Uncertainties of spatial data analysis introduced by selected sources of error

Monghyeon Lee[1], Yongwan Chun[2], and Daniel A. Griffith[3]

School of Economic, Political and Policy Sciences, The University of Texas at Dallas,
[1]Email: monghyeon.lee@utdallas.edu
[2]Email: ywchun@utdallas.edu
[3]Email: dagriffith@utdallas.edu

Abstract

Spatial data analysis uncertainty has been examined with various sources of error through simulation experiments. The uncertainty may be occurred by sampling error, measurement error, specification error, or location error. The location error may exist because spatial data have locational characteristics of features and it might be deviate from the true locations. We simulate spatial data analysis with different levels of location and measurement errors and compare the simulation results. Aggregated pediatric blood lead level point data in Syracuse, NY are utilized for the simulation with simultaneous autoregressive model. The results show that even with different levels of error, regression coefficients are quite robust and consistent. However, more deviate coefficient standard errors are with higher level of location error and small administration unit such as census block.

Keywords: Spatial data analysis, Spatial data analysis uncertainty, Location error, Measurement error.

1. Introduction

Various sources of error lead to spatial data analysis uncertainty. Like aspatial data analysis, sampling error (i.e., deviations of sample statistics from their corresponding population parameter values) is one major source of uncertainty. The scoring of attributes also contains measurement error (i.e., differences between pairs of true and measured values). Another major source is specification error, which is the difference between reality and a model's representation of it. Furthermore, due to the locational aspect of spatial data, location error is another source of uncertainty. Spatial data are georeferenced data that consist of aspatial information and spatial information. Aspatial information refers to the non-locational characteristics of features, whereas spatial information describes relative and/or absolute positioning of these features. Here, location error may affect a spatial data analysis because it might introduce deviations from true locations. These four sources of error interact and affect the quality of a spatial data analysis. In addition, features or geographical units can be merged or aggregated, perhaps due to confidentiality, data management, and/or representational concerns. Spatial aggregation can exacerbate and propagate error in spatial data from all four of these sources.

This paper summarizes our investigation through simulation experiments about how these errors impact spatial data analyses. The simulation deals with location error and measurement error with geographically aggregated variables.

2. Data and simulation experiments

The data that we use in our simulation experiments are pediatric blood lead level (BLL) measurements collected with lead poisoning tests (capillary, via finger prick, or venous, via a blood draw) for children in Syracuse, NY during 1992-1996. This database, which originally was obtained from the Onondaga County Health Department, also was utilized by Griffith et al. (2007). This database contains 16,691 blood test results for children who resided in the City of Syracuse during the six-year period, as well as the residential addresses for these children (Figure 1). The (x, y) coordinates of the residential addresses were generated through a rigorous geocoding process, and have undergone extensive cleaning. These geographic points with their individual data can be aggregated into census blocks, census block groups, and census tracts for ecological regression analysis purposes. These data serve as the points that are perturbed in the simulation experiments.

The simulation experiments were conducted with 1,000 replications of each error type, individually and then combined. The response variable is the mean BLL. The simultaneous autoregressive (SAR) model is utilized to describe mean BLLs. The independent variables vary according to the level of administration unit aggregation. Table 1 lists the variables.

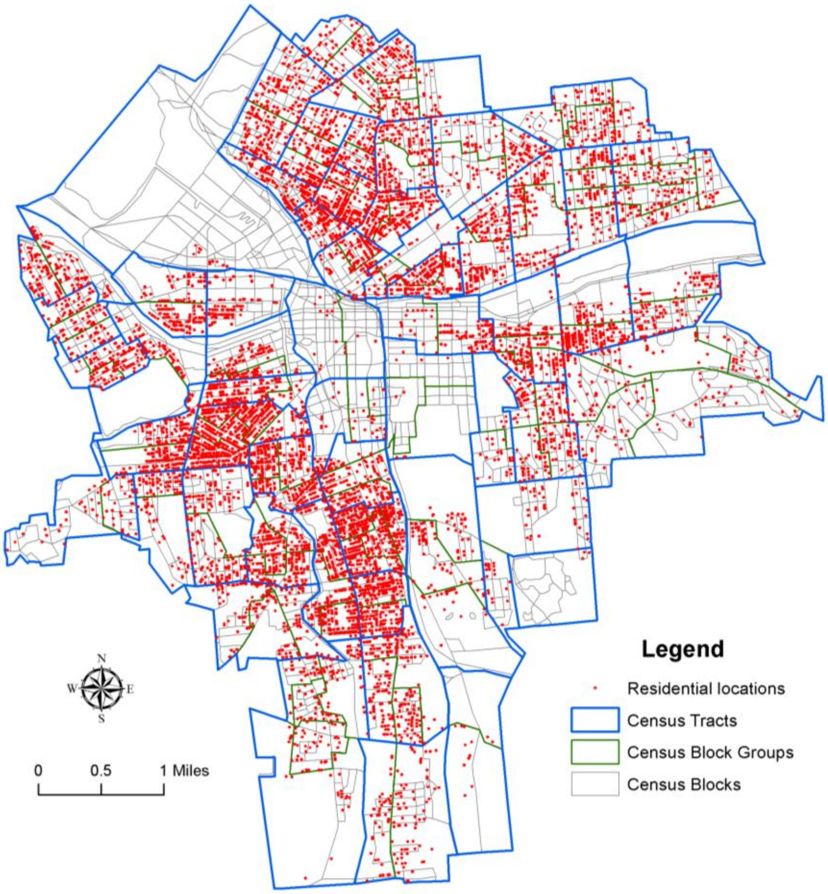


Figure 1. The distribution of pediatric BLL locations across the city of Syracuse, NY.

Administration unit	Independent variables
Census block	average house value percentage in cohort under 5 years of age percentage black percentage Hispanic population density zero indicator
Census block group	average house value percentage black population density east-west coordinate logarithm of number of cases
Census track	average house value percentage in cohort under 18 years of age

Table 1. Independent variables by different census administration levels.

2.1 Location error simulation experiment design

With the development of geographic information system (GIS) technology, geocoding is utilized to match address labelled spatial data with census or other geographic area data (Cromley and McLafferty, 2002). During this process, the locational and ecological accuracy becomes a critical concern in spatial data analyses. For example, Cayo and Talbot (2003) show that some positional errors caused by selected geocoding methods are unacceptably large.

Our simulation experiment design adds location error as follows:

1) Randomly sample 10% of BLL points
2) Assign 10m of location error with a random direction to the sampled points, constraining them to remain within the City of Syracuse
3) Conduct regressions based on the changed points' location with socio-economic variables of the areal units
4) Return to step 2) and assign 25m, 50m, 75m, and 100m of distance errors
5) When step 4) is completed, return to step 1) and sample 20%, 30%, 40%, and 50% of the points, repeating steps 2) to 4) each time

These steps were executed for the three different administration units of census block, census block group, and census track.

2.2 Measurement error simulation experiment design

Measurement error within georeferenced data leads to spatial data uncertainty (Griffith et al., 2009). We added measurement error to the response variable, BLL, following guidelines from the Centers for Disease Control and Prevention (CDC). Federal regulations allow laboratories that perform blood lead testing to operate with a total allowable error of either ±4 µg/dL or ±10%[1]. Like the location error process, the measurement error simulation experiment was conducted for five different sample sizes: 10%, 20%, 30%, 40%, and 50%. The errors follow an approximately bell-shaped Beta distribution with a range from -5 to 5. After adding measurement error, we re-calculate mean BLLs for the response variables, and repeated spatial regression analyses.

[1] http://www.cdc.gov/mmwr/preview/mmwrhtml/rr5608a1.htm

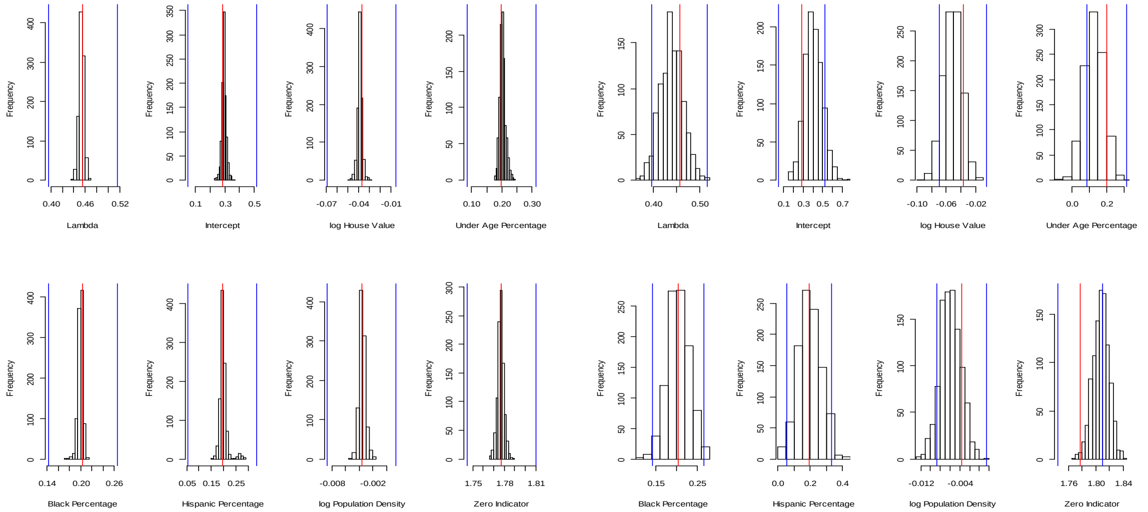
2.3 Combined location error and measurement error

The two preceding types of errors were combined and analysed with the same simulation experiment design. To maximize the total amount of error, all the points that had added location error also had measurement error.

3. Results

3.1 Location error

Figure 2 portrays the census block level regression coefficients of independent variables (red vertical lines) and their 95% confidence intervals (blue vertical lines). The histogram bars represent the coefficients of 1,000 replicates with distance errors. According to the results, almost all of the simulated coefficients are inside of the confidence intervals, even when 100m of distance error is added on 50% of observations (Figure 2b).

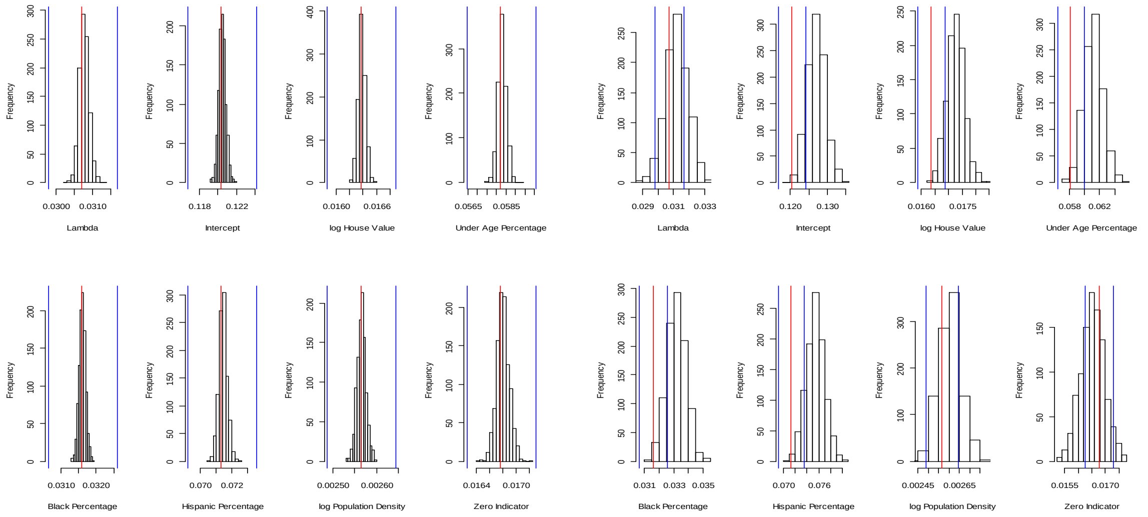


(a) 10m distance error added to 10% of the observations

(b) 100m distance error added to 50% of the observations

Figure 2. SAR coefficients for census blocks

Figure 3 portrays the coefficient standard errors. The red lines are the values of the original data without any error added and the blue lines are the confidence intervals of these original values. Histogram bars represent the coefficients from 1,000 replicates. Most of the coefficient standard errors from the replication have been inflated by larger distance errors (Figure 3b).



(a) 10m distance error added to 10% of the observations

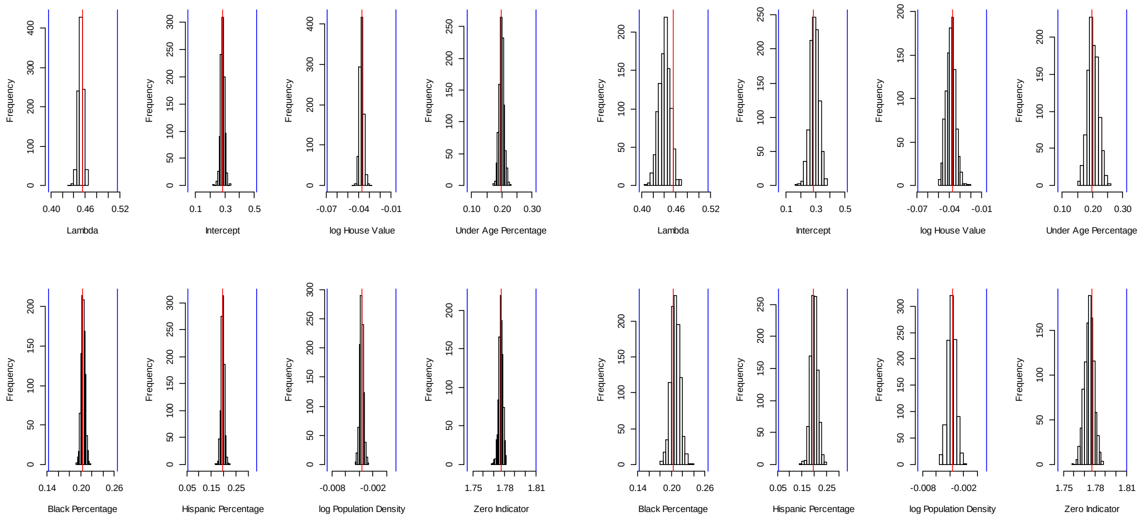
(b) 100m distance error added to 50% of the observations

Figure 3. SAR coefficient standard errors for census blocks

Results from the coarser geographic resolutions—census block groups and census tracts—have smaller uncertainty than census blocks, because the larger regions contain more observations and have fewer observation points that cross the region boundaries when adding the distance error.

3.2 Measurement error

Results from this simulation experiment are similar to those from the location error experiment. These results show that measurement errors do not cause the SAR results to deviate significantly from the original SAR results. In all cases, every parameter from 1,000 replicates is located inside its confidence intervals (Figure 4).

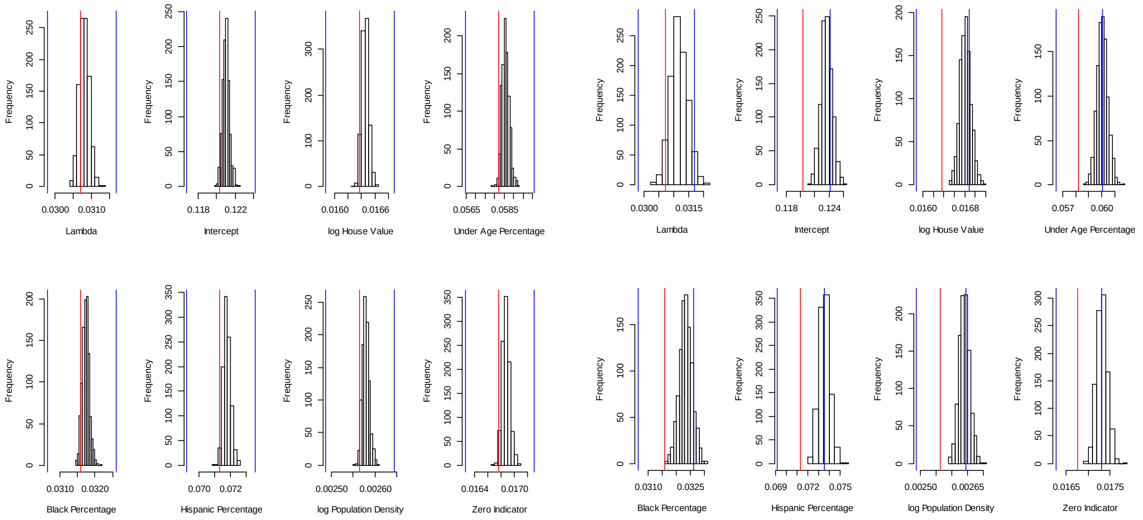


(a) Measurement error added to 10% of the observations

(b) Measurement error added to 50% of the observations

Figure 4. SAR coefficients for census blocks

Figure 5 portrays the coefficient standard errors from the measurement error simulation experiment. These results indicate that measurement error also inflates the standard error of these coefficients.



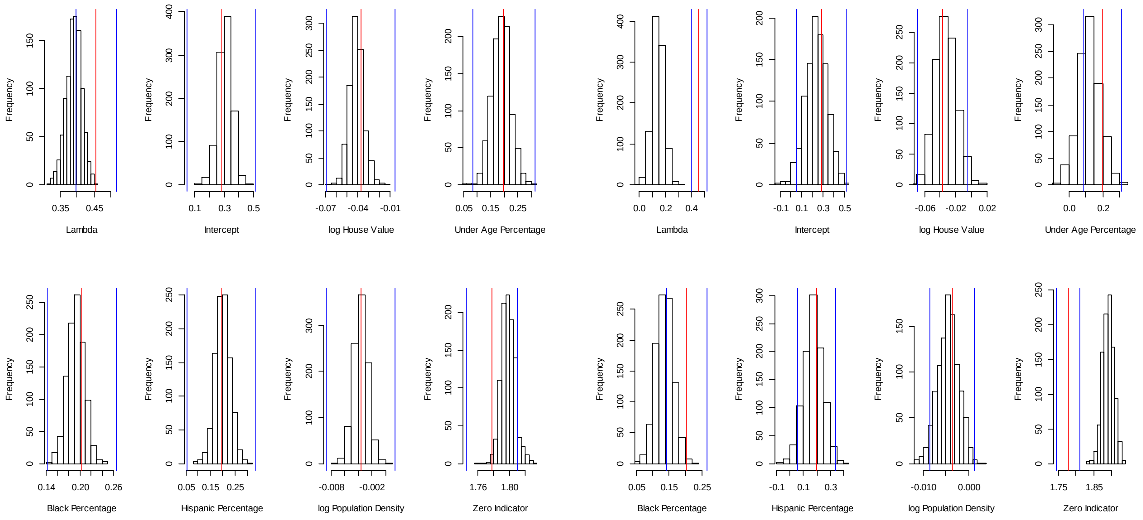
(a) Measurement error added to 10% of the observations

(b) Measurement error added to 50% of the observations

Figure 5. SAR coefficient standard errors for census blocks

3.3 Combined location error and measurement error

Estimates deviate more from their respective parameters with combined location and measurement error than when the errors are added separately (Figure 6). However, the majority of values still are within their confidence intervals.

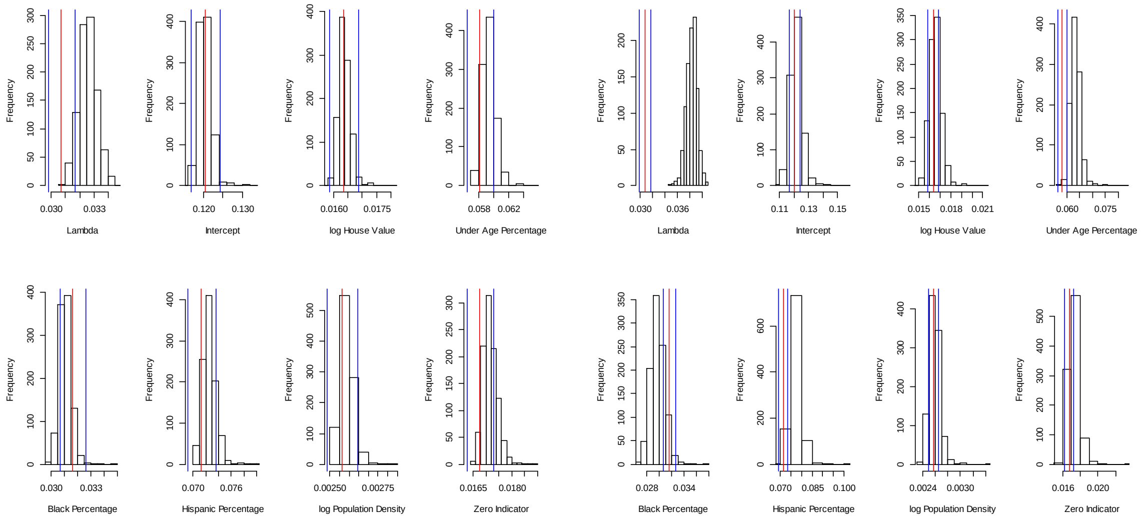


(a) 10m distance and measurement errors added to 10% of the observations

(b) 100m distance and measurement errors added to 50% of the observations

Figure 6. SAR coefficients for census blocks

Figure 7 portrays the coefficient standard errors from the combined error simulation experiment. These results indicate that most standard errors are inflated by these errors.



(a) 10m distance and measurement error added to 10% of the observations

(b) 100m distance and measurement error added to 50% of the observations

Figure 7. SAR coefficient standard errors for census blocks

4. Findings

Ecological spatial regression analyses of mean BLLs appear to be robust in the presence of relatively severe but realistic levels of locational and measurement error. Majority of coefficients from the simulation with different level of location and measurement errors are inside of original coefficient's confidence intervals, but coefficient standard errors are more inflated with higher level of location errors. Furthermore, even though we did not show the results from other administration unit such as census block group, and census tracts, the smallest unit, census block, has more significant deviation of coefficients and standard errors from their original results than larger units.

5. Acknowledgements

We would like to thank Dr. Parmanand Sinha for comments and concerning this research during the earlier phase of the project. This research was supported by the National Institutes of Health, grant 1R01HD076020-01A1; any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors, and do not necessarily reflect the views of the National Institutes of Health.

6. References

Cayo, M. R., & Talbot, T. O. (2003). "Positional error in automated geocoding of residential addresses." *International Journal of Health Geographics*, Vol.2(1), 10.

Cromley, E. K., and S. L. McLafferty. (2002), *GIS and public health*, New York.

Griffith, D. A., Millones, M., Vincent, M., Johnson, D.L., Hunt, A. (2007), "Impacts of Positional Error on Spatial Regression Analysis: A Case Study of Address Locations in Syracuse, New York." *Transactions in GIS*, Vol. 11(5):655-679.

Griffith, D. A., Johnson, D. L., & Hunt, A. (2009). "The geographic distribution of metals in urban soils: the case of Syracuse, NY." *GeoJournal*, Vol.74(4), 275-291.