# Using spatial modeling to address covariate measurement error

Susanne M. Schennach
Vincent Starck

The Institute for Fiscal Studies
Department of Economics, UCL

**cemmap** working paper CWP58/20

# Using spatial modeling to address covariate measurement error

Susanne M. Schennach* and Vincent Starck
Brown University

December 8, 2020

## Abstract

We propose a new estimation methodology to address the presence of covariate measurement error by exploiting the availability of spatial data. The approach uses neighboring observations as repeated measurements, after suitably controlling for the random distance between the observations in a way that allows the use of operator diagonalization methods to establish identification. The method is applicable to general nonlinear models with potentially nonclassical errors and does not rely on a priori distributional assumptions regarding any of the variables. The method's implementation combines a sieve semiparametric maximum likelihood with a first-step kernel conditional density estimator and simulation methods. The method's effectiveness is illustrated through both controlled simulations and an application to the assessment of the effect of pre-colonial political structure on current economic development in Africa.

**Keywords**: Errors-in-variables, Economic development, Operator methods, Spatial statistics.

[*] Support from NSF grant SES-1950969 is gratefully acknowledged. The authors would like to thank Florian Gunsilius and Stelios Michalopoulos for useful comments.

# 1 Introduction

With the increasing availability of Graphical Information System (GIS) data (Zhou et al., 2017) and network data (de Paula, 2017), spatial econometrics (Pinkse and Slade, 2010*b*; Redding and Rossi-Hansberg, 2017) is becoming an increasingly influential field. Further, spatial setups readily generalize to more abstract spaces, with the spatial dimensions representing individual or product characteristics, and the increasing availability of rich datasets with suitable covariates enables this avenue of research.

This paper identifies another advantage provided by the use for spatial datasets. The inherent redundancy provided by numerous nearby observations in spatial frameworks generates information that can be used to correct for covariate measurement error and achieve consistency without requiring additional information such as validation data or the knowledge of the measurement error distribution. The method is very generally applicable, as it allows for nonlinear models as well as non-classical measurement error (Schennach, 2016). This is made possible in part by leveraging identification results from Hu and Schennach (2008) and Hu (2008) and in part by devising a scheme to generate "virtual" observations that can act as repeated measurements, from the information provided by the observed sample.

Our approach is to be contrasted to others developed within the Kriging literature (Krige (1951), Chilès and Desassis (2018)). Kriging is a common method to carry out inference regarding spatial quantities in between available measurements. While this approach has been extended to allow for measurement error (e.g. Cressie (1993)), most of this line of research does not consider the implications of using the mismeasured data as a covariate. Methods that do consider covariates tend to rely on distributional assumptions and linearity (e.g., Szpiro, Sheppard and Lumley (2011)) or achieve bias reduction but not consistency (e.g. Alexeeff, Carroll and Coull (2016)).

While the approach we take is reminiscent of using lags or leads as repeated measurements in the context of time series or panel data econometrics (Hu and Shum, 2012; Cunha, Heckman and Schennach, 2010; Griliches and Hausman, 1986), a corresponding approach in a spatial framework is not currently available, due to significant conceptual and algorithmic challenges. Unless the spatial data happens to lie on a fixed grid (a rare occurrence), there is no spatial analog of a fixed time-shift, since the spacing between data points is a random quantity.[1] This randomness generally invalidates the use of neighboring observations as proper repeated measurements.

We propose to overcome this challenge by expressing the joint density of the dependent variable, the mismeasured variable, and its value at a neighboring point, conditional on the distance to the neighboring point. This approach enables us to condition on a fixed distance to generate a virtual repeated measurement with statistical properties suitable to play the role of the counterpart of a fixed lag repeated measurement. We show that any given fixed distance permits the identification of the model, but efficiency considerations suggest the use

[1] Although there is long tradition of using neighboring observations as instruments in the spatial literature (e.g., Kelejian and Prucha (1998)), it is well-known that instruments cannot be used to correct for measurement error in general nonlinear models (Amemiya, 1985). Furthermore, such instruments cannot simply be converted into suitable repeated measurements, because the variable distance between the observations causes an unknown bias in the measurement error that is difficult to account for.

of a weighted average of estimators coming from different distances. The effectiveness and feasibility of this approach is demonstrated through a controlled simulations study.

The estimator is applied to provide further corroboration to an important study (Michalopoulos and Papaioannou, 2013) seeking to quantify whether pre-existing political structures of ethnic groups in the pre-colonial Africa still have a significant impact on contemporary economic development. The main descriptor of the political structure is a measure of centralization of political power (i.e., whether decisions are made at a very local level in a decentralized fashion or at a broader level in a centralized fashion).

The conclusions of this study, however, rest on the accuracy of such estimated centralization measures. Our approach specifically enables us to quantify the relevant error distributions and obtain measurement error-robust estimates by exploiting the spatial nature of the data to construct repeated measurements of centralization using data points in the geographical vicinity of each observation. Remarkably, our results reinforce those of the authors by uncovering an even stronger relationship between pre-colonial centralization and contemporary development. This points to a significant potential for our method to circumvent measurement error issues in a broader range of similar applications.

The paper is organized as follows. Section 2 describes the setup, its motivation and establishes identification, section 3 discusses the estimator and its implementation, section 4 provides simulations to assess the performance of the estimator, section 5 applies our estimator to the study of political complexity on current economic development, and section 6 concludes.

# 2 Setup and Identification

Throughout the text, we denote random variables (or random functions) by upper case letters, while the corresponding lower case letter denotes specific values. We also denote (conditional) densities by $f$ with suitable random variable subscripts and assume their existence, relative to a suitable dominating measure.

We consider a spatial setup, denoting (potentially abstract[2]) locations by $S$. The model of interest is

$$Y(S) = g(X^*(S)) + U(S) \tag{1}$$

where $Y(S)$ is the dependent variable, $X^*(S)$ is an unobserved regressor, $U(S)$ is the model error. We observe a sample $(S_i, X_i \overset{\text{def}}{=} X(S_i), Y_i \overset{\text{def}}{=} Y(S_i), i = 1, ..., n)$ where $X(S)$ is an error-contaminated version of $X^*(S)$:

$$X(S) = X^*(S) + V(S). \tag{2}$$

Although, for simplicity, we do not make this explicit in the notation, covariates could be included by making all assumptions and densities conditional on the covariates in what follows.

[2] Abstract location examples could include product or individual characteristics. In "big data" settings, low-dimensional abstract location variables could be extracted from high-dimensional covariates through linear (Jolliffe, 1986) or nonlinear (Gunsilius and Schennach, 2019) principal component analysis.

We are interested in the conditional distribution $f_{Y(s)|X^*(s)}(y|x^*)$, which will allow us to recover the function $g$. Since $X^*(s)$ is unobserved due to measurement error, this density is not directly revealed by the data and its identification will be secured through availability of repeated measurements. Here we observe that spatial processes provide natural candidates for repeated measurements for $X(s)$ through neighboring observations $X(s+\Delta s)$, where $\Delta s$ is some fixed vector-valued shift. Our identification argument relies on one specific value of $\Delta s$, but, in fact, there are potentially an infinite number of repeated measurements (for different $\Delta s$), which can be used to improve efficiency.

In our approach, the disturbances satisfy the following:

**Assumption 2.1** (Exclusion restrictions)**.** *The random variables $V(s), U(s), V(s+\Delta s)$ are mutually independent conditional on $X^*(s)$ for any $s$ and any $\Delta s$ such that $\|\Delta s\| > d$ for some given known $d \geq 0$.*

The fact that the assumption involves a spatial shift $\Delta s$ will allow us to consider a neighboring observation as repeated measurements. Note that, while Assumption 2.1 places restrictions on the spatial dependence of the measurement error process $V(s)$, we place no such restrictions on the generating processes of $X^*(s)$, $U(s)$ and thus $Y(s)$.

To precisely state our identification results, we first require some basic regularity conditions about the distributions.

**Assumption 2.2** (Existence of bounded densities)**.** *For a given $\Delta s$, the joint distribution of $Y(s)$ and $X(s)$, $X(s+\Delta s)$ and $X^*(s)$, admits a bounded density $f_{Y(s),X(s),X(s+\Delta s),X^*(s)}$ with respect to a dominating measure of the form $\mu_Y \times \mu_X \times \mu_X \times \mu_X$ where $\mu_Y$ is unrestricted while $\mu_X$ could be either the Lebesgue measure or a discrete measure supported on a finite set of points. All marginal and conditional densities are also bounded.*

These conditions on the density allow us to cover both continuous and discrete $X(s)$ (and $X^*(s)$), thus covering either measurement error or misclassification. Although our presentation abstracts away from the differences in these two cases, they demand significantly different treatments both on a theoretical and implementation level (see (Hu and Schennach, 2008) and (Hu, 2008), for the continuous and discrete cases, respectively). Few restrictions are placed on the nature of the distribution $Y(s)$.

We also impose

**Assumption 2.3** (Centering)**.** *For known functionals $M_x, M_y$, we have $M_x[f_{X(s)|X^*(s)}(\cdot|x^*)] = x^*$ and $M_y[f_{Y(s)|X^*(s)}(\cdot|x^*)] = g(x^*)$ for any $x^*$.*

This type of assumption is commonly made in the context of nonclassical measurement error models (Hu and Schennach, 2008) and extends standard conditional mean assumptions to more general centering concepts (e.g. mode, median or general quantiles). For conciseness, we state here a condition that is sufficient to transparently cover both the discrete and continuous cases, although it could be relaxed in the discrete case (see Hu (2008)).

We also require nonparametric analogues of rank conditions, which have a long history in the nonparametric instrumental variable literature (Newey and Powell, 2003; Hall and Horowitz, 2005; Hu and Schennach, 2008)

**Assumption 2.4** (Injectivity of operators)**.** *The operators $L_{X(s)|X^*(s)}$ and $L_{X(s+\Delta s)|X^*(s)}$ are injective, where $L_{B|A}$ is defined through its action on a function $h$ by $[L_{B|A}h](b) \overset{\text{def}}{=} \int f_{B|A}(b|a)h(a)d\mu_X(a)$.*

In the discrete case, this condition reduces to a familiar full rank condition on the matrices of conditional probabilities $f_{X(s)|X^*(s)}(x|x^*)$ and $f_{X(s+\Delta s)|X^*(s)}(x|x^*)$ (indexed by $x$ and $x^*$).

For the outcome variable $Y(s)$, a weaker rank-like condition is sufficient:

**Assumption 2.5** (Outcome variation)**.** *For all $x_1^* \neq x_2^*$, the set $\{y : f_{Y(s)|X^*(s)}(y|x_1^*) \neq f_{Y(s)|X^*(s)}(y|x_2^*)\}$ has positive probability.*

Hu and Xiao (2018) observe that, in the discrete case, these conditions provide easily verifiable conditions that reach Kruskal's minimum rank bounds for the identification of discrete probability models defined in terms of three-way arrays (Kruskal, 1977). As noted in Schennach (2016), in the continuous case, these two conditions also reach a continuous analog of Kruskal's minimum rank bounds.

We are now ready to state our main identification result (proven in the Appendix):

**Theorem 2.1** (Identification)**.** *Under assumptions 2.1 to 2.6, the (conditional) densities $f_{Y(s)|X^*(s)}$, $f_{X(s)|X^*(s)}$, $f_{X(s+\Delta s)|X^*(s)}$, and $f_{X^*(s)}$ are identified (almost everywhere) from the observed joint density $f_{Y(s),X(s),X(s+\Delta s)}$.*

From this result, any model (such as Equation (1)) that seeks to determine a relation between $Y$ and $X^*$ is also identified. The practical use of this identification result obviously requires the determination of the density $f_{Y(s),X(s),X(s+\Delta s)}$. When locations are regularly spaced, $\Delta s$ can be fixed so that knowledge of the sample $(Y(S_i), X(S_i), X(S_i + \Delta s))$ is sufficient for estimation. However, as noted earlier, if locations $S_i$ have random spacings, there may not be pairs of observations exactly $\Delta s$ apart from each other. In this case, we propose to view the density of interest as a conditional density:

$$f_{Y(s),X(s),X(s+\Delta s)}(y,x,z) \overset{\text{def}}{=} f_{Y(s),X(s),X(s+\Delta S)|\Delta S}(y,x,z|\Delta s) \tag{3}$$

$$= \frac{f_{Y(s),X(s),X(s+\Delta S),\Delta S}(y,x,z,\Delta s)}{f_{\Delta S}(\Delta s)} \tag{4}$$

where the numerator and denominator can be estimated by kernel smoothing over all the continuous variables, including $\Delta S$, under the assumptions that locations are drawn from some continuous density over space. Naturally, this approach relies on a stationarity assumption for estimation:

**Assumption 2.6** (Stationarity)**.** *The joint distribution of $Y(s), X(s), X(s+\Delta s)$ does not depend on $s$.*

While stationarity assumptions have been criticized in spatial applications (Pinkse and Slade, 2010*a*) due to inherent geographic inhomogeneities, this assumption can be weakened by instead considering a conditional density

$$f_{Y(s),X(s),X(s+\Delta s)|T,\Delta S}(y,x,z|t,\Delta s) \tag{5}$$

where $T$ is a position-dependent variable that controls for the source of the lack of stationarity. All above assumptions and results are then understood to be conditional on $T$ (which is suppressed in the notation, for simplicity). For instance, $T$ could be the distance to the nearest body of water, the degree of a node in graph/network applications or controls for treatment status or law enactments.[3]

It is even possible, in principle, to fully relax stationarity by partitioning the space of $S$ through a grid of resolution $b$ and letting $T$ denote which grid "box" point $S$ belongs to. If we let $b \to 0$ as $n \to \infty$, stationarity conditional on $T$ will hold asymptotically under suitable regularity conditions regarding the generating process. Consistency can be maintained if $b \to 0$ sufficiently slowly as sample size grows. It is also possible to replace partitioning into boxes by suitable kernel smoothing. We leave a formal analysis of these extensions for future work, however, to avoid obscuring the main ideas.

# 3 Estimator and Implementation

Estimation is based on the identity

$$\begin{aligned} & f_{Y(s),X(s),X(s+\Delta s)}(y,x,z) \\ = & \int f_{Y(s)|X^*(s)}(y|x^*) f_{X^*(s)}(x^*) f_{X(s)|X^*(s)}(x|x^*) f_{X(s+\Delta s)|X^*(s)}(z|x^*,\Delta s) d\mu_X(x^*), \end{aligned} \tag{6}$$

implied by conditional independence (Assumption 2.1). Theorem 2.1 implies that this integral equation, for a given left-hand side density, has a unique solution. Hence, we can use the right-hand side of (6) to construct an estimator analogous to a maximum likelihood estimator (MLE) in terms of 4 unknown densities to be estimated. In the misclassification case ($\mu_X$ discrete), the densities $f_{X(s)|X^*(s)}(x|x^*)$, $f_{X(s+\Delta s)|X^*(s)}(z|x^*)$ and $f_{X^*(s)}(x^*)$ can be parametrized as a matrix (or a vector) of probabilities, as in Hu (2008). In the continuous $\mu_X$ case, the densities are represented by a sieve approximation, as in Hu and Schennach (2008).

One important aspect of our approach that is distinct from earlier work (such as Hu and Schennach (2008)) is the fact that $X(s+\Delta s)$ is not a repeated measurement in the usual sense, because we only have access to its estimated density, not its specific value of each sample point. We address this by sampling pseudo-observations from the density

$$f_{X(s+\Delta S)|Y(s),X(s),\Delta S}(z|y,x,\Delta s) = \frac{f_{Y(s),X(s),X(s+\Delta S),\Delta S}(y,x,z,\Delta s)}{\int f_{Y(s),X(s),X(s+\Delta S),\Delta S}(y,x,z,\Delta s) d\mu_X(z)}$$

where the right-hand side can be estimated from kernel smoothing, as in Equation (3), for some pre-specified $\Delta s$. For estimation purposes, our sample then consists of $Y_i \overset{\text{def}}{=} Y(S_i)$, $X_i \overset{\text{def}}{=} X(S_i)$ and $Z_i$ drawn from an estimate of $f_{X(s+\Delta S)|Y(s),X(s),\Delta S}(z|Y_i,X_i,\Delta s)$ for $i = 1,\ldots,n$. One could, of course, draw multiple pseudo-observations per data point to reduce

[3] It should be stated that high-dimensionality of $T$ may have an impact on estimation accuracy, due to the data needs associated with high-dimensional density estimations. In practice, dimensionality of $T$ may thus be limited by the size of the available data.

the simulation noise, although we did not find this to be necessary in our application and simulations study.

We then use a semiparametric sieve maximum likelihood estimator (MLE)(Shen, 1997) of the form:

$$(\hat{\theta}, \hat{\eta}, \hat{f}_1, \hat{f}_2, \hat{f}_3) = \underset{(\theta,\eta,f_1,f_2,f_3)}{\arg\max} \sum_{i=1}^{n} \ln L(Y_i, X_i, Z_i; \theta, \eta, f_1, f_2, f_3) \tag{7}$$

where the maximization if performed under suitable constraints detailed below and where

$$L(y, x, z; \theta, \eta, f_1, f_2, f_3) \overset{\mathsf{def}}{=} \int_{\mathcal{X}^*} f(y|x^*; \theta, \eta) f_1(x^*) f_2(x|x^*) f_3(z|x^*, \Delta s) dx^* \tag{8}$$

where $\mathcal{X}^*$ denotes the support of $X^*$. In (8), the density $f_{Y(s)|X^*(s)}(y|x^*)$ is indexed by $\theta$, the parameter of interest and $\eta$, some nuisance parameter. In our setup, $\theta$ could specify the shape of the function $g$ in Equation (1), while $\eta$ could specify the density of the disturbance $U(S)$. (Other ways to separate $\delta$ and $\eta$ are possible: for instance, $\theta$ could represent an average derivative, while $\eta$ includes both the density of $U(S)$ and degrees of freedom of $g$ which do not affect the average derivative. See Hu and Schennach (2008) for more details.) No such separation is imposed on the remaining densities $(f_1, f_2, f_3)$, which are all considered nuisance parameters. Note that only $f_3$ depends on the shift $\Delta s$. The parameter of interest $\theta$ is considered finite dimensional, while all other parameters are infinite dimensional and approximated through sieves in finite samples. This setup reflects most empirical studies and will enable the development of an asymptotic theory for asymptotic normality and root-$n$ consistency (in the next section).

The optimization in Equation (7) must be performed under some constraints in order to enforce the assumptions needed for identification as well as basic properties of densities. To enforce nonnegativity constraints, we actually model the square root of densities, so that their respective squares are automatically positive:

$$f_1^{\frac{1}{2}}(x^*) = \sum_{i=1}^{i_n+1} \alpha_i p_{i,1}(x^*) = \boldsymbol{p}_1(x - x^*)'\boldsymbol{\alpha} \tag{9}$$

$$f_2^{\frac{1}{2}}(x|x^*) = \sum_{i=1}^{i_n+1} \sum_{j=1}^{j_n+1} \beta_{ij} p_{i,2}(x - x^*) q_j(x^*) = \boldsymbol{p}_2(x - x^*)'\boldsymbol{\beta}\boldsymbol{q}(x^*) \tag{10}$$

$$f_3^{\frac{1}{2}}(z|x^*) = \sum_{i=1}^{i_n+1} \sum_{j=1}^{j_n+1} \gamma_{ij} p_{i,3}(x - x^*) q_j(x^*) = \boldsymbol{p}_3(z - x^*)'\boldsymbol{\gamma}\boldsymbol{q}(x^*) \tag{11}$$

Let $x^* \in [0, l_x]$, $(x - x^*) \in [-l_1, l_1]$, and $(z - x^*) \in [-l_2, l_2]$. If we use Fourier series, we have $p_{k,1}(a) =$ $\cos(\frac{k2\pi}{l_x}a)$ or $p_{k,1}(a) =$ $\sin(\frac{k2\pi}{l_x}a) \forall k > 1$, $p_{k,m}(a) =$ $\cos(\frac{k\pi}{l_m}a)$ or $p_{k,m}(a) =$ $\sin(\frac{k\pi}{l_m}a) \forall k > 1$ and $m \in \{1, 2\}$, and $q_k(a) =$ $\cos(\frac{k\pi}{l_x}a)$. $f(y_i|x^*)$ can be specified similarly or be fully parametric. In the following, we use both cosines and sines in numbers $\frac{i_n}{2}$ each (the first of the $(i_n + 1)$ terms being the constant).

Since non-negativity constraints are automatically satisfied by squaring, $M_x[f_2(\cdot|x^*)] = x^*$, $M_y[f_1(\cdot|x^*)] = g(x^*)$ and densities integrating to 1 remain to enforce. We proceed as follows. Consider the density $f = (\sum_{i=1}^{i_n+1} \sum_{j=1}^{j_n+1} p_i \Lambda_{ij} q_j)^2$. In matrix form, $f = (p^{'}\Lambda q)^2 = q^{'}\Lambda^{'} pp^{'}\Lambda q$. When considering the constraint that densities integrate to 1, the use of an orthonormal basis yields

$$\int f = q^{'}\Lambda^{'} I \Lambda q = (q' \otimes q')vec(\Lambda^{'}\Lambda) \tag{12}$$

For the vector of orthogonal functions $B(x^*) = [1 \ \cos(x^*) \ ... \ \cos(2j_n x^*)]'$ and the transformation T that satisfies $TB(x^*) = q(x^*) \otimes q(x^*)$, we obtain the restriction $B'(x^*)T'vec(\Lambda^{'}\Lambda) = 1$, i.e. $[T'vec(\Lambda^{'}\Lambda)]_{11} = 1$ and $[T'vec(\Lambda^{'}\Lambda)]_{k1} = 0$ for $k > 1$. Other constraints can be treated similarly; after a bit of algebra, it is simple to implement the constraint brought by the functional, whether the expected value, the median, the mode, or a percentile.

Solving the optimization problem (7) subject to the constraints delivers $\hat{\theta}_{\Delta s}$ for the chosen $\Delta s$. Although any single nonzero value of $\Delta s$ delivers a consistent estimator, its efficiency can be improved by combining the information provided by all other distances. Since kernel estimates at two nearby points are asymptotically uncorrelated, an asymptotically optimal linear combination of the different $\hat{\theta}_{\Delta s}$ simply involves weights inversely proportional to the variance of the corresponding estimators. This approach is supported by our simulation experiments in finite sample, which reveal only weak correlation between the estimation errors of estimators based on different distances. Naturally, to ensure that this asymptotic behavior is reached, it is recommended that the spacing between the different $\Delta s$ be selected so that it converges to zero slower than the bandwidth does, as sample size grows.

# 4 Inference

Our estimator's hybrid nature (i.e. with $Z_i$ drawn from a kernel density estimator fed into a sieve semiparametric MLE) makes its asymptotic analysis much more involved than an application of standard results on sieve MLE and complicates an explicit calculation of its asymptotic variance. To address this, we establish that the construction of the $Z_i$ still yields an estimator that admits an asymptotically linear representation, provided that the corresponding (infeasible) sieve estimator with observed $Z_i$ has that property. This result, stated formally below, will simultaneously ensure asymptotic normality, root $n$ consistency, and asymptotic validity of the bootstrap for our estimator.

To state our main asymptotic result, we define a profiled likelihood that focuses on the parameter $\theta$ of interest:

$$\mathcal{L}(\theta, f) = E\left[\ln L\left(Y, X, Z; \theta, \omega\left(\theta\right)\right)\right] \tag{13}$$

for

$$\omega\left(\theta\right) = \arg\max_{\omega \in \Omega} E\left[\ln L\left(Y, X, Z; \theta, \omega\right)\right]$$

with $Z$ distributed according to the conditional density $f \equiv f_{Z|X,Y,\Delta s}$ of the repeated measurement, and where $\omega \equiv (\eta, f_1, f_2, f_3)$ denotes all the nuisance parameters, which belong to some set $\Omega$ imposing suitable regularity conditions. Let $\theta_0$ and $f_0$ denote the true values of $\theta$ and $f$, respectively.

The empirical counterpart of (13) is:

$$\hat{\mathcal{L}}\left(\theta, \hat{f}\right) = \frac{1}{n}\sum_{i=1}^{n} \ln L\left(Y_i, X_i, Z_i; \theta, \hat{\omega}\left(\theta\right)\right) \tag{14}$$

for

$$\hat{\omega}\left(\theta\right) = \arg\max_{\hat{\omega}\in\Omega_n} \frac{1}{n}\sum_{i=1}^{n} \ln L\left(Y_i, X_i, Z_i; \theta, \hat{\omega}\right)$$

with $Z_i$ drawn from the density $\hat{f} \equiv \hat{f}_{Z|X,Y,\Delta s}$ and the maximum is taken over a sample-size dependent sieve space $\Omega_n$ (as described in the previous section). We define $\hat{\theta} = \arg\max_\theta \hat{\mathcal{L}}\left(\theta, \hat{f}\right)$, for some estimated $\hat{f}$.

In accordance with the definition of a profiled likelihood, all gradients with respect to $\theta$ below (denoted by $\nabla$) incorporate the effect of simultaneous changes in the nuisance parameters through the function $\omega\left(\theta\right)$ or $\hat{\omega}\left(\theta\right)$. This approach provides a simple way to formally abstract away the nuisance parameters from the expansion relevant to the asymptotics of $\hat{\theta}$. Let $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ denote the support of $X, Y, Z$, respectively[4], while $\Theta$ is the parameter space for $\theta$. Let $\mathcal{F}$ denote a neighborhood of $f_0$ (where the sup-norm is used for $f$).

With these definitions in mind, we can now state our key assumptions.

**Assumption 4.1** (Consistency). *(i) $\mathcal{L}\left(\theta, f_0\right)$ is uniquely maximized at $\theta = \theta_0$, (ii) $\sup_{\theta\in\Theta} \sup_{f\in\mathcal{F}} \left|\hat{\mathcal{L}}\left(\theta, f\right) - \mathcal{L}\left(\theta, f\right)\right| \xrightarrow{p} 0$ and (iii) $\mathcal{L}\left(\theta, f\right)$ is continuous in $f$ at $f_0$ uniformly for $\theta \in \Theta$.*

**Assumption 4.2** (Limiting distribution). *(i) $\sup_{\theta\in\Theta} \left|\nabla\nabla'\hat{\mathcal{L}}\left(\theta, f_0\right) - \nabla\nabla'\mathcal{L}\left(\theta, f_0\right)\right| \xrightarrow{p} 0$, (ii) $H = \nabla\nabla'\mathcal{L}\left(\theta_0, f_0\right)$ is invertible, (iii) $\sup_{\theta\in\Theta, f\in\mathcal{F}} \left|\nabla\nabla'\hat{\mathcal{L}}\left(\theta, f\right) - \nabla\nabla'\mathcal{L}\left(\theta, f\right)\right| \xrightarrow{p} 0$, (iv) $\nabla\nabla'\mathcal{L}\left(\theta, f\right)$ is continuous in $f$ at $f_0$ uniformly for $\theta \in \Theta$.*

We deliberately phrase Assumptions 4.1 and 4.2 in a high-level form because they arise in the asymptotic analysis of a conventional sieve MLE estimator and a number of different possible sufficient conditions are already available in the literature (e.g. Hu and Schennach (2008)). Assumption 4.1(i) merely restates the conclusion of our earlier identification argument. Assumptions 4.1(ii), 4.2(i) and (iii) only require uniform consistency and thus follow from uniform laws of large numbers for spatial data (see Jenish and Prucha (2009), who establish laws of large numbers under mixing and moment conditions and turn them into uniform laws of large numbers by adding stochastic equicontinuity and dominance). These conditions are slightly strengthened here (relative to a standard sieve MLE) to account for an estimated $f_0$. Assumptions 4.1(iii), 4.2(ii) and (iv) do not involve random quantities, hence the spatial nature of the data is of no consequence. Assumption 4.1(iii) and 4.2(iv) ensures that estimation of $f_0$ won't degrade the estimator's properties. We now use a more

[4] The assumption of rectangular support of $(X, Y, Z)$ is made purely for notational convenience and can be trivially relaxed.

primitive formulation for the assumptions that are specific to our estimator, for instance, those related to the fact that the distribution of $Z$ is estimated and that $Z_i$ are simulated draws.

**Assumption 4.3** (Support). *(i) $f_{Y,X}(y,x) \geq \varepsilon > 0$ and (ii) $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ are compact.*

This assumption is commonly made in the analysis of semiparametric estimators with estimated densities, but could be relaxed through tail trimming arguments. We do not consider this extension here, because the intricate details needed would obscure the main ideas.

**Assumption 4.4** (Density estimation). *(i) $\sup_{y,x \in \mathcal{Y}\times\mathcal{X}} \left| \hat{f}_{Y,X}(y,x) - f_{Y,X}(y,x) \right| = o_p\left(n^{-1/4}\right)$, (ii) $\sup_{y,x,z \in \mathcal{Y}\times\mathcal{X}\times\mathcal{Z}} \left| \hat{f}_{Y,X,Z}(y,x,z) - f_{Y,X,Z}(y,x,z) \right| = o_p\left(n^{-1/4}\right)$, (iii) $\sup_{y,x \in \mathcal{Y}\times\mathcal{X}} \left| E\left[\hat{f}_{Y,X}(y,x)\right] - f_{Y,X}(y,x) \right| = o\left(n^{-1/2}\right)$ and (iv) $\sup_{y,x,z \in \mathcal{Y}\times\mathcal{X}\times\mathcal{Z}} \left| E\left[\hat{f}_{Y,X,Z}(y,x,z)\right] - f_{Y,X,Z}(y,x,z) \right| = o\left(n^{-1/2}\right)$.*

Sufficient conditions for Assumptions 4.4(i) and (ii) in spatial contexts can be found in Carbon, Tran and Wu (1997). Assumptions 4.4(iii) and (iv) are not affected by spatial dependence and standard conditions implying them can be found in Andrews (1995).

**Assumption 4.5** (Generated $Z_i$). *(i) $\nabla \ln L(y,x,z;\theta,\omega(\theta))$ is bounded and Lipschitz in $z$ and (ii) $f_{Z|YX}(z|y,x)$ is bounded.*

These conditions are needed to account for the simulated nature of $Z_i$ and are simple to verify by inspection.

**Theorem 4.1** (Asymptotically linear representation). *Under Assumptions 4.1, 4.2, 4.3, 4.4 and 4.5,*

$$\begin{aligned} n^{1/2}\left(\hat{\theta} - \theta\right) &= n^{-1/2}\sum_{i=1}^{n} \psi_{MLE}(Y_i, X_i, Z_i) \\ &\quad + n^{-1/2}\sum_{i=1}^{n} \psi_{kernel}(Y_i, X_i, Z_i) + o_p(1) \end{aligned}$$

*where*

$$\psi_{MLE}(y,x,z) = -H^{-1}\nabla \ln L(y,x,z;\theta,\omega(\theta))$$

*is the usual influence function of a standard sieve semiparametric MLE with observed $Z_i$, while*

$$\begin{aligned} \Psi_{kernel}(y,x,z) &= H^{-1}\left(\nabla \ln L(y,x,z;\theta,\omega(\theta)) - E\left[\nabla \ln L(Y,X,Z;\theta,\omega(\theta))\right]\right) \\ &\quad + H^{-1}\left(E\left[\nabla \ln L(Y,X,Z;\theta,\omega(\theta))\,|Y=y, X=x\right]\right. \\ &\quad \left. - E\left[\nabla \ln L(Y,X,Z;\theta,\omega(\theta))\right]\right) \end{aligned}$$

*is the correction term due to constructing the measurement $Z_i$.*

The conclusion of Theorem 4.1 is stated in a way such that any central limit theorems for sample averages involving spatial data (see, e.g., Bolthausen (1982); Lahiri (2003); Jenish and Prucha (2009, 2012) for CLT under various types of mixing and moment conditions) can be freely used to obtain the limiting distribution. If a resampling approach is preferred, a block bootstrap (Hall, Horowitz and Jing, 1995; Nordman, Lahiri and Fridley, 2007) approach should be used to account for the possible spatial dependence.

# 5 Simulations

We conduct simulations to test the performance of our spatial estimator that corrects for measurement error. We generate a random field $X^*(S)$ on a subset of $\mathbb{R}^2$, on which we then construct $Y(S) = g(X^*(S)) + U(S)$ and $X(S) = X^*(S) + V(S)$ which are observed for $S = S_i, i = 1, ..., n$. Sampled location are randomly chosen and thus unevenly spaced.

We specify $g(x^*) = \theta_1 + \theta_2 x^*$, and $(\theta_1, \theta_2) = (-3.5, 2)$. The error terms, $U$ and $V$, are normally distributed independently of $x^*$ with standard deviations of 1.3 and 0.8, respectively.

We parametrically specify $f(y_i|x^*)$ in the optimization problem and analyze results for $(\theta_1, \theta_2, \sigma_u)$. Two forms of the estimators are tested: a simple, unweighted average over all distances, and an inverse-variance-weighted average scheme.

The number of Sieve terms has been chosen by examining the resulting densities and ensuring small variations in the number of Sieves do not cause the resulting estimator to vary much. This is in line with the suggestion in Hu and Schennach (2008) ”that a valid smoothing parameter can be obtained by scanning a range of values in search of a region where the estimates are not very sensitive to small variations in the smoothing parameter”. With Section 3’s notations, this leads to $i_n = 6$, $j_n = 4$ for $f(x|x^*)$, $i_n = 4$, $j_n = 4$ for $f(z|x^*)$, and $i_n = 4$ for $f(x^*)$. In the appendix, we report additional simulations for a range of sieves truncation choices - $i_n = j_n = 2k, k = 1, 2, 3$ for all densities - which suggest performance does not depend strongly on the number of Sieves terms within the range $4-6$. Lower values appear too rough and reduce the performance of the estimator, while higher values add too much variance and let the number of parameters explode, which also increases computational burden. These two versions of the estimators are compared to the infeasible OLS that uses the unobserved regressor and to the biased OLS estimator that regresses on the mismeasured regressor. Results are displayed in the following tables.

Table 1: $\theta_1 = -3.5$

| | Mean | Standard deviation | RMSE |
|---|---|---|---|
| Infeasible OLS | $-3.50$ | 0.12 | 0.12 |
| OLS | $-0.77$ | 0.15 | 2.74 |
| Unweighted Spatial | $-3.58$ | 0.10 | 0.13 |
| Weighted Spatial | $-3.57$ | 0.08 | 0.11 |

It is seen — as expected in presence of substantial measurement error — that the biased OLS regression using the mismeasured regressor performs poorly, displaying strong

$\theta_2 = 2$

| | Mean | Standard deviation | RMSE |
|---|---|---|---|
| Infeasible OLS | 2.00 | 0.03 | 0.03 |
| OLS | 1.22 | 0.04 | 0.78 |
| Unweighted Spatial | 2.04 | 0.04 | 0.06 |
| Weighted Spatial | 2.03 | 0.03 | 0.05 |

$\sigma_u = 1.3$

| | Mean | Standard deviation | RMSE |
|---|---|---|---|
| Infeasible OLS | 1.30 | 0.02 | 0.02 |
| OLS | 1.80 | 0.03 | 0.50 |
| Unweighted Spatial | 1.26 | 0.03 | 0.05 |
| Weighted Spatial | 1.26 | 0.03 | 0.05 |

Simulations with 1500 observations. Infeasible OLS refers to the infeasible OLS estimator using the unobserved true regressor; OLS is the biased OLS estimator using mismeasured covariate; Unweighted Spatial is our unweighted average spatial estimator; Weighted Spatial is the optimally weighted average spatial estimator.

attenuation bias.

Our estimator exhibits significant improvement over biased OLS. For all parameters, it also attains a RMSE that is no more than three times that of the efficient, infeasible OLS estimator that uses the actual covariate. Of course, a finite-sample bias is expected, especially given the slight misspecification induced by the truncation of the Sieve expansion, but its magnitude remains reasonable. Furthermore, the estimator performs similarly to OLS in terms of variance.

Further simulations exploring the link between the distance $\Delta s$ that determines the choice of instrument and estimation accuracy reveal a non-trivial relationship. Using the RMSE of the estimated $\theta_2$ as a figure of merit, we find that using $\Delta s = 1.5$ provides the best result with a RMSE of 0.04, beating both the closer distance of 0.75 (RMSE 0.07) and the larger distance of 1.25 (RMSE 0.9). The reason for this non-monotone behavior is likely that closer observations improve the instrument's strength, while larger distances induces a higher count of observations, which allows a more precise estimate of the conditional density. The analysis of the $\Delta s$-dependence is done here for illustration purposes — when using a weighted average over a range of $\Delta s$ (as we shall do in our application), there is no need to select a specific $\Delta s$.

While estimators from individual distances can exhibit heavy tails, the presence of outlier estimates is alleviated thanks to averaging over different estimates. For instance in estimating $\theta_2 = 2$, the first percentile is 1.98 and the $99^{th}$ reaches 2.08. In this example, the intercept is somewhat more prone to outliers; in estimating $\theta_1 = -3.5$, the corresponding percentile figures are -4.15 and -3.51.

# 6 Application

We revisit the influential study of Michalopoulos and Papaioannou (2013) to demonstrate how our approach can effectively deliver measurement-error robust estimation and inference in the context of spatial data, without necessitating additional auxiliary variables, such as instruments or validation data. In this particular application, the possibility of significant measurement error in a key regressor is an important concern that existing methods have been unable to fully address.

This study investigates the relationships between pre-colonial ethnic political centralization and contemporary development. The underlying motivation is to confirm anecdotal observations that the pre-existence of a complex large-scale political structure within ethnic groups appears to strongly impact economic development, independently of political structures put in place during colonization. The pre-colonial political structure is captured by measures of the extent of jurisdictional hierarchy beyond the local level developed by Murdock (1969). Obtaining such measures is challenging, as it involves subjective assessments, and is thus prone to misclassification errors, as discussed by Michalopoulos and Papaioannou (2013). As this quantity appears as a regressor in the analysis, the possibility of measurement error induced bias must be considered.

The dependent variable in this study is economic activity. Given unavailability of comparable economic indicators across African ethnic homelands, the authors employ nighttime artificial light intensity as a proxy for economic activity, in the spirit of Henderson, Storeygard and Weil (2012), Elvidge et al. (1997) and Doll, Muller and Morley (2006), among others.

Their main regression takes the form:

$$y_i = \beta_0 + \beta_1 x_i^* + w_i' \beta_W + \epsilon_i \tag{15}$$

where $y_i$ denotes light density at night, $x_i^*$ is the (correctly-measured) level jurisdictional hierarchy or "complexity", taking value in $\{0, 1, 2, 3, 4\}$, and $w_i$ is a vector of covariates including population density, location controls (distance to the capital city, distance to the border, and distance to the coast), geographic features (land suitability for agriculture, malaria stability index, land area under water, and petroleum and diamond dummies), and income per capita. Country fixed effects are also considered.

Results from Table 2 and 3 in Michalopoulos and Papaioannou (2013), which are partially reproduced in Table 2, suggest that a one unit increase in the jurisdictional hierarchy index — roughly corresponding to a one standard deviation increase — leads to an increase in light luminosity of 20 % (with all controls and country fixed effects) to 40% (without controls) — corresponding to a 0.1 to 0.2 standard deviation increase. See Michalopoulos and Papaioannou (2013) for details.

These results suggest a strong relationship between pre-colonial political complexity and current economic development, and here we seek to ensure that this finding is robust the presence of misclassification errors. It is also of independent interest to quantify how prevalent classification errors are in such frameworks.

While the correctly classified variable is unobserved, it can be argued that the misclassification is mode-preserving (Schennach, 2018), i.e. for any true underlying level of complexity,

Table 2: Replicated results

| | Coefficient | se | 95% CI lb | 95% CI ub |
|---|---|---|---|---|
| No controls | 0.41 | 0.12 | 0.17 | 0.66 |
| Controls | 0.2 | 0.05 | 0.1 | 0.29 |
| Controls and FE | 0.18 | 0.05 | 0.08 | 0.27 |

OLS estimate for hierarchy index coefficient on (log) light luminosity; standard errors (se); lower bound (lb) and upper bound (ub) of 95% confidence interval (CI). FE refers to country fixed effects.

Table 3: Measurement error robust estimates

| | Coefficient | se | 95% CI lb | 95% CI ub |
|---|---|---|---|---|
| No control | 1.28 | 0.31 | 0.68 | 1.87 |
| Controls | 0.77 | 0.16 | 0.45 | 1.08 |
| Controls and FE | 0.27 | 0.13 | 0.01 | 0.52 |

Measurement-error corrected estimate for hierarchy index coefficient on (log) light luminosity; standard errors (se) are estimated using a block bootstrap; lower bound (lb) and upper bound (ub) of 95% confidence interval (CI). FE refers to country fixed effects.

the most likely observed assessed value is the correct appraisal. Combined with repeated measurements provided by the spatial structure, this identifies the distribution $f(y_i|x^*)$.

In the sample, the highest level of complexity ($x_i = 4$) occurs less than 1% of the time, thus making it difficult to estimate probabilities involving that event with good accuracy. To alleviate the issue, we pool outcomes $X = 3$ and $X = 4$ together at the value 3.[5]

We estimate the spatial autocorrelation of the hierarchical complexity to vary from 0.35 to 0.25 for distances between 10 and 150 kilometers. This supports our identification strategy and we consider $\Delta s = j \times 10$ km for $j = 1, ..., 15$ as instruments.

To account for the numerous covariates $W$, we consider a first-step regression of $y$ on $x$ and $W$ and subtract the effects of the covariates to obtain $\tilde{y} \overset{\text{def}}{=} y - W\beta_W$ to which we apply our procedure. This is justified if the explanatory power of the controls does not fundamentally differ when using the true rather than the mismeasured variable as a regressor. As the strongest correlation between X and a control is -0.25, we believe the assumption is plausible. Since the main estimator is nonparametric, its slower convergence rate dominates the asymptotics, and the noise from this parametric first-step is neglected for the computation of asymptotic variances. We exploit our asymptotically linear representation result and estimate standard errors via a block bootstrap.

Applying our measurement-error robust, inverse variance-weighted spatial estimator yields the results shown in Table 3.

A regression without additional controls yields a statistically significant estimate of 1.28,

[5] Alternative strategies would be to use the weighted-average value (3.1) or to drop observations with a 4. These options do not materially change the results, as can be expected by the very low frequency of 4s.

a much larger finding than the OLS counterpart. The coefficient decreases as controls are added, though measurement error robust estimates still point to a stronger influence of political complexity on development than the biased OLS coefficients do. The use of our measurement error robust estimator also does not come at the cost of a significant decrease in statistical significance.

Our method also identifies the misclassification matrix, which is reported in tables 4, 5, and 6. There appears to be substantial misclassification in all specifications. While extreme misclassifications are less frequent, subjective assessments can often deviate to nearby categories and this is reflected in the estimated probabilities. These results support the view that measurement error is a major concern in such applications and that our method offers a viable avenue to address this issue.

Table 4: $\mathbb{P}[X = i|X^* = j]$ (no control)

| | | | |
|---|---|---|---|
| 0.35 | 0.30 | 0.15 | 0.12 |
| 0.35 | 0.35 | 0.31 | 0.28 |
| 0.21 | 0.26 | 0.32 | 0.22 |
| 0.08 | 0.09 | 0.21 | 0.38 |

Table 5: $\mathbb{P}[X = i|X^* = j]$ (controls)

| | | | |
|---|---|---|---|
| 0.42 | 0.07 | 0.16 | 0.19 |
| 0.42 | 0.38 | 0.30 | 0.27 |
| 0.12 | 0.38 | 0.31 | 0.27 |
| 0.04 | 0.17 | 0.23 | 0.27 |

Table 6: $\mathbb{P}[X = i|X^* = j]$ (controls and FE)

| | | | |
|---|---|---|---|
| 0.47 | 0.10 | 0.15 | 0.13 |
| 0.40 | 0.39 | 0.27 | 0.29 |
| 0.09 | 0.39 | 0.32 | 0.29 |
| 0.05 | 0.12 | 0.27 | 0.29 |

Misclassification probability matrices (i: row; j: column

Overall, our results reinforce those of Michalopoulos and Papaioannou (2013) and, if anything, uncover an even stronger relationship between pre-colonial centralization and current development. Not only are the point estimates of the coefficients larger, but their statistical significance also remains very high. Our proposed approach thus seems to provide a practical and feasible way to address measurement error issues at no extra data collection cost in spatial settings. This capability should prove especially useful in the context of noisy historical data and, more broadly, in any noisy data setting where observation pairs can be assigned a quantifiable notion of “proximity”. This not only includes geographically tagged data, but also more abstract spaces, such product or consumer characteristics or network data.

# 7 Conclusion

We have show that the use of spatial data provides a formal and effective way to correct for the presence of potentially nonclassical covariate measurement error in general nonlinear model without relying on distributional assumptions. Using neighboring observations as repeated measurements requires carefully controlling for the nonuniform spacing between observations by constructing the joint distribution of all measurements conditional on the distance between observations, in order to ensure that the resulting measurement system satisfies the appropriate conditional independence restrictions needed to establish identification of the model.

The method’s implementation combines a sieve semiparametric maximum likelihood with a first-step kernel conditional density estimator and simulation methods. Monte Carlo simulations suggest that this implementation performs well at typically available sample sizes.

The method’s effectiveness is further illustrated by revisiting a well-known study of the effect of pre-colonial political structure on current economic development in Africa. Our estimator support the authors’ original findings by showing that their results are robust to allowing for the likely possibility that political structure is measured with error. Our results suggest that the studied relationship could even be stronger than previously thought.

Our approach opens the way to considering much broader classes of repeated measurements than previously thought possible, as long as a well-defined notion of proximity between pairs of observations can be defined. Beyond geographical data, this could be applicable to network data as well as more abstract spaces of consumer or product characteristics.

# References

**Alexeeff, Stacey E., Raymond J. Carroll, and Brent Coull.** 2016. “Spatial measurement error and correction by spatial SIMEX in linear regression models when using predicted air pollution exposures.” *Biostatistics*, 17: 377–389.

**Amemiya, Y.** 1985. “Instrumental Variable Estimator for the Nonlinear Errors-in-Variables Model.” *Journal of Econometrics*, 28: 273–289.

**Andrews, D. W. K.** 1995. “Nonparametric Kernel Estimation for Semiparametric Models.” *Econometric Theory*, 11: 560–596.

**Bolthausen, Erwin.** 1982. “On the central limit theorem for stationary mixing random fields.” *The Annals of Probability*, 1047–1050.

**Carbon, Michel, Lanh Tat Tran, and Berlin Wu.** 1997. “Kernel density estimation for random fields (density estimation for random fields).” *Statistics & Probability Letters*, 36(2): 115–125.

**Chilès, J. P., and N. Desassis.** 2018. "Fifty Years of Kriging." In *Handbook of Mathematical Geosciences.* , ed. B. Daya Sagar, Q. Cheng and F. Agterberg. Springer.

**Cressie, N.** 1993. *Statistics for Spatial Data.* New York:Wiley Interscience.

**Cunha, Flavio, James J Heckman, and Susanne M Schennach.** 2010. "Estimating the technology of cognitive and noncognitive skill formation." *Econometrica*, 78(3): 883–931.

**de Paula, A.** 2017. "Econometrics of Network Models." In *Advances in Economics and Econometrics: Eleventh World Congress.* , ed. B. Honoré, A. Pakes, M. Piazzesi and L. Samuelson, Chapter 8, 268–323. Cambridge University Press.

**Doll, Christopher NH, Jan-Peter Muller, and Jeremy G Morley.** 2006. "Mapping regional economic activity from night-time light satellite imagery." *Ecological Economics*, 57(1): 75–92.

**Elvidge, Christopher D, Kimberley E Baugh, Eric A Kihn, Herbert W Kroehl, Ethan R Davis, and Chris W Davis.** 1997. "Relation between satellite observed visible-near infrared emissions, population, economic activity and electric power consumption." *International Journal of Remote Sensing*, 18(6): 1373–1379.

**Griliches, Z., and J. A. Hausman.** 1986. "Errors in Variables in panel data." *Journal of Econometrics*, 31: 93–118.

**Gunsilius, F., and S. M. Schennach.** 2019. "Independent Principal Component Analysis." Cemmap Working Paper CWP46/19.

**Hall, P., and J. L. Horowitz.** 2005. "Nonparametric Methods for Inference in the Presence of Instrumental Variables." *Annals of Statistics*, 33: 2904–2929.

**Hall, Peter, Joel L Horowitz, and Bing-Yi Jing.** 1995. "On blocking rules for the bootstrap with dependent data." *Biometrika*, 82(3): 561–574.

**Henderson, J Vernon, Adam Storeygard, and David N Weil.** 2012. "Measuring economic growth from outer space." *The American Economic Review*, 102(2): 994–1028.

**Hu, Y., and M. Shum.** 2012. "Nonparametric identification of dynamic models with unobserved state variables." *Journal of Econometrics*, 171: 32–44.

**Hu, Y., and R. Xiao.** 2018. "Global estimation of finite mixture and misclassication models with an application to multiple equilibria." CeMMAP CWP32/18.

**Hu, Yingyao.** 2008. "Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution." *Journal of Econometrics*, 144(1): 27–61.

**Hu, Yingyao, and Susanne M Schennach.** 2008. "Instrumental variable treatment of nonclassical measurement error models." *Econometrica*, 76(1): 195–216.

**Jenish, Nazgul, and Ingmar R Prucha.** 2009. "Central limit theorems and uniform laws of large numbers for arrays of random fields." *Journal of econometrics*, 150(1): 86–98.

**Jenish, Nazgul, and Ingmar R Prucha.** 2012. "On spatial processes and asymptotic inference under near-epoch dependence." *Journal of econometrics*, 170(1): 178–190.

**Jolliffe, I. T.** 1986. *Principal component analysis.* New York:Spinger-Verlag.

**Kelejian, H. H., and I. R. Prucha.** 1998. "A Generalized Spatial Two-Stage Least Squares Procedure for Estimating a Spatial Autoregressive Model with Autoregressive Disturbances." *Journal of Real Estate Finance and Economics*, 17: 99–121.

**Krige, D. G.** 1951. "A statistical approach to some mine valuations and allied problems at the Witwatersrand." Master's diss. University of Witwatersrand.

**Kruskal, J. B.** 1977. "Three-Way Arrays: Rank and Uniqueness of Trilinear Decompositions, with Applications to Arithmetic Complexity and Statistics." *Linear Algebra and its Applications*, 18: 95–138.

**Lahiri, SN.** 2003. "Central limit theorems for weighted sums of a spatial process under a class of stochastic and fixed designs." *Sankhyā: The Indian Journal of Statistics*, 356–388.

**Michalopoulos, Stelios, and Elias Papaioannou.** 2013. "Pre-colonial ethnic institutions and contemporary African development." *Econometrica*, 81(1): 113–152.

**Murdock, George Peter.** 1969. *Ethnographic atlas.* University of Pittsburgh Press.

**Newey, W.** 1994. "The Asymptotic Variance of Semiparametric Estimators." *Econometrica*, 62: 1349–1382.

**Newey, W., and D. McFadden.** 1994. "Large Sample Estimation and Hypothesis Testing." In *Handbook of Econometrics.* Vol. IV, , ed. R. F. Engel and D. L. McFadden. Elsevier Science.

**Newey, W. K., and J. L. Powell.** 2003. "Instrumental Variable Estimation of Nonparametric Models." *Econometrica*, 71: 1565–1578.

**Nordman, Daniel J, Soumendra N Lahiri, and Brooke L Fridley.** 2007. "Optimal block size for variance estimation by a spatial block bootstrap method." *Sankhyā: The Indian Journal of Statistics*, 468–493.

**Pinkse, J., and M. E. Slade.** 2010*a*. "The future of spatial econometrics." *Journal of Regional Science*, 50: 103–117.

**Pinkse, Joris, and Margaret E Slade.** 2010*b*. "The future of spatial econometrics." *Journal of Regional Science*, 50(1): 103–117.

**Redding, Stephen J, and Esteban Rossi-Hansberg.** 2017. "Quantitative spatial economics." *Annual Review of Economics*, 9: 21–58.

**Schennach, S. M.** 2016. “Recent Advances in the Measurement Error Literature.” *Annual Reviews of Economics*, 8: 341–377.

**Schennach, S. M.** 2018. “Mismeasured and unobserved variables.” In *Handbook of Econometrics.* Vol. 7A, invited, under (minor) revision. Elsevier Science.

**Shen, X.** 1997. “On Methods of Sieves and Penalization.” *Annals of Statistics*, 25: 2555–2591.

**Szpiro, Adam A., Lianne Sheppard, and Thomas Lumley.** 2011. “Efficient measurement error correction with spatially misaligned data.” *Biostatistics*, 12: 610–623.

**Zhou, Chenghu, Fenzhen Su, Francis Harvey, and Jun Xu.** 2017. *Spatial Data Handling in Big Data Era.* Springer.

# A Proofs

*Theorem 2.1.* We handle the case of discrete and continuous $\mu_X$ separately. We let $=^d$ denote equality in distribution.

For the continuous case, we show assumptions 1 to 5 in Hu and Schennach (2008) are satisfied in our framework. Identification then follows from their Theorem 1.

First, assumption 2.1 implies their assumption 2, both (i) and (ii). For (i), we observe that $Y(s)|X(s), X^*(s), X(s+\Delta s) =^d g(X^*) + U(S)|X(s), X^*(s), X^*(s+\Delta s) + V(s+\Delta s)$ by our assumptions about the generating process, equations (1) and (2). The first term is known conditional on $X^*$ and thus by assumption 2.1

$$\begin{aligned} Y(s)|X(s), X^*(s), X(s+\Delta s) &=^d g(X^*) + U(S)|X^*(s) \\ &=^d Y(s)|X^*(s). \end{aligned} \tag{16}$$

Next, we have

$$\begin{aligned} X(s)|X^*(s), X(s+\Delta s) &=^d X^*(s) + V(s)|X^*(s), X^*(s+\Delta s) + V(s+\Delta s) \\ &=^d X^*(s) + V(s)|X^*(s) \\ &=^d X(s)|X^*(s), \end{aligned} \tag{17}$$

so (ii) holds as well.

Assumptions 2.2, 2.4, and 2.5 are direct counterparts of assumptions 1, 3, and 4 in Hu (2008) adapted to our spatial setup. Finally, the existence of $M_x$ in assumption 2.3 establishes their assumption 5.

Hence, by Theorem 1 in Hu (2008), the knowledge of $f_{Y(s),X(s),X(s+\Delta s)}(y,x,z)$ identifies $f_{Y(s)|X^*(s)}$, $f_{X(s)|X^*(s)}$, $f_{X(s+\Delta s)|X^*(s)}$, and $f_{X^*(s)}$.

For the discrete case, we first show that our assumptions imply the assumptions 1, 2, 2.1, 2.2 of Hu (2008). Note that their assumptions explicitly include possible conditioning on a covariate $w$, while we our notation leaves such conditioning implicit, for simplicity.

Our assumption 2.1 implies their assumption 1 and 2, by the same reasoning that lead to Equations (16) and (17) above. Next, our assumption 2.4 reduces to their assumptions 2.1 and 2.2 in the discrete case, since the integral operators reduce to matrix multiplications when $\mu_X$ is discrete: $[L_{B|A}h](b) = \int f_{B|A}(b|a)h(a)d\mu_X(a) = \sum_a F_{B|A}(b|a)h(a)\mu(\{a\})$.

Finally, although none of our assumptions imply one of their set of alternative assumptions 2.3 through 2.7, these assumptions are only needed to secure the proper ordering of the possible values of the latent discrete variable $X^*$. Any re-ordering of it implies a re-ordering of the column of the matrix $f_{X(s)|X^*(s)}(x|x^*)$. However, any ordering other than the correct one would lead to a violation of our assumption 2.3: $M_x[f_{X(s)|X^*(s)}(\cdot|x^*)] = x^*$. Hence our assumption 2.3 has the same effect as their set of alternative assumptions 2.3 through 2.7. (Note that in the special case where $M_x$ is the mode functional, our assumption 2.3 regarding $X(s)|X^*(s)$ is the same as their assumption 2.7.)

From the above consideration, we can invoke their Theorem 1 to establish identification of our model in the discrete case.

□

*Proof of Theorem 4.1.* We take the following convention to ensure that the $Z_i$ vary smoothly as $f$ is changed in the expression $\hat{\mathcal{L}}(\theta, f)$ for $f \neq \hat{f}$. Letting $F^{-1}(\cdot|x,y)$ denotes the inverse of the cdf of $Z$ given $X$ and $Y$ with respect to the first argument, we set $Z_i = \hat{F}^{-1}_{Z|X,Y,\Delta s}(U_i|X_i, Y_i)$ (in the unidimensional case[6]) where $U_i$ is drawn from a uniform and the $U_i$ are kept fixed as $f$ varies. This is purely a device of proof and a harmless convention because $\hat{\mathcal{L}}(\theta, f)$ is only evaluated at $f = \hat{f}$ in the estimator. However, the structure of the proof (which uses constructs involving $\hat{\mathcal{L}}(\theta, f)$ for $f \neq \hat{f}$) is considerably simplified with this convention.

We first show consistency. Assumptions 4.4(i),(ii) and 4.3(i) imply that $\left\|\hat{f} - f_0\right\| \xrightarrow{p} 0$. To show that $\hat{\theta} \xrightarrow{p} \theta$, we observe that, by the triangular inequality,

$$\left|\hat{\mathcal{L}}\left(\theta, \hat{f}\right) - \mathcal{L}(\theta, f_0)\right| \leq \left|\hat{\mathcal{L}}\left(\theta, \hat{f}\right) - \mathcal{L}\left(\theta, \hat{f}\right)\right| + \left|\mathcal{L}\left(\theta, \hat{f}\right) - \mathcal{L}(\theta, f_0)\right|.$$

The first term satisfies $\left|\hat{\mathcal{L}}\left(\theta, \hat{f}\right) - \mathcal{L}\left(\theta, \hat{f}\right)\right| \xrightarrow{p} 0$ by Assumption 4.1(ii) and the fact that eventually $\hat{f} \in \mathcal{F}$ since $\hat{f} \xrightarrow{p} f_0$. The second term is also such that $\left|\mathcal{L}\left(\theta, \hat{f}\right) - \mathcal{L}(\theta, f_0)\right| \xrightarrow{p} 0$ by Assumption 4.1(iii) and $\hat{f} \xrightarrow{p} f_0$. Since $\hat{\mathcal{L}}\left(\theta, \hat{f}\right)$ converges uniformly to a function that is uniquely maximized at $\theta_0$ (by Assumption 4.1(i)), it follows that $\hat{\theta} = \arg\max_{\theta\in\Theta} \hat{\mathcal{L}}\left(\theta, \hat{f}\right) \xrightarrow{p} \arg\max_{\theta\in\Theta} \mathcal{L}(\theta, f_0) = \theta_0$, by Theorem 2.1 in Newey and McFadden (1994).

By a standard expansion of the first order conditions $\nabla\hat{\mathcal{L}}\left(\hat{\theta}, \hat{f}\right) = 0$ around the true value $\theta = \theta_0$, we have:

$$\nabla\hat{\mathcal{L}}\left(\theta_0, \hat{f}\right) + \nabla\nabla'\hat{\mathcal{L}}\left(\bar{\theta}, \hat{f}\right)\left(\hat{\theta} - \theta_0\right) = 0$$

[6] In the the multivariate $Z_i$ case, one proceeds iteratively, starting with $Z_{i,1} = F^{-1}_{Z_1|X,Y}(U_{i,1}|X_i, Y_i)$ and continuing with $Z_{i,k} = F^{-1}_{Z_k|Z_1,\ldots,Z_{k-1},X,Y}(U_{i,k}|Z_{i,1},\ldots,Z_{i,k-1},X_i,Y_i)$ for $k = 2,\ldots,\dim Z_i$ and with all $U_{i,k}$ mutually independent.

where $\bar{\theta}$ is mean value between $\theta_0$ and $\hat{\theta}$. Rearranging, we have

$$\begin{aligned} & n^{1/2}\left(\hat{\theta}-\theta_0\right) \\ = \; & -n^{1/2}\left(\nabla\nabla'\hat{\mathcal{L}}\left(\bar{\theta},\hat{f}\right)\right)^{-1}\nabla\hat{\mathcal{L}}\left(\theta_0,\hat{f}\right) \\ = \; & -n^{1/2}\left(\nabla\nabla'\hat{\mathcal{L}}\left(\bar{\theta},\hat{f}\right)\right)^{-1}\left(\nabla\hat{\mathcal{L}}\left(\theta_0,\hat{f}\right)-\nabla\mathcal{L}\left(\theta_0,\hat{f}\right)+\nabla\mathcal{L}\left(\theta_0,\hat{f}\right)-\nabla\mathcal{L}\left(\theta_0,f_0\right)\right) \\ = \; & \hat{\Psi}_{\text{MLE}}+\hat{\Psi}_{\text{kernel}}+R_1 \end{aligned}$$

where we have inserted $-\nabla\mathcal{L}\left(\theta_0,\hat{f}\right)+\nabla\mathcal{L}\left(\theta_0,\hat{f}\right)=0$ and $\nabla\mathcal{L}\left(\theta_0,f_0\right)=0$ (by definition) and where

$$\begin{aligned} \hat{\Psi}_{\text{MLE}} &= -n^{1/2}\hat{H}^{-1}\left(\nabla\hat{\mathcal{L}}\left(\theta_0,f_0\right)-\nabla\mathcal{L}\left(\theta_0,f_0\right)\right) \\ \hat{\Psi}_{\text{kernel}} &= -n^{1/2}\hat{H}^{-1}\left(\nabla\mathcal{L}\left(\theta_0,\hat{f}\right)-\nabla\mathcal{L}\left(\theta_0,f_0\right)\right) \\ \hat{R} &= -n^{1/2}\hat{H}^{-1}\left(\left(\nabla\hat{\mathcal{L}}\left(\theta_0,\hat{f}\right)-\nabla\mathcal{L}\left(\theta_0,\hat{f}\right)\right)-\left(\nabla\hat{\mathcal{L}}\left(\theta_0,f_0\right)-\nabla\mathcal{L}\left(\theta_0,f_0\right)\right)\right) \\ \hat{H} &= \nabla\nabla'\hat{\mathcal{L}}\left(\bar{\theta},\hat{f}\right). \end{aligned}$$

We first show that $\hat{H}\xrightarrow{p}H\equiv\nabla\nabla'\mathcal{L}\left(\theta_0,f_0\right)$ as follows:

$$\hat{H}-H=\left(\nabla\nabla'\hat{\mathcal{L}}\left(\bar{\theta},f_0\right)-\nabla\nabla'\mathcal{L}\left(\theta,f_0\right)\right)+\left(\nabla\nabla'\hat{\mathcal{L}}\left(\bar{\theta},\hat{f}\right)-\nabla\nabla'\hat{\mathcal{L}}\left(\bar{\theta},f_0\right)\right)$$

where the first term is such that $\left(\nabla\nabla'\hat{\mathcal{L}}\left(\bar{\theta},f_0\right)-\nabla\nabla'\mathcal{L}\left(\theta,f_0\right)\right)\xrightarrow{p}0$ from Assumption 4.2(i), while the second term can be written as:

$$\begin{aligned} \nabla\nabla'\hat{\mathcal{L}}\left(\bar{\theta},\hat{f}\right)-\nabla\nabla'\hat{\mathcal{L}}\left(\bar{\theta},f_0\right) &= \left(\nabla\nabla'\hat{\mathcal{L}}\left(\bar{\theta},\hat{f}\right)-\nabla\nabla'\mathcal{L}\left(\bar{\theta},\hat{f}\right)\right) \\ &\quad -\left(\nabla\nabla'\hat{\mathcal{L}}\left(\bar{\theta},f_0\right)-\nabla\nabla'\mathcal{L}\left(\bar{\theta},f_0\right)\right) \\ &\quad +\left(\nabla\nabla'\mathcal{L}\left(\bar{\theta},f_0\right)-\nabla\nabla'\mathcal{L}\left(\bar{\theta},\hat{f}\right)\right). \end{aligned}$$

The two first term converge in probability to zero by Assumption 4.2(iii) and the fact that eventually $\hat{f}\in\mathcal{F}$, by Assumptions 4.4(i),(ii) and 4.3(i). The last term converges in probability to 0 since, by Assumption 4.2(iv),

$$\text{plim}_{n\longrightarrow\infty}\nabla\nabla'\mathcal{L}\left(\theta,\hat{f}\right)=\nabla\nabla'\mathcal{L}\left(\theta,\text{plim}_{n\longrightarrow\infty}\hat{f}\right)=\nabla\nabla'\mathcal{L}\left(\theta,f_0\right)\text{ uniformly for }\theta\in\Theta.$$

It follows that $\hat{H}\xrightarrow{p}H$. By assumption 4.2(ii), we also have $\hat{H}^{-1}\xrightarrow{p}H^{-1}$, so that $\hat{\Psi}_{\text{MLE}}-\Psi_{\text{MLE}}\xrightarrow{p}0$, $\hat{\Psi}_{\text{kernel}}-\tilde{\Psi}_{\text{kernel}}\xrightarrow{p}0$ and $\hat{R}-R\xrightarrow{p}0$ for

$$\begin{aligned} \Psi_{\text{MLE}} &= -n^{1/2}H^{-1}\left(\nabla\hat{\mathcal{L}}\left(\theta_0,f_0\right)-\nabla\mathcal{L}\left(\theta_0,f_0\right)\right)=-n^{-1/2}H^{-1}\sum_{i=1}^{n}\psi_{MLE}\left(Y_i,X_i,Z_i\right) \\ \tilde{\Psi}_{\text{kernel}} &= -n^{1/2}H^{-1}\left(\nabla\mathcal{L}\left(\theta_0,\hat{f}\right)-\nabla\mathcal{L}\left(\theta_0,f_0\right)\right) \\ R &= -n^{1/2}H^{-1}\left(\left(\nabla\hat{\mathcal{L}}\left(\theta_0,\hat{f}\right)-\nabla\mathcal{L}\left(\theta_0,\hat{f}\right)\right)-\left(\nabla\hat{\mathcal{L}}\left(\theta_0,f_0\right)-\nabla\mathcal{L}\left(\theta_0,f_0\right)\right)\right). \end{aligned}$$

where $\psi_{\text{MLE}}(y,x,z) = \nabla \ln L(y,x,z;\theta,\omega(\theta))$ is the usual influence function of a sieve MLE estimator of $\theta$, while

$$\begin{aligned}
& \tilde{\Psi}_{\text{kernel}} \\
= \; & -n^{1/2}H^{-1}\int\int\int \left(\hat{f}(z|y,x) - f(z|y,x)\right) f_{YX}(y,x)\nabla \ln L(y,x,z;\theta,\omega(\theta))\, dydxdz \\
= \; & -n^{1/2}H^{-1}\int\int\int \left(\frac{\hat{f}_{ZYX}(z,y,x)}{\hat{f}_{YX}(y,x)} - \frac{f_{ZYX}(z,y,x)}{f_{YX}(y,x)}\right) f_{YX}(y,x)\nabla \ln L(y,x,z;\theta,\omega(\theta))\, dydxdz
\end{aligned}$$

where we have set $\hat{f}(z|y,x) \equiv \hat{f}(z|y,x,\Delta s)$ and $f(z|y,x) \equiv f(z|y,x,\Delta s)$ to simplify the notation.

$\tilde{\Psi}_{\text{kernel}}$ can be further linearized by using the fact that:

$$\begin{aligned}
\left(\frac{\hat{a}}{\hat{b}} - \frac{a}{b}\right) &= \left(\frac{\hat{a}-a}{b} - \frac{a}{b}\frac{\hat{b}-b}{b}\right) + \left(1 + \frac{\hat{b}-b}{b}\right)^{-1}\left(\frac{a}{b}\left(\frac{\hat{b}-b}{b}\right)^2 - \frac{\left(\hat{b}-b\right)}{b}\frac{(\hat{a}-a)}{b}\right) \\
&= \left(\frac{\hat{a}-a}{b} - \frac{a}{b}\frac{\hat{b}-b}{b}\right) + o_p\left(n^{-1/2}\right)
\end{aligned}$$

if $\|\hat{a}-a\| = o_p\left(n^{-1/4}\right)$, $\left\|\hat{b}-b\right\| = o_p\left(n^{-1/4}\right)$ and $b \geq \varepsilon > 0$. Setting $b = f_{ZYX}(z,y,x)$ and $a = f_{YX}(y,x)$ and invoking Assumption 4.4(i) and (ii) yields:

$$\begin{aligned}
& \tilde{\Psi}_{\text{kernel}} \\
= \; & -n^{1/2}H^{-1}\int\int\int \frac{\hat{f}_{ZYX}(z,y,x) - f_{ZYX}(z,y,x)}{f_{YX}(y,x)} f_{YX}(y,x)\nabla \ln L(y,x,z;\theta,\omega(\theta))\, dydxdz \\
& + n^{1/2}H^{-1}\int\int\int \frac{f_{ZYX}(z,y,x)}{f_{YX}(y,x)}\frac{\left(\hat{f}_{YX}(y,x) - f_{YX}(y,x)\right)}{f_{YX}(y,x)} f_{YX}(y,x) \\
& \times \nabla \ln L(y,x,z;\theta,\omega(\theta))\, dydxdz + n^{1/2}o_p\left(n^{-1/2}\right) \\
= \; & \tilde{\Psi}^1_{\text{kernel}} + \tilde{\Psi}^2_{\text{kernel}} + o_p(1)
\end{aligned}$$

where

$$\begin{aligned}
\tilde{\Psi}^1_{\text{kernel}} &= -n^{1/2}H^{-1}\int\int\int \left(\hat{f}_{ZYX}(z,y,x) - f_{ZYX}(z,y,x)\right)\nabla \ln L(y,x,z;\theta,\omega(\theta))\, dydxdz \\
\tilde{\Psi}^2_{\text{kernel}} &= n^{1/2}H^{-1}\int\int \left(\hat{f}_{YX}(y,x) - f_{YX}(y,x)\right) E\left[\nabla \ln L(Y,X,Z;\theta,\omega(\theta))\,|Y=y,X=x\right] dydx.
\end{aligned}$$

Using standard semiparametric correction terms for density estimation (Newey (1994)) and under the small bias Assumption 4.4(iii) and (iv), these terms can be shown to be asymptotically equivalent to sample averages:

$$\tilde{\Psi}^k_{\text{kernel}} = n^{-1/2}\sum_{i=1}^{n}\psi^k_{\text{kernel}}(Y_i,X_i,Z_i) + o_p(1)$$

for $k = 1, 2$, where

$$\begin{aligned}
\psi^1_{\text{kernel}}(y,x,z) &= H^{-1}\left(\nabla \ln L(y,x,z;\theta,\omega(\theta)) - E\left[\nabla \ln L(Y,X,Z;\theta,\omega(\theta))\right]\right) \\
\psi^2_{\text{kernel}}(y,x,z) &= H^{-1}\left(E\left[\nabla \ln L(Y,X,Z;\theta,\omega(\theta))\,|Y=y,X=x\right] - E\left[\nabla \ln L(Y,X,Z;\theta,\omega(\theta))\right]\right)
\end{aligned}$$

and $\psi_{\text{kernel}}(y,x,z) = \psi^1_{\text{kernel}}(y,x,z) + \psi^2_{\text{kernel}}(y,x,z)$ is thus as given in the statement of the Theorem.

To show that the remainder term $R$ is $o_p(1)$, we need to show that $n^{1/2}((\nabla\hat{\mathcal{L}}(\theta_0, f) - \nabla\mathcal{L}(\theta_0, f)) - (\nabla\hat{\mathcal{L}}(\theta_0, f_0) - \nabla\mathcal{L}(\theta_0, f_0)$ is stochastically equicontinuous in $f$ at $f = f_0$ for all sufficiently large $n$. This standard property follows from (a) $\nabla\mathcal{L}(\theta_0, f)$ being linear in $f$ with bounded prefactor by Assumption 4.5(i), (b) $\nabla\hat{\mathcal{L}}(\theta_0, f)$ being Lipschitz in each of the $Z_i$ by Assumption 4.5(i) and (c) the $Z_i$ being Lipschitz in $f$ (in the sup norm $\|\cdot\|_\infty$). The third assertion can be shown by observing that changes $F - F_0$ in the conditional cdf of $Z_i$ are bounded by $C\|f - f_0\|_\infty$ for some $C < \infty$. Since both $f_0$ and $f$ are bounded by Assumption 4.5(ii), the change $F^{-1} - F_0^{-1}$ is also bounded by $C'\|f - f_0\|_\infty$ for some $C' < \infty$. Thus the $Z_i$ are Lipschitz in $f$. □

# B Additional simulations

Table 7: $\theta_1 = -3.5$

| | Mean | Standard deviation | RMSE |
|---|---|---|---|
| Unweighted Spatial; low Sieve | $-4.38$ | 0.24 | 0.91 |
| weighted Spatial; low Sieve | $-4.32$ | 0.23 | 0.85 |
| Unweighted Spatial; medium Sieve | $-3.58$ | 0.19 | 0.20 |
| weighted Spatial; medium Sieve | $-3.58$ | 0.19 | 0.20 |
| Unweighted Spatial; high Sieve | $-3.60$ | 0.11 | 0.15 |
| weighted Spatial; high Sieve | $-3.59$ | 0.08 | 0.12 |

$\theta_2 = 2$

| | Mean | Standard deviation | RMSE |
|---|---|---|---|
| Unweighted Spatial; low Sieve | 2.26 | 0.06 | 0.27 |
| weighted Spatial; low Sieve | 2.25 | 0.06 | 0.26 |
| Unweighted Spatial; medium Sieve | 2.05 | 0.06 | 0.07 |
| weighted Spatial; medium Sieve | 2.05 | 0.06 | 0.08 |
| Unweighted Spatial; high Sieve | 2.04 | 0.05 | 0.07 |
| weighted Spatial; high Sieve | 2.04 | 0.04 | 0.06 |

$\sigma_u = 1.3$

| | Mean | Standard deviation | RMSE |
|---|---|---|---|
| Unweighted Spatial; low Sieve | 1.17 | 0.06 | 0.14 |
| weighted Spatial; low Sieve | 1.23 | 0.05 | 0.08 |
| Unweighted Spatial; medium Sieve | 1.34 | 0.03 | 0.05 |
| weighted Spatial; medium Sieve | 1.34 | 0.03 | 0.05 |
| Unweighted Spatial; high Sieve | 1.24 | 0.05 | 0.08 |
| weighted Spatial; high Sieve | 1.23 | 0.04 | 0.07 |

Unweighted Spatial: unweighted average spatial estimator; Weighted Spatial: optimally weighted estimator. With Section 3's notations, low Sieve means $i_n = j_n = 2$; medium Sieve means $i_n = j_n = 4$; high Sieve means $i_n = j_n = 6$