$$
r_{it}(\psi) = \left(\frac{1}{NT}\sum_i^N \sum_t^T \ddot{z}_{it}'\ddot{z}_{it}\right)^{-1} \left[(NT)^{-\frac{1}{2}} \sum_i^N \sum_t^T \ddot{z}_{it}'\ddot{v}_{it}\right]
$$

Then we can write

$$
\sqrt{NT}(\hat{\theta}-\theta) = B^{-1} \left\{(NT)^{-\frac{1}{2}} \sum_i^N \sum_t^T (\ddot{z}_{it}\psi)'(\ddot{e}_{it} + \ddot{v}_{it}\theta) - G \cdot r_{it}(\psi)\theta\right\} + o_p(1)
$$

And therefore, by the Central Limit Theorem,

$$\sqrt{NT}(\hat{\theta}-\theta) \overset{a}{\sim} N\left\{0, B^{-1}MB^{-1}\right\}$$

where $M = \text{Var}\left[\sum_i^N \sum_t^T (\ddot{z}_{it}\psi)'(\ddot{e}_{it} + \ddot{v}_{it}\theta) - G \cdot r_{it}(\psi)\theta\right] = \text{Var}\left[\sum_i^N \sum_t^T m_{it}\right]$.
B can be estimated with

$$\hat{B} = \frac{1}{NT}\sum_i^N \sum_t^T \hat{\ddot{a}}'_{it}\hat{\ddot{a}}_{it}$$

To estimate M, let

$$\hat{m}_{it} = (\ddot{z}_{it}\hat{\psi})'(\hat{\ddot{e}}_{it} + \hat{\ddot{v}}_{it}\hat{\theta}) - \hat{G} \cdot \hat{r}_{it}(\hat{\psi})\hat{\theta}$$

where,
$\hat{\ddot{e}}_{it}$ are the residuals from the second stage.
$\hat{\ddot{v}}_{it}$ are the residuals from the first stage (note that v_{it} is a vector).
$\hat{G} = \frac{1}{NT}\sum_i^N \sum_t^T (\ddot{z}_{it}\hat{\psi})'\ddot{z}_{it}$.
$\hat{r}(\hat{\psi}) = \left(\frac{1}{NT}\sum_i^N \sum_t^T \ddot{z}'_{it}\ddot{z}_{it}\right)^{-1}\left[(NT)^{-\frac{1}{2}}\sum_i^N \sum_t^T \ddot{z}'_{it}\hat{\ddot{v}}_{it}\right]$.
With these quantities defined, the (r,s)-th element of M can be estimated as

$$\hat{M}_{rs} = \frac{1}{NT}\sum_{i=1}^N \sum_{j=1}^N \sum_{t=1}^T \sum_{l=1}^T \hat{\ddot{m}}_{it,r}\hat{\ddot{m}}_{jl,s}K\left[\frac{\rho^*(i,j)}{\rho_b}\right]$$

where once again the kernel function $K(\cdot)$ is operationalizing the weak spatial dependence assumption.

Additional Simulation Results

Table A1: Average of the estimated variance of β_1 over the 1,000 replications using a rook type weighting matrix, N = 400, T=5.

		CF	CF_no1	True value	HACSC	SHAC	Cluster	Non-Robust	True value
ρ	ψ			CF					2SLS
	0.0	0.041	0.037	0.0386	0.082	0.068	0.086	0.069	0.0866
0.0	0.3	0.037	0.034	0.0352	0.076	0.062	0.078	0.063	0.0726
	0.7	0.035	0.032	0.0323	0.071	0.058	0.073	0.059	0.0706
	0.0	0.043	0.039	0.0428	0.087	0.071	0.089	0.072	0.0906
0.3	0.3	0.040	0.036	0.0364	0.079	0.065	0.082	0.066	0.079
	0.7	0.037	0.033	0.0359	0.074	0.061	0.076	0.062	0.0829
	0.0	0.062	0.055	0.0558	0.111	0.091	0.115	0.092	0.1092
0.7	0.3	0.057	0.051	0.0535	0.103	0.085	0.106	0.085	0.1118
	0.7	0.054	0.048	0.0488	0.096	0.079	0.099	0.080	0.0954

*True value computed as the variance of β_1 across the 1,000 replications.

All the numbers were multiplied by 100 for readability.

Table A2: Average of the estimated variance of β_2 over the 1,000 replications using a rook type weighting matrix, N = 400, T=5.

		CF	CF_no1	True value	HACSC	SHAC	Cluster	Non-Robust	True value
ρ	ψ			CF					2SLS
	0.0	0.069	0.066	0.0724	0.080	0.066	0.083	0.067	0.0803
0.0	0.3	0.063	0.060	0.0644	0.074	0.061	0.076	0.061	0.0738
	0.7	0.060	0.057	0.0623	0.069	0.056	0.071	0.057	0.0704
	0.0	0.074	0.071	0.0756	0.086	0.070	0.089	0.071	0.0894
0.3	0.3	0.068	0.065	0.0761	0.078	0.065	0.081	0.065	0.0862
	0.7	0.063	0.060	0.0646	0.073	0.061	0.076	0.061	0.0793
	0.0	0.119	0.114	0.1237	0.131	0.108	0.136	0.109	0.1376
0.7	0.3	0.108	0.104	0.1125	0.120	0.099	0.125	0.100	0.1297
	0.7	0.101	0.097	0.1004	0.113	0.093	0.116	0.094	0.113

*True value computed as the variance of β_2 across the 1,000 replications.

All the numbers were multiplied by 100 for readability.

Table A3: Average of the estimated variance of β_3 over the 1,000 replications using a rook type weighting matrix, N = 400, T=5.

ρ	ψ	CF	CF_no1	True value CF	HACSC	SHAC	Cluster	Non-Robust	True value 2SLS
	0.0	0.275	0.242	0.24	0.742	0.615	0.772	0.620	0.79
0.0	0.3	0.252	0.220	0.22	0.680	0.558	0.700	0.563	0.64
	0.7	0.239	0.206	0.21	0.631	0.522	0.654	0.526	0.63
	0.0	0.291	0.252	0.27	0.769	0.636	0.796	0.640	0.81
0.3	0.3	0.271	0.232	0.24	0.705	0.581	0.730	0.587	0.71
	0.7	0.254	0.215	0.23	0.660	0.545	0.681	0.549	0.72
	0.0	0.377	0.314	0.33	0.917	0.758	0.949	0.763	0.90
0.7	0.3	0.350	0.290	0.29	0.860	0.709	0.884	0.712	0.91
	0.7	0.329	0.270	0.29	0.794	0.657	0.824	0.662	0.79

*True value computed as the variance of β_3 across the 1,000 replications.

All the numbers were multiplied by 100 for readability.

CF_no1 refers to the HACSC estimator ignoring the first stage estimation using a CF approach.

Table A4: Rejection probabilities for the null hypothesis $H_0 : \beta_1 = 0.7$ at a 5% of significance using a t-test over the 1,000 replications with a rook type weighting matrix, N = 400, T=5.

ρ	ψ	CF	CF_no1	HACSC	SHAC	Cluster	Non-Robust
	0.0	0.050	0.062	0.068	0.088	0.055	0.081
0.0	0.3	0.047	0.057	0.056	0.072	0.052	0.076
	0.7	0.043	0.054	0.053	0.068	0.046	0.068
	0.0	0.054	0.066	0.068	0.089	0.058	0.088
0.3	0.3	0.048	0.062	0.057	0.073	0.042	0.072
	0.7	0.045	0.067	0.059	0.083	0.060	0.083
	0.0	0.049	0.067	0.065	0.079	0.057	0.079
0.7	0.3	0.047	0.062	0.071	0.093	0.064	0.095
	0.7	0.051	0.064	0.050	0.070	0.040	0.070

Table A5: Rejection probabilities for the null hypothesis $H_0 : \beta_2 = 0.6$ at a 5% of significance using a t-test over the 1,000 replications with a rook type weighting matrix, N = 400, T=5.

ρ	ψ	CF	CF_no1	HACSC	SHAC	Cluster	Non-Robust
	0.0	0.076	0.067	0.069	0.096	0.061	0.093
0.0	0.3	0.078	0.064	0.077	0.101	0.066	0.102
	0.7	0.079	0.072	0.079	0.104	0.074	0.105
	0.0	0.082	0.072	0.089	0.108	0.076	0.105
0.3	0.3	0.082	0.075	0.079	0.106	0.076	0.104
	0.7	0.081	0.069	0.089	0.119	0.078	0.108
	0.0	0.071	0.068	0.075	0.099	0.073	0.099
0.7	0.3	0.077	0.066	0.092	0.111	0.080	0.110
	0.7	0.069	0.064	0.063	0.080	0.053	0.077

Figures

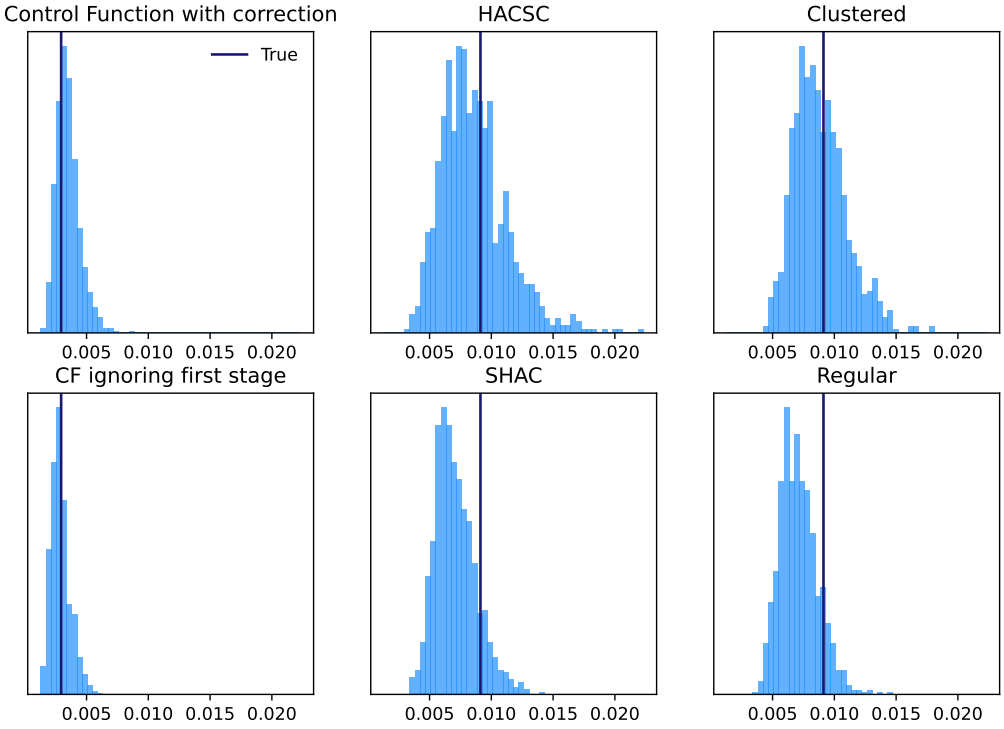

Figure A1: Distribution of the computed variances of $\hat{\beta}_3$ obtained for the case with e following a spatial AR(1) process ($\rho = 0.7$), and a following an AR(1) ($\psi = 0.3$), $N = 400$, $T = 5$.
*True value computed as the variance of β_3 across the 1,000 replications.

References

Amemiya, Takeshi (1985). *Advanced Econometrics*. Harvard University Press.

Arellano, Manuel (1987). "Computing robust standard errors for within-groups estimators". In: *Oxford Bulletin of Economics and Statistics* 49.4, pp. 431–434.

Baltagi, Badi and Long Liu (2011). "Instrumental Variable Estimation of a Spatial Autoregresive Panel Model with Random Effects". In: *Economic Letters* 111, pp. 135–137.

Basile, Roberto (2009). "Productivity Polarization across Regions in EuropeThe Role of Nonlinearities and Spatial Dependence". In: *International Regional Science Review*, pp. 92–115.

Basile, Roberto et al. (2014). "Modeling regional economic dynamics: Spatial dependence, spatial heterogeneity and nonlinearities". In: *Journal of Economic Dynamics and Control*, pp. 229–245.

Bester, C. Alan, Timothy G. Conley, and Christian B. Hansen (2011). "Inference with dependent data using cluster covariance estimators". In: *Journal of Econometrics* 165.2, pp. 137–151.

Bester, C. Alan et al. (2016). "Fixed-b Asymptotics for Spatially Dependent Robust Nonparametric Covariance Matrix Estimators". In: *Econometric Theory* 32.1, pp. 154–186.

Blundell, Richard and James Powell (2003). "Endogeneity in nonparametric and semiparametric regression models". In: *Econometric society monographs* 36, pp. 312–357.

Cameron, Colin and Pravin Trivedi (2005). *Microeconometrics: Methods and Applications*. Cambridge University Press.

Conley, Timothy G. (1999). "GMM estimation with cross sectional dependence". In: *Journal of Econometrics* 92.1, pp. 1–45.

Conley, Timothy G. and Francesca Molinari (2007). "Spatial Correlation Robust Inference with Errors in Location or Distance". In: *Journal of Econometrics* 140, pp. 76–96.

Debarsy, Nicholas (2012). "The Mundlak Approach in the Spatial Durbin Panel Data Model". In: *Spatial Economic Analysis* 7.1, pp. 109–131.

Driscoll, John C. and Aart C. Kraay (1998). "Consistent Covariance Matrix Estimation with Spatially Dependent Panel Data". In: *The Review of Economics and Statistics* 80.4, pp. 549–560.

Greene, William (2007). *Econometric Analysis*. 7th ed. Prentice Hall.

Joshi, Riju and Jeffrey M. Wooldridge (2019). "Correlated Random Effects Models with Endogenous Explanatory Variables and Unbalanced Panels". In: *Annals of Economics and Statistics* 134.

Kapoor, Mudit, Harry H. Kelejian, and Ingmar R. Prucha (2007). "Panel data models with spatially correlated error components." In: *Journal of Econometrics* 140.140, pp. 97–130.

Kelejian, Harry and Ingmar Prucha (2007). "HAC estimation in a spatial framework". In: *Journal of Econometrics* 140.1, pp. 131–154.

Kelejian, Harry H and Ingmar R Prucha (1998). "A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances". In: *The Journal of Real Estate Finance and Economics* 17.1, pp. 99–121.

Kelejian, Harry H, Ingmar R Prucha, and Yevgeny Yuzefovich (2004). "Instrumental variable estimation of a spatial autoregressive model with autoregressive disturbances: Large and small sample results". In: *Spatial and spatiotemporal econometrics*. Emerald Group Publishing Limited.

Kim, Min Seong and Yixiao Sun (2011). "Spatial heteroskedasticity and autocorrelation consistent estimation of covariance matrix". In: *Journal of Econometrics* 160, pp. 346–371.

– (2013). "Heteroskedasticity and spatiotemporal dependence robust inference for linear panel models with fixed effects". In: *Journal of Econometrics* 177, pp. 85–108.

Kunsch, Hans R. (1989). "The Jackknife and the Bootstrap for General Stationary Observations". In: *Annals of Statistics* 17.3, pp. 1217–1241.

Lee, Lung-fei (2003). "Best spatial two-stage least squares estimators for a spatial autoregressive model with autoregressive disturbances". In: *Econometric Reviews* 22.4, pp. 307–335.

LeSage, James and R. Kelley Pace (2009). *Introduction to Spatial Econometrics*. CRC Press.

Li, Liyao and Zhenlin Yang (2020). "Spatial dynamic panel data models with correlated random effects". In: *Journal of Econometrics*.

Müller, Ulrich K. and Mark W. Watson (2022a). "Spatial Correlation Robust Inference". In: *Econometrica* 90.6, pp. 2901–2935.

– (2022b). "Spatial Correlation Robust Inference in Linear Regression and Panel Models". In: *Journal of Business and Economics Statistics* 00.0, pp. 1–15.

Mundlak, Yair. (1978). "On the pooling of time series and cross section data". In: *Econometrica* 46.1, pp. 69–85.

Mutl, Jan and Michael Pfaffermayr (2010). "The Hausman Test in a Cliff and Ord Panel Model". In: *Econometrics Journal* 10, pp. 1–30.

Nazgul, Jenish and Ingmar R. Prucha (2009). "Central limit theorems and uniform laws of large numbers for arrays of random fields". In: *Journal of Econometrics* 150, pp. 86–98.

Newey, Whitney K. and Kenneth D. West (1987). "A simple, positive semi-definite, heteroskedasticityand autocorrelation consistent covariance matrix". In: *Econometrica* 55, pp. 703–798.

Papke, Leslie (2005). "The effects of spending on test pass rates: evidence from Michigan". In: *Journal of Public Economics* 89.5-6, pp. 821–839.

Papke, Leslie E. and Jeffrey M. Wooldridge (2008). "Panel data methods for fractional response variables with an application to test pass rates". In: *Journal of Econometrics* 145.1-2, pp. 121–133.

Politis, Dimitris N. and Halbert White (2004). "Automatic Block-Length Selection for the Dependent Bootstrap". In: *Econometric Reviews* 23.1, 53–70.

Vogelsang, Timothy (2012). "Heteroskedasticity, autocorrelation, and spatial correlation robust inference in linear panel models with fixed-effects". In: *Journal of Econometrics* 166, pp. 303–319.

White, Halbert (1980). "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." In: *Econometrica* 48.4, pp. 817–838.

Wooldridge, Jeffrey (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT Press.