# Spatial Correlation Robust Inference with Imperfect Distance Information

Timothy G. Conley[*]
Graduate School of Business
University of Chicago

Francesca Molinari
Department of Economics
Northwestern University

April 1, 2003
Very Preliminary and Incomplete

**Abstract**

This note presents preliminary results from a Monte Carlo study concerning inference with spatially dependent data. We investigate the impact of location/distance measurement problems upon the accuracy of parametric and nonparametric estimators of asymptotic variances. We consider measurement errors in distances, and locations that are known only up to broad areas like zip codes, SMSAs, or counties.

## 1 Introduction

Spatial econometric models have proven useful in many areas of economics.[1] Economic models underpinning empirical work in urban, environmental, development, industrial organization, and growth frequently suggest that observed agents will have outcomes that are not independent. Often these models suggest a suitable metric or a set of locations in some space that characterizes the dependence structures among agents. A spatial model is simply a data generating model that

[*] The authors would like to thank Jeff Russell and Federico Bandi for helpful comments. Conley gratefully acknowledges funding support from the National Science Foundation.

[1] An incomplete list of relevant papers is: Case (1991), Case, Hines, and Rosen (1993), Elliott (1993), Moreno and Trehan (1997), Kelejian and Prucha (1999), Topa (2001), Conley and Topa (2000). A more complete description of the extent of this literature will appear in a future draft.

utilizes such a set of locations or distances to define the relationships between agents' variables. The notion of space can be general and is certainly not confined to physical or geographic space.

Typical spatial models are parametric models of the dependence between agents, e.g. those used by Ord (1975), Anselin and Griffith (1988), Case (1991), and Kelejian and Prucha (1999). The most prevalent models are for Gaussian data with a covariance structure that is a parametric function of known locations. More recently, nonparametric methods for estimating covariance structure have been proposed for estimating covariance structure both as a direct object of interest and to conduct inference about conditional mean estimates, see e.g. Hall, et al. (1992), Conley (1999).

The key ingredient in any spatial model is the choice of locations for the observed agents, the space or metric. There are clearly some applications where the agents locations are known with certainty. For example, in an environmental application where local weather conditions constitute the major source of dependence, agents' physical locations may be available and be the most appropriate coordinates to use in a spatial model. However, it is routinely the case that agents' locations are not known with certainty. It is very common for information about agents' physical locations to be imprecise, e.g. locations to be known only within an area– census tract, zip code, county, or SMSA. Moreover, in many applications the most appropriate metric may not be physical distance. For example, the travel time between locations, an object that must be estimated and cannot be known with certainty. Thus it is common for the econometrician's measurements of distance to be imprecise or measured with error.

Imperfect distance measurements create problems for parametric models of spatial covariance. Unless they include an explicit treatment of the measurement error process, parametric models will generally be misspecified and inconsistent when distances/locations are measured with error.[2] To our knowledge, such an explicit modeling of measurement errors has not been done, perhaps because of the lack of guidance about such errors from economics. In contrast, the nonparametric inference procedure in Conley (1999) is consistent with bounded measurement errors and generally robust in practice to either measurement errors or imprecision in distances. Imprecise locations will pose less of a problem for parametric estimators, provided the appropriate calculations are done to infer the properties of the aggregated process from that assumed for individuals. Spatial

[2] See Griffith and Lagona (1998) for results on the inconsistency of MLE estimators of spatial correlations when locations are misspecified.

aggregation will undoubtedly reduce the available information and thus the precision of parametric estimators. Of course, a reduction in precision will occur for nonparametric estimators as well, though it may be less severe for estimators that require only broad definitions of near and far sets of observations.

This paper presents a Monte Carlo study that investigates the impact of these two types of location/distance measurement problems upon the accuracy of estimators of the asymptotic variances of sample averages. In particular we compare the performance of parametric asymptotic variance estimators to the nonparametric asymptotic variance estimator of Conley (1999) when agents' locations are measured with error and measured imprecisely (up to census tracts or zip codes). We consider a stationary, mixing data generation model and use an increasing domain asymptotic approach. Asymptotic covariances for averages of spatial data are sums of spatial autocovariances, analogous to the asymptotic variance of averages of covariance stationary time series. The parametric estimators rely on estimating the autocovariance function and then inferring its sum, an approach that is derailed by misspecification of the covariance function arising from distance errors. The nonparametric estimator can be viewed as directly estimating the sum of covariances and this is what allows it to remain consistent and robust in practice with distance errors.

We address three main questions in the experiments we conduct. First we address the question of how much measurement error in distances/locations is required for the parametric model to perform worse than the nonparametric one. Despite the fact that the parametric model will be inconsistent, we expect it to outperform the nonparametric model, e.g. in terms of mean squared error, in finite samples with amounts of measurement error that are small enough. Next, we investigate how estimators' precision varies with neighborhood size when locations are only measured up to neighborhood of residence. We expect that precision of both estimators will decline with an increase in neighborhood size, holding constant the dependence across individuals. The relevant question is the magnitude of these decreases for each estimator and their relative performance. Finally, we investigate the potential of a specification test that compares the nonparametic versus parametric estimators of asymptotic variances. Our simulations provide us with the small sample distribution of test statistics under the null hypothesis of no distance/location error (and a correctly specified covariance model). We construct critical values using the simulations under the null and use them to calculate power for alternatives given by the measurement error models in our experiments. We conjecture that this may be in some sense a best-case scenario for power versus these alternatives that might be obtained from a test using critical values coming from a large sample approximation.

The remaining sections of the paper are organized as follows. Section two presents the data generating model and our two estimators. Section three presents our design of data generating processes for data and location/distance errors, as well as the specific forms for estimators we use in our simulations. Section four concludes this paper by presenting our preliminary results. We defer concluding remarks for a future draft.

## 2 Econometric Model and Estimation Problem

The econometric model we use assumes there is a population of agents residing at d-dimensional integer lattice locations with one individual per location. We focus on an expectation zero process $X_s$ indexed on this lattice that is assumed to be mixing ($X_{s_i}$ and $X_{s_j}$ approach independence as the distance between $s_i$ and $s_j$ grows). For simplicity, we also assume the process is stationary: the joint distribution of $X_s$ for a collection of locations is invariant to translation and so, assuming second moments exist, $E\{X_s X_{s+h}\} = C(h)$. The econometrician's sample consists of realizations of agents' random variables $X_s$ at a collection of locations $\{s_i\}$ inside a sample region $\Lambda_\tau$. We use the notation $|\Lambda_\tau|$ to denote the number of agents in our sample region and, for simplicity, assume that all locations in $\Lambda_\tau$ are sampled. When taking limits, we view $\Lambda_\tau$ as one of a sequence of regions that grow to $Z^d$, an increasing domain approach to asymptotic approximations.

We are interested in conducting inference about $EX$ using the usual large-sample distribution approximations for the sample average of points in $\Lambda_\tau$ : $\bar{X} = \frac{1}{|\Lambda_\tau|}\sum_{i=1}^{|\Lambda_\tau|} X_{s_i}$. To do this, we need to estimate the asymptotic variance of a normalized sample mean. Using, for example, the central limit theorem due to Bolthausen (1982) for stationary, mixing random fields on regular lattices, we know that (under mixing and moment conditions) the normalized sample mean has a limiting normal distribution:

$$\frac{1}{\sqrt{|\Lambda_\tau|}}\sum_{i=1}^{|\Lambda_\tau|} X_{s_i} \to N\left(0, V\right). \tag{1}$$

The general form for the asymptotic covariance $V$ is as an infinite sum of an autocovariance function

$C(h)$. Referring to the entries of the vector $h$ individually as $k_1, k_2, ...., k_d$ , $V$ has the form:

$$V = \sum_{k_1=-\infty}^{\infty} ... \sum_{k_d=-\infty}^{\infty} C\left(k_1, k_2, ...., k_d\right).$$

Thus if $d = 1$, the expression for $V$ coincides with the asymptotic variance of a sample mean for a covariance stationary time series: $V = \sum_{k_1=-\infty}^{\infty} C\left(k_1\right)$.

We are interested in comparing the performance of parametric and nonparametric estimators of $V$ when locations are imperfectly measured. We examine a parametric estimator that corresponds to an assumption that the covariance function is known up to a finite-dimensional parameter vector so it can be written as $C(h;\theta)$. We compute a minimum distance estimator $\hat{\theta}$ and then compute a parametric estimator $\hat{V}_P$ by plugging in the estimate $\hat{\theta}$, and calculating the sum of $C(h;\hat{\theta})$. When using specific locations, we minimize the distance from a vector of sample analog covariance estimates for displacements $h$ and $C(h;\theta)$. When we have only neighborhood locations, we choose $\hat{\theta}$ by minimizing the distance between the parametric expression for covariances between neighborhood aggregates and their sample analogs. In the presence of measurement error in distances, $C(h;\theta)$ will generally be misspecified and the resulting estimator for $V$ inconsistent. However, the parametric estimator using neighborhood aggregates is properly specified and so of course remains consistent.

Our nonparametric estimator of $V$ is that proposed by Conley (1999). This method is a straightforward generalization of spectral density estimators known since at least Bartlett (1950). With specific location/distance measurements, we estimate $V$ as:

$$\hat{V}_N = \frac{1}{|\Lambda_\tau|} \sum_{i=1}^{|\Lambda_\tau|} \sum_{j=1}^{|\Lambda_\tau|} K\left(s_i - s_j\right) \cdot \left(X_{s_i} - \bar{X}\right) \cdot \left(X_{s_j} - \bar{X}\right) \tag{2}$$

where the $N$ subscript refers to nonparametric and dependence on sample size is suppressed. $K(\cdot)$ is a kernel which will be used to weight the observations, and is such that $K(0) = 1$, $K(h)$ is uniformly bounded, and $K(h) \to 1$ for all $h$ as $\tau \to \infty$, slowly enough so that the variance of $\hat{V}_N$ collapses to zero. The kernel $K$ can be chosen so that $\hat{V}_N$ will be positive in sample, however we use a uniform kernel for simplicity. $\hat{V}_N$ will remain consistent in the presence of all but extreme specifications of measurement error because all locations' displacements $h$ will eventually have a weight approaching 1 (see Conley (1999) for a proof with bounded location measurement error). This estimator will also be robust to moderate location/distance measurement errors in practice as mismeasured locations weightings under the kernel will generally be close to the weight they would get with perfectly measured locations. The only exceptions will occur for pairs of observations near the edge of the support of the kernel $K$. For example, if $K$ is uniform kernel equal to one only for displacements with length less than $L$, only those pairs of observations whose true displacement lengths are in a neighborhood around $L$ will have different weights from those for true displacements. Typically these misweighted observations are small fraction of the total with moderate measurement errors. With zip-code level location information, we use (2) but replace the weight $K\left(s_i - s_j\right)$ with a uniform kernel weight that depends only on the distance between the centers of neighborhoods for observations $i$ and $j$. This estimator remains consistent for weights that converge to one for all neighborhood distances at an appropriate rate.

# 3 Data Generating Processes for Simulations

This section describes the data generating processes (DGP) for $X_s$ and the measurement error process for locations that we use in our simulation experiments. In this preliminary draft we consider a process indexed in only one dimension, future drafts will use DGPs with at least a two-dimensional index. We run three sets of experiments each with the same DGP for $X_s$, but with differing structures for the location information. In all cases we simulate 1000 Monte Carlo samples with $|\Lambda_\tau| = 500$.

## 3.1 DGP for X

The DGP we consider for $X_s$ is a finite-order two-sided moving average with geometrically declining weights:

$$X_s = \rho^m u_{s-m} ... + \rho^2 u_{s-2} + \rho u_{s-1} + u_s + \rho u_{s+1} + \rho^2 u_{s+2} + ... \rho^m u_{s+m} \tag{3}$$

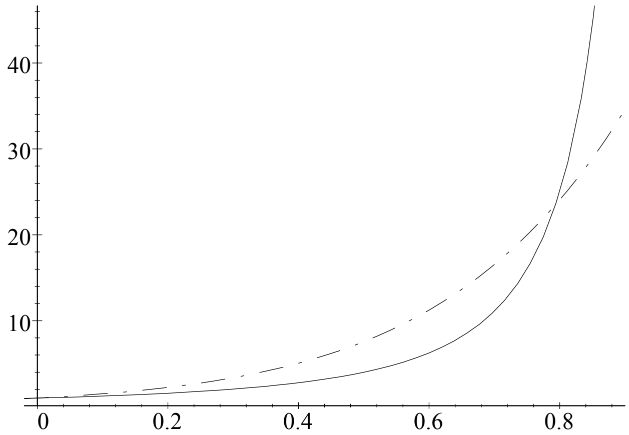
where $u_s$ is IID $N(0,\sigma^2)$. As this process is a finite-order moving average, $V = \sum_{k=-2m}^{2m} C(k)$ with $C(k)$ being given by:

$$C(k;\rho,\sigma) = \begin{cases} \sigma^2 \left(1 + 2\sum_{j=1}^{m} \rho^{2j}\right) & \text{if } k = 0 \\ \sigma^2 \left[(|k|+1)\rho^{|k|} + 2\rho^{2+|k|} \sum_{j=0}^{m-|k|-1} \rho^{2j}\right] & \text{if } |k| \leq m-1 \\ \sigma^2 \left[(m+1)\rho^m\right] & \text{if } |k| = m \\ \sigma^2 (2m - |k| + 1)\rho^{|k|} & \text{if } m+1 \leq |k| \leq 2m \\ 0 & \text{otherwise} \end{cases} . \tag{4}$$

In the Monte Carlo simulations which follow, we choose $m = 3$. In this case, the explicit expression for $V$ is:

$$V = \sigma^2 \left(4\rho^6 + 8\rho^5 + 12\rho^4 + 12\rho^3 + 8\rho^2 + 4\rho + 1\right). \tag{5}$$

To illustrate how $V$ varies with the decay parameter $\rho$, we take $\sigma^2 = 1$ and plot $V$ below with the dot-dash line, for different values of $\rho$. For comparison we also plot with a solid line the asymptotic variance for a first-order autoregression with correlation parameter $\rho$ and an innovation variance of one: $\frac{1}{(1-\rho)^2}$.



$V$ versus $\frac{1}{(1-\rho)^2}$

## 3.2 Measurement Error/ Neighborhood Designs

We run three sets of simulation experiments with the same DGP for $X_s$ but different specifications/treatments for location information. In the first experiment, we assume that the exact locations are known and used in estimation. In the second and third experiments, we analyze how the competing methods perform when locations or distances are measured with error, and when they are measured correctly but imprecisely. We model errors in economic distances as erroneously measured positions $\{s_i\}$. When locations are measured imprecisely, we assume that all is known about each observation is that it resides in a given neighborhood, but the relative positions of observations within a neighborhood are unknown.

**Locations Error**

We tried to choose a specification for location measurement error that would correspond to measured locations being in the right ballpark but perhaps often not exactly correct . Having potentially many locations a little bit off but few if any dramatically off seems to us the most empirically relevant situation.

We model location measurement errors by perturbing locations with the following algorithm. First assign each agent's true location integer coordinates from 1 to $|\Lambda_\tau|$. Each agent's integer location is independently perturbed by adding a random amount $\xi$ from a uniform distribution on [-v,v]. In other words agent $i$ is given a perturbed location $\tilde{s}_i = s_i + \xi_i$. Then, each agent's measured location is defined by a re-labeling of the perturbed locations $\{\tilde{s}_i\}$ from 1 to $|\Lambda_\tau|$ , according to the rank order of the $\{\tilde{s}_i\}$ from smallest to largest. We vary the amount of reshuffling of agents' locations by changing the magnitude of $v$.

We present results for seven different levels of $v$ which we will refer to as levels one through seven. Tables 1 and 2 are meant to provide some sense of how much change in locations is induced by each level of measurement error. The percentages of agents' measured locations that are different from their true locations by 1 to 6 units are given by Table 1. Table 2, contrasts the true autocorrelations for $X_s$ with an approximation for the autocorrelations under each level of measurement error (calculated as the average across 1000 Monte Carlo simulations of length 1000 each).

Table 1: Degree of Deviations of Measured Locations from True Locations

| | $v$ | Percentage at True Location | 1 Unit Off (%) | 2 Units Off (%) | 3 Units Off (%) | 4 Units Off (%) | 5 Units Off (%) | 6 Units Off (%) |
|---|---|---|---|---|---|---|---|---|
| Level 1 | 1.0 | 75.0% | 25.0% | 0% | 0% | 0% | 0% | 0% |
| Level 2 | 1.5 | 51.8% | 42.0% | 6.2% | 0% | 0% | 0% | 0% |
| Level 3 | 2.0 | 36.8% | 45.5% | 15.9% | 1.8% | 0% | 0% | 0% |
| Level 4 | 2.5 | 27.5% | 42.7% | 23.1% | 6.2% | 0.5% | 0% | 0% |
| Level 5 | 3.0 | 21.5% | 37.9% | 26.5% | 11.5% | 2.4% | 0.2% | 0% |
| Level 6 | 3.5 | 17.6% | 33.0% | 26.9% | 15.9% | 5.6% | 1.0% | 0% |
| Level 7 | 4.0 | 14.7% | 28.9% | 25.7% | 18.5% | 9.1% | 2.7% | 0.4% |

Table 2: True Correlations vs. Correlations with Different Levels of Location Errors

| | $v$ | $\frac{C(1)}{C(0)}$ | $\frac{C(2)}{C(0)}$ | $\frac{C(3)}{C(0)}$ | $\frac{C(4)}{C(0)}$ | $\frac{C(5)}{C(0)}$ | $\frac{C(6)}{C(0)}$ |
|---|---|---|---|---|---|---|---|
| True Correlations | 0 | 0.599 | 0.287 | 0.119 | 0.030 | 0.007 | 0.001 |
| Level 1 | 1.0 | 0.522 | 0.317 | 0.137 | 0.044 | 0.010 | 0.000 |
| Level 2 | 1.5 | 0.445 | 0.321 | 0.169 | 0.068 | 0.021 | 0.004 |
| Level 3 | 2.0 | 0.382 | 0.304 | 0.192 | 0.096 | 0.038 | 0.010 |
| Level 4 | 2.5 | 0.330 | 0.278 | 0.202 | 0.123 | 0.059 | 0.024 |
| Level 5 | 3.0 | 0.290 | 0.253 | 0.199 | 0.136 | 0.081 | 0.039 |
| Level 6 | 3.5 | 0.255 | 0.230 | 0.191 | 0.144 | 0.096 | 0.057 |
| Level 7 | 4.0 | 0.229 | 0.209 | 0.180 | 0.143 | 0.106 | 0.071 |

**Grouping of Locations by Neighborhood**

In order to examine the impact of imprecise location information, we divide the line into neighborhoods (that is, in this case, intervals). Estimation will proceed as though agents' locations are known only up to these neighborhoods. We index neighborhoods of $M$ agents each with $j = 1, \ldots, J$. We let $Y_j$ denote the spatially aggregated version of $X_s$, i.e.: $Y_j$ is the sum of $X_s$ for the $M$ agents within neighborhood $j$.

We consider three different neighborhood sizes: $M = 3, 6, 9$. As $M$ increases from 3 to 9, the correlation across neighborhoods decreases. We define a distance between neighborhoods using the difference between their integer indices $j$. Letting covariance function for $Y$ be denoted by $C_Y(\ell) = E\left(Y_j Y_{j+\ell}\right)$, it takes the form in terms of $\rho$ and $\sigma$ of:

$$M = 3 \Rightarrow \quad C_Y(\ell) = \sigma^2 \cdot \begin{cases} 3 + 12\rho^2 + 10\rho^4 + 6\rho^6 + 8\rho + 8\rho^3\left(1+\rho^2\right) & \text{if } |\ell| = 0 \\ 2\rho + 2\rho^3\left(1+\rho^2\right) + 6\rho^2 + 10\rho^4 + 12\rho^3 + 2\rho^5 & \text{if } |\ell| = 1 \\ 3\rho^4 + 4\rho^5 + 3\rho^6 & \text{if } |\ell| = 2 \\ 0 & \text{if } |\ell| \geq 3 \end{cases}$$

$$M = 6 \Rightarrow \quad C_Y(\ell) = \sigma^2 \cdot \begin{cases} 6 + 36\rho^2 + 40\rho^4 + 12\rho^6 + 20\rho + 20\rho^3\left(1+\rho^2\right) + 24\rho^3 + 4\rho^5 & \text{if } |\ell| = 0 \\ 2\rho + 2\rho^3\left(1+\rho^2\right) + 6\rho^2 + 16\rho^4 + 12\rho^3 + 10\rho^5 + 6\rho^6 & \text{if } |\ell| = 1 \\ 0 & \text{if } |\ell| > 1 \end{cases}$$

$$M = 9 \Rightarrow \quad C_Y(\ell) = \sigma^2 \cdot \begin{cases} 9 + 60\rho^2 + 76\rho^4 + 24\rho^6 + 32\rho + 32\rho^3\left(1+\rho^2\right) + 48\rho^3 + 16\rho^5 & \text{if } |\ell| = 0 \\ 2\rho + 2\rho^3\left(1+\rho^2\right) + 6\rho^2 + 16\rho^4 + 12\rho^3 + 10\rho^5 + 6\rho^6 & \text{if } |\ell| = 1 \\ 0 & \text{if } |\ell| > 1 \end{cases} \tag{6}$$

## 3.3 Specific Estimators Used in Simulations

**Parametric Estimators**

The parametric estimators used in the simulations are based on a minimum distance approach. For the case of locations exactly measured or measured with error, we first estimate the sample analogs of the correlation functions of $X : \Gamma(\rho) = \left[\frac{C(1;\rho,\sigma)}{C(0;\rho,\sigma)}, \ldots, \frac{C(6;\rho,\sigma)}{C(0;\rho,\sigma)}\right]$, obtaining a vector $\hat{\Gamma} =$

$\left[\frac{\hat{C}(1)}{\hat{C}(0)}, \ldots, \frac{\hat{C}(6)}{\hat{C}(0)}\right]$. We then estimate $\hat{\rho}$ as:

$$\hat{\rho} = \arg\min_{\rho} \left[\Gamma(\rho) - \hat{\Gamma}\right]' \left[\Gamma(\rho) - \hat{\Gamma}\right].$$

Once we have the estimate $\hat{\rho}$, we can estimate $\sigma^2$ by means of equation (4). In particular, we can use the sample variance of $X_s$ as an estimate of $C(0)$, and then solve for $\sigma^2$ as follows:

$$\hat{\sigma}^2 = \frac{\hat{C}(0)}{1 + 2\sum_{j=1}^{m} \hat{\rho}^{2j}} \tag{7}$$

Once we have these estimates, we can get $\hat{V}_P$ by plugging $\hat{\rho}$ and $\hat{\sigma}^2$ in (5).

The estimator used in the case of grouping of locations by neighborhood follows a similar strategy. Again, we first estimate the sample analogs of the correlation functions of $Y : \Delta(\rho) = \left[\frac{C_Y(1)}{C_Y(0)} \frac{C_Y(2)}{C_Y(0)}\right]$ for $M = 3$, and $\Delta(\rho) = \left[\frac{C_Y(1)}{C_Y(0)}\right]$ if $M = 6, 9$. We then get our estimate $\hat{\rho}$ by the minimum distance estimator:

$$\hat{\rho} = \arg\min_{\rho} \left[\Delta(\rho) - \hat{\Delta}\right]' \left[\Delta(\rho) - \hat{\Delta}\right]$$

Once we have the estimate $\hat{\rho}$, we again estimate $\sigma^2$ by means of equation (4), and then plug $\hat{\rho}$ and $\hat{\sigma}^2$ in (5) to get a parametric estimator which, with some abuse of notation, we will also refer to as $\hat{V}_P$ as the estimator refered to will be clear from context.

**Nonparametric Estimators**

The nonparametric estimator used in the simulations with individual locations measured (perhaps with error) is that in expression (2), with

$$K(s_i - s_j) = \begin{cases} 1 & \text{if } |s_i - s_j| < L = 8 \\ 0 & \text{otherwise} \end{cases}. \tag{8}$$

In the simulatons with neighborhood-level location information, the weight function $K$ is taken to be a uniform kernel that is one for neighborhoods that contain individuals with distances less than or equal to 8. We again abuse notation and use $\hat{V}_N$ to refer to all versions of these nonparametric estimators.

# 4 Results

This Section reports the results of 1000 repetitions of a Monte Carlo experiment based on a sample of size $|\Lambda_\tau| = 1000$ and with the following parameters: $\sigma^2 = 1$, $\rho = 1/3$, which imply a value of $V = 3.8532$. We present results on the performance of our two estimators and for power of two potential specification tests based on discrepancies between $\hat{V}_P$ and $\hat{V}_N$.

## 4.1 V Estimator Performance

Tables 3 and 4 collect the results obtained when the locations are perfectly measured and when we have location errors of level one through seven. Table 3 reports Bias and Root Mean Squared Error (RMSE) for $\hat{V}_P$ and $\hat{V}_N$, as well as coverage probabilities for 95% confidence intervals for $EX$ constructed using the alternative variance estimators. Table 4 reports the 10th, 30th, 50th, 70th and 90th deciles of the distribution of $\hat{V}_P$ and $\hat{V}_N$ with true and error-ridden locations.

Table 3 shows that when the true locations are used, the bias associated with $\hat{V}_N$ is bigger than that associated with $\hat{V}_P$. However, as the level of the location errors ranges from one to seven, the bias of the parametric estimator increases sharply (in absolute terms), and already at level 2 is bigger than that of $\hat{V}_N$. As the level of location errors increases from three to seven, the bias of the parametric estimator grows almost linearly. The bias of the nonparametric estimator is relatively constant with respect to the different levels of location errors. A similar pattern can be observed when looking at the RMSE: the RMSE associated with the nonparametric estimator is higher than that associated with the parametric one with true locations. However, if locations are incorrectly measured, the nonparametric estimator's performance varies little as the level of location errors increases. In contrast, the RMSE of the parametric estimator deteriorates rapidly. As soon as location errors of level 4 and higher are introduced, both the bias and the RMSE of the parametric estimator get worse than that of the nonparametric estimators. The behavior of the RMSE of the two estimators can be in part explained looking at the deciles of their distributions, reported in Table 4. When the locations are accurately measured, the distribution of the parametric estimates is less spread out and it is centered closer to the true value of $V$ than the nonparametric one. The introduction of location errors implies a shift of the distribution of the parametric estimates towards the left. Although the distribution does not spread out much, the increased bias drives up the RMSE. The distribution of the nonparametric estimator is relatively unaffected by the introduction of the location errors.

Perhaps the best measure of these estimators' performance is their coverage probabilities. The 95% confidence intervals for $EX$ constructed using $\hat{V}_N$ cover the zero in approximately 95% of the Monte Carlo draws for all levels of location errors. In contrast, the coverage probabilities of the 95% confidence intervals constructed using $\hat{V}_P$ deteriorate with a rise in the level of the location errors. While the coverage probability is slightly higher than 95% with location errors of level one, it goes down to 88.4% with location errors of level seven.

Tables 5 and 6 collect the results obtained when the locations are perfectly measured, and when we have imprecisely measured locations, with neighborhoods of sizes 3, 6, and 9. Table 5 reports Bias and Root Mean Squared Error (RMSE) for the parametric and nonparametric estimators of $V$, as well as coverage probabilities for 95% confidence intervals for $EX$ constructed using the alternative variance estimators. Table 6 reports the 10th, 30th, 50th, 70th and 90th deciles of the distribution of the estimators of $V$ using neighborhood aggregates.

As the neighborhood size increases, the bias of the parametric estimator does not increase much, while that of the nonparametric estimator is more sensitive to locations grouping. However, the RMSE of the nonparametric estimator stays roughly constant over the different levels of aggregation, while that of the parametric estimator keeps increasing, and ends up being, for neighborhoods of size 9, almost three times bigger than that of the nonparametric estimator. Looking at the deciles of the distributions of the parametric and nonparametric estimators, reported in Table 6, we observe that the selected deciles of the nonparametric estimator don't change much as the neighborhood size increases. On the other hand, those of the parametric estimates get more and more spread out and the median tends to decrease.

In terms of coverage probabilities, the 95% confidence intervals for $EX$ constructed using the nonparametric estimates of $V$ cover the zero in approximately 94.6% to 93.2% of the cases as the neighborhood size increases from 3 to 9. The coverage probabilities of the 95% confidence intervals constructed using the parametric estimates deteriorate much faster with the size of the neighborhoods. In particular, while it is slightly below 95% with neighborhoods of size 3, it goes down to 77.3% with neighborhoods of size 9.

## 4.2 Investigation of Potential for Specification Tests

When locations/distances are potentially measured with error, a good specification test for the joint null hypothesis that distances and the parametric model are correct would clearly be useful. In this subsection, we use our simulations to investigate the potential performance of specification tests based on discrepancies between $\hat{V}_P$ and $\hat{V}_N$. We investigate the power of one and two-sided tests based on $(\hat{V}_P - \hat{V}_N)$. Figures 1 and 2 plot kernel density estimates of the sampling distribution of $(\hat{V}_P - \hat{V}_N)$ with true locations and locations resulting from our measurement error models. Likewise, Figures 3 and 4 plot analogous density estimates for $(\hat{V}_P - \hat{V}_N)^2$ for this set of models. We view these exercises as being an optimistic scenario for test performance since in practice critical values would likely need to come from a distribution approximation in order for the test to be useful in situations without a full DGP specification.

We constructed a 5% critical value for a two-sided test of the null hypothesis that the distances and covariance model are correct using the 95th percentile of the simulated distribution of $(\hat{V}_P - \hat{V}_N)^2$ with correct distances. We then calculated the proportion of $(\hat{V}_P - \hat{V}_N)^2$ statistics that were greater than this critical value under each alternative measurement error level. The resulting estimates of the probability of (correctly) rejecting the null under these seven alternative hypotheses are collected in the first column of Table 7. The results are not particularly encouraging as power remains less than 1/2 until level 6.

However, we think that in practice it may be useful to conduct a specification test under the assumption of partial information on location of the distribution of $(\hat{V}_P - \hat{V}_N)$ under the relevant alternatives. In particular, it may be plausible to assume that the $\hat{V}_P$ will underestimate $V$ more than $\hat{V}_N$ under the alternative distribution, i.e. that the distribution of $(\hat{V}_P - \hat{V}_N)$ under the alternative will be largely to the left of its distribution under the null. Therefore, we investigate a one-sided test with critical value obtained as the 5th percentile of the simulated distribution of $(\hat{V}_P - \hat{V}_N)$ with correct distances. The resulting estimates of the probability of rejecting the null under these seven alternative hypotheses are collected in the second column of Table 7. Unsurprisingly, power performance is much better with the one-sided test. Future drafts will contain an attempt to formally characterize the set of alternatives where one-sided tests will be desireable.

## References

[1] Anselin L., *Spatial Econometrics: Methods and Models*, Kluwer Academic Publishers, 1988.

[2] Anselin L. and R.J.G.M. Florax editors, *New Directions in Spatial Econometrics*, Springer, 1995.

[3] Anselin L. and Griffith D. A., "Do Spatial Effects Really Matter in Regression Analysis?", *Papers of the Regional Science Association*, 65, 11-34.

[4] Bartlett M. S., "Periodogram Analysis and Continuous Spectra, *Biometrika*, Vol. 37, No. 1/2. (June 1950), pp. 1-16.

[5] Bolthausen E., "On the Central Limit Theorem for Stationary Mixing Random Fields", *The Annals of Probability*, Volume 10, 1982, 1047-1050.

[6] Case A., "Spatial Patterns in Household Demand", *Econometrica*, Vol. 59, No. 4. (Jul., 1991), pp. 953-965.

[7] Case A., Rosen H.S., and Hines J.R., "Budget Spillovers and Fiscal Policy Interdependence: Evidence from the States", *Journal of Public Economics*, 52(3), October 1993, pages 285-307.

[8] Conley T.G., "GMM Estimation with Cross Sectional Dependence", *Journal of Econometrics*, Volume 92, 1999, 1-45.

[9] Conley T.G. and G. Topa, "Socio-economic Distance and Spatial Patterns in Unemployment," *Journal of Applied Econometrics,* Volume 17, Issue 4, (July/August 2002) 303-327.

[10] Cressie N., *Statistics for Spatial Data*, Wiley.

[11] Elliott G., "Spatial Correlations and Cross-Country Regressions", Manuscript, Harvard University, 1993.

[12] Griffith G.D. and F. Lagona, "On the Quality of Likelihood-Based Estimators in Spatial Autoregressive Models when the Data Dependence Structure Is Mispecified," *Journal of Statistical Planning and Inference*, Volume 69, 1998, 153-174.

[13] Hall P, Fisher N. I., and Hoffman B., “On the Nonparametric Estimation of Covariance Functions”, Working Paper, Australian National University, 1992.

[14] Kelejian H.H. and I.R. Prucha, “A Generalized Moments Estimator for the Autoregressive Parameter in a Spatial Model”, *International Economic Review*, Volume 40, 1999, 509-533.

[15] Lee L., “Asymptotic Distributions of Quasi-Maximum Likelihood Estimators for Spatial Econometric Models I: Spatial Autoregressive Processes”, Working Paper, 2001.

[16] Moreno R and B. Trehan, “Location and the Growth of Nations”, *Journal of Economic Growth*, 2(4), December 1997, pages 399-418.

[17] Moulton B.R., “An Illustration of Pitfall in Estimating the Effects of Aggregate Variables on Micro Units”, *The Review of Economics and Statistics*, Volume 72, 1990, 334-338.

[18] Ord K., “Estimation Methods for Models of Spatial Interaction”, *Journal of the American Statistical Association*, 70(349), March 1975, pages 120-26.

[19] Topa G., “Social Interactions, Local Spillovers and Unemployment”, *Review of Economic Studies*, 68(2), April 2001, pages 261-95.

[20] White H. and I. Domowitz, “Nonlinear Regression with Dependent Observations”, *Econometrica*, Volume 52, 143-161.

**Table 3: Bias, Root MSE and 95% CI Coverage Probabilities for V Estimators with True and Error-ridden Locations**

| | Bias | | Root MSE | | 95% CI Coverage Probability | |
|---|---|---|---|---|---|---|
| | Parametric | Nonparametric | Parametric | Nonparametric | Parametric | Nonparametric |
| True Locations | -0.0330 | -0.1149 | 0.5790 | 0.8785 | 0.9540 | 0.9470 |
| Level 1 Location Errors | -0.0669 | -0.0441 | 0.5985 | 0.8734 | 0.9530 | 0.9520 |
| Level 2 Location Errors | -0.2354 | -0.0434 | 0.6576 | 0.8669 | 0.9490 | 0.9510 |
| Level 3 Location Errors | -0.4853 | -0.0552 | 0.7879 | 0.8564 | 0.9380 | 0.9520 |
| Level 4 Location Errors | -0.7583 | -0.0590 | 0.9734 | 0.8542 | 0.9270 | 0.9500 |
| Level 5 Location Errors | -1.0443 | -0.0794 | 1.1956 | 0.8417 | 0.9070 | 0.9500 |
| Level 6 Location Errors | -1.2975 | -0.1178 | 1.4110 | 0.8423 | 0.8940 | 0.9500 |
| Level 7 Location Errors | -1.4961 | -0.1724 | 1.5784 | 0.8351 | 0.8840 | 0.9500 |

Table notes: sample size = 500, rho = 1/3, and the true value of V = 3.8532.

**Table 4: Selected Deciles for V Estimators with True and Error-ridden Locations**

| | Deciles: | 10% | 30% | 50% | 70% | 90% |
|---|---|---|---|---|---|---|
| True Locations | Parametric | 3.1338 | 3.4852 | 3.7561 | 4.0872 | 4.5967 |
| | Nonparametric | 2.6777 | 3.2075 | 3.6743 | 4.1648 | 4.8844 |
| Level 1 Location Errors | Parametric | 3.0957 | 3.4320 | 3.7229 | 4.0582 | 4.5742 |
| | Nonparametric | 2.7544 | 3.2862 | 3.7415 | 4.2337 | 4.9303 |
| Level 2 Location Errors | Parametric | 2.8957 | 3.2656 | 3.5697 | 3.9123 | 4.4430 |
| | Nonparametric | 2.7217 | 3.3069 | 3.7373 | 4.2424 | 4.9689 |
| Level 3 Location Errors | Parametric | 2.6571 | 3.0092 | 3.3127 | 3.6620 | 4.1963 |
| | Nonparametric | 2.7445 | 3.2853 | 3.7398 | 4.2022 | 4.9742 |
| Level 4 Location Errors | Parametric | 2.3705 | 2.7266 | 3.0503 | 3.3680 | 3.9039 |
| | Nonparametric | 2.7473 | 3.2959 | 3.7623 | 4.2224 | 4.9153 |
| Level 5 Location Errors | Parametric | 2.1027 | 2.4585 | 2.7706 | 3.0751 | 3.5638 |
| | Nonparametric | 2.7757 | 3.2838 | 3.6882 | 4.1784 | 4.8618 |
| Level 6 Location Errors | Parametric | 1.9137 | 2.2282 | 2.4933 | 2.7946 | 3.2829 |
| | Nonparametric | 2.7332 | 3.2789 | 3.6561 | 4.1346 | 4.8389 |
| Level 7 Location Errors | Parametric | 1.7623 | 2.0577 | 2.3032 | 2.5705 | 3.0178 |
| | Nonparametric | 2.6686 | 3.1970 | 3.6316 | 4.0618 | 4.7432 |

Table notes: sample size = 500, rho = 1/3, and the true value of V = 3.8532.

**Table 5: Bias, Root MSE, and Coverage Probabilities for V Estimators with Neighborhood-level Locations**

| | Bias | | Root MSE | | 95% CI Coverage Probability | |
|---|---|---|---|---|---|---|
| | Parametric | Nonparametric | Parametric | Nonparametric | Parametric | Nonparametric |
| Exact, True Locations | -0.0330 | -0.1149 | 0.5790 | 0.8785 | 0.9540 | 0.9470 |
| Neighborhood Group Size = 3 | -0.0634 | -0.1841 | 0.9721 | 1.0787 | 0.9480 | 0.9430 |
| Neighborhood Group Size = 6 | -0.1593 | -0.1630 | 2.5666 | 0.9777 | 0.8630 | 0.9460 |
| Neighborhood Group Size = 9 | -0.0020 | -0.2244 | 3.4497 | 1.2214 | 0.7730 | 0.9320 |

Table notes: sample size = 500, rho = 1/3, and the true value of V = 3.8532.

**Table 6: Selected Deciles for V Estimators with Neighborhood-level Locations**

| | Deciles: | 10% | 30% | 50% | 70% | 90% |
|---|---|---|---|---|---|---|
| Exact, True Locations | Parametric | 3.1338 | 3.4852 | 3.7561 | 4.0872 | 4.5967 |
| | Nonparametric | 2.6777 | 3.2075 | 3.6743 | 4.1648 | 4.8844 |
| Neighborhood Group Size = 3 | Parametric | 2.5839 | 3.2047 | 3.7386 | 4.2559 | 5.0946 |
| | Nonparametric | 2.3616 | 3.0631 | 3.5987 | 4.1737 | 5.0124 |
| Neighborhood Group Size = 6 | Parametric | 0.6450 | 1.7571 | 3.2875 | 4.9251 | 7.5874 |
| | Nonparametric | 2.5122 | 3.1436 | 3.6276 | 4.1389 | 4.9318 |
| Neighborhood Group Size = 9 | Parametric | 0.2324 | 0.6258 | 2.6149 | 6.9130 | 8.8112 |
| | Nonparametric | 2.1975 | 2.9249 | 3.4845 | 4.2088 | 5.2269 |

Table notes: sample size = 500, rho = 1/3, and the true value of V = 3.8532.

**Table 7: Power for One and Two-sided Specification Tests**

| Alternative Hypothesis | v | Power of the Two-sided Test at 5% | Power of the One-sided Test at 5% |
|---|---|---|---|
| Level 1 Location Errors | 1.0 | 0.036 | 0.064 |
| Level 2 Location Errors | 1.5 | 0.038 | 0.100 |
| Level 3 Location Errors | 2.0 | 0.085 | 0.217 |
| Level 4 Location Errors | 2.5 | 0.236 | 0.448 |
| Level 5 Location Errors | 3.0 | 0.448 | 0.698 |
| Level 6 Location Errors | 3.5 | 0.654 | 0.834 |
| Level 7 Location Errors | 4.0 | 0.751 | 0.899 |

Figure 1: Densities of $V_P - V_N$ for Different Levels of Location Errors

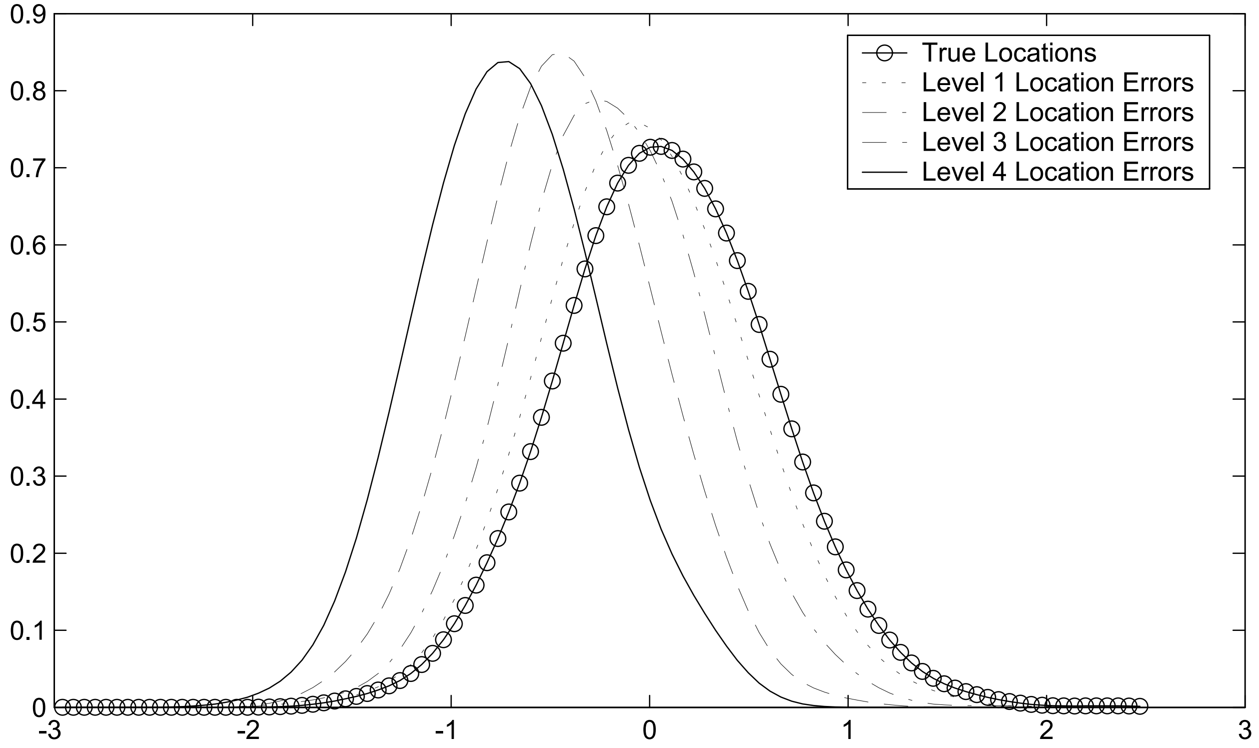


Figure 2: Densities of $V_P - V_N$ for Different Levels of Location Errors

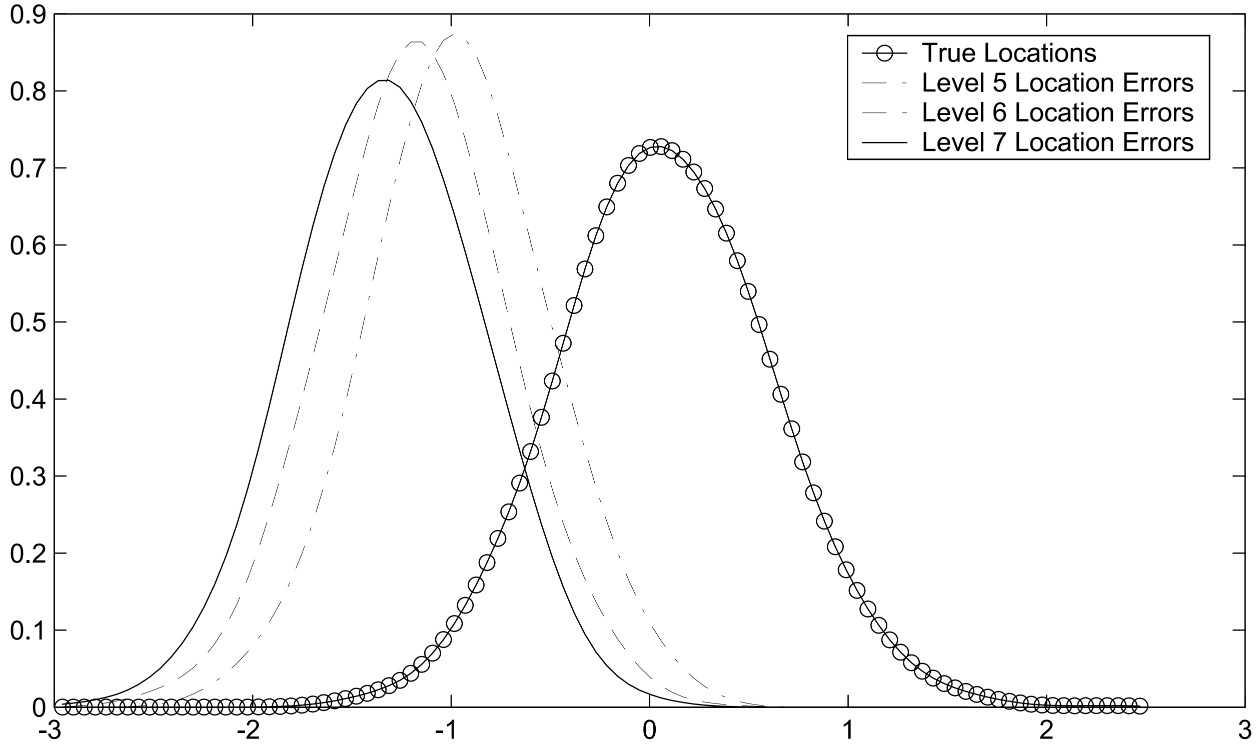


Figure 3: Densities of $(V_P - V_N)^2$ for Different Levels of Location Errors

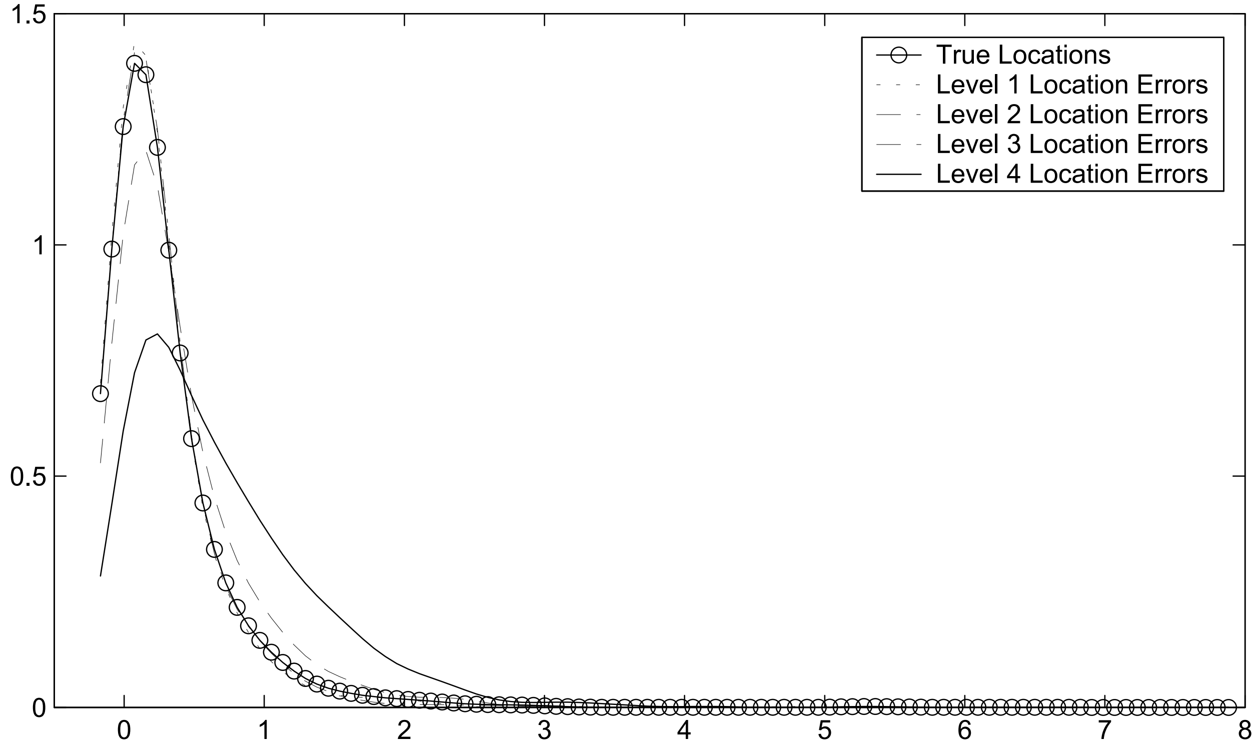
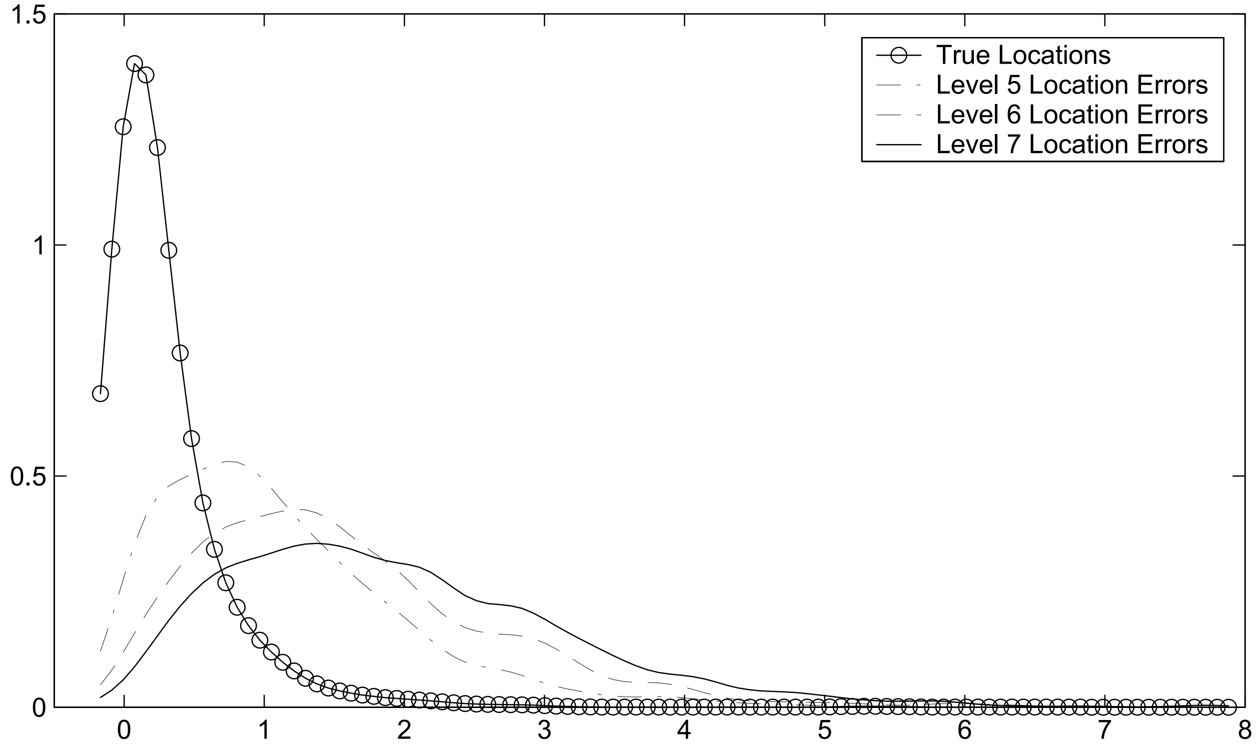


Figure 4: Densities of $(V_P - V_N)^2$ for Different Levels of Location Errors


True Locations
Level 5 Location Errors
Level 6 Location Errors
Level 7 Location Errors