Published in final edited form as:

Ann Epidemiol. 2016 March ; 26(3): 204–211. doi:10.1016/j.annepidem.2016.01.002.

Potential Selection Bias Associated With Using Geocoded Birth Records for Epidemiological Research

Sandie Ha¹, Hui Hu¹, Liang Mao², Dikea Roussos-Ross³, Jeffrey Roth⁴, and Xiaohui Xu⁵

¹Department of Epidemiology, College of Public Health and Health Professions, College of Medicine, University of Florida. Gainesville, Florida, USA

²Department of Geography, University of Florida, Gainesville, Florida, USA

³Department of Obstetrics and Gynecology, University of Florida, Gainesville, Florida, USA

⁴Department of Pediatrics, College of Medicine, University of Florida, Gainesville, Florida, USA

⁵Department of Epidemiology & Biostatistics, School of Public Health, Texas A&M Health Science Center

Abstract

Purpose—There is an increasing use of geocoded birth registry data in environmental epidemiology research. Ungeocoded records are routinely excluded.

Methods—We used classification and regression tree analysis (CART) and logistic regression to investigate potential selection bias associated with this exclusion among all singleton Florida births in 2009 (N=210,285).

Results—The rate of unsuccessful geocoding was 11.5% (n=24,171). This ranged between 0% to 100% across zip codes. Living in a rural zip code was the strongest predictor of being ungeocoded. Other predictors for geocoding status varied with urbanity status. In urban areas, maternal race [adjusted odds ratio (aOR) ranging between 1.08 for Hispanic to 1.18 for Black compared to White], maternal age [aOR: 1.16 (1.10-1.23) for ages 20-34 compared to <20], maternal nativity [aOR: 1.20 (1.15, 1.25) for Non-US vs. US born], delivery at a birth center [aOR: 1.72(1.49, 2.00 compared to hospital delivery)], multiparity [aOR: 0.91 (0.88, 0.94)], maternal smoking [aOR: 0.82 (0.76-0.88)] and having non-private insurance [aOR: 1.25 (1.20-1.30) for Medicaid vs. private insurance] were significantly associated with being ungeocoded. In rural areas, births delivered at birth center [aOR: 2.91(1.80-4.73)] or home [aOR: 1.94(1.28-2.95)] had increased odds compared to hospital births. The characteristics predictive of being ungeocoded were also significantly associated with adverse birth outcomes such as low birthweight and preterm delivery, and the association for maternal age was different when ungeocoded births were included and excluded.

Address correspondence to: Xiaohui Xu, PhD, MS | Associate Professor, Department of Epidemiology & Biostatistics, School of Public Health | Texas A&M Health Science Center, 205A SRPH Administration Building | MS 1266, 212 Adriance Lab Road, College Station, Texas 77843-1266, phone 979-436-9500 | fax 979-458-1877, xiaohui.xu@sph.tamhsc.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Conclusions—Geocoding status is not random. Women with certain exposure-outcome characteristics may be more likely to be ungeocoded and excluded, indicating potential selection bias.

Keywords

birth certificates; selection bias; geocode; environmental epidemiology

Birth certificates in state vital statistics systems are widely used in epidemiological research (1-4). The transition from paper-based to digital format has tremendously improved the timeliness and quality of birth records (5). By applying geographic information system (GIS) software to the electronic birth certificate, many state vital statistics programs are able to document maternal addresses at delivery through an automated geocoding technique. The software assigns geographic coordinates to an address based on spatial reference data, such as digital street maps. This automated technique has provided more geo-referenced information for birth data and has enabled sophisticated spatial analyses in epidemiological research, especially in environmental epidemiology (6). As a result, there has been an increasing application of geocoded birth registry in studies of environmental risk factors for adverse birth outcomes (7-11).

The limitations of geocoded addresses including positional accuracy have been well studied in epidemiological research (12-15). As geocoding technique is a process through probability matching between addresses and spatial reference data, a proportion of unmatched records remains as a major problem. For example, the lack of address standardization, misspelling of street addresses, and the limited quality of spatial reference map (e.g. no updated street information) create conditions for misclassification or missing information. In many epidemiological studies, addresses that fail to geocode have to be excluded from studies due to missing geographic information (8, 16-18). A study by Zimmerman et al (19) showed that approximately 10-30% of records, even higher in some subgroups, would have to be excluded if only geocoded records were considered. This exclusion not only reduces the study sample size, but possibly introduces issues related to generalizability and selection bias.

Generalizability is the extent to which the results in a given study pertain to a broader population. If ungeocoded and geocoded populations are significantly different, then results may not be generalizable from one to the other. In addition, selection bias indicates a situation where the result in a study is not valid because of different sampling probabilities related to exposure-outcome cells. For example, if individuals with the exposure and outcome are less likely to be sampled than other cross-classified cells in a 2×2 table, then the measure of association will be biased towards the null. Without consideration of these issues, any epidemiological study that is solely based on geocoded information may end up with unreliable conclusions. To the best of our knowledge, issues related to generalizability and potential selection bias arising from differential ungeocoding have received little attention in epidemiological research.

In this retrospective cohort study, we employed classification and regression tree (CART) analysis and logistic regression to explore generalizability related to exclusion of

ungeo-coded births by examining whether there were significant differences between geo-coded and ungeo-coded birth records. We further assessed whether there is potential selection bias by a direct analysis that involved determining a) whether births with certain exposures (e.g. characteristics) were more likely to be ungeo-coded (or excluded); b) whether these exposures were associated with important adverse birth outcomes; and c) whether the exposure-birth outcome relationships among the geo-coded population and the entire population are different.

METHODS

All singleton births in Florida in 2009 (N=210,285) were identified from the Florida Birth Vital Statistics (BVS). Singleton births were chosen to avoid duplication of addresses. Latitude and longitude for all maternal addresses at delivery were provided by the state vital statistics program. These births were independently geo-coded at street address level using North American Locator in ArcGIS 9.3 (ESRI, Redlands, California, USA) by the Florida Department of Health (FDoH) Vital Statistics. Spelling sensitivity was 80, and minimum matching score was 90 on the first round. Addresses that were not matched on the first round were screened and edited for spelling and random character issues, and were re-matched using the same criteria. After this round, addresses that were assigned latitude and longitude were defined as geo-coded; the remaining addresses were defined as ungeo-coded. We used only geo-codes provided by the FDoH, which were based on street address for several reasons. First, a majority of those who weren't geo-coded based on home address during delivery in medical records often had missing address. Therefore, we could not obtain other geographic information for a different method of geo-coding. Second, for those with only zip-codes available, they may be systematically different from those with full address available. Therefore, geo-coding births using two methods may introduce some information bias. Third, although using zip codes for all births may increase matching rates, this introduces another major issue involving positional accuracy. Specifically geo-coding to zip code centroid can improve the match rate but this information in some studies may not be very useful if exact location of an address is required (e.g. distance to highway calculation).

Characteristics such as demographics, behavioral factors, and adverse birth outcomes were used as potential predictors of geo-coding status. For demographic factors, we assessed infant sex, maternal race, maternal age, maternal education, parental marital status, parity, maternal nativity, birth facility, and private vs. public medical insurance as a proxy for socioeconomic status. For behavioral factors, we assessed tobacco and alcohol use during pregnancy, adequacy of prenatal care assessed by Kotelchuck index, and pre-pregnancy body mass index (BMI). As markers of adverse birth outcomes we included low birth weight (LBW) and preterm delivery (PTD). LBW was defined as birth that was born less than 2,500 grams. PTD was defined as a birth that occurred before 37 weeks of gestation. We determined the proportion of each ZIP Code Tabulation Area (ZCTA) that falls within the urban areas defined by the 2010 US census (20). We further defined urbanity of each ZIP code based on the following cutoff proportion: rural: <5%, urban: 5%. We selected the cutoff of 5% because this proportion indicates the probability of the address located in the urban area within the specific ZCTA, and 5% is commonly used as a cutoff to indicate small probability events.

To examine the differences in the characteristics of geocoded and ungeocoded participants, we used CART. The details of this method have been previously described (21, 22). Briefly, CART is a non-parametric regression method that sequentially splits the data into dichotomous groups, such that each resulting group contains increasingly similar responses for the outcome. The end product of a typical CART analysis is a tree diagram illustrating the paths of dichotomous splits. Every tree starts with a root node, which contains all data from which the tree will be generated. Next, the data is split into two child nodes based on the values of an independent variable in a way that the observations within the two groups have the most similar responses for the outcome (i.e. minimizing residual sums of squares). The resulting child nodes contain a subset of the observations and are further split in the same manner until a pre-set stopping point is reached, in this analysis a p-value greater than 0.05 was set as statistical significant. The smallest resulting nodes are called terminal nodes. For each terminal node, the CART gives an estimation of the conditional probability of observations in each node having the given outcome (in this study, being ungeocoded). The CART offers several advantages. First, it makes no assumption about monotonic or parametric relationship between predictors and outcomes. Second, it can identify complex interactions among predictors without *a priori* specification. It also provides results that are easy to interpret. CART analyses were performed using the PARTY package in R.

We also used univariate and multivariable logistic regression to determine the odds ratios (OR) and 95% confidence intervals (CI) for the association between selected characteristics and geocoding status, and whether these differences persist after typical adjustment that is common in studies. We stratified our analyses by urbanity status due to the strong evidence of interaction between this variable and other predictors from the CART analyses. We also used logistic regression to determine the association between exposures predictive of geocoding status and common adverse birth outcomes including LBW and PTD. We repeated these analyses for both the geocoded group and the entire study sample. Logistic regression models were performed using SAS 9.4 (Cary, NC).

RESULTS

Table 1 describes study participants' characteristics by geocoding status. During the study period, 11.5% of the study population was ungeocoded. This prevalence varied from 0% to 100% across different zip codes, with the highest rates located in rural areas (Figure 1). It is important to notice that some ungeocoded records may have had valid zip codes, but were PO boxes or for general delivery only and therefore were unable to geocode. Overall, there were slight differences between geocoded and ungeocoded births. Compared to the geocoded group, there was a lower percentage of White participants (40.7% vs. 44.3%), and higher percentage of Black (20.2% vs. 18.6%) and Hispanic participants (34.1% vs. 31.9%) among the ungeocoded group (Table 1). The percentages of women with lower education, unmarried, on Medicaid, or were born outside of the US were also slightly higher in the ungeocoded group. The biggest and most significant difference between the two groups was in the percent of women in urban zip codes who were in the geocoded group compared to those who were in the ungeocoded group (94.1% vs 85.1%).

The conditional probabilities of being ungeocoded for participants with specific characteristics are in Figure 2. The numbers in the terminal node (gray boxes) represent the number of participants in each node (N) and the conditional probability of participants in that node being ungeocoded (Y). The boxes with bold font letters are significant predictors of being ungeocoded. According to the CART tree diagram, urbanity was the most important predictor of geocoding status, thus appearing on the top of the tree. Among those in the rural areas (first splitting on the right), no predictor was significant based on minimized sums of squares. However, in urban areas, there were more predictors, indicating potential interaction between rural/urban status and the covariates. For example, among the 20,829 non-smoking women with insurance other than Medicaid or private insurance, the prevalence of being ungeocoded was 12.9% (Figure 2). This was the group with the highest prevalence of being ungeocoded among women living in urban ZCTAs. Overall, the CART also indicated that insurance, maternal age, maternal race, marital status, and smoking during pregnancy were important predictors of being ungeocoded in urban areas (Figure 2).

Table 2 presents the unadjusted and adjusted odds ratios (OR) and 95% confidence intervals (95% CI) for the associations between participant characteristics and geocoding status by urbanity status. We stratified the analyses by urbanity status due to strong evidence of interaction between this variable and other covariates in the CART analysis. The logistic regression results were generally consistent with the CART analysis. Specifically, among the urban population, women who were non-White, Non-US born, had non-private insurance, had delivery at a birth center, or ages between 20 and 34 had increased odds of being ungeocoded compared to their counterparts. Women living in rural ZCTAs were less likely to be successfully geocoded. Different from the CART model, which showed no association between geocoding status and any covariates, the logistic regression model showed that in rural ZCTAs, giving birth in a birth center or home was significantly associated with increased odds of being ungeocoded compared to hospital.

We further assessed whether the exposures predictive of geocoding status were also associated with LBW and PTD, two common adverse birth outcomes. The results are presented in Table 3. Overall, all characteristics predictive geocoding status, especially in the urban areas, were strongly associated with LBW and PTD. When analyzed with and without the ungeocoded population separately, the associations were similar for both populations except for that between PTD and maternal age. A significantly higher odds of PTD (OR: 1.08, 95% CI: 1.02, 1.15) were observed among women aged 20-34 years compared to <20 in the geocoded only analyses. However, when analyzing the entire sample, this association was no longer significant (OR: 1.06, 95% CI: 0.99, 1.12).

DISCUSSION

Geocoded birth data with geographical information provide a valuable resource for research. However, these data have certain limitations with respect their positional accuracy as well as the non-random distribution of missing geographic information. Our study found that the proportion of ungeocoded births in 2009 among singleton Florida birth records was 11.5%, which was a considerable portion of the study population. More importantly, large spatial heterogeneities were observed as this proportion ranged from 0% to 100% across zip codes.

We sought to determine whether there were issues related with generalizability and potential selection bias associated with inclusion of only geocoded births by showing whether a) births with certain exposures (e.g. characteristics) were more likely to be ungeocoded and subsequently excluded; b) these exposures were associated with LBW and PTD; and c) the exposure-birth outcome relationships between the geocoded population and the entire population are different.

Our analyses showed that there were significant differences in the characteristics of ungeocoded and geocoded women, even after typical adjustment that is common in other studies, suggesting that this would remain beyond usual modeling. These findings suggested limited generalizability in studies that excluded ungeocoded births. Specifically, maternal addresses in the Florida birth data located in rural areas were more likely to be ungeocoded. Moreover, the predictors for being ungeocoded were different between women who lived in rural or urban areas.

Furthermore, it also appears that the exclusion of ungeocoded births could also potentially induce a systematic selection bias. First, as previously discussed, the participants appeared to be ungeocoded (and excluded) dependent on certain characteristics including urbanity status, nativity, parity, birth facility, insurance status, or maternal race. Moreover, our results also suggested that exclusion may also be outcome-dependent. Specifically, the exposures predictive of being geocoded were also significantly associated with LBW and PTD. When we determined the relationships between these exposures and the outcomes LBW and PTD these associations were different for maternal age and PTD when we only included geocoded population compared to the entire population. Together, these findings may suggest that by including only geocoded births, studies may inevitably select births with different probability for different exposure-outcome cells. More specifically, they are more likely to select births with certain characteristics (e.g., urban), who also are less likely to have adverse birth outcomes. Although the associations between adverse birth outcomes and most of the characteristics included in our analyses were similar between the entire population and the geocoded population, we did find significant differences in the associations between maternal age and PTD. Since our analyses were based on only a few characteristics, there may be other exposures (e.g. environmental) or characteristics that may have different associations when the ungeocoded population is excluded.

The results from our CART analysis and traditional logistic regression analysis, both of which aimed to determine differences in characteristics between geocoded and ungeocoded births, were also consistent. Specifically, in the CART analysis there was evidence that the predictors of geocoding status were different from those from rural vs. urban areas. This suggests interaction between rural/urban status and participant characteristics on the prevalence of being ungeocoded, which was confirmed in the logistic regression analyses.

Although the exposures considered in this direct analysis were not environmental factors, previous studies have found that they are directly associated with many environmental exposures. For example, many studies have found that minority groups tend to be exposed to higher levels of environmental hazards (23, 24). In addition, women from urban and rural areas may have very different environmental exposures (25, 26). These differences may

imply that exclusion of women with certain characteristics from an epidemiological study may significantly affect the results, especially when they have also been found to have higher risk of adverse health outcomes as previously discussed. Exclusion of participants with potentially higher exposures and poorer outcomes may underestimate the associations between environmental exposures and birth outcomes.

The successful geocoding rate of 88.5% is comparable to other studies that involve geocoding birth certificate data (27, 28). However, a significantly lower proportion of successful geocoded addresses among the rural population was not optimal. The low successful geocoding rate in rural areas may be explained by a two main reasons. First, women living in rural areas in Florida may be more likely to report postal office box (PO Box) or general delivery only addresses, which are often not in the reference map file. Secondly, rural addresses often use rural route, which are also often not referenced in the reference file. To our knowledge, few studies have been conducted to compare matching rates or positional errors between geocoding methods (13). None have explored potential selection bias associated with unsuccessful geocoding. One study on residential geocoding methods and the effects of air quality on birth defects found that geocoding status modified the association between ethnicity and birth defect (30). The study also found that geocoding status was independently associated with air pollution exposures. In addition, rural addresses are more difficult to locate due to the use of rural roads and post office boxes (15, 29). This pattern is consistent with our study which found that the majority of ungeocoded addresses were in rural areas.

Because of potential selection bias resulting from excluding ungeocoded subjects, future studies will need to try and mitigate this threat to validity by several methods. Perhaps the method most currently available would be to use the population centroid of a zip code or census tract as a substitute for those who fail to geocode. This method can be more effective in increasing matching rate. However, it is important to notice potential positional errors when using zip code centroid to approximate the address of the ungeocoded individuals. If the geocodes for the successfully geocoded individuals are based on street address, and the unsuccessfully geocoded individuals are based on zip code, this may introduce differential errors in positional accuracy. This method is more appropriate when the geographic/environmental factors we want to link are provided on zip-code level. Therefore, although using zip code centroids can substantially reduce the number of ungeocoded individuals to address the selection bias, it can also increase the positional errors and cause differential measurement bias (e.g. there is a tradeoff between successful geocoding rate and positional accuracy using a fixed method). Another possible method is development of innovative and reliable geo-statistical models to predict locations of ungeocoded participants, based on their known characteristics, a strategy similar to propensity score matching.

Since failure to geocode may partly be the result of errors associated with reporting addresses (e.g. spelling, issues in recalling, data entry errors, etc.), future studies may also benefit from smart phones geo-locationing applications or other GIS-based mobile devices to pinpoint location during pregnancy and delivery. These devices would allow accurate collection of spatial locations without the need for geocoding addresses. Recording of location would reduce the potential bias associated with unsuccessful geocoding of

residence at time of birth. At a minimum, studies should adjust for variables predictive of sampling (e.g. urbanity, marital status, race). A similar method would involve changes during the vital registration process. For example, vital registration processes can be changed to require that birth or death certificate should not be accepted for filing unless the residential addresses provided can be geocoded.

This study has several strengths and limitations. First, we believe this is the first study to investigate the association between demographic, behavioral and health factors with geocoding status. Second, our study population includes all singleton births in the state of Florida in 2009. This large sample supports the generalizability of our findings. Third, the CART analyses strengthened our regression models by accounting for the potential complex interactions between predictors. Despite strengths, the study has several limitations. First, we were able to examine only factors that exist in the Florida birth certificate. Other factors important for environmental studies such as occupation and physical activities, which were not available in the birth record could not be analyzed. Second, information available on birth certificates is not always accurate (31, 32). However, existing studies have demonstrated that demographic and birth outcome information on birth certificate data are generally reliable (32, 33). Nevertheless, information on lifestyle choices may not be quite as reliable, which may have affected the results of our study. Moreover, the parameter settings of automated geocoding may also cause selection bias, which was not taken into account in this study. We could not test geocoding method stringency because we were provided with data independently geocoded. Lastly, due to the unavailability of accurate geographical information for the ungeocoded population, we were unable to specifically direct our analysis towards an actual environmental exposure. However, as previously described, our exposures have been found to be correlated with environmental exposures in previous literature.

CONCLUSIONS

Studies that rely on successful geocoding of birth certificate data may suffer from potential selection bias and limited generalizability. Such studies may systematically exclude participants based on both exposure and outcomes status. Excluded individuals are more likely to live in rural areas, and are lower socioeconomic status, or ethnic minorities. These subgroups are also at higher risk of poor birth outcomes. They may also have higher environmental exposures thus their exclusion may underestimate the associations between environmental risk factors and adverse birth outcomes. Future efforts need to pay particular attention to the inclusion of rural areas and more disadvantaged groups in study populations, and new approaches need to be explored to assess the potential selection bias associated with excluding subjects who fail to geocode. In addition, there is also a need for improvements in address reporting/geocoding practices in the field since geocoding quality is limited by current post-hoc geocoding methods, which are the only options available to simultaneously reduce selection bias and improve positional accuracy.

ACKNOWLEDGEMENTS AND FUNDING

This work was supported by Grant Number K01ES019177 from the National Institute of Environmental Health Sciences (Xu) and the University of Florida Graduate School Fellowship (Ha). The authors also wish to thank the

Florida Department of Health Office of Vital Statistics for allowing us to study birth certificate data. The conclusions of this study are solely the responsibility of the authors and do not necessarily represent the official views of the funding agency or the data provider.

List of abbreviations

BMI	Body mass index
CART	Classification and Regression Tree
LBW	low birth weight
PTD	preterm delivery
OR	odds ratio
CI	confidence interval

References

1. Kachoria R, Oza-Frank R. Trends in Breastfeeding Initiation in the NICU by Gestational Age in Ohio, 2006-2012. *Birth*. 2015 Epub 2015/01/17.
2. Fulda KG, Kurian AK, Balyakina E, Moerbe MM. Paternal race/ethnicity and very low birth weight. *BMC pregnancy and childbirth*. 2014; 14(1):385. Epub 2014/11/20. [PubMed: 25406725]
3. Hamilton BE, Hoyert DL, Martin JA, Strobino DM, Guyer B. Annual summary of vital statistics: 2010-2011. *Pediatrics*. 2013; 131(3):548-58. Epub 2013/02/13. [PubMed: 23400611]
4. Brauer M, Lencar C, Tamburic L, Koehoorn M, Demers P, Karr C. A cohort study of traffic-related air pollution impacts on birth outcomes. *Environmental health perspectives*. 2008; 116(5):680-6. Epub 2008/05/13. [PubMed: 18470315]
5. Starr P, Starr S. Reinventing Vital Statistics: The Impact of Changes in Information Technology, Welfare Policy, and Health Care. *Public Health Reports*. 1995; 110:534-44. [PubMed: 7480607]
6. Nuckols JR, Ward MH, Jarup L. Using geographic information systems for exposure assessment in environmental epidemiology studies. *Environmental health perspectives*. 2004; 112(9):1007-15. Epub 2004/06/17. [PubMed: 15198921]
7. Dadvand P, Ostro B, Figueras F, Foraster M, Basagana X, Valentin A, et al. Residential proximity to major roads and term low birth weight: the roles of air pollution, heat, noise, and road-adjacent trees. *Epidemiology*. 2014; 25(4):518-25. Epub 2014/05/03. [PubMed: 24787556]
8. Vinikoor-Imler LC, Davis JA, Meyer RE, Messer LC, Luben TJ. Associations between prenatal exposure to air pollution, small for gestational age, and term low birthweight in a state-wide birth cohort. *Environmental research*. 2014; 132:132-9. Epub 2014/04/29. [PubMed: 24769562]
9. Hannam K, McNamee R, Baker P, Sibley C, Agius R. Air pollution exposure and adverse pregnancy outcomes in a large UK birth cohort: use of a novel spatio-temporal modelling technique. *Scandinavian journal of work, environment & health*. 2014; 40(5):518-30. Epub 2014/03/22.
10. Padula AM, Tager IB, Carmichael SL, Hammond SK, Yang W, Lurmann FW, et al. Traffic-related air pollution and selected birth defects in the San Joaquin Valley of California. *Birth defects research Part A, Clinical and molecular teratology*. 2013; 97(11):730-5. Epub 2013/10/11.
11. Pereira G, Belanger K, Ebisu K, Bell ML. Fine particulate matter and risk of preterm birth in Connecticut in 2000-2006: a longitudinal study. *American journal of epidemiology*. 2014; 179(1): 67-74. Epub 2013/09/27. [PubMed: 24068199]
12. Krieger N, Waterman P, Lemieux K, Zierler S, Hogan JW. On the wrong side of the tracts? Evaluating the accuracy of geocoding in public health research. *American journal of public health*. 2001; 91(7):1114-6. Epub 2001/07/10. [PubMed: 11441740]
13. Zhan FB, Brender JD, De Lima I, Suarez L, Langlois PH. Match rate and positional accuracy of two geocoding methods for epidemiologic research. *Annals of epidemiology*. 2006; 16(11):842-9. Epub 2006/10/10. [PubMed: 17027286]

14. Whitsel EA, Rose KM, Wood JL, Henley AC, Liao D, Heiss G. Accuracy and repeatability of commercial geocoding. *American journal of epidemiology*. 2004; 160(10):1023–9. Epub 2004/11/04. [PubMed: 15522859]
15. Bonner MR, Han D, Nie J, Rogerson P, Vena JE, Freudenheim JL. Positional accuracy of geocoded addresses in epidemiologic research. *Epidemiology*. 2003; 14(4):408–12. Epub 2003/07/05. [PubMed: 12843763]
16. Ha S, Hu H, Roussos-Ross D, Haidong K, Roth J, Xu X. The effects of air pollution on adverse birth outcomes. *Environmental research*. 2014; 134:198–204. Epub 2014/09/01. [PubMed: 25173052]
17. Porter TR, Kent ST, Su W, Beck HM, Gohlke JM. Spatiotemporal association between birth outcomes and coke production and steel making facilities in Alabama, USA: a cross-sectional study. *Environmental health : a global access science source*. 2014; 13:85. Epub 2014/10/25. [PubMed: 25342170]
18. Gray SC, Edwards SE, Schultz BD, Miranda ML. Assessing the impact of race, social factors and air pollution on birth outcomes: a population-based study. *Environmental health : a global access science source*. 2014; 13(1):4. Epub 2014/01/31. [PubMed: 24476365]
19. Zimmerman DL. Estimating the intensity of a spatial point process from locations coarsened by incomplete geocoding. *Biometrics*. 2008; 64(1):262–70. Epub 2007/08/08. [PubMed: 17680833]
20. Census, U. 2010 Census Urban and Rural Classification and Urban Area Criteria. US Census. 2010. [cited 2014 February 01]; Available from: <https://www.census.gov/geo/reference/ua/urban-rural-2010.html>
21. Loh WYSY. Split selection methods for classification trees. *Statistica Sinica*. 1999; 7:815–40.
22. Strobl C, Malley J, Tutz G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological methods*. 2009; 14(4):323–48. Epub 2009/12/09. [PubMed: 19968396]
23. Gray SC, Edwards SE, Miranda ML. Race, socioeconomic status, and air pollution exposure in North Carolina. *Environmental research*. 2013; 126:152–8. Epub 2013/07/16. [PubMed: 23850144]
24. Jones MR, Diez-Roux AV, Hajat A, Kershaw KN, O'Neill MS, Guallar E, et al. Race/ethnicity, residential segregation, and exposure to ambient air pollution: the Multi-Ethnic Study of Atherosclerosis (MESA). *American journal of public health*. 2014; 104(11):2130–7. Epub 2014/09/12. [PubMed: 25211756]
25. Kundu S, Stone EA. Composition and sources of fine particulate matter across urban and rural sites in the Midwestern United States. *Environmental science Processes & impacts*. 2014; 16(6):1360–70. Epub 2014/04/17. [PubMed: 24736797]
26. Hendryx M, Fedorko E, Halverson J. Pollution sources and mortality rates across rural-urban areas in the United States. *The Journal of rural health : official journal of the American Rural Health Association and the National Rural Health Care Association*. 2010; 26(4):383–91. Epub 2010/10/30.
27. Edwards SE, Strauss B, Miranda ML. Geocoding large population-level administrative datasets at highly resolved spatial scales. *Transactions in GIS : TG*. 2014; 18(4):586–603. Epub 2014/11/11. [PubMed: 25383017]
28. Wang Y, Le LH, Wang X, Tao Z, Druschel CD, Cross PK, et al. Development of web-based geocoding applications for the population-based Birth Defects Surveillance System in New York state. *Journal of registry management*. 2010; 37(1):16–21. Epub 2010/08/28. [PubMed: 20795565]
29. Vine MF, Degnan D, Hanchette C. Geographic information systems: their use in environmental epidemiologic research. *Environmental health perspectives*. 1997; 105(6):598–605. Epub 1997/06/01. [PubMed: 9288494]
30. Gilboa SM, Mendola P, Olshan AF, Harness C, Loomis D, Langlois PH, et al. Comparison of residential geocoding methods in population-based study of air quality and birth defects. *Environmental research*. 2006; 101(2):256–62. Epub 2006/02/18. [PubMed: 16483563]
31. Northam S, Knapp TR. The reliability and validity of birth certificates. *Journal of obstetric, gynecologic, and neonatal nursing : JOGNN / NAACOG*. 2006; 35(1):3–12. Epub 2006/02/10.

32. Zollinger TW, Przybylski MJ, Gamache RE. Reliability of Indiana birth certificate data compared to medical records. *Annals of epidemiology*. 2006; 16(1):1–10. Epub 2005/07/26. [PubMed: 16039875]
33. Vinikoor LC, Messer LC, Laraia BA, Kaufman JS. Reliability of variables on the North Carolina birth certificate: a comparison with directly queried values from a cohort study. *Paediatric and perinatal epidemiology*. 2010; 24(1):102–12. Epub 2010/01/19. [PubMed: 20078836]

Figure 1.
Geographical Distribution of The Prevalence of Ungeocoded Births in Florida Vital Statistics, 2009.

Figure 2.
Classification Tree of the Association Between Selected Participant Characteristics and Geocoding status Among Florida Singleton Birth, 2009. Legend: Gray boxes are terminal nodes with N representing sample size and Y representing the conditional probability of being ungeocoded.

Table 1

Selected Characteristics of Singleton Florida Births by Geocoding Status in 2009 (n= 210,285).

Selected Characteristics	Geo coded		Ungeo coded	
	N	%	N	%
Total	186,114	88.5	24,171	11.5
Maternal race				
White	82,461	44.3	9,835	40.7
Black	34,525	18.6	4,871	20.2
Hispanic	59,352	31.9	8,235	34.1
Asian/PI	5,434	2.9	698	2.9
Other	3,145	1.7	408	1.7
Unknown	1,197	0.6	124	0.5
Maternal age				
<20	18,895	10.2	2,456	10.2
20-34	139,745	75.1	18,970	78.5
35	27,472	14.8	2,745	11.4
Unknown	2	0	0	0
Maternal education				
<High school	34,245	18.4	4,807	19.9
High School	93,157	50.1	12,649	52.3
College or more	57,844	31.1	6,588	27.3
Unknown	868	0.5	127	0.5
Infant sex				
Female	90,995	48.9	11,815	48.9
Male	95,114	51.1	12,354	51.1
Unknown	5	0	2	0
Marital status				
Married	98,190	52.8	11,973	49.5
Unmarried	87,918	47.2	12,196	50.5
Unknown	6	0.0	2	0.0
Urbanity				
Urban	175,158	94.1	20,575	85.1
Rural	10,956	5.9	3,596	14.9
Smoking				
Yes	12,112	6.5	1,462	6.1
No	173,861	93.4	22,369	92.5
Unknown	141	0.1	14	0.1
Kotelchuck Index				
Inadequate	46,800	25.2	6,253	25.9
Adequate	117,433	63.1	14,468	59.9

Selected Characteristics	Geo coded		Ungeo coded	
	N	%	N	%
Unknown	21,881	11.8	3,450	14.3
Pre-pregnancy BMI				
Underweight	8,792	4.7	1,191	4.9
Normal weight	87,284	46.9	11,026	45.6
Overweight or obese	78,446	42.2	10,443	43.2
Unknown	11,592	6.2	1,511	6.3
Insurance source				
Medicaid	87,333	46.9	12,463	51.6
Private	76,279	41.0	8,234	34.1
Other	21,953	11.8	3,403	14.1
Unknown	549	0.3	71	0.3
Birthing facility				
Hospital	183,879	98.8	23,762	98.3
Birth Center	1,123	0.6	242	1.0
Home	990	0.5	149	0.6
Other	122	0.1	18	0.1
Parity				
0	78,738	42.3	10,472	43.3
1	105,626	56.8	13,607	56.3
Unknown	1,750	0.9	92	0.4
Nativity				
US	128,049	68.8	15,836	65.5
Non-US	58,065	31.2	8,335	34.5
LBW				
Yes	12,779	6.9	1,682	7.0
No	173,320	93.1	22,488	93.0
Unknown	15	0	1	0
PTD				
Yes	16,107	8.7	2,127	8.8
No	169,844	91.3	22,014	91.1
Unknown	163	0.1	30	0.1

Associations Between Selected Characteristics and Geocoding Status by Urbanity Among Singleton Florida Birth in 2009.

Table 2

Characteristics	Urban				Rural			
	Unadjusted		Adjusted ^a		Unadjusted		Adjusted ^a	
	OR	95% CI	OR	95% CI	OR	95% CI	OR	95% CI
Maternal race								
White	1.00 (reference)		1.00 (reference)		1.00 (reference)			
Black	1.29	1.24, 1.34	1.18	1.13, 1.23	1.10	0.99, 1.22		
Hispanic	1.30	1.25, 1.34	1.08	1.03, 1.13	0.96	0.88, 1.06		
Asian/PI	1.24	1.14, 1.35	1.08	0.99, 1.18	0.91	0.58, 1.42		
Other	1.20	1.08, 1.35	1.08	0.96, 1.21	0.78	0.57, 1.07		
Maternal age								
<20	1.00 (reference)		1.00 (reference)		1.00 (reference)		1.00 (reference)	
20-34	1.09	1.04, 1.15	1.16	1.10, 1.23	1.11	1.00, 1.24	1.11	0.99, 1.23
35	0.82	0.77, 0.88	0.91	0.85, 0.98	0.93	0.79, 1.09	0.92	0.78, 1.08
Maternal education								
<High school	1.13	1.08, 1.18	0.92	0.88, 0.97	1.00	0.89, 1.12		
High School	1.15	1.12, 1.19	1.02	0.98, 1.06	0.98	0.88, 1.08		
College or more	1.00 (reference)		1.00 (reference)		1.00 (reference)			
Infant sex								
Female	1.00 (reference)				1.00 (reference)			
Male	1.00	0.97, 1.03			1.02	0.95, 1.10		
Marital status								
Married	1.00 (reference)		1.00 (reference)		1.00 (reference)			
Unmarried	1.13	1.10, 1.17	1.00	0.96, 1.04	1.00	0.92, 1.07		
Maternal smoking								
No	1.00 (reference)		1.00 (reference)		1.00 (reference)			
Yes	0.79	0.74, 0.84	0.82 (0.76, 0.88)		0.98	0.88, 1.10		
Kotelchuck index								

Characteristics	Urban				Rural			
	Unadjusted		Adjusted ^a		Unadjusted		Adjusted ^a	
	OR	95% CI	OR	95% CI	OR	95% CI	OR	95% CI
Inadequate	1.07	1.04, 1.11	1.02	0.99, 1.06	1.09	0.99, 1.18		
Adequate		1.00 (reference)		1.00 (reference)		1.00 (reference)		
Pre-pregnancy BMI								
Underweight	1.05	0.98, 1.12			1.15	0.96, 1.37		
Normal weight		1.00 (reference)				1.00 (reference)		
Overweight or obese	1.03	0.99, 1.06			1.05	0.97, 1.13		
Insurance source								
Private		1.00 (reference)		1.00 (reference)		1.00 (reference)		
Medicaid	1.30	1.25, 1.34	1.27	1.22, 1.32	0.95	0.87, 1.04		
Other	1.45	1.39, 1.52	1.32	1.25, 1.39	1.11	0.97, 1.26		
Birth facility								
Hospital		1.00 (reference)		1.00 (reference)		1.00 (reference)		1.00 (reference)
Birth Center	1.65	1.42, 1.91	1.72	1.49, 2.00	2.89	1.78, 4.70	2.91	1.80, 4.73
Home	1.04	0.85, 1.26	1.07	0.88, 1.31	1.95	1.28, 2.96	1.94	1.28, 2.95
Other	1.07	0.59, 1.95	0.97	0.52, 1.82	0.71	0.29, 1.73	0.70	0.29, 1.72
Parity								
0		1.00 (reference)		1.00 (reference)		1.00 (reference)		
1	0.94	0.91, 0.96	0.92	0.89, 0.95	1.08	0.99, 1.17		
Nativity								
US		1.00 (reference)		1.00 (reference)		1.00 (reference)		
Non-US	1.26	1.22, 1.30	1.20	1.15, 1.25	1.06	0.96, 1.17		
LBW								
Yes	1.01	0.95, 1.07			1.08	0.93, 1.25		
No		1.00 (reference)				1.00 (reference)		
PTD								
Yes	1.03	0.98, 1.08			0.98	0.85, 1.12		
No		1.00 (reference)				1.00 (reference)		

Adjusted for significant variables in the unadjusted models

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 3
Association Between Characteristics Predictive of Sampling and Adverse Birth Outcomes by Populations.

Characteristics	LBW				PTD			
	Geocoded only		Entire sample		Geocoded only		Entire sample	
	OR	95% CI	OR	95% CI	OR	95% CI	OR	95% CI
White	1.00 (reference)		1.00 (reference)		1.00 (reference)		1.00 (reference)	
Black	2.38	2.26, 2.50	2.33	2.23, 2.45	1.71	1.64, 1.79	1.69	1.62, 1.77
Hispanic	1.28	1.21, 1.36	1.27	1.20, 1.34	1.13	1.07, 1.18	1.12	1.07, 1.18
Asian/PI	1.75	1.56, 1.97	1.76	1.57, 1.96	1.21	1.08, 1.35	1.21	1.09, 1.34
Other	1.50	1.30, 1.73	1.49	1.31, 1.70	1.11	0.97, 1.27	1.12	0.99, 1.27
Maternal age								
<20	1.00 (reference)		1.00 (reference)		1.00 (reference)		1.00 (reference)	
20-34	1.03	0.97, 1.10	1.02	0.96, 1.08	1.08	1.02, 1.15	1.06	0.99, 1.12
35	1.46	1.35, 1.59	1.47	1.36, 1.59	1.49	1.39, 1.61	1.49	1.38, 1.59
Maternal education								
<High school	1.31	1.24, 1.41	1.33	1.25, 1.42	1.28	1.20, 1.36	1.28	1.21, 1.36
High School	1.22	1.15, 1.28	1.22	1.17, 1.28	1.16	1.11, 1.21	1.17	1.12, 1.22
College or more	1.00 (reference)		1.00 (reference)		1.00 (reference)		1.00 (reference)	
Marital status								
Married	1.00 (reference)		1.00 (reference)		1.00 (reference)		1.00 (reference)	
Unmarried	1.15	1.10, 1.20	1.14	1.10, 1.20	1.08	1.03, 1.12	1.08	1.04, 1.12
Maternal smoking								
No	1.00 (reference)		1.00 (reference)		1.00 (reference)		1.00 (reference)	
Yes	1.93	1.81, 2.06	1.95	1.83, 2.07	1.37	1.28, 1.46	1.35	1.27, 1.43
Kotelchuck index								
Inadequate	0.79	0.76, 0.83	0.80	0.77, 0.84	0.77	0.74, 0.81	0.78	0.75, 0.81
Adequate	1.00 (reference)		1.00 (reference)		1.00 (reference)		1.00 (reference)	
Insurance source								
Private	1.00 (reference)		1.00 (reference)		1.00 (reference)		1.00 (reference)	
Medicaid	1.08	1.03, 1.14	1.08	1.03, 1.13	1.02	0.98, 1.07	1.01	0.97, 1.05

Characteristics	LBW				PTD			
	Geocoded only		Entire sample		Geocoded only		Entire sample	
	OR	95% CI	OR	95% CI	OR	95% CI	OR	95% CI
Other	1.05	0.98, 1.13	1.02	0.96, 1.10	0.99	0.93, 1.05	0.97	0.92, 1.03
Birth facility								
Hospital	1.00 (reference)		1.00 (reference)		1.00 (reference)		1.00 (reference)	
Birth Center	0.27	0.17, 0.44	0.26	0.17, 0.41	0.04	0.02, 0.12	0.04	0.01, 0.10
Home	0.80	0.89, 1.10	0.81	0.61, 1.08	0.38	0.26, 0.55	0.43	0.31, 0.60
Other	3.51	2.25, 5.48	3.34	2.20, 5.09	2.35	1.47, 3.76	2.52	1.64, 3.86
Parity								
0	1.00 (reference)		1.00 (reference)		1.00 (reference)		1.00 (reference)	
1	0.70	0.67, 0.73	0.70	0.68, 0.73	0.93	0.90, 0.96	0.93	0.90, 0.97
Nativity								
US	1.00 (reference)		1.00 (reference)		1.00 (reference)		1.00 (reference)	
Non-US	0.91	0.86, 0.96	0.92	0.87, 0.96	0.89	0.84, 0.93	0.88	0.84, 0.92
Urbanity								
Urban	1.00 (reference)		1.00 (reference)		1.00 (reference)		1.00 (reference)	
Rural	0.97	0.90, 1.05	0.99	0.93, 1.06	0.99	0.92, 1.06	0.98	0.92, 1.04