Impacts of positional error on spatial statistics confidence intervals

Yongwan Chun and Daniel A. Griffith

School of Economic, Political and Policy Sciences, The University of Texas at Dallas,
ywchun@utdallas.edu, dagriffith@utdallas.edu

Abstract

Spatial statistical analyses inevitably are affected by positional errors embedded in georeferenced data. Positional errors further propagate when observations need to be geographically aggregated. This data accuracy issue often results in unreliable statistical conclusion about attribute variables in spatial statistical modeling. This paper investigates how positional errors propagate to variables based on geographically aggregated units in terms of impacts on regression coefficient confidence intervals.

Keywords: Positional error, spatial regression, confidence interval, geographic aggregation.

1. Introduction

Positional errors embedded in georeferenced data inevitably affect spatial statistical analyses. They arise from differences between true locations and recorded locations of geographical features. Although technical developments, such as global positioning system (GPS) equipment, have improved global georeferenced data accuracy, positional errors continue to occur during data gathering. For example, in a geocoding procedure, positional errors can arise because of incorrect and/or incomplete address information in a database as well as the quality of street network geographic information system (GIS) files. Also, positional errors can lead to errors in geometric calculations, such as area (Chrisman and Yandell 1988) and distance (Griffith 1989). Furthermore, positional errors can propagate through the often necessary geographic aggregation of georeferenced data, for example, to preserve confidentiality of the individuals involved in a health study.

Griffith *et al*. (2007) investigate empirically how positional errors are spatial clustered and, subsequently, how positional errors impact on spatial regression coefficient estimate confidence intervals. Conducting geocoding for the same set of addresses in Syracuse, NY with the Census TIGER line files and a digital cadastral parcel map, they found that positional errors potentially can have a substantial impact on parameter estimates in a regression analysis. The purpose of this paper is to summarize a more comprehensive investigation, based on simulation experiments, of the impacts of positional errors on spatial statistics confidence intervals. Specifically, it focuses on how positional errors propagate through geographically aggregated variables and distort confidence intervals of regression model parameter estimates.

2. Simulation experiments

Exploratory simulation experiments are designed employing a bivariate normal probability model of (x, y) geocode errors for random displacements (Griffith 1989). Two different types of positional errors are considered. First, random location errors are introduced to the (x, y) geocoded point events. Each point event represents the location of an individual observation, such as the residential location of a cancer patient. These random positional errors are generated with different levels of correlation between their x and y components, ranging from near-zero to 0.9. Second, positional errors are generated with a systematic shift. Eastward shifts change only x coordinates of individual events. Similarly, northward shifts change only y coordinates. Additionally, positional errors are introduced with concomitant north and east shifts, which produce changes in both the x and y coordinates.

These simulation experiments use a database of 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 28,512 blood test results for 13,708 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 simulated event locations containing induced positional errors are geographically aggregated to the three census geographies, and then are described with regression models employing the 2000 census covariates reported in Griffith *et al.* (2007). The following different model specifications are studied here: linear regression, and eigenvector spatial filtering (ESF).

2.1. Model specification

Griffith *et al*. (2007) furnish regression model specifications in terms of the following ecological covariates.

Census tracts (n = 57): population density, average (median) house value, percentage in cohort ≤ 18 years of age

Census block groups (n = 147): population density, average house value, percentage black, east-west coordinate, logarithm of number of cases

Census blocks (n = 2,025): average house value, percentage in cohort ≤ 18 years of age, percentage black, north-south coordinate, east-west coordinate, inverse square root of number of cases

The conventional linear specification included only these covariates. The ESF linear specification also included stepwise selected eigenvectors to account for spatial autocorrelation. These covariates account for roughly 50% to 80% of the geographic variability in average BLLs across the census geography, with the additional ESF term tending to increase this amount by approximately 10%.

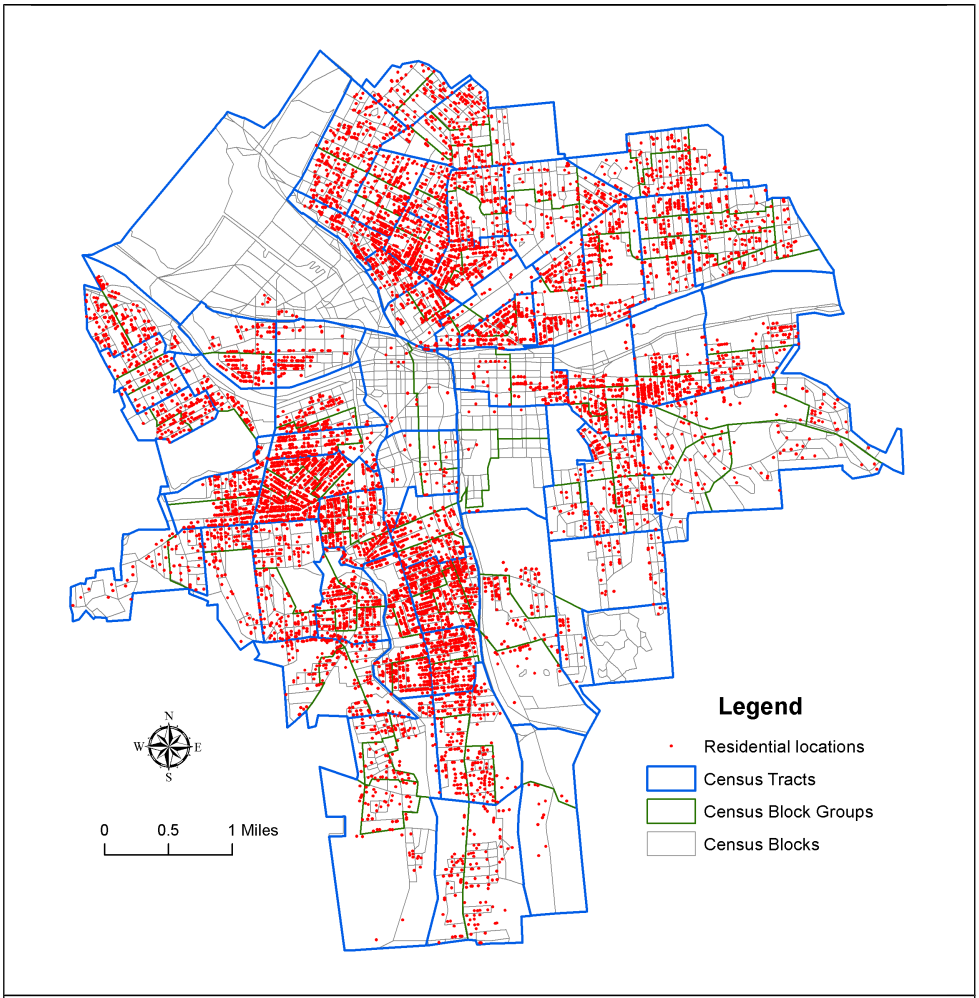

Figure 1. The distribution of pediatric BLL locations across the City of Syracuse, NY.

2.2. Random displacement

For exploratory purposes, 100 replications of randomly selected perturbations of each individual point were generated from a bivariate normal distribution, added to the point, and then the perturbed points were aggregated into census areal units to compute average BLLs. The variance was set to 5^2 and 10^2 meters to mimic the accuracy of common handheld GPS units (Wing, 2011).

Figure 2 portrays selected results (for substantive covariates appearing in the model specifications for at least two geographic resolutions) of the random perturbation simulation experiment. Correlation between the north-south and east-west displacements fails to produce any differences in the confidence intervals. ESF differences are observable between the variance levels for average house price at the census block and census block group resolutions, although these differences do not persist for the census tract resolution. Differences across all cases occur between the conventional linear and the ESF results; nevertheless, the corresponding confidence intervals overlap. These results characterize the nonportrayed covariate confidence intervals, too.

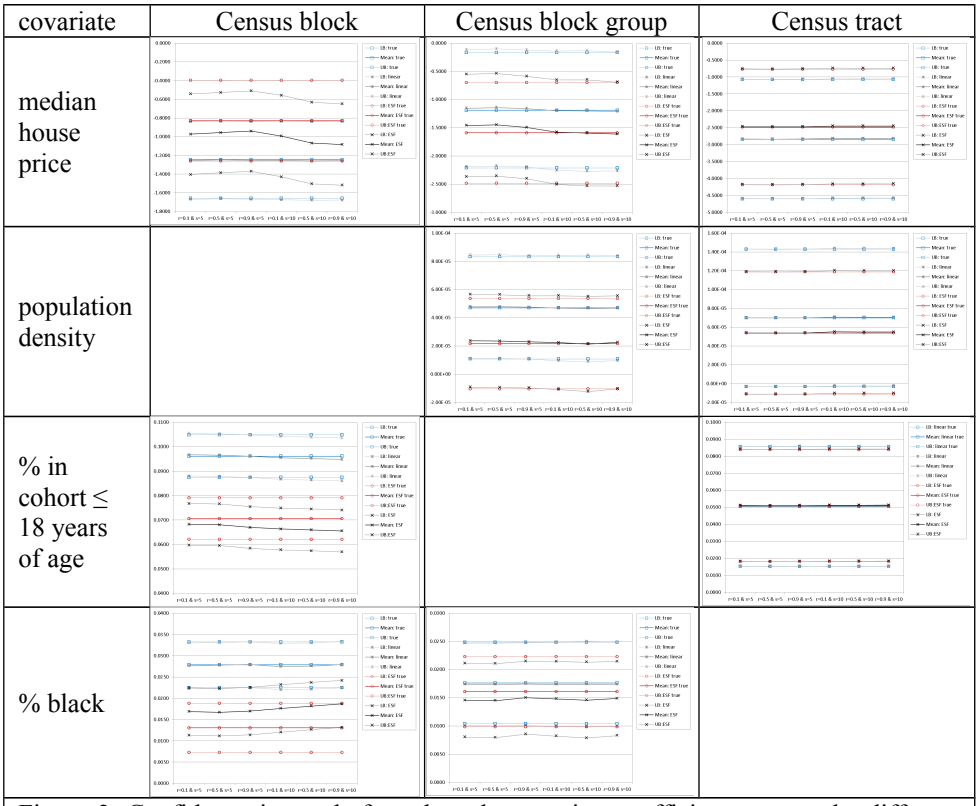

Figure 2. Confidence intervals for selected regression coefficients across the different variances, bivariate correlated random displacements, and census geography resolutions.

2.3. Systematic displacement

For exploratory purposes, displacements of 10 and 20 meters were added to the BLL locations in the east-west, in the north-south, and jointly in the east-west and north-south directions. Figure 3 portrays selected results (in parallel with Figure 2) from these simulation experiments. As for the random perturbations, differences in ESF results are observable for average house price at the census block and census block group resolutions that do not persist at the census tract resolution. In contrast with the random perturbation results, ESF differences also appear for the other substantive covariates at the two finer geographic resolutions. In all cases, differences essentially do not appear for the linear regression results. Again, the linear regression and ESF confidence intervals overlap. These results characterize the nonportrayed covariate confidence intervals, too.

3. Implication

Although, in a very general sense, uncertainty in spatial data is a well-research topic, impacts of positional errors on statistical inference about attribute variables have not been well investigated. Findings summarized here help to fill that gap in the literature, corroborate those reported in Griffith *et al.* (2007), and demonstrate that positional error affiliated with digitizing of point data in an urban area tends not to impact dramatically

upon spatial statistical results. It also illustrates, similar to most geospatial data analysis situations, the need to account for spatial autocorrelation, which does make a difference in confidence interval construction. The output of this type of research is needed to formulate fundamental guidelines for studies that need to deal with geographically aggregated point attributes, such as cancer rates. Subsequent research motivated by findings summarized in this paper will include the following: (1) a more comprehensive set of simulation experiments (e.g., 10,000 replications, larger displacements, such as 15 meters), (2) investigation of results based upon spatial autoregressive models, and (3) simulation experiments for rates of BLLs (i.e., percentage data, which requires comparisons with generalized linear models, such as binomial regression specifications).

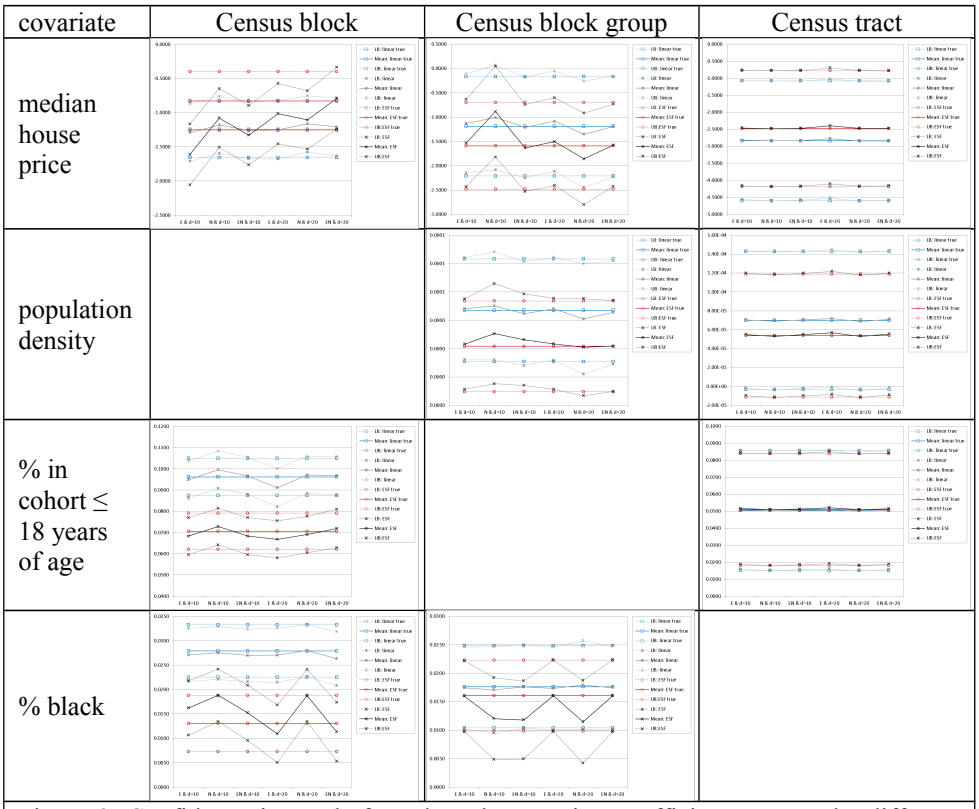

Figure 3. Confidence intervals for selected regression coefficients across the different variances, systematic displacement magnitudes, and census geography resolutions.

References

Chrisman, N.R., Yandell, B.S. (1988), "Effects of point error on area calculations: A statistical model." *Surveying and Mapping*, Vol. 48:241-246.

Griffith, D.A. (1989), "Distance calculation and errors in geographic database." In: Goodchild, M., Gopal, S. (eds.). The Accuracy of spatial databases, London, UK: Taylor & Francis Ltd, pp. 53-58.

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.

Wing, M. G. (2011). "Consumer-Grade GPS Receiver Measurement Accuracy in Varying Forest Conditions." *Research Journal of Forestry*, *5*(2):78-88.