Locational Error Impacts on Local Spatial Autocorrelation Indices: A Syracuse Soil Sample Pb-level Data Case Study

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

[1] University of Texas at Dallas, USA

*Corresponding author: dagriffith@utdallas.edu

Abstract

This paper focuses on propagation of location errors in spatial data analysis. Specifically, it investigates how location errors impacts local spatial autocorrelation indices that often are used to identify spatial clusters. Results of a simulation experiment using heavy metal soil sample points in Syracuse, NY, are summarized. In the simulation experiment, artificial location errors were introduced to perturb points, and then local Moran's *I* and Getis-Ord local statistics were calculated. The results show that location errors have an impact on the identification of spatial clusters. Some significant spatial clusters with no location error become insignificant ones with location errors, and some insignificant ones with no location error become significant ones with location errors. More severe deviations from the true results are observed with larger location error, as expected.

Keywords

Location error, local spatial autocorrelation, heavy metal soil sample, uncertainty

I INTRODUCTION

Location error occurs when the spatial information of a geotagged observation deviates from its true locational information. If spatial data have locational error, the error may increase any uncertainty associated with a spatial data analysis. Although uncertainties may render only slightly incorrect modelling results, they also can be completely fatal to an analysis of georeferenced data, and undermine the outcome of the spatial data analysis (Fisher, 1999). Any spatial model output may have incorrect results that have been corrupted by uncertainty propagated to the output. Local indicators of spatial association (LISA; Anselin, 1995) and Getis-Ord statistics (G_i^*; Ord and Getis, 1995) are well-known indices measuring local spatial autocorrelation and identifying spatial clusters. This paper summarizes simulation experiment results demonstrating how location error affects these spatial autocorrelation indices.

II DATA

Simulation experiments furnishing the basis of the analysis summarized in this paper utilized soil samples collected from across the City of Syracuse, NY over three years, mainly during the summers of 2003 and 2004. These samples were used to measure a suite of six heavy metals (Fe, Pb, Rb, Sr, Zn, and Zr), assayed in a chemistry laboratory using a NITON XL-700-series X-ray fluorescence (XRF) instrument (Griffith, 2008). These samples were assayed based on a 120-s testing time and NIST 2,711 standard reference materials (SRM); measurements are in milligrams of heavy metal per kilogram of soil, or ppm (Griffith, 2008). We focus on Pb because lead poisoning is one of the critical issues in public health today, and soil Pb levels tend to be strongly related to human Pb poisoning. The total number of collected soil sample points is 3,574. We exclude five points that are outside of the city boundary, and averaged 284 duplicate sample assays by location. Consequently, this study utilized 3,290 distinct soil sample points (see Figure 1). The Pb levels were subjected to a logarithm transformation—$\mathrm{LN}(\mathrm{Pb}/s_{\mathrm{Pb}} - 12)$—so that their frequency distribution better conforms to a bell-shaped curve (Griffith, 2008).

The simulation experiments involved the following steps: (1) randomly adding location error to a subset of the soil sample location points, (2) aggregating the transformed Pb measures by census geography (i.e., census tracts and census block groups), and (3) calculating local Moran's *I* and G_i^* statistics. The City of Syracuse has 57 census tracts and 147 block groups in its 2000 census geography maps. One census block group (ID: 29002) has been merged with its neighborhood block group (ID: 29001), which is in the same census tract as it, because this census block group does not contain a soil sample point.

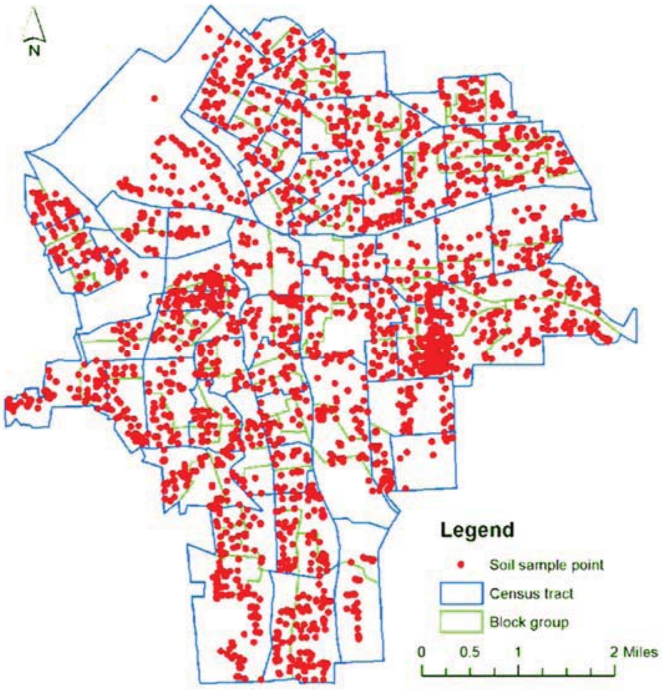


Figure 1: Soil sample points, census tracts, and block group in Syracuse, NY.

III METHOD

After extensive data cleaning, we artificially and randomly added locational deviations to the soil sample points in the simulation experiments. The simulation experimental design may be described as follows:

1) Randomly sample a selected percentage (beginning with 10%) of the 3,290 soil sample points;
2) Assign a specified amount of location error (beginning with 10m) with a random direction to each sampled point, constraining the perturbed locations to remain within the city limits;

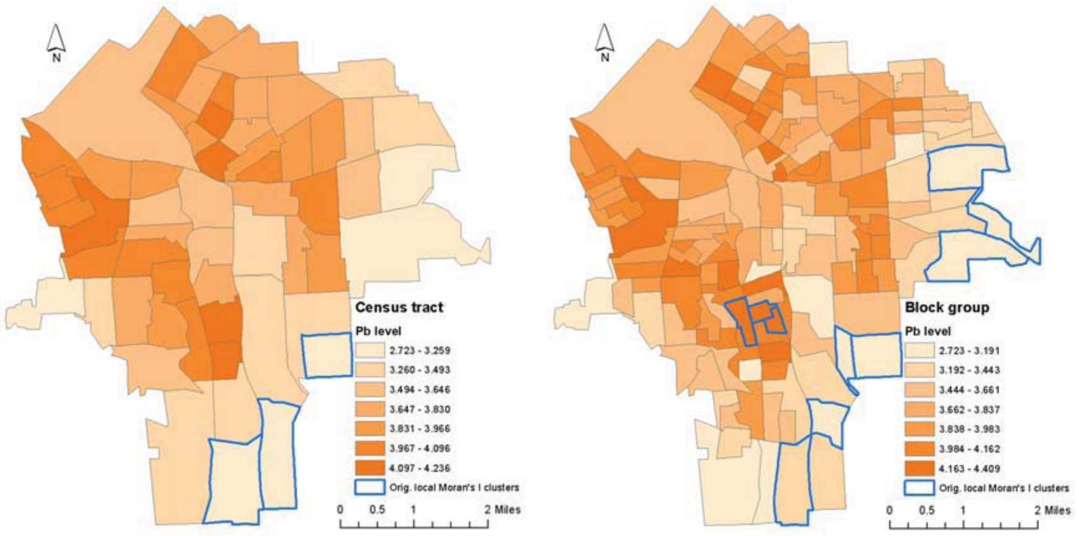
3) Aggregate transformed Pb measures in each census geography administrative unit (i.e., census tract or census block group), and then calculate local Moran's *I* and G_i^* statistics, recording clusters (e.g., hotspots and coldspots) for each simulation replicate;
4) Repeat Steps 1) to 3) 1,000 times (the number of replicates);
5) Repeat Steps 2) to Step 4) for 25m, then 50m, then 75m, and finally 100m of location error; and,
6) Repeat Steps 1) to 4) for 20%, then 30%, then 40%, and finally 50% of the soil sample points.

This simulation experimental design generates 25 different sample sizes (five percentages by five levels of location error), each with 1,000 replicates (to exploit the Law of Large Numbers).

IV RESRULTS

This section summarizes aspects of the two extreme simulation cases: results for the minimum (a 10% sample size, and 10m of added location error) and maximum (a 50% sample size, and 100m of added location error) error levels. Intermediate error level results are between these minimum and maximum results. Statistical significant levels for evaluating the local Moran's *I* have been adjusted using a Bonferroni correction based upon an effective sample size that adjusts for latent spatial autocorrelation (Chun and Griffith, 2013).

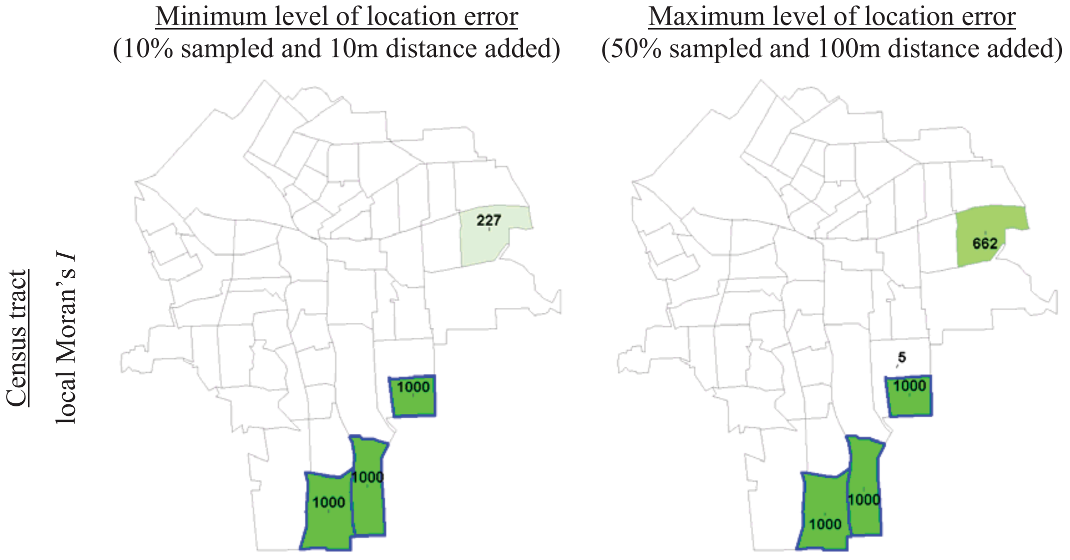
Figure 2 represents the original aggregated Pb level maps for 2000 census tracts and census block groups. The eastern and southern areas of Syracuse have relatively low Pb levels, whereas northwest and downtown areas of the City have relatively higher Pb levels.



(a) (b)

Figure 2: Pb level and its local Moran's *I* cluster maps for 2000 census tracts (a), and census block groups (b) in Syracuse, NY (Note: no G_i clusters exist).

In Figures 3 and 4, red and blue boundary regions represent original (i.e., true) statistically significant positive High-High (HH; a high local Moran's I value region is surrounded by high local Moran's I value regions) and Low-Low (LL; a low local Moran's I value region is surrounded by low local Moran's I value regions) local Moran's I geographic clusters. Meanwhile, the G_i^* maps reveal no original geographic clusters. Polygons filled with green, rather than white, are geographic clusters; the numbers that overlay them represent their aggregate detection frequencies as significant geographic clusters across the simulation experiments. For example, if a region has a red boundary, is filled with dark green, and 1,000 overlays it, this region is an original HH cluster identified in every one of the 1,000 simulation replications; in other words, location error does not affect its detection. A blue boundary region filled with light green and overlaid with 200 is an original LL cluster identified in only 200 of the 1,000 simulation replications; location error obscures or hides this geographic cluster in eighty percent of the simulation replicates. Another possible scenario is that a region has a grey boundary, which means this region is not an original cluster, but filled with green and overlaid with a number; this is a false cluster that location error creates in a certain percentage of the simulation replicates. Finally, many regions have a grey boundary and are filled with white; these are neither original nor location error created clusters.


Minimum level of location error
(10% sampled and 10m distance added)
Maximum level of location error
(50% sampled and 100m distance added)
Census tract
local Moran's I
227
1000
1000
1000
662
5
1000
1000
1000

(a) (b)

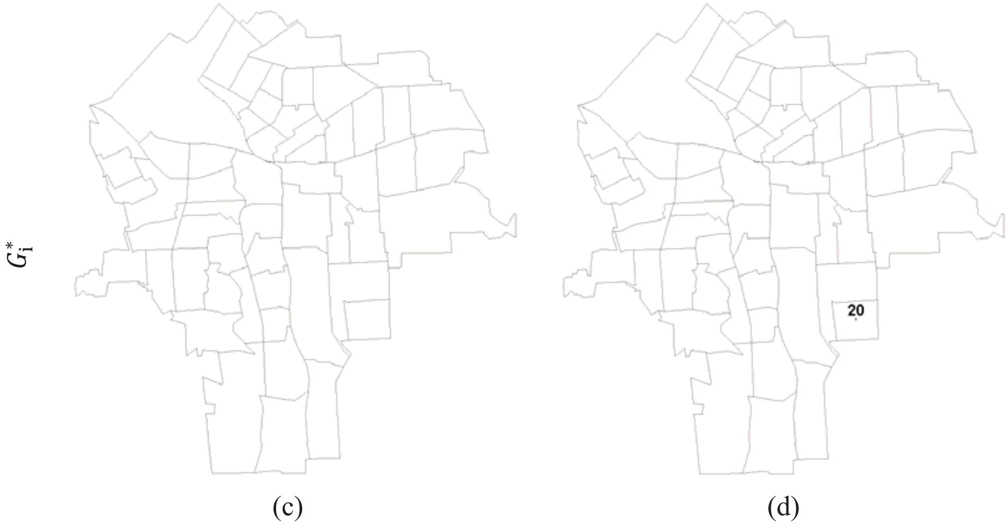
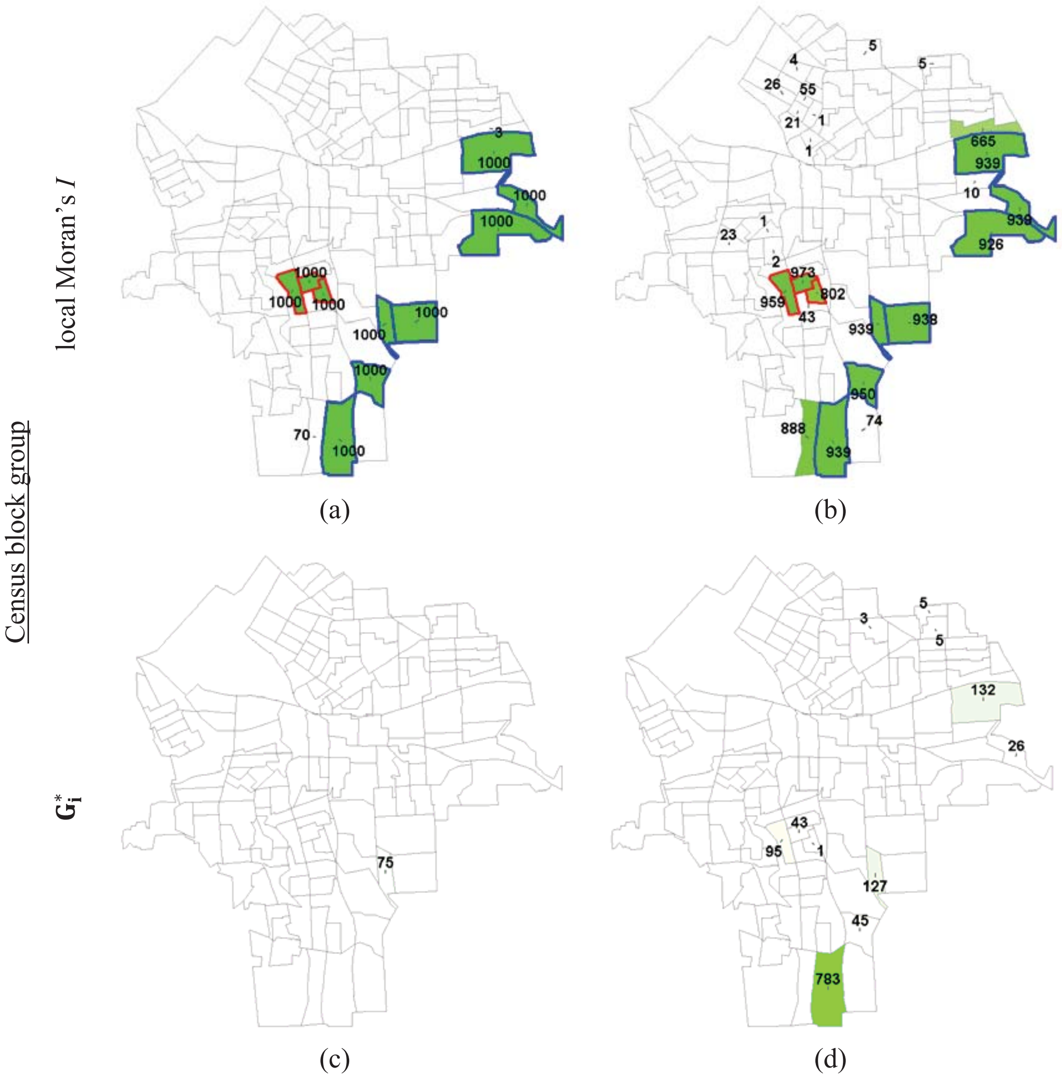


Figure 3: Local Moran's *I* and G_i^* location error simulation results for census tracts.

Figure 3 portrays census tract simulation results. When the minimum level of location error is introduced (see Figure 3a), three original LL local Moran's *I* cluster regions are colored with dark green and overlaid with 1,000. However, one non-original cluster region has been detected 227 times out of 1,000 replicates. In this latter case, location error affects the final outcome. When we introduce the maximum level of location error (see Figure 3b), even though the three original cluster census tracts are not affected, one more non-original cluster region emerges five times as a cluster. Furthermore, the census tract that has been identified as a cluster 227 times with the minimum level of location error now has been detected 662 times with the maximum level of location error. One implication here is that as the amount of location error increases, uncertainty propagation also increases and impacts spatial analyses and output. Statistically significant original hotspots or coldspots in terms of G_i^* are not identified at the census tract geographic resolution. However, G_i^* identifies a census tract as a significant coldspot 20 times in the 1,000 simulation replicates (see Figure 3d).

Minimum level of location error
(10% sampled and 10m distance added)

Maximum level of location error
(50% sampled and 100m distance added)



(a) (b)

(c) (d)

Figure 4: Local Moran's *I* and $\mathrm{G}_{\mathrm{i}}^{*}$ location error simulation results for census block groups.

Figure 4 portrays the census block group simulation results. Significant $\mathrm{G}_{\mathrm{i}}^{*}$ cluster regions do not exist at this geographic resolution, either. However, the local Moran's *I* simulation results clearly show location uncertainty propagation. In Figures 4a and 4b, the City has 10 significant original local Moran's *I* cluster regions: three of them are HH cluster regions, and seven of them are LL cluster regions. With the minimum level of location error, all of the original clusters are identified in every one of the 1,000 simulation replicates. In other words, the location error does not affect identification of the original clusters. However, two non-original clusters have been detected three and 70 times in the 1,000 replicates. They emerge because of location error. Uncertainty propagation tends to increase as the level of location error increases. With the largest level of location error, none of the original cluster regions is identified as a meaningful cluster in all 1,000 replicates. Nevertheless, detection of these clusters still occurs over 900 times; location error impacts cluster detection in this case. In terms of false positives, 16 census block groups are detected as significant geographic clusters even though they are not original clusters. In addition, the census block groups that have been detected as significant cluster regions with the minimum level of location error are identified only 665 and 888 times out of 1,000 replicates. These findings imply that location error can generate severely biased outcomes. Similarly, G_i^* results also have been affected by location error. One non-original cluster census block group has been detected 75 times with the minimum level of location error, whereas 11 originally nonsignificant census block groups have become significant geographic clusters when introduction of the maximum level of location error.

Table 1 summarizes the number of detected significant geographic clusters in the City, both the original ones and the location error created ones. This table includes both false positives and false negatives: location error creating geographic clusters that are not true clusters; and, location error obscuring or hiding geographic clusters that are true cluster. All red numbers in Table 1 denote that location error affects local spatial autocorrelation cluster detection results. Because the census block group geographic resolution tends to have units smaller than census tracts, in terms of area, location error impacts tend to be more severe at this geographic resolution. Census block groups have more soil sample points that potentially can be perturbed across unit boundaries with a certain magnitude of location error.

		Original data		**Location error added to data (1,000 replicates)**			
				Minimum (10%; 10m)		Maximum (50%; 100m)	
				Cluster	Not cluster	Cluster	Not cluster
Census tract	local Moran's *I*	Cluster	3	3	0	3	0
		Not cluster	54	1	53	2	52
	$\mathbf{G_i^*}$	Cluster	0	0	0	0	0
		Not cluster	57	0	57	1	56
Block group	local Moran's *I*	Cluster	10	10	0	0	10
		Not cluster	136	2	134	16	120
	$\mathbf{G_i^*}$	Cluster	0	0	0	0	0
		Not cluster	146	1	145	11	135

Table 1. A cross-tabulation of the number of significant soil Pb cluster regions in Syracuse, NY.

V FINDINGS

A majority of areal units remain unchanged in terms of geographic cluster detection in the presence of location error. Nevertheless, location error does tend to change a considerable number of true geographic clusters, and it also artificially creates geographic clusters that actually do not exist. The principal implication is that location error can propagate to the final output of a spatial analysis, compromising its quality and precision. Furthermore, location error seems to affected local Moran's *I* results more than G_i^* results.

VI ACKNOWLEDGEMENTS

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.

References

Anselin, L. (1995). Local indicators of spatial association—LISA. *Geographical analysis* 27, 93–115.

Chun, Y., Griffith, D.A. (2013). *Spatial Statistics and Geostatistics: Theory and Applications for Geographic Information Science and Technology*. Thousand Oaks: SAGE Publications.

Fisher, P.F. (1999). Models of uncertainty in spatial data. *Geographical information systems* 1, 191–205.

Griffith, D.A. (2008). Geographic sampling of urban soils for contaminant mapping: How many samples and from where. *Environmental Geochemistry and Health* 30, 495–509.

Ord, J.K., Getis, A. (1995). Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. *Geographical Analysis* 27, 286–306.