Spatially-Explicit Imputation of Missing Data in Small Area Demographic Estimates

Jack Baker*, Adelamar Alcantara, Xiaomin Ruan, Srini Vasan, Nathan Crouse

*Correspondence to:

Jack Baker
Geospatial and Population Studies
1 University of New Mexico
MSC06 3510
Albuquerque, NM 87131
(505) 277-2212
kali@unm.edu

Abstract. Small-area estimates of age-specific fertility often rely upon incompletely geocoded microdata that is likely subject to largely unignorable missingness. In spite of this challenge, little if any discussion of methods for remediating geocoded data inputs prior to making ASFR estimates has been presented in the literature. This paper presents a unified, spatially-explicit approach to correcting small area ASFR estimates. The proposed method is based on estimation of counts of unobserved birth events in light of zip-code level estimates of geocoding success rates and imputation of mother's age onto these missing data based on American Community Survey results. Unremediated and remediated ASFR estimates are used to make April 1, 2010 estimates of the 0 to 1 population for Census tracts in the Bernalillo County, New Mexico. These estimates are then compared to observed 0 to 1 age counts in the 2010 Census. Results are discussed in light of the growing need for accurate small-area demographic estimates and the challenges associated with producing them.

Introduction

Small-area demographic estimates are often founded on geocoded microdata based on addresses associated with both physical locations (xy coordinates) and demographic events such as births or deaths (Drummond, 1995; Baker et. al., 2012). This practice permits the location of demographic events within the specific areal units for which demographers typically make estimates or projections (Voss et. al., 1999; Voss, 2007) and for which demographic characteristics (such as teen pregnancy rates) important to public health practitioners are computed. While geocoding of demographic data potentially makes extension of standard demographic methods to smaller geographies possible (Jarosz, 2008; Popoff and Judson, 2004; Swanson and Pol, 2005; Baker et. al., 2012), it is not without pitfalls as it is well-known that such datasets are subject to systematic incompleteness that inevitably down-biases demographic estimates (Zandbergen, 2009; Oliver et. al., 2005; Gilboa et. al., 2006; Baker, 2010; Baker et. al., 2012). Such incompleteness likely introduces non-ignorable missing data in small area demographic estimates (Haining, 2003; Frotheringham et. al., 2002; Le Sage and Pace, 2004; Little and

Rubin, 1987; Little and Schenker, 1993; Shaffer, 1999). Existing studies have repeatedly suggested incomplete geocoding is a spatially-heterogeneous and dependent process (Baker et.al., 2012) that biases estimates of specific demographic sub-groups (Cavanaugh, 1981) along racial and ethnic and rural/urban lines (Zandbergen, 2009; Oliver et. al., 2005; Gilboa et. al., 2006). Incomplete geocoding leads to down-bias in small-area counts of demographic events that is similar to incomplete surveillance (Shyrock and Siegel, 1980; UN, 1983; Brass, 1964; Brass et.al., 1968; Baker et. al., 2011) and that is spatially patterned (Baker et. al., 2012). Because these estimates involve the use of count data, these down-biases are especially important to demographers and are expected to have a greater impact upon rate calculation than upon the estimation of sample characteristics such as mean and variance (Little and Rubin, 1987; Little and Schenker, 1994). Treatment of missing data in small-area demography may even be more important than in standard statistical models.

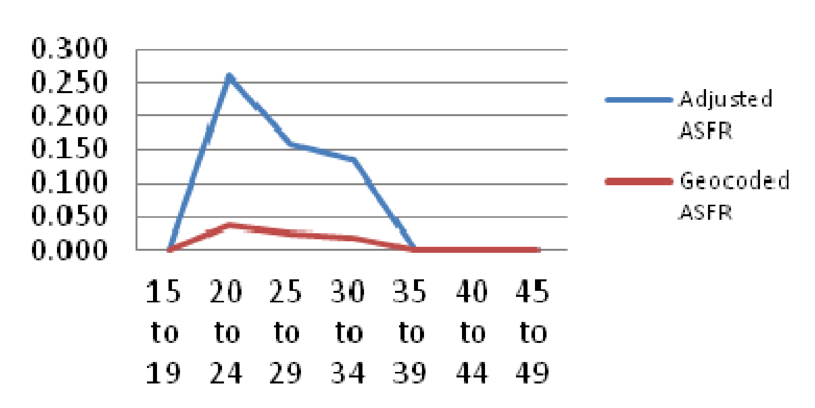

Figure 1. Geocoded vs Adjusted Age-Specific Fertility Schedules

The potentially large impact of missing data on small-area estimates of age-specific fertility (ASFR) is observed in Figure 1. Here a census-tract level set of ASFRs based on geocoded data (red line) is observed to be dramatically lower than a remediated set of estimates utilizing the well-known Brass PF ratio method (blue line--see Brass, 1964, 1968; UN, 1983; Baker et. al., 2011 for a complete description of the method). The Brass PF ratio method up-adjusts estimates of ASFR based on information on children-ever-born by age, correcting the magnitude of the ASFR estimates while preserving the shape of the ASFR curve (Brass, 1964; Baker et. al., 2011). The large-scale difference here highlights the impact of missing data in demographic estimates--and how under-reporting of counts of data has a much larger impact on estimation of demographic quantities than it may have for estimates of sample-based characteristics where the mean and variance of estimates may sometimes be adequately estimated even with missing data (Little and Rubin, 1987; Little and Schenker, 1994) or may be adequately remediated using resampling methods such as bootstrapping (Efron, 1979, 1981, 1985) or monte-carlo or statistics-based imputation (Lemieux, 2009; Schaffer, 1999). In demographic estimates based on geocoded microdata, incompleteness of data down-biases estimates and few probabilistic methods for remediation have been presented in the literature (Baker et. al., 2011, 2012).

This paper reports research on spatial patterns of missingness in geocoded data, how these patterns may impact estimates of ASFR, and how ASFR estimates made using geocoded microdata on births may be remediated to improve these estimates. Estimates of ASFRs are made for 2010 census tracts within Bernalillo County, NM for the year 2005 using geocoded microdata on births and intercensal estimates from the Geospatial and Population Studies program at the University of New Mexico. Estimates of missing birth counts were made using previous research on spatially-explicit patterns of missingness at the zip-code level reported in Baker et. al., (2012). American Community Survey (ACS) data were utilized to impute ages to these missing events at the Census tract level and a second set of ASFR estimates were made using this two-step remediation procedure (missing data estimated and age-characteristics imputed-see discussion in materials and methods below).

Materials and Methods

Data and Variables

Microdata on births for 2005 were made available by the New Mexico Bureau of Vital Records. Births data were georeferenced at the street level in ESRI's ARC-GIS 10.0. Microdata were referenced against multiple electronic road network sources, with the residual of records remaining ungeocoded being rematched sequentially against these sources in the following order: (1) Teleatlas' DynaMaps product (Vintage 2008), (2) E911 road networks provided by the New Mexico Division of Finance Administration, (3) the Census Bureau's TIGER (vintage 2009) road networks, and (4) local sources provided by jurisdictions such as the City of Albuquerque. Of the 9,291 births recorded in Bernalillo County in 2005, 7,763 were successfully geocoded at the street level (match rate = 83.55%). Data on population at risk were taken from the intercensal estimates series produced by the Geospatial and Population Studies group at the University of New Mexico. Following the convention of Mazumdar et. al. (2008), we envision this geocoding process as a specific probability model (G_i) whose success rate corresponds to a census tract-specific Bernoulli random process (Grinstead and Snell, 2006) that is spatially-heterogeneous and specific to the algorithm described here.

Estimates of "Missing Data" and Imputation of its Age-Structure

Estimates of missing data were taken from previous research by the GPS group at UNM, made in conjunction with construction of a statewide Master Address File (GPS-MAF) in preparation for the 2010 Census. This file included the collation and integration of numerous forms of address data including E911 files, assessor parcel sources, driver's license files, building permits, and vital records. Geocoded microdata were used to make zip-code level summary counts using a combination of the most recent ZCTA files provided by the US Census Bureau (2010) and the Zip-code level file available from ESRI's framework dataset for Arc-GIS version 10.0 (2010). These summary counts of successfully geocoded microdata were then compared to the original input data at the zip-code level and used to create estimates of the probability of successfully geocoding an address at the zip-code level. These estimates of "success" were modeled as a Bernoulli random process (Grinstead and Snell, 2006; Samuels and Witmer, 1998) with an associated mean (p) and variance (pq) that was utilized in further modeling described below. Horvitz-Thompson style raising factors (Horvitz and Thompson, 1952; Popoff and Judson, 2004) were estimated at the zip-code level as:

$$\mathbf{R_i = [\ 1/probability\ of\ geocoding\ success \mid G_{ij}\ ]}$$

where G_{ij} is the geocoding process. This algorithm is described in further detail in Baker et. al. (2012), to which the reader is referred. In an extension of the methodology described there, in this research uncertainty in estimates of the zip-code specific levels of missingness are made using stochastic simulation (Lemieux, 2009; Kulkarni, 2011; Fishman, 1986; ; Baker et. al., 2011) and appropriate 95% certainty bounds were constructed and utilized in modeling missing data. Estimates of the probability of successful geocoding were used to make estimates of missing data counts. Each observed birth was weighted according to its zip-code specific raising factor, census tract identifiers were assigned spatially in Arc-GIS 10.0, and the estimates of missing data resulting from these adjustments were aggregated to the Census tract level. Precedents in the literature on population ecology, where observability of species siting or occupancy is explicitly modeled, exist (Caswell, 2001; Linstrom, 2011; Morris and Doak, 2005).

As made, these estimates of missing birth counts in each census tract included no data on the age-structure of missingness. Logistic regression did not suggest that age of the mother was a significant predictor of the probability of geocoding success (p=.359), suggesting that age of the mother might safely be assumed to be "missing" at random *sensu* Little and Rubin (1987), Little and Schenker (1993), Schafer(1999) with respect to age. As such, imputation of age to the estimated counts of missing data were made using data from the American Community Survey--available as a period estimate for 2006-2010 at the Census tract level. These imputations were accomplished using monte-carlo simulation (Lemieux, 2009; Gardiner, 1983; Kulkarni, 2011) based on the suggested multinomial distribution of age at birth reported in the 2006-2010 ACS data. Multinomial probabilities are underpinned by individual binomial probabilities in each category comprising the distribution (Grinstead and Snell, 2006). These probabilities were modeled using the ACS data to estimate births within age groups including 15-19 years, 20-34 years, and 35-49 years. This broad age-structure was imputed to the missing births data, then broken out into five-year age groups (15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49) using the proportional distribution estimated in the ACS (assuring an identical data source) to arrive at five-year interval based estimates. The probability of a given birth being within each age group was estimated using simulation (Lemieux, 2009; Fishman, 1986; Kulkarni, 2011) by randomly resampling the estimates of births and age within each broad group and resampling their distribution to the finer-grained categories. All simulations were programmed in the R package (www.cran.org) assuming a normal distribution and using the ACS-reported counts as a mean and by converting the reported margin of errors in the ACS data to standard deviations through algebraic rearrangement. Since resampling algorithms accomplished using random number generations display serial autocorrelation that can bias estimates (Lemieux, 2009), here we utilize only each 100th sample (Lemieux, 2009; Fishman, 1986; Kulkarni, 2011). In the case of 11 census tracts (n=154 in Bernalillo County as of 2010, with 143 used here), the margins of error reported in the ACS were too large to permit a stable analysis (they suggested negative numbers in over 66% of realizations) and these census tracts were ejected from the final evaluation. In all other cases, a simulation-based estimate of mean and variance of the ages of mothers was possible.

ASFR Estimates

Age-specific fertility was estimated at the census tract level (with i representing the five-year age-interval and j the specific census tract) as:

$$\mathbf{ASFR_{ij} = Births_{ij} / Population_{ij}}$$

Two sets of ASFR estimates were made. The first set relied solely upon the 2005 geocoded birth counts by Census tract, divided by the July 1, 2005 GPS intercensal population estimate for each age-group. The second ASFR estimate used the same denominator estimate, but utilized a census-tract level birth count for each age-category that included the estimate of missing data described above, with age imputed based on the information available in the 2006-2010 ACS.

The estimating equation associated with this process is:

$$\mathbf{ASFR_{ij} = B_{Gij} + (B_{miss,ij} * m(age_{x\ to\ x+5,\ ij})) / n_{2005,\ x\ to\ x+5,\ ij}}$$

where B_{Gij} represents births geocoded at the census tract level, $B_{miss,ij}$ the estimate of births unobserved in the census tract and $m(age_{x\ to\ x+5,\ ij})$ is the value of the multinomial distribution of age of mothers giving birth in the last twelve months based on the ACS data, and $n_{2005,x\ to\ x+5,\ ij}$ represents population.

The procedure is detailed graphically in Figure 2.

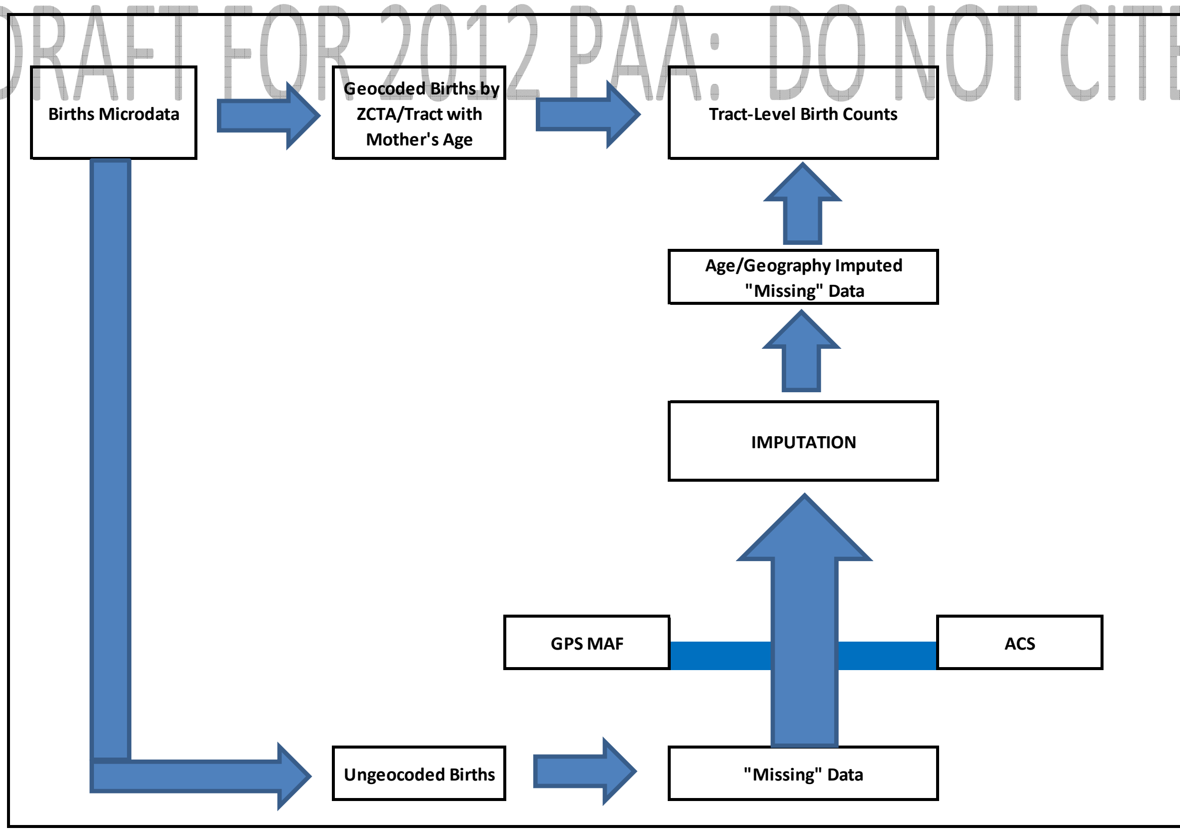

Figure 2. Spatial-Remediation Algorithm

Evaluation

No gold-standard measure of age-specific fertility exists at the census tract level. This research took a pragmatic approach to evaluation of the proposed method in light of an inability to directly compare estimated ASFRs to an observed "truth". Age-specific fertility rates form the basis of population renewal (Lotka, 1911; Sharpe, 1926; Shyrock and Siegel, 1980; Preston et. al., 2003; Keyfitz and Caswell, 2005). When the vector of age-specific fertility rates is an accurate estimate of the birth hazard by age, then the inner product of this vector and a vector of census counts of women in corresponding age-intervals should produce an accurate estimate of the number of children aged 0 to 1 for the corresponding year. Errors in estimates of the number of children aged 0 to 1 on April 1, 2010 based on multiplying counts of women by age against ASFRs should be proportional to the accuracy associated with the ASFR estimates. In this research, we compare observed Census 2010 counts of children aged 0 to 1 year with estimates made by multiplying Census 2010 counts of women aged 15-49 against the corresponding estimates of ASFR. These estimates and counts are used to estimate the algebraic errors (both numeric and percentage) associated with each set of ASFR estimates (remediated and non-remediated), to compute the root mean-squared error (RMSE), and estimate the marginal percentage-point improvement (the reduction in error as percentage points) associated with the spatially-explicit remediation of ASFR estimates described here.

Results

Table 1 provides a summary of the results in terms of algebraic accuracy measures (numeric and percentage), of variance associated with the estimates (Root Mean Squared Error), and the marginal improvements in estimation of 0 to 1 year olds in the 2010 Census associated with using spatially-

Table 1. Results: 0 to 1 Population Estimate Using Geocoded and Remediated ASFR Estimates

Measure	Geocoded		Remediated		Marginal Improvement Numeric	Marginal Improvement Percentage Points
	Numeric	Percentage	Numeric	Percentage		
Mean	-15	-25.15	-2	-3.95	13	21.20
Median	-14	-28.77	-4	-9.71	10	19.06
RMSE	25	4.07	21	4.06	5	0.01

remediated ASFR estimates. Utilization of unremediated geocoded birth counts to estimate ASFRs resulted in a significantly down-biased estimate of the number of 0 to 1 year olds. Since the root-mean squared error in both cases suggests important outlier cases are present in the data (the RMSE is numerically higher than either mean or median estimates of the average error in the set), we here emphasize the median error in the discussion. The Median Algebraic Percentage Error associated with the unremediated ASFR set was nearly 30.0 percent (Median=-28.77 percent). Numerically, this meant an underestimate of 14 infants on average at the census tract level. Spatially-remediated estimates resulted in errors over 3 times smaller, with a median just under 10 percent (Median= -9.71 percent). Both sets of estimates continue to produce down-biased estimates of 0 to 1 year olds and the variation associated with these errors (RMSE) is nearly identical. The average marginal improvement associated with the use of the spatially-remediated estimates, however, is nearly 20 percentage points (Median= 19.06). A further consideration of the mean errors associated with each set is informative. While the median and mean are similar for the unremediated set (suggesting that fewer significant outliers exist), the mean and median are quite different for the spatially-remediated ASFR-based estimates. On average, using the remediated ASFR estimates would result in an error of only -2 persons (-3.95 percent), suggesting that some influential outlying estimates (perhaps for tracts with less reliable ACS data or estimates of the amount of missing data) may be driving the higher average error associated with the medians.

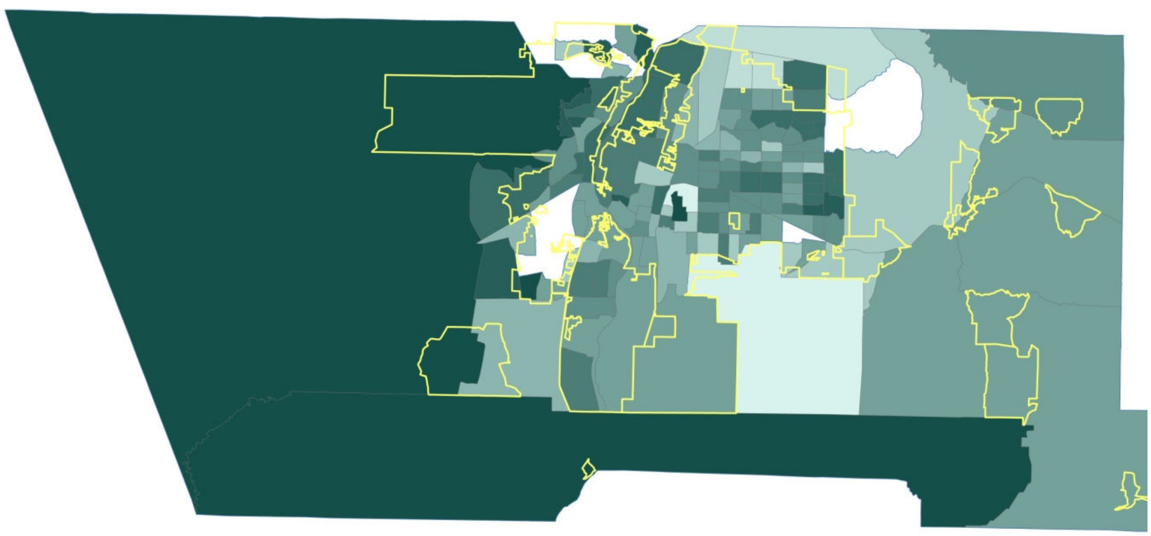

Figure 3. County-Wide Patterning in Marginal Improvements

Figure 3 illustrates a clearly-spatially-dependent patterning of estimate improvements associated with the procedure described in this paper. It also provides hints into the dynamics of demographic estimation that are worthy of further consideration. To the far left of this Figure, census tracts are characterized by dramatic improvements in the accuracy of estimates for the population aged 0 to 1. These areas are characterized by significant, recent development of housing with construction of new roads and conversion of existing street names. While zip-codes have remained constant in this area, the ability to geocode microdata to street names and positional intervals based on house number might easily be anticipated to be negatively impacted by these challenges (Baker et. al., 2012; Zandbergen, 2009). Other small-scale heterogeneities reveal surprises. Figure 4 illustrates one example--in the

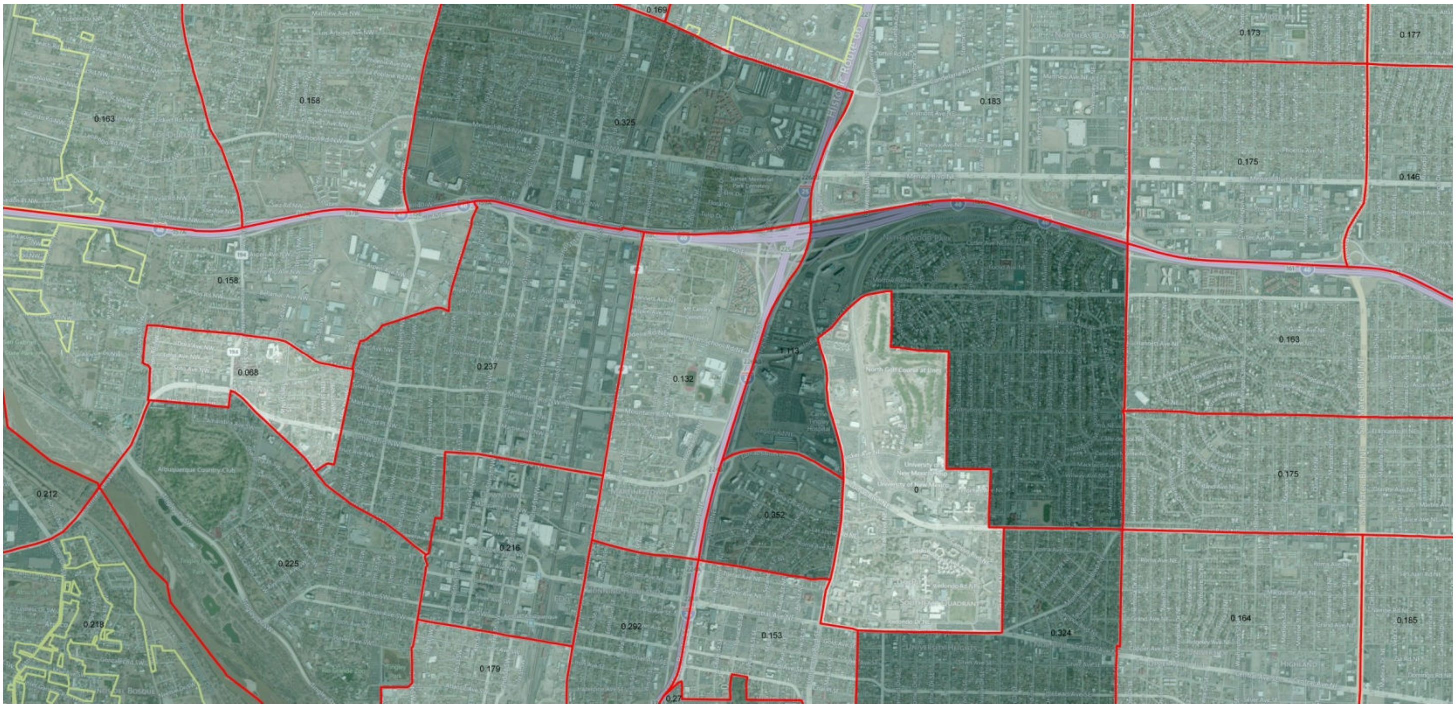

Figure 4. Missing Data and Marginal Improvements by Census Tract (UNM Circled)

vicinity of the University of New Mexico--where large-scale improvements were observed (ranging between 35 percent and 113 percent) by utilizing spatially-remediated ASFRs to estimate the 0 to 1 population. The presence of these improvements in areas characterized by older and much more stable street networks or housing-unit changes is surprising and raises questions about the sorts of characteristics that might be associated with marginal percentage-point improvements in these estimates. Ideally, a highly-effective set of estimates should not be associated with improvements for particular sub-groups, but with aspects of the geocoding process itself. An exploratory follow-up regression analysis was conducted to identify potential correlates of marginal improvement with the adjusted method, as well as those variables associated with algebraic percentage errors in both remediated and unadjusted (geocoded observations only) estimates. Tables 2-4 (next page) detail these results. Few predictors were statistically significantly associated with either marginal improvements or errors in either set of estimates. Marginal improvements were associated with the share of the county's housing units (beta=7.837, p= 0.0402) and with the magnitude of corrections made in births to women aged 20 to 24 (beta=0.06855, p=0.044) and 25 to 29 (beta=0.09059, p=.000). Only housing unit occupancy rates predicted algebraic errors for either the unremediated (beta=3.184, p=0.004) or remediated (beta=2.8313, p=0.001) estimates. Neither marginal improvements nor algebraic errors were associated with a youthful age-structure (proportion 0 to 18), the proportion of the population that is of Hispanic origin, or any of the other age-specific ASFR adjustments.

Discussion

Spatially-explicit remediation of geocoded births data can improve small-area estimates of age-specific fertility dramatically. Accurate estimates of age-specific fertility should produce accurate estimates of the 0 to 1 population for April 1, 2010 when multiplied against the 2010 population of women of child-

Table 2. Correlates of Marginal Improvements				
Predictor	Coeff	SE Coeff	T	P-value
Constant	*	*	*	*
Race/Ethnicity Structure				
Proportion 0 to 18	*	*	*	*
Proportion Hispanic	*	*	*	*
Housing Characteristics				
Share County HU	7.837	3.811	2.06	0.0402
Occupancy Rate	*	*	*	*
Age-Specific Corrections				
Correction 15 to 19	*	*	*	*
Correction 20 to 24	0.06855	0.03375	2.03	0.044
Correction 25 to 29	0.09059	0.02457	3.69	0.000
Correction 30 to 34	*	*	*	*
Correction 35 to 39	*	*	*	*
Correction 40 to 44	*	*	*	*
Correction 45 to 49	*	*	*	*

Table 3. Correlates of Algebraic Percent Error (Adjusted Estimates)				
Predictor	Coeff	SE Coeff	T	P-value
Constant	-3.121	1.311	-2.38	0.019
Race/Ethnicity Structure				
Proportion 0 to 18	*	*	*	*
Proportion Hispanic	*	*	*	*
Housing Characteristics				
Share County HU	*	*	*	*
Occupancy Rate	3.184	1.074	2.97	0.004
Age-Specific Corrections				
Correction 15 to 19	*	*	*	*
Correction 20 to 24	*	*	*	*
Correction 25 to 29	*	*	*	*
Correction 30 to 34	*	*	*	*
Correction 35 to 39	*	*	*	*
Correction 40 to 44	*	*	*	*
Correction 45 to 49	*	*	*	*

Table 4. Correlates of Algebraic Percent Error (Unadjusted Estimates)				
Predictor	Coeff	SE Coeff	T	P-value
Constant	-2.574	1.027	-2.51	0.013
Race/Ethnicity Structure				
Proportion 0 to 18	*	*	*	*
Proportion Hispanic	*	*	*	*
Housing Characteristics				
Share County HU	*	*	*	*
Occupancy Rate	2.8313	0.8411	3.37	0.001
Age-Specific Corrections				
Correction 15 to 19	*	*	*	*
Correction 20 to 24	*	*	*	*
Correction 25 to 29	*	*	*	*
Correction 30 to 34	*	*	*	*
Correction 35 to 39	*	*	*	*
Correction 40 to 44	*	*	*	*
Correction 45 to 49	*	*	*	*

bearing age. In the current study, estimates of the 0 to 1 year population are improved by over 19.0 percentage points. While this analysis is preliminary, these improvements are not associated with any demographic characteristics, but only with the share of housing units in a given census tract. Errors that remain in the adjusted estimates are only linked to occupancy rates and, again, are not predicted by any demographic characteristics examined in this research. These results suggests that errors associated with these estimates are linked to geocoding processes, and work well systematically across census tracts with differing ethnic structures , rates of housing vacancy, and any number of unexamined population and housing characteristics. Marginal improvements observed here appear to apply equally well across these heterogeneities and do not appear to penalize specific sub-groups (Cavanaugh, 1981).

In spite of the encouraging results reported here, there are a number of limitations to this study. First, comparing predicted births to the 0 to 1 population assumes zero migration has occurred. In reality, the dynamics of the 0 to 1 population will include stable residents, those who have moved (thus subtracting births that occurred within the tract), and those that have moved in from other census tracts. Though these facts place limitations on the results of this study, little accurate data is available on residency from the American Community Survey--or any other source--at the Census tract level. Although residence one-year ago is collected in the ACS, and permits an analysis of the stability of residency (for those who reside in the same house as one year ago--this, of course--averaged over a five-year period 2006-2010), it does not provide data on tract-to-tract migration flows that would be necessary to examine the effects of net-migration on these estimates. One possibility would be to discount the 2010 population, restricting the age-specific vector to only to the number of women estimated to be present April 1, 2010, *and to have been at the same residence for the entire prior year,* then similarly discount the 0 to 1 Census 2010 population and make comparisons based on these adjusted totals. To do so, however, would simply resolve to subtracting a constant proportion from both quantities--*resulting in estimates of accuracy that are precisely those already reported in these comparisons.* Little marginal utility in understanding the accuracy associated with estimates based on remediated ASFRs is to be gained from this and it is true that the bias in comparison would be associated with both unremediated and remediated ASFR estimates. As such, the *marginal improvements associated with the method would be unaffected by this bias in estimating accuracy: such bias is common to both sets of estimates.* In the end, biases in the absolute measures of accuracy reported here should not be anticipated to bias the estimates of marginal improvement. In spite of these limitations in the current study, it appears that remediated ASFR estimates represent dramatic improvements over the naive use of geocoded data.

Mis-reporting of zip-codes in microdata inputs should be expected to introduce subtle and largely unmeasurable biases into the results reported here. Here, zip-code specific raising factors were employed that relied upon the ratio of geocoded to reported addresses at the zip-code level. We are aware of no study that systematically reports the level of zip-code misreporting one might expect in an address-based dataset. Adjustment factors could be inflated in cases where mis-reporting increases the number of births self-reported to be within a zip-code above the true level (making the adjustment ratio higher than in reality) and could be down-biased in the alternative scenario of under-reporting. If we assume that most zip-code mis-reporting happens when the neighboring zip-code is incorrectly listed in a birth record, then in cases where zip-codes are larger than census tracts we should expect a greater level of error (Simpson, 2002) while in cases where the tracts are similarly-sized or larger than zip-codes we should expect smaller levels of bias (Fisher and Langford, 1995; Sadahiro, 2000). At present, with no additional information on the magnitude or spatial direction of such mis-reporting we are unable to account for this form of bias in evaluating our results.

In spite of the limitations reported here, the clear and compelling result of this study is that spatially-explicit remediation and imputation of geocoded births data is likely to substantially improve estimates of age-specific fertility for small geographic areas such as census tracts. Applied demographers and population geographers interested in applying the method will be primarily constrained by availability of data inputs for estimating missingness by geographic area, the necessity of utilizing simulation modeling procedures in using ACS data, and by availability of GIS expertise. Significant financial and time-investment was associated with construction of the GPS MAF, totaling over $1,000,000.00 in direct costs and several years of person-effort by the authors of this paper and a core set of junior staff and student research assistants sometimes totaling 10 or more persons. It is possible that estimation of missingness based on smaller datasets may be possible, but is unknown whether the use of multiple data sources, rather than treating building permits for example as a sample, improved the estimates of missingness made in this research. For those demographers and geographers interested in applying this method, significant improvements in estimation of age-specific fertility will come at the cost of intensive data processing and modeling effort.

References

Baker J. Estimating New Mexico Municipalities: The Devil is in the Details (of Data). New Mexico Business: Current Report. August. 2010.

Baker J, Alcantara A, Ruan XM (2011) A Stochastic Version of the Brass PF Ratio Adjustment of Age-Specific Fertility Schedules. PLoS One. 6(8):e23222. doi.10.1371/journalpone00232222.

Baker J, Alcantara A, Ruan XM, Watkins K (2012) The Impact of Incomplete Geocoding on Small Area Population Estimates. *Journal of Population Research*. 29:91-112.

Brass W (1964) Uses of Census and Survey Data for the Estimation of Vital Rates. United Nations, E/CN.14/CAS/7. November.

Brass W, Coale AJ, Demeny P, Heisel DF, Lorimer F, Romaniuk A, Van de Walle E (1968) The Demography of Tropical Africa. Princeton: Princeton University Press.

Cavanaugh F (1981) The Census Bureau's 1980 Census Test of Population Estimates. In: *Small Area Population Estimates--Methods and their Accuracy and New Metropolitan Area Definitions and Their Impact on the Private and Public Sector, Series GE-141, No. 7. Washington, DC: Government Planning Office.*

Drummond WJ (1995) Address Matching: GIS Technology for Mapping Human Activity Patterns. *Journal of the American Planning Association*. 61(2):240-251.

Efron B (1979) Bootstrap Methods: Another Look at the Jackknife. Annals of Statistics. 7(1):1-26.

Efron B (1981) Nonparametric Estimates of Standard Error: The Jackknife, the Boostrap, and Other Methods. Biometrika. 68(3):589-599.

Fishman GS (1986) Monte Carlo: Concepts, Algorithms, and Applications. New York: Springer.

Frotheringham AS, Brudson C, Charlton M (2002) *Geographically-Weighted Regression: The analysis of spatially-varying relationships*. West Sussex:Wiley.

Gilboa SM (2006) Comparison of Residential Geocoding Methods in a Population-Based Study of Air Quality and Birth Defects. *Environmental Research*. 101:256-262.

Haining R (2003) *Spatial Data Analysis: Theory and Practice*. New York: Cambridge.

Horvitz DG and Thompson DJ (1952) A Generalization of Sampling Without Replacement from a Finite Universe. *Journal of the American Statistical Association.* 47:663-685.

Jarosz B (2008) Using Assessor Parcel Data to Maintain Housing Unit Counts for Small Area Population Estimates. In: S Murdock and D Swanson (editors), *Applied Demography in the 21st Century*. (pp. 89-101). New York: Springer.

Kulkarni V (2011) *Introduction to Modeling and Analysis of Stochastic Systems*. 2nd edition. New York: Springer.

Lemieux C (2009) *Monte Carlo and Quasi Monte Carlo Sampling*. New York: Springer.

LeSage J and Pace KR (2004) *Journal of Real Estate Finance and Economics*. 29(2):233-254.

Little R and Rubin DB (1987) *Statistical Analysis with Missing Data*. New York: Wiley.

Little R and Schenker N (1994) Missing Data. In G. Arminger, CC. Clogg, and ME Sobel (Editors). *Handbook of Statistical Modeling in the Social and Behavioral Sciences*. (pp.39-75). New York: Plenum.

Mazumdar S, Rushton G, Smith B, Zimmerman D, and Donham K (2008) Geocoding Accuracy and Recovery of Relationships Between Environmental Exposures and Health. *International Journal of Health Geographics*. 2008:7:13. doi:10.1186/1476-072X-7-13.

Oliver MN (2005) Geographic Bias Related to Geocoding in Epidemiologic Studies. *International Journal of Health Geographics. 4(29):Online.*

Popoff C and Judson D (2004) Selected General Methods. In: JS Siegel and D Swanson (Editors). *The Methods and Materials of Demography*. 2nd Edition. (pp.644-675). New York: Springer.

Preston S, Hueverline P, and Guillot M (2003) *Demography: Measuring and Modeling Population Processes.* New York: McGraw-Hill.

Sadahiro Y (2000) Accuracy of Count Data Transferred Through the Areal Weighting Interpolation Method. *International Journal of Geographical Information Science.* 14:25-50.

Shyrock H and Siegel J (1980) The Methods and Materials of Demography. Volume II. Washington DC: US Bureau of the Census.

Simpson L (2002) Geography Conversion Tables: A Framework for Conversion of Data Between Geographical Units. *International Journal of Population Geography.*

Swanson D and Pol L (2005) Contemporary Developments in Applied Demography Within the United States. *Journal of Applied Sociology*. 21(2):26-56.

Voss P, Long DD, Hammer RB (1999) When Census Geography Doesn't Work: Using Ancillary Information to Improve the Spatial Interpolation of Demographic Data. Center for Demography and Ecology, University of Wisconsin, Madison. Working Paper NO. 99-26.

Voss P (2007) Demography as Spatial Social Science. *Population Policy and Research Review. 26:457-476.*

Zandbergen P (2009) Geocoding Quality and Implications for Spatial Analysis. *The Geography Compass.* 3(2):647-680.