The Impact of Incomplete Geocoding on Small Area Population Estimates

Article *in* Journal of Population Research · March 2012

DOI: 10.1007/s12546-011-9077-y

CITATIONS
10

READS
190

4 authors, including:

Jack Baker
Farmers Insurance Group
69 PUBLICATIONS **362** CITATIONS
SEE PROFILE

Adelamar Alcantara
University of New Mexico
27 PUBLICATIONS **298** CITATIONS
SEE PROFILE

Xiaomin Ruan
Portland State University
15 PUBLICATIONS **146** CITATIONS
SEE PROFILE

The impact of incomplete geocoding on small area population estimates

Jack Baker, Adelamar Alcantara, Xiaomin Ruan & Kendra Watkins

Journal of Population Research

ISSN 1443-2447
Volume 29
Number 1

J Pop Research (2012) 29:91-112
DOI 10.1007/s12546-011-9077-y

Springer

J Pop Research (2012) 29:91–112
DOI 10.1007/s12546-011-9077-y

The impact of incomplete geocoding on small area population estimates

Jack Baker · Adelamar Alcantara · Xiaomin Ruan · Kendra Watkins

Published online: 23 December 2011

Abstract Small-area population estimates are often made using geocoded address data in conjunction with the housing-unit method. Previous research, however, suggests that these data are subject to systematic incompleteness that biases estimates of race, ethnicity, and other important demographic characteristics. This incompleteness is driven largely by an inability to complete georeference address-based datasets. Given these challenges, small-area demographers need further, and to date largely unavailable, information on the amount of error typically introduced by using incompletely geocoded data to estimate population. More specifically, we argue that applied demographers should like to know if these errors are statistically significant, spatially patterned, or systematically related to specific population characteristics. This paper evaluates the impact of incomplete geocoding on accuracy in small-area population estimates, using a Vintage 2000 set of block-group estimates of the household population for the Albuquerque, NM metro area. Precise estimates of the impact of incomplete geocoding on the accuracy of estimates are made, associations with specific demographic characteristics are considered, and a simple potential remediation based on Horvitz-Thompson theory is presented. The implications of these results for the practice of applied demography are reviewed.

Keywords Small area estimation · Housing-unit method · Geocoding

J. Baker (✉) · X. Ruan
Geospatial and Population Studies, University of New Mexico, MSC06 3510,
Albuquerque, NM 871331, USA
e-mail: jbaker4762@gmail.com

A. Alcantara
Department of Geography, Geospatial and Population Studies, University of New Mexico,
Albuquerque, NM, USA

K. Watkins
Mid-Region Council of Governments, Albuquerque, NM, USA

Introduction

The introduction of Geographic Information Systems software to the practice of applied demography holds the promise to revolutionize the field as a new generation of 'geodemographers' extends methods of population estimation to ever-smaller geographic levels (Jarosz 2008; Swanson and Pol 2005). One source of small-area population estimates is geocoded address data, which may be used to represent demographic events such as births, deaths, or residential construction within specific geographic units such as census-block groups (Voss 2007). In a simple sense, geocoding involves matching address strings (such as 1313 Mockingbird Lane) from tabular data to equivalent strings that are georeferenced or contain specific information about their spatial location that allows them to be placed on a map as in Fig. 1 (Drummond 1995). Map placement is approximated through linear interpolation along known address ranges within a given street vector, assuming equal spacing between integers along the street (Drummond 1995; Ratcliffe 2001). Geocoded data may be aggregated within geographic units to potentially capture population change over time, allowing apparently simple extensions of basic methods of population estimation to small geographic units. For example, geocoded data on residential construction or address parcels might form the basis for housing unit-based estimates (Baker 2010; Bryan 2004; Jarosz 2008; Ruan et al. 2008).

Unfortunately, while the potential of this practice is without question, inherent difficulties in geocoding introduce challenges in the use of these data for small-area population estimates. In some instances, use of geocoded data to track housing unit stock may introduce positive errors (overcoverage). This may happen when data on housing unit loss are inadequate, mobile homes form a primary component of the housing stock and are moved without administrative data capture, or duplicate addresses are not thoroughly removed from input datasets (National Research Council 2010; Perrone 2008). Assuming that duplicate addresses are removed, the primary source of overcoverage errors is inadequate surveillance of housing unit loss or mobile home replacements. Fortunately, most previous studies suggest that the former source of error should be rather small, with rates of housing unit loss quite modest (perhaps 1.5% per year overall) and related primarily to age and housing unit type (Baer 1990; Brown 2008; Perrone 2008). Thus, the overall biasing effect of overcoverage errors is likely to be relatively small except in geographic units characterized by a large proportion of mobile homes with high turnover rates. While not unimportant, the magnitude of these errors is likely to pale in comparison to the potential biases introduced by undercoverage driven by incomplete geocoding. Because of inherent limitations in the georeferencing process, address datasets will without exception contain a fraction of addresses that will remain ungeocoded. This may potentially introduce gaps in coverage (Fig. 1) that have important consequences for the accuracy and precision of small-area population estimates. Recent reviews suggest that under optimal conditions, geocoding success rates in urban areas will vary between 75 and 90%, while success in rural areas may be substantially lower: about 40% or even less (Goldberg et al. 2007; Rushton 2006; Zandbergen 2009). These are not small potential undercoverage errors and they may obviously introduce very large inaccuracies or biases in estimates of population

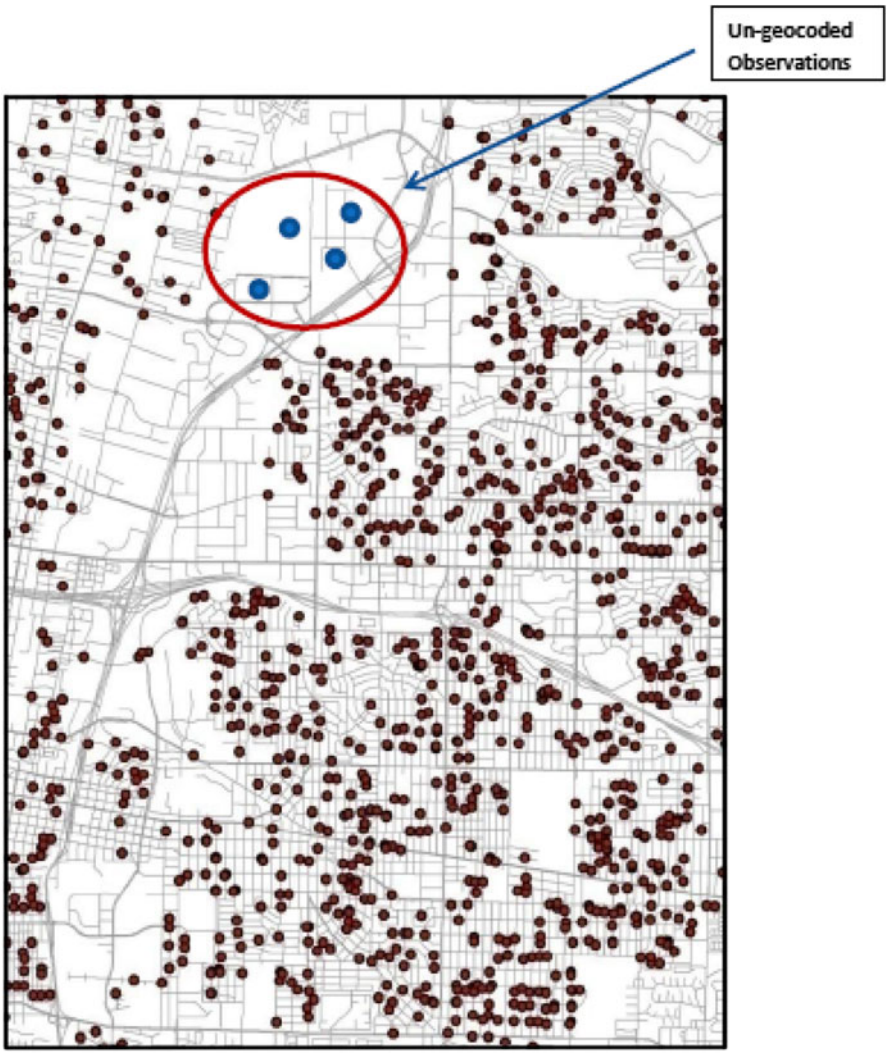

Fig. 1 Geocoded and ungeocoded address point data

(Baker 2010; Ruan et al. 2008). This paper evaluates the impact of incomplete geocoding on small area population estimates made using the housing-unit method (Bryan 2004).

Geocoding-based undercoverage is known to be systematic, spatially-dependent, and associated with biased estimates of important demographic characteristics such as race, ethnicity, and urban/rural residence (De Bruin and Bregt 2001; Gilboa 2006; Oliver 2005; Zandbergen 2009). Haining (2003) has pointed out that incompleteness of spatial datasets cannot be ignored (Little and Schenker 1994): spatial clusters of incomplete data will have a large influence on the fit of any model designed to describe spatial relationships. Such incompleteness of data should be presumed to

adversely influence estimates of spatial distributions of all data types equally (Belsley et al. 1980), including small-area estimates of population (Baker 2010; Ruan et al. 2008). Approaches to the remediation of spatially dependent missing data (Le Sage and Pace 2004) are often focused on 'filling in' gaps in data using imputation (Little and Rubin 1987), often based on interpolation (Haining 2003), Kriging or other regression-based models (Haining 2003; Le Sage and Pace 2004; Little and Rubin 1987), algorithms such as sub-area allocation models (Smith et al. 1999), or even more sophisticated 'stochastic cellular automaton' models (Sprott 2004). These models are most often preceded, necessarily, by identification of areas that are under-represented in spatial data (Haining 2003) and therefore probably stem from incomplete geocoding. For example, many small-area demographers use alternative data sources such as aerial photographic imagery to check for errors and deficiencies in geocoding.

Incomplete geocoding is driven by a number of straightforward yet difficult-to-remedy factors (Goldberg et al. 2007; Karimi and Durcik 2004). For example, failure to geocode a specific address can stem from a bad address (typos, transpositions in data entry such as 'st' being rendered 'ts' in the word 'street'), incomplete road networks (missing streets on the electronic map); or errors in the range of addresses specified in the electronic road network (Drummond 1995; Goldberg et al. 2007; Ruan et al. 2008; Rushton 2006). Manual remediation of these data can be both costly and time-consuming, suggesting that an evaluation of the merits of using unadjusted data may be worthwhile. The reviewed research on the potential shortcomings of geocoded data suggest caution to applied demographers wishing to use these data; however, to date no study has specifically evaluated to what extent the phenomenon might effect the accuracy of small-area population estimates. While the unqualified use of geocoded data as representative is clearly inappropriate given its known shortcomings (Gilboa 2006; Goldberg et al. 2007; Haining 2003; Karimi and Durcik 2004; Oliver 2005; Rushton 2006; Zandbergen 2009), without further specific evaluations, it remains unclear just how inappropriate such an application may be. Applied demographers making small-area population estimates with geocoded data should know the magnitude of the impact of geocoding-based undercoverage on accuracy. They should be informed about whether these areas are spatially clustered and if observed undercoverage is statistically significant. Moreover, they should be aware of whether incomplete geocoding is associated with important demographic and socio-economic factors, perhaps leading to biased estimates for specific demographic groups. At present, there is no published research on these relationships to guide applied demographers in the use of georeferenced data.

This paper evaluates the impact of incomplete geocoding on the accuracy of a set of Vintage 2000 block-group-level population estimates for the Albuquerque, NM metro area made using the housing unit procedure (Bryan 2004; Murdock and Ellis 1991; Smith and Mandell 1984; Starcynik and Zitter 1968). Estimates are made for the household population only, using geocoded data on building permits in conjunction with the true Census 2000 values of occupancy and average household size to isolate the effect of geocoding errors on accuracy of the estimates (an approach suggested by Stan Smith, personal communication). This also avoids the

additional challenge of estimating the group quarters population (National Research Council 2010), for which estimation errors related to surveillance and geocoding are best considered separately. Accuracy is evaluated as the discrepancy between estimates and observed Vintage 2000 household populations, measured as the mean numeric and percentage errors across all block groups (Shahidullah and Flotow 2005; Smith and Mandell 1984; Smith et al. 1999; Smith and Shahidullah 1995). The spatial patterning of these errors is explored using heat maps, in which increasingly dark coloration of block groups represents increasing magnitude of error (Berke 2005; Fotheringham et al. 2002). The relative importance of these errors is analysed in a descriptive fashion using map-based loss-function analysis (Bryan 2000; Hough and Swanson 2006), a method for assessing the relative importance of specific block-group-level errors in light of variation in the population sizes associated with each areal unit. The block-group-level statistical significance of these errors is estimated using Kuldorff's spatial scan statistic (Berke 2005; Kuldorff 1997; Neill 2009; Pollack et al. 2006) and the association of geocoding-based undercoverage with a number of demographic and socio-economic characteristics of block groups is explored using Poisson regression (Coleman 1964; Long 1997; Neter et al. 1999). Last, this paper presents and evaluates the effect of a simple potential method for remediating geocoded data on residential construction using Horvitz-Thompson adjustments (Horvitz and Thompson 1952; Judson and Popoff 2004; Samford 1967) based on a previous locally specific analysis of geocoding success at small geographic levels by the Geospatial and Population Studies at the University of New Mexico. The implications of this research for the practice of small-area population estimation are reviewed and the limitations of the current study are discussed.

Materials and methods

Hypotheses and variable measurement

This research tests the alternative hypotheses that (1) use of geocoded data on residential construction reduces the accuracy of population estimates at the block-group level, (2) these reductions are statistically significant for specific block-groups, (3) these reductions are associated with specific demographic and socio-economic population characteristics at the block-group level, and (4) these errors may be remediated using Horvitz-Thompson adjustments. Differences between estimated and observed 2000 household populations are used to measure the accuracy of estimates in terms of average overall numeric and percentage errors across all block groups. Because error distributions are asymmetric and tend to be right-skewed (Shahidullah and Flotow 2005; Tayman and Swanson 1999), both mean and median errors were considered in evaluating the accuracy of estimates. Differences in these error distributions between unremediated and remediated estimates were used to evaluate improvements associated with the application of Horvitz-Thompson adjustments. Analysis of the relationship of geocoding-based undercoverage with demographic and socio-economic characteristics of block

groups was measured as the partial coefficients associated with the inclusion of specific predictors in a Poisson regression model (Neter et al. 1999). The coefficient associated with each characteristic is a direct measure of the effect of a one-unit increase in the value of the variable on the Poisson count of undercoverage, after controlling for all other predictors in the regression model (Coleman 1964; Long 1997). Negative coefficients should be interpreted as indicating reductions in undercoverage while positive relationships suggest increases. Table 1 reviews the list of covariates considered for inclusion, which are exploratory rather than definitive.

Study area and database development

For this study, we defined the Albuquerque Metro area liberally, to include the municipalities of Albuquerque, Rio Rancho, Los Ranchos de Albuquerque, and Bernalillo as well as areas of Census 2000 block groups whose boundaries were at

Table 1 Factors potentially associated with geocoding-based undercoverage

Variable	Dataset
Demographic	
Proportion of population Urban	SF1
Proportion of population Hispanic	SF1
Median age	SF1
Average household size	SF1
Average family size	SF1
Proportion of population foreign born	SF3
Proportion with year of entry 1990–2000	SF3
Proportion of population that moved, 1995–2000	SF3
Proportion of population with HS diploma	SF3
Proportion of population graduated college	SF3
Economic	
Proportion of housing units owned	SF3
Proportion of population with commute 0–15 min	SF3
Proportion of population with commute 15–30 min	SF3
Proportion of population with commute 30–45 min	SF3
Proportion of population with commute 45–60 min	SF3
Proportion of population with commute >60 min	SF3
Proportion of population owning vehicle	SF3
Median household income, 1999	SF3
Proportion of households receiving public assistance	SF3
Proportion of households below poverty level, 1999	SF3
Median year housing unit built	SF3
Average rent	SF3
Proportion of population employed in 1999	SF3
Proportion of houses without complete kitchen	SF3
Proportion of houses without complete plumbing	SF3

least partly within these jurisdictional limits. The overall study area (Fig. 2) is the most rapidly growing section of the state of New Mexico between 1990 and 2000. During this time, the study area's housing stock grew by 45,466 units, with the permitting process capturing 93.03% of this activity, or 42,299 units. Data were provided by the Cities of Albuquerque and Rio Rancho as well as the New Mexico Construction Industries Division and a total of 33,873 records were successfully geocoded. These data were processed by the Mid-Region Council of Governments (www.mrcog.gov) and the Geospatial and Population Studies program at the University of New Mexico. This estimated 'match rate' of 80.07% is in line with the most recent reviews on geocoding success levels in urban areas (Zandbergen 2009).

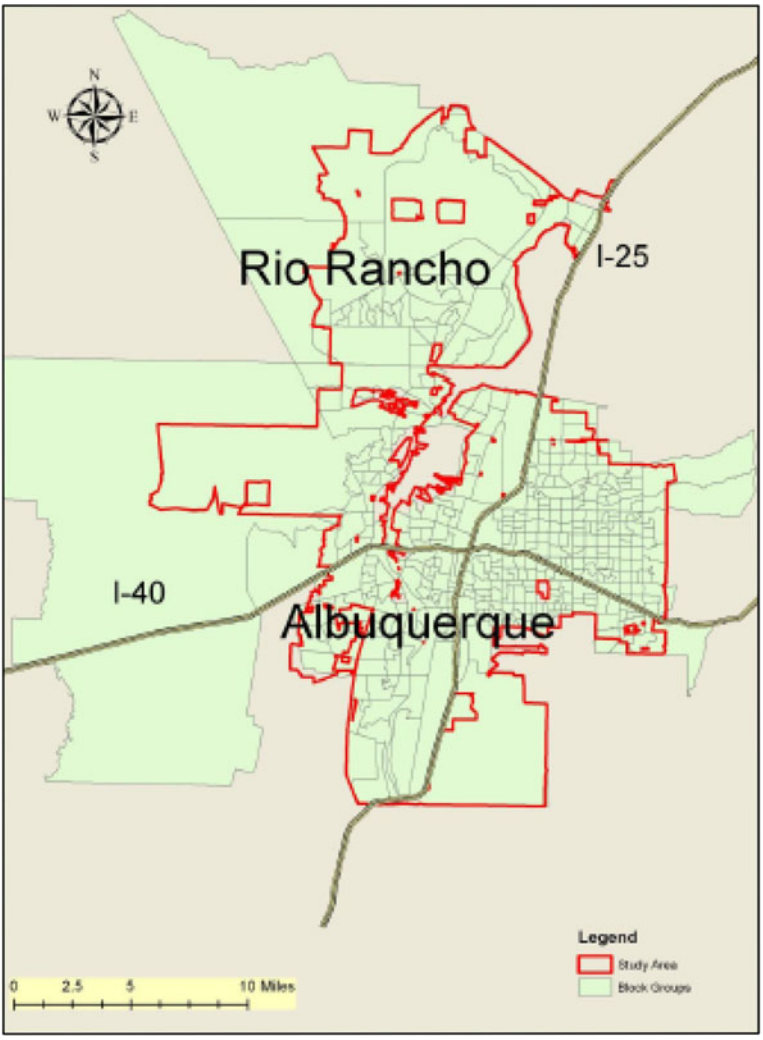

Fig. 2 Albuquerque metro study area

All permits were geocoded using the Arc GIS 9.3 'composite address locator' feature (ESRI 2009), which sequentially matched records against a number of road networks including those provided in the Census Tiger Files, as well as commercial sources including the ESRI and Teleatlas (Dynamaps) products, and locally provided networks from the City of Albuquerque (http://www.cabq.com/GIS). Importantly, this study did not rely on additional 'hand-matching' algorithms (ESRI 2009) that many researchers think enhance the success rate. In this manner, the study focuses on the effects of electronic address matching and avoids the complexities involved in evaluating individual variation in judgement-based decision associated with hand-matching and the resulting issues related to inter-observer error. Thus it should be recognized that an unknown amount of either improvement or increase in bias may be associated with this common practice and this error is beyond the scope of this paper. Each of the 33,873 successfully geocoded records were assigned to their appropriate Census block group boundary, then summed to arrive at an aggregate number of permits accounted for within each geographic unit. A total of 434 Census 2000 block groups were included in the analysis. Of these, construction was completely accounted for in 113 block groups. Of the 113 block groups for which all housing unit growth was captured in building permits, 52 of the 113 displayed positive errors, whose average was 19 housing units or 49 persons when estimated using the mean within an obviously skewed distribution. The median error was only 10 persons with outliers clearly driving larger measures of positive error. The majority of errors were under 10 persons. The amount of undercoverage error in the remaining 321 block groups varied in magnitude, as described below.

The housing-unit-based estimation procedure

The housing unit-based method of estimating population relies upon the assumption that growth in housing reflects changes in population size (Bryan 2004; Murdock and Ellis 1991; Smith and Mandell 1984; Starcynik and Zitter 1968). The method depends upon accurate tracking of residential construction and demolition over time and precise estimates of related parameters including the rate at which housing units are occupied (the occupancy rate) and the average number of persons living within each occupied housing unit (the average household size). To estimate the total population, data on the group-quarters population are also necessary; however, in this study the focus is upon estimating household population, since this will allow a more precise assessment of the effect of incomplete geocoding on the accuracy and precision of estimates without consideration of complex errors associated with estimating the group-quarters population (National Research Council 2010). The estimating equation involved is:

$$\text{Housing Units} * \text{Occupancy Rate} * \text{Average Household Size} = \text{Household Population}$$

The housing unit method should perform strongly in areas characterized by accurate surveillance of residential construction, demolition, or housing-unit movement as seen in mobile homes (Smith and Mandell 1984; Bryan 2004;

Starcynik and Zitter 1968). This may make estimates for urbanized areas more accurate than those for rural areas (Baker 2010; Ruan et al. 2008).

Loss function analysis

Loss function analysis has been used previously in applied demography to consider observed discrepancies between estimates and decennial census counts (Bryan 2000; NRC 1980; Hough and Swanson 2006). Loss function analysis involves the scaling of numeric discrepancy statistics in a fashion that allows geographic units of widely varying population sizes (as is the case in this study) to be compared while accounting for these relative differences in size. Loss function scores are computed as (Bryan 2000; Hough and Swanson 2006):

$$L_i = |\text{Estimate} - \text{True value}|/\text{Estimate}^{\wedge}\text{weight}$$

where the weight is computed as (Bryan 2000):

$$W_i = 1 - [\ln(\text{range of true values})/25]$$

Individual block-group-level loss function values are discrepancy scores, with a computed statistic that is not conceptually different from the well known Chi-squared statistic (Christensen 1996; Witmer and Samuels 1998). Their ability to enable comparisons where absolute differences for larger population units may tend to overshadow those for smaller ones, without any credible statistical evidence that they differ other than in magnitude, is directly applicable to this research. In this manner, loss function scores avoid the type II error pitfall associated with the Chi-squared statistic, which is prone to reject the null hypothesis when discrepancies are numerically large, but to retain it when they are small (Christensen 1996). In this paper, the scores are used to explore the relative importance of differences across block groups, extending the suggestion of Bryan (2000) to an exploratory spatial data analysis using heat maps. While descriptive maps of raw numeric differences are informative, the use of loss function scores in this context allows size-related biases in visual assignment of importance to be avoided. The approach is complemented by the application of spatial scanning to assess actual statistical significance.

Kuldorff's Spatial Scan statistic

Kuldorff's Spatial Scan statistic was used to identify clustering of geocoding-based undercoverage and to classify block groups according to their relative risk of undercoverage (Berke 2005; Kuldorff 1997; Neill 2009; Pollack et al. 2006). This procedure identifies a baseline intensity associated with a one-dimensional point process that is proportional to a population at risk that is distributed in a spatially heterogeneous manner (Kuldorff 1997; Kuldorff and Nagarwala 1995). Once this baseline intensity is estimated, the procedure then compares each block group to this baseline to identify geographic units that exceed an expected number of events, in this case a missed geocode (Kuldorff and Nagarwala 1995). The approach may be applied to any variety of probabilistic outcomes, including binomial, Poisson,

exponential decay, or time-to-failure distributions (Jung et al. 2007; Kuldorff 1997; Kuldorff et al. 2005). The method revolves around the flexible application of a 'scanning window' through which events are viewed with counts of relevant events made at each iteration (Boscoe et al. 2003; Kuldorff 1997). The flexible scanning window is accomplished through random placement and replacement of circles of randomly-varying radii (from 0 to some upper limit) over sets of xy co-ordinates representing the centroid of the geographic unit of interest, here block groups (Naus 1965; Turnbull et al. 1990; Wallenstein et al. 1993; Weinstock 1981). The procedure is repeated many times, perhaps 10,000 iterations, enabling the application of an MCMC-based comparison (Dwass 1957) through the Gibbs sampler method (Casella and George 1992) that compares event occurrence against the baseline intensity (Boscoe et al. 2003; Kuldorff 1997, 1999; Kuldorff et al. 2005; Turnbull et al. 1990). A likelihood ratio test is used to assess statistical significance for all sampled scan windows, and thereby estimates clustering of events when a particular geography is highlighted in the top 500 Monte Carlo trials (Kuldorff 1997). In addition, the procedure allows computation of relative risk for each geography in relation to the estimated background intensity (Berke 2005; Kuldoff 1997; Neill 2009; Pollack et al. 2006), allowing classification of block groups according to risk of experiencing incomplete geocoding. Because relative risk measures may be defined in statistical terms, with available derivations of asymptotic confidence intervals about the ratio (Christensen 1996; Zhang and Yu 1998), this procedure allows assessment of statistical significance at the block-group level in instances where the relative risk ratio's confidence interval does not overlap one. In this study, a Poisson distribution was used to model geocoding-based undercoverage. The spatial scan statistical algorithms were performed using open-source software known as SatScan, which was developed by Kuldorff and collaborators for implementation of the procedure and is freely available at www.satscan.org.

Estimation of association of incomplete geocoding with demographic and economic population characteristics

The implication of incomplete geocoding for underestimation of specific demographic groupings was assessed through simple Poisson regression analysis. A large number of demographic and socio-economic variables from the Census 2000 Summary Files 1 and 3 (Table 1) were extracted, then regressed upon the block-group counts of ungeocoded observations using the technique of Poisson regression (Coleman 1964; Long 1997; Neter et al. 1999). Backward selection was used, in which all potential independent variables are included in the first run of the model, with the predictor with the highest p-value eliminated at one-step iterations until the only predictors remaining in the model are statistically significant at the alpha less than or equal to 0.05 level. The final model was considered in light of appropriate regression diagnostics including a lack of systematic relationship between the independent variables in the model, a lack of correlation between residuals and predicted values, normality of both input variables and standardized residuals, and homogeneity of variance. The assumption of independence of predictors was assessed using graphical review of bivariate plots as well as the computation of

Variance Inflation Factors (Neter et al. 1999). Residual plots were used to assess the normality of residuals and the relationship of fitted values and standardized residuals. The final model selected met all regression diagnostics appropriately, indicating a well specified and statistically valid analysis. Since many of the observed coverage errors were zero, a zero-inflated variant of the Poisson Regression was used (Neter et al. 1999).

Estimation of Horvitz-Thompson raising factors

The theory of sampling introduced by Horvitz and Thompson (1952) was applied as a potential method for remediating incomplete geocoding in making small-area housing unit-based estimates. The basis of all sampling theory is an equal opportunity for selection, which in the case of geocoding does not exist since certain records are systematically excluded through failed georeferencing. Horvitz-Thompson estimators are raising factors that reflect the probability of non-selection and permit corrections for selection bias (Judson and Popoff 2004; Samford 1967) associated with the observed incomplete geocoding. Raising factors were computed for each block group using an independent database developed by the Geospatial and Population Studies unit at the University of New Mexico. Collated from a variety of administrative data sources, this database represents over 1,000,000 independent attempts to geocode addresses. Since over 90% of these records contain zip-code identifiers, the probability of success of geocoding was estimated for each zip-code. Horvitz-Thompson estimators were then computed as one divided by the geocoding success rate found within the zip-codes that corresponded to each block group. Each successfully georeferenced permit was then raised by the factor associated with its block group in an attempt to remediate geocoding-based undercoverage at these specific and geographically fine-grained levels.

Results

Housing-unit-based estimates of block-group populations were found to be surprisingly accurate with an overall average percentage error of 96 persons or 12.94% (Table 2). The difference between the magnitude of overcoverage errors and those associated with incomplete geocoding is striking, and clearly justifies the emphasis of this study upon the impact of incomplete geocoding in population estimation. The average of observed positive errors was only 48 persons, or 4.71% (only 10 persons when the median was used), while the average of geocoding-based undercoverage errors reached—135 persons or—15.64%. When geocoding-based undercoverage errors were considered separately, the average error within this group was—103 persons or—8.8%. The variance of these errors, however, was large, with 44 out of the 432 block groups displaying an undercoverage error exceeding 20% (10.19% of the block groups) and 16 exceeding an error of 50.0%. The observed block-group-level patterns of undercoverage are clearly spatially patterned, with significant visual adjacency in colours representing increasing undercoverage as the colour deepens (Fig. 3). While the loss function analysis did

Table 2 Errors associated with housing-unit based estimates made using geocoded data

Measure	Error type			
	Algebraic	Absolute	Algebraic percentage	Absolute percentage
Mean	−46	96	−0.0347	0.1294
Median	0	21	0.000	0.0182
Positive mean	48	*	0.0471	*
Negative mean	135	*	15.64	*

* Algebraic and absolute errors will be the same when comparisons are made only within positive or negative errors

not suggest glaring impacts of incompletely geocoded data (Fig. 4), the more sensitive spatial scan analysis (Fig. 5) indicates that much of the undercoverage is statistically significant, with clear clusterings of significance within specific sections of the study area.

Table 3 reports the results of the Poisson regression analysis performed to examine links between incomplete geocoding and demographic and socio-economic summary characteristics for each block group. The vast majority of predictors were statistically significant at the $p < 0.0001$–0.005 level; however, in many cases the magnitude of the coefficients does not indicate effect sizes that are likely to produce strong bias in the estimates reported here. Since the Poisson outcomes represent accumulation of undercoverage, it is important to remember that covariates that predict increases in geocoding-based undercounts are increasing underestimates, while negative coefficients indicate marginal improvements in geocoding-based undercoverage. Coefficients for a number of demographic and socio-economic indicators suggest that inflation or decreases in incomplete geocoding are of little importance. For example, a 1-year increase in median age predicts a statistically significant increase in undercoverage; however, with a coefficient of 0.028 (rounded), it would require a 33-year increase in median age to result in a single-person increase in error. Clearly, the effect of increase in median age in this analysis is demographically unimportant. Similarly, in another example, a 1% increase in the proportion of the population that is foreign-born and entered the country in the previous 10 years is here estimated to result in an increased undercoverage of less than one person. A 10% increase would result in only a 6.28 persons increase in geocoding-based undercoverage. The vast majority of predictors that were associated with Poisson counts of undercoverage were of similar importance, or lack thereof.

A few predictors of geocoding-based undercoverage, however, are likely to introduce more important effects. Examples include the proportion of houses without complete kitchen or plumbing facilities, the proportion of the population commuting to work at various time intervals less than 60 min, the proportion of persons completing high school and college-level education, and the proportions of the population that are of Hispanic origin or Foreign-born. A 10% increase in the percentage of the block-group population that is Hispanic would result in an increase in geocoding-based undercoverage of 15.72 persons. On the other hand, a

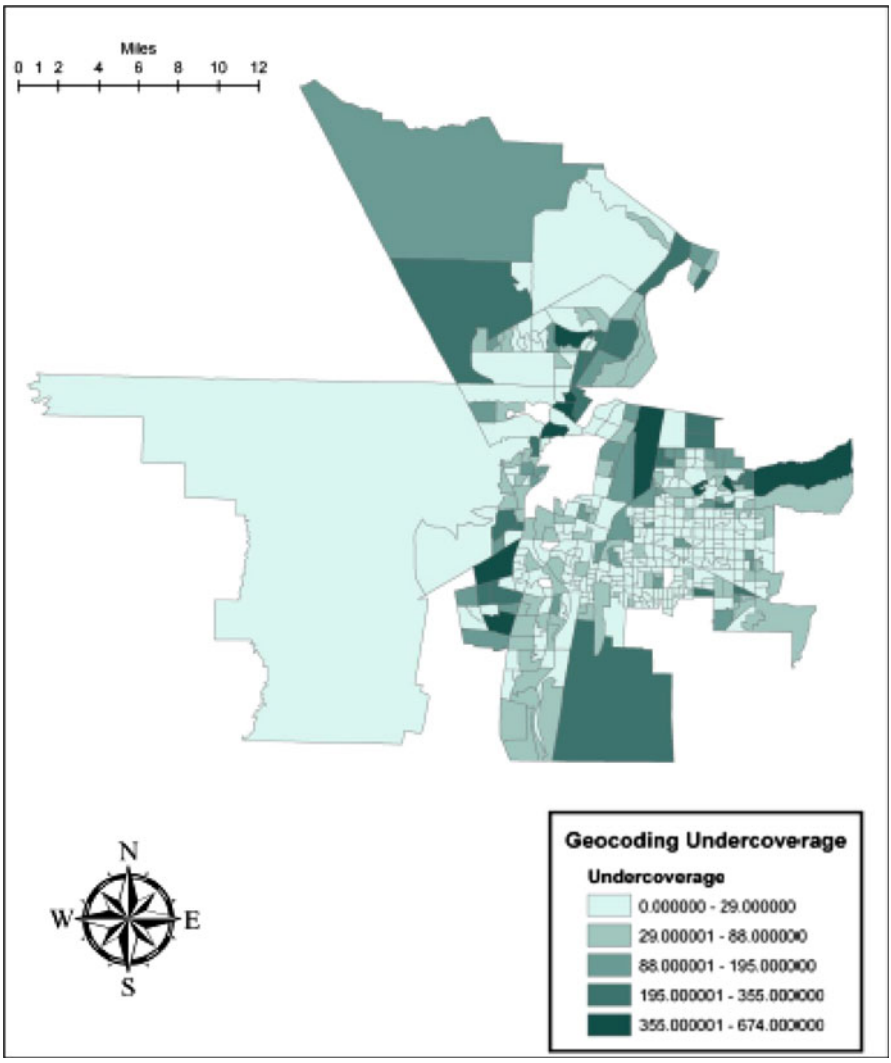

Fig. 3 Geocoding-based undercoverage counts by block group

10% increase in the proportion of the population that was foreign-born would be estimated to result in a decrease in geocoding-based undercoverage of 16.41 persons. A 10% increase in the proportion of the population with a high school education would, oddly enough, result in a 33.39 persons increase in geocoding-based undercoverage, while a ten-person increase in the proportion of persons with a college education in this study would result in approximately a 30 person increase. The largest effects of all, however, are related to increases in the percentage of houses without complete kitchen facilities or plumbing. In the former case a 10% increase results in an anticipated decrease in undercoverage of 41.23 persons while in the latter an observed decrease of 77.25 persons would be expected.

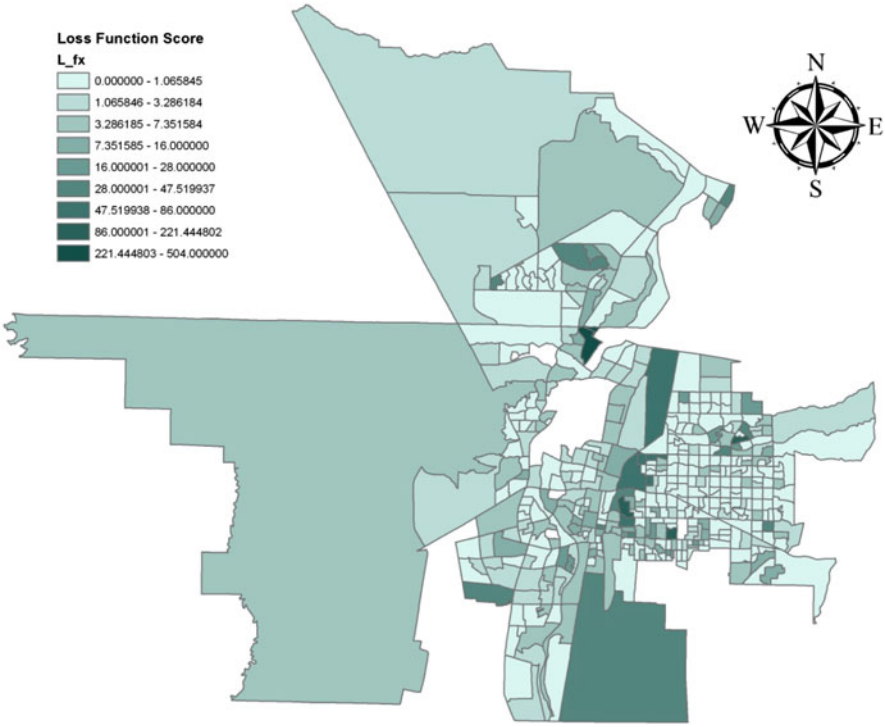

Fig. 4 Mapped loss function of geocoding-based undercoverage in housing unit-based population estimates for albuquerque metro area block groups

Adjusted estimates computed through estimation of Horvitz-Thompson raising factors had mixed effects upon the magnitude of geocoding-based undercoverage. Table 4 reports the results, which strongly suggest that locally-specific adjustment of input data reduces undercoverage-based error significantly. Horvitz-Thompson adjustments based on a block-group specific set of raising factors dramatically improved the accuracy of these estimates. The mean undercoverage associated with these estimates was approximately half of that observed with no adjustment of the geocoded data (52 persons vs. 103) and also resulted in a two per cent reduction in the average error (6.9% vs. 8.8%). The large variation and presence of outliers in the adjusted estimates was also reflected in the discrepancy between the mean and median numeric and percentage errors; however, the median numeric and percentage errors associated with the adjusted estimates was zero.

Discussion

This research suggests that incomplete geocoding potentially introduces substantial error into small-area population estimates. Indeed, it is hard to imagine a single block-group estimate with a 50% error being acceptable to users of small area

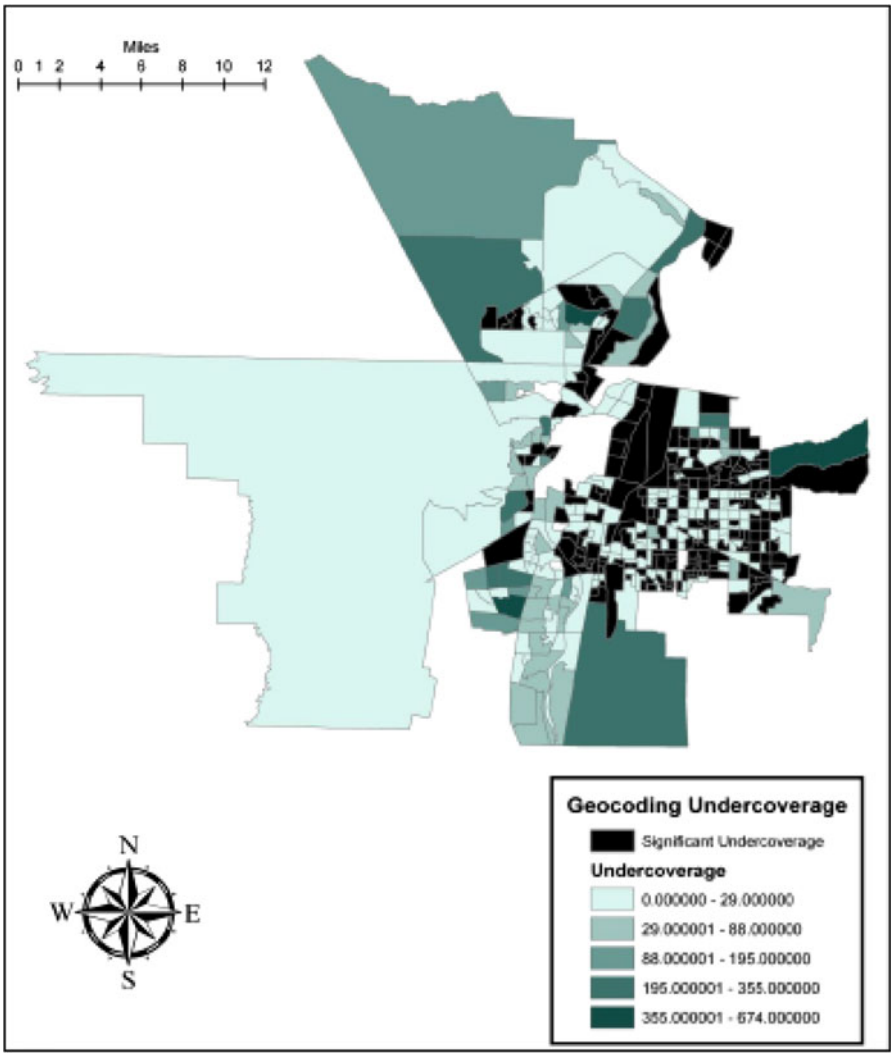

Fig. 5 Statistically significant undercoverage by block group (relative risk determined using the spatial scan statistic)

population estimates. The overall results, however, may be somewhat similar to those reported in previous assessments of the accuracy of estimates for populations of similar sizes. Previous studies have suggested that as population sizes decrease, errors increase dramatically (Smith and Shahidullah 1995; Smith et al. 1999). Only two studies systematically report error in populations of similar size to the block groups focused upon here. Smith and Shahidullah (1995) estimated errors associated with census tract estimates to vary between 20 and 60%, indicating substantial uncertainty in the quality of such estimates. Lunn et al. (1998) reports an error of approximately 10% in total population estimates made using a regression-based

Table 3 Poisson-regression results

Variable	Coefficient	*p* value
Demographic		
Proportion of population Urban	*	ns
Proportion of population Hispanic	1.572139	0.000
Median age	0.0278315	0.000
Average household size	*	ns
Average family size	0.5056548	0.000
Proportion of population foreign born	−1.641407	0.000
Proportion with year of entry 1990–2000	0.6281606	0.000
Proportion of population that moved, 1995–2000	−1.353814	0.000
Proportion of population with HS diploma	3.339139	0.000
Proportion of population graduated college	2.955759	0.000
Economic		
Proportion of housing units owned	0.8085026	0.000
Proportion of population with commute 0–15 min	−5.216059	0.000
Proportion of population with commute 15–30 min	−1.720679	0.000
Proportion of population with commute 30–45 min	−2.479673	0.000
Proportion of population with commute 45–60 min	*	ns
Proportion of population with commute >60 min	0.8484942	0.007
Proportion of population owning vehicle	0.008878	0.000
Median household income, 1999	−0.0000357	0.000
Proportion of households receiving public assistance	0.0043727	0.000
Proportion of households below poverty level, 1999	0.3433768	0.005
Median year housing unit built	0.0214132	0.000
Average rent	0.0015242	0.000
Proportion of population employed in 1999	−0.7624098	0.000
Proportion of houses without complete kitchen	−4.122818	0.000
Proportion of houses without complete plumbing	−7.725226	0.000

Table 4 Undercoverage improvements associated with Horvitz-Thompson adjustments

Method	Data source	Mean undercoverage	Median undercoverage	Mean percent undercoverage	Median percent undercoverage
No adjustment	Geocoded permits	103	29	0.088	0.027
Block-group	Zip code analysis	52	0	0.069	0.000

procedure in England, again with significant variability in accuracy between block groups. The overall magnitude of error in the estimates reported here does not seem, therefore, to differ substantially from those observed in other, similar studies and may suggest that incomplete data capture at smaller geographic levels, perhaps

stemming in all cases from issues of georeferencing success, is a primary driver of accuracy in small-area population estimates.

In the end, the introduction of an average error of over 100 persons due to geocoding-based undercoverage (an average underestimate of 8.8%) is not trivial. Moreover, it is worth observing that these estimates also contained a significant number of outliers, as over 10% of these errors were greater than 20%. It should not be forgotten that this study attempts to quantify the impact of incomplete geocoding on the accuracy of small-area population estimates, not to report the accuracy of housing-unit-based estimates made using geocoded data. To achieve this end, the true Census 2000 values for the associated parameters of occupancy and average household size were used. When estimates replace these values, it is possible that the variance in accuracy of these estimates may increase in unexpected ways since small errors in either occupancy or average household size can have large impacts upon overall estimates. Missing an occupied housing unit in a block group with an average household size of 2.5 persons could be magnified by 10% if there was just a 0.25-person overestimate of the true household size. Across 4 housing units, this very plausible level of error would mean an additional misestimate of 1 person, based solely on error in the estimate of this associated quantity. In a block group missing 20 housing units, this amounts to an error of 5 additional persons, in addition to the already present error of 50 persons associated with geocoding-based undercoverage. In real-life practice, we should expect small-area-housing unit-based estimates made using geocoded data to suffer significantly greater inaccuracy than the overall estimates observed here. From this point of view, it seems clear why many of the block-group-level undercoverage errors were found to be statistically significant using the spatial scan statistic (Fig. 5). It is very likely that small-area remediation based on various forms of 'ground-truthing' such as windshield surveys or aerial photographic reviews are here to stay (Swanson and Pol 2005; Zandbergen and Ignizio 2010). The use of geocoded data for population estimates will leave a need for careful diagnostic review of individual block-group estimates and often these methods are the only accurate option for evaluation; many block groups (Fig. 5) displayed statistically significant undercoverage that is clearly not trivial.

Results linking specific demographic and socio-economic block-group characteristics to the level of observed undercoverage suggest that systematic underestimation of specific demographic groupings may be rather unproblematic. In cases where factors appear to be strongly associated, it is unclear from a behavioural point of view why these relationships may exist. For example, the observed increases in error associated with the proportion of the population with greater levels of education and the observed reductions in error associated with a lack of basic facilities such as complete kitchens or plumbing seem to fly in the face of commonsense. These effects are strong, and even relatively small changes in these factors could potentially create large errors in small-area population estimates at the block-group level. It is possible that market forces may underlie the associations with education as newly-built housing, and the ability to afford it, may be linked to higher levels of education. In a follow-up regression of the median year built of housing on the proportion of the population with a college education, a small but statistically significant effect was identified (beta = 0.0028117, $p = 0.0000$),

lending some support to this idea. Newly-built housing is more likely to occur in areas for which electronic road networks remain incomplete as the incorporation of recently-paved streets into these maps can certainly lag by as much as two to 3 years, even in the most up-to-date sources available. If so, however, one would expect to see stronger associations between incomplete geocoding and median income or poverty measures, which are not observed in this study. The reduction in geocoding-based undercoverage in block groups with incomplete kitchens or plumbing facilities occurs infrequently in these data. It may be possible that this is also associated with newer housing; however, neither incomplete plumbing nor kitchen facilities were found to be statistically significantly associated with the median year in which housing was built in a follow-up regression.

Part of the reason for this lack of clear links with demographic and socio-economic characteristics may be the use of aggregate data for the analysis, which is clearly subject to ecological bias (Aschengrau and Seage 2003). While it does not appear that these factors are strongly related to undercoverage errors, a much greater bias has been observed in previous studies conducted with individual address data linked to characteristics (Gilboa 2006; Oliver 2005), suggesting that further research is merited before any conclusion is reached that geocoded data do not bias estimates of any particular population category.

Adjustment of estimates made for block groups subject to geocoding-based undercoverage dramatically improved population estimates. With a reduction from 103 persons to 52 (from 8.8% average error to only 6.9%), it is hard to deny the importance of such adjustment as a potential solution for errors introduced by incomplete geocoding. Similar adjustments could be applied in instances of overcoverage as well, for which adjustments for assumed incomplete coverage would only serve to magnify positive errors. The process described here relied upon the application of zip-code-level estimates of geocoding success to estimates for block groups, suggesting that this procedure could be improved upon if more fine-grained raising factors could be estimated accurately. Spatial heterogeneity across block groups within a single zip code most certainly exists, and indiscriminate application of zip-code level raising factors may lead to less precise estimates of the effect. In fact, in some cases an inaccurate adjustment could lead to overcoverage instead of undercoverage. In the current study, this limitation is introduced because while most addresses contain a zip code, before geocoding there is no equivalent indicator for block group codes. The limitation of an inability to estimate geocoding success rates at the same geographic unit as that for which estimates are to be made (applying zip-code level analysis to block-group geocoding results) is a principal shortcoming of the procedure reported here. A fruitful avenue for such research has been previously suggested by the National Academies of Science Panel on Coverage Measurement in the 2010 Census (National Research Council 2010), which recommended analysis of the life-course of specific addresses within the Census Bureau's Master Address File. Since the Bureau's Master Address File is formulated using geocoded data and is tied directly to coverage errors, the suggestion is also relevant here. An individual-address-level analysis on the characteristics associated with failure to geocode specific addresses could inform post hoc weighting algorithms in a much more powerful way than an aggregate zip-code-level

analysis. Further research on the dynamics of geocoding success is merited and may significantly improve small-area population estimation methodologies.

The implication of these findings for small-area population estimates is generally promising, but several additional important limitations of this study are worth mentioning. First, it should be remembered that in the current study area, nearly 95% of housing unit growth was captured within a well regulated and enforced permitting process. Under such conditions, construction surveillance with associated address data should be largely complete, suggesting that a failure to geocode will be a principal source of undercoverage errors. In many areas, especially rural ones, a lack of a well defined process for regulating residential construction may limit the relevance of these findings. In areas characterized by a large proportion of mobile housing that is easily moved or by poor surveillance of construction, the role of incomplete geocoding may be much smaller than these data-capture issues. The role of surveillance was intentionally limited here through selection of the Albuquerque Metro area in order to isolate the effects of incomplete geocoding. It should be remembered that approximately 5% undercapture of 1990–2000 housing-unit change was estimated for this area and it remains unknown how this undercapture is distributed across the study area. While this limited focus does not invalidate the findings of this study, it does introduce limits on extrapolation of the observed estimates of the effect of geocoding on the accuracy of small-area population estimates in this urbanized setting to other settings characterized by poor surveillance or rural residence.

Another important limitation of the current study is an inability to capture variation in the rate of completion of building permits. Although it is known that not all permits are eventually constructed even in an urbanized area such as the focus of this study, no current data exist on permit completion at the local level for the study area. One national-level study suggested a completion rate of 95% for issued permits in general (Perrone 2008); and anecdotal observations in the study area suggests that this is reasonable. Lack of completion of permits should result in a tendency to overcoverage in block groups where construction is completely accounted for. In block groups subject to geocoding-based undercoverage, however, low rates of permit completion would simply make undercoverage look smaller than it is. If anything, variation in completion rates may lead to underestimates of the true degree of undercoverage. From this perspective, errors could be marginally greater than those reported in the study. While it would be unusual for any given block group to experience large-scale incompletion of issued permits, a small percentage of error would be plausible, even up to the 5% observed at the national level.

A common practice in applied demography is to control small-area population estimates to larger ones, which is thought to minimize errors as well as to allow uniformly nested sets of estimates to be made across geographic levels (Bryan 2004; Murdock and Ellis 1991). While not explicitly addressed within this study, the spatial heterogeneity in the accuracy of estimates (Figs. 4, 5) reported here is in many cases statistically significant and suggests that some forms of controlled estimates may potentially be subject to greater bias if geocoded data are employed. When small-area estimates are controlled to larger-scale ones, misestimation of the shares of population in individual block groups can lead to magnified errors that are

often not considered. To picture this, imagine a fictitious county with two block groups displaying true shares of 54% and 46%, respectively. If the second block group suffers geocoding-based undercoverage on the order of 10% while the first is characterized by complete coverage, it would be estimated at 36%, while the proportion of the population in the first block group would rise to 6%. This two-sided magnification of error could lead to large-scale misestimation in the context of controlled estimates, suggesting that in these situations geocoding-based undercoverage might have a much larger effect than that observed in the current study.

The results of this study should be interpreted with caution; however, they strongly suggest that further research is worthwhile on the potential and pitfalls of using geocoded data to make small-area population estimates. While lingering outliers will probably continue to be a problem, adjustments made using simple Horvitz-Thompson raising factors have significantly reduced observed estimates of geocoding-based undercoverage to a level similar to those observed due to surveillance-based overcoverage. Further research aimed at investigating geocoding success at small geographic levels, and the factors associated with it, may permit the development of more accurate and specific adjustment factors that could drastically improve the accuracy of small-area population estimates. In the meantime, the current study suggests that geocoding-based undercoverage is likely to introduce an important source of error that under the best-case scenario is likely to approach or exceed 10%. In most cases, this study suggests that it is not likely, however, to result in systematic underestimation of specific demographic groupings. This research, then, provides both an estimate of the magnitude of this issue and a potential guide to future research.

References

Aschengrau, A., & Seage, G. (2003). *Essentials of epidemiology in public health* (2nd ed.). Sudbury: Jones-Bartlett.

Baer, W. C. (1990). Aging of the housing stock and components of inventory change. In D. Myers (Ed.), *Housing demography: Linking demographic structure and housing markets* (pp. 249–273). Madison: University of Wisconsin.

Baker, J. (2010). Estimating New Mexico municipalities: The devil is in the details (of data). *New Mexico Business: Current Report*. August.

Belsley, D. A., Kuh, E., & Welch, R. (1980). *Regression diagnostics: Identifying influential data and source of collinearity*. New York: Wiley.

Berke, O. (2005). Exploratory spatial relative risk mapping. *Preventative Veterinary Medicine, 71*, 173–182.

Boscoe, F. P., McLaughlin, C., Shymura, M. J., & Kelb, C. L. (2003). Visualization for the spatial scan statistic using nested circles. *Health and Place, 9*, 273–277.

Brown, W. (2008). *Changes to the housing unit stock: Loss of housing units. Presentation at the New York State Data Center Affiliate Meeting. May 15, 2008*. New York: West Point.

Bryan, T. (2000). US Census Bureau population estimates and evaluations with loss functions. *Statistics in Transition, 4*(4), 537–548.

Bryan, T. (2004). Population estimates. In J. Siegel & D. Swanson (Eds.), *The methods and materials of demography*. New York: Springer.

Casella, G., & George, E. (1992). Explaining the Gibbs sampler. *The American Statistician, 46*(3), 167–174.

Christensen, R. (1996). *Log-linear models and logistic regression* (2nd ed.). New York: Springer.
Coleman, J. S. (1964). *Introduction to mathematical sociology*. New York: Free Press.
De Bruin, S., & Bregt, A. (2001). Assessing fitness for use: The expected value of spatial datasets. *International Journal of Geographical Information Science, 15*(5), 457–471.
Drummond, W. J. (1995). Address matching: GIS technology for mapping human activity patterns. *Journal of the American Planning Association, 61*(2), 240–251.
Dwass, M. (1957). Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics, 28*, 181–187.
ESRI. (2009). Creating a composite address locator. Online at: http://help.arcgis.com/en/arcgisdesktop/10.0/help/index.html#//00250000003r000000.htm.
Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2002). *Geographically-weighted regression: The analysis of spatially-varying relationships*. West Sussex: Wiley.
Gilboa, S. M. (2006). Comparison of residential geocoding methods in a population-based study of air quality and birth defects. *Environmental Research, 101*, 256–262.
Goldberg, D. W., Wilson, J. P., & Knoblock, C. A. (2007). From text to geographic coordinates: The current state of geocoding. *URISA Journal, 19*(1), 33–46.
Haining, R. (2003). *Spatial data analysis: Theory and practice*. New York: Cambridge.
Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association, 47*, 663–685.
Hough, G., & Swanson, D. (2006). An evaluation of the American community survey: Results from the oregon test site. *Population Research and Policy Review*, *25*, 257–273.
Jarosz, B. (2008). Using assessor parcel data to maintain housing unit counts for small area population estimates. In S. Murdock & D. Swanson (Eds.), *Applied demography in the 21st century*. (pp. 89–101). New York: Springer.
Judson, D., & Popoff, C. A. (2004). Selected general methods. In J. S. Siegel & D. Swanson (Eds.), *The methods and materials of demography* (pp. 644–675). New York: Springer.
Jung, I., Kuldorff, M., & Klassen, A. (2007). A spatial scan statistic for ordinal data. *Statistics in Medicine, 26*, 1594–1607.
Karimi, H. A., & Durcik, M. (2004). Evaluation of uncertainties associated with geocoding techniques. *Computer-Aided Civil and Infrastructural Engineering, 19*, 170–185.
Kuldorff, M. (1997). A spatial scan statistic. *Communication in Statistics: Theory and Methods, 26*, 1481–1496.
Kuldorff, M. (1999). An isotonic spatial scan statistic for geographical disease surveillance. *Journal of the National Institute of Public Health, 48*, 94–101.
Kuldorff, M., Heffernan, R., Hartman, J., Assuncao, R. M., & Mostashari, F. (2005). A space-time permutation scan statistic for the early detection of disease outbreaks. *PloS Medicine, 2*, 216–224.
Kuldorff, M., & Nagarwala, N. (1995). Spatial disease clusters: Detection and inference. *Statistics in Medicine, 14*, 799–810.
Le Sage, J., & Pace, K. R. (2004). Models for spatially-dependent missing data. *Journal of Real Estate Finance and Economics, 29*(2), 233–254.
Little, R., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.
Little, R., & Schenker, N. (1994). Missing data. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook for statistical modeling in the social and behavioral sciences* (pp. 39–75). New York: Plenum.
Long, J. C. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks: Sage.
Lunn, D., Simpson, S., Diamond, I., & Middleton, L. (1998). The accuracy of age-specific population estimates for small areas in Britain. *Population Studies, 52*, 327–344.
Murdock, S., & Ellis, D. (1991). *Applied demography: An introduction to basic concepts, methods, and data*. Boulder: Westview Press.
National Research Council. (1980). *Estimating population and income of small areas*. Washington, DC: National Academy Press.
National Research Council. (2010). Coverage measurement in the 2010 Census. In Robert Bell & Michael Cohen (Eds.). Washington DC: National Academies of Science.
Naus, J. L. (1965). Clustering of random points in two dimensions. *Biometrika, 52*, 263–267.
Neill, D. B. (2009). An empirical comparison of spatial scan statistics for outbreak detection. *International Journal of Health Geographics, 8*(20), 1–16.

Neter, J., Kutner, M., Wasserman, M., & Nachtshem, C. (1999). *Applied linear statistical models* (4th ed.). New York: McGraw-Hill.

Oliver, M. N. (2005). Geographic bias related to geocoding in epidemiologic studies. *International Journal of Health Geographics.* 4(29):Online.

Perrone, S. (2008). Address coverage improvement and evaluation program—2005 National estimate for coverage of the master address file. In S. H. Murdock & D. Swanson (Eds.), *Applied demography in the 21st century* (pp. 37–85). New York: Springer.

Pollack, L. A., Gotway, C. A., Bates, J. H., Parihk-Patel, A., Richards, T., Seef, L. C., et al. (2006). Use of the spatial scan statistic to identify geographic variations in late-stage colorectal cancer in California (United States). *Cancer Causes and Control, 17*, 449–457.

Ratcliffe, J. H. (2001). On the accuracy of Tiger-type geocoded address data in relation to cadastral and census area units. *International Journal of Geographical Information Science, 15*(5), 473–485.

Ruan, X. M., Alcantara, A., Baker, J. (2008). Potential and pitfalls of geocoding for spatial demography and population estimates. *The Map Legend*. December.

Rushton, G. (2006). Geocoding in cancer research: A review. *American Journal of Preventitive Medicine, 30*(2S), S16–S24.

Samford, M. R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika, 45*, 499–513.

Shahidullah, M., & Flotow, M. (2005). Criteria for selecting a suitable method for producing post-2000 county population estimates: A case study of population estimates in Illinois. *Population Research and Policy Review, 24*, 215–229.

Smith, S., & Mandell, M. (1984). A comparison of population estimation methods: Housing unit versus component II, ratio correlation, and administrative records. *Journal of the American Statistical Association, 79*(386), 282–289.

Smith, S., & Shahidullah, M. (1995). An evaluation of projection errors for census tracts. *Journal of the American Statistical Association, 90*(429), 64–71.

Smith, S., Tayman, J., & Swanson, D. (1999). *State and local population projections: Methodology and analysis*. New York: Plenum.

Sprott, J. C. (2004). A method for approximating missing data in spatial patterns. *Computers and Graphics, 28*, 113–117.

Starcynik, D., & Zitter, M. (1968). Accuracy of the housing unit method in preparing population estimates for cities. *Demography, 5*, 475–484.

Swanson, D., & Pol, L. (2005). Contemporary developments in applied demography within the United States. *Journal of Applied Sociology, 21*(2), 26–56.

Tayman, J., & Swanson, D. (1999). On the validity of MAPE as a measure of population forecast accuracy. *Population Research and Policy Review, 18*, 299–322.

Turnbull, B. W., Iwano, E. J., Burnett, W. S., Howe, H. L., & Clark, L. C. (1990). Monitoring for clustering of disease: Application to leukemia incidence in upstate New York. *American Journal of Epidemiology, 67*, 425–428.

Voss, P. (2007). Demography as a spatial social science. *Population Research and Policy Review*, *26*, 457–476.

Wallenstein, S., Naus, J., & Glas, J. (1993). Power of the scan statistic for detection of clustering. *Statistics in Medicine, 12*, 1829–1843.

Weinstock, M. A. (1981). A generalized scan statistic test for the detection of clusters. *International Journal of Epidemiology, 10*, 289–293.

Witmer, J. A., & Samuels, M. L. (1998). *Statistics for the life sciences*. New York: Sinauer.

Zandbergen, P. (2009). Geocoding quality and implications for spatial analysis. *The Geography Compass*, *3*(2), 647–680.

Zandbergen, P., & Ignizio, D. (2010). Comparison of dasymetric mapping techniques for small area population estimates. *Cartography and Geographic Information Science, 37*(3), 199–214.

Zhang, J., & Yu, K. F. (1998). What's the relative risk? A method for correcting the odds ratio in cohort studies of common outcomes. *Journal of the American Medical Association, 280*(19), 1690–1691.