Predicting Coverage Error on the Master Address File using Spatial Modeling Methods at the Block Level

Krista Heim* Andrew M. Raim†

Abstract

This paper explores methods of spatial modeling to identify opportunities for reduced fieldwork in census Address Canvassing operations. The purpose of Address Canvassing is to improve the coverage and quality of the Census Bureau's address list, the Master Address File (MAF), prior to census enumeration. Various modeling techniques such as zero-inflated negative binomial regression have been explored in the past to predict areas with many coverage errors on the MAF and identify blocks which would likely contain change (and those which would not). Such information could inform a reduction to the in-field canvassing workload and reduced field costs. We use a recently developed spatial mixed effects model with dimension reduction, and take New York County as an example. It is seen that accounting for spatial dependence has a large effect on the estimated coefficients, including which predictors are significant. The impact to predicted values is more subtle, with the spatial model producing slightly more accurate predictions.

Key Words: Listing; Coverage Error Models; Address Canvassing; Census; Spatial Statistics, Address Based Sampling.

1. Introduction

The purpose of census Address Canvassing operations is to ensure that the addresses in the Master Address File (MAF) are as accurate as possible before Census enumeration. In 2010, a full-scale Addressing Canvassing operation was conducted in the field (Address List Operations Implementation Team, 2012). The use of statistical modeling has been researched as a part of the effort to redesign Address Canvassing operations in preparation for the 2020 Census (Boies et al., 2012). Statistical models were explored to identify which blocks would benefit from a field canvassing and which blocks already had sufficiently accurate complete addresses on the MAF.

In 2014, the MAF Model Validation Test (MMVT) was conducted to assess the performance of select statistical models as part of the 2015 Address Validation Test (U.S. Census Bureau, 2016). The MMVT was based on a nationally representative sample of 10,100 blocks — 10,000 blocks with preexisting housing units and 100 blocks without preexisting housing units — drawn to conduct an operation where listers made changes to the address list by verifying, correcting, adding and deleting addresses.[1] The results of this dependent listing were used to evaluate the performance of various statistical models, including both logistic and zero-inflated negative binomial regression. From this test, it was determined that the statistical models put forward did not perform well enough to identify areas of change to use operationally.

*Decennial Statistical Studies Division, U.S. Census Bureau, Washington, DC, 20233, U.S.A.

†Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, DC, 20233, U.S.A

Disclaimer: This paper is released to inform interested parties of ongoing research and to encourage discussion of work in process. Any views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

[1]The 100 blocks without preexisting housing units were excluded from this analysis because these were obtained as a convenience sample and do not have sampling weights.

In this paper, we explore improving previous modeling techniques by incorporating a spatial component. Traditional statistical modeling techniques assume the independence of observations in space; however, this may not always be a correct assumption, as spatial patterning (autocorrelation) may occur. Previous models are seen to feature residuals with large spatial autocorrelation, which motivates shifting focus to spatial statistical modeling methods. Spatial models use effects from neighboring areas to account for spatial autocorrelation among observations. Spatial modeling techniques will potentially give more accurate predicted values, give smaller residual values, and reduce spatial autocorrelation among these residuals. They will also help to distinguish which predictors are actually significant and which may only be significant/not significant due to their spatial dependence. We explore Census tabulation block level modeling using Bayesian methodology. Conditional autoregressive (CAR) models (Lee, 2013) have long been used to represent spatial dependence between observations in regression models, but become computationally intensive when there are many areal units, as required by this application. Therefore, we consider the sparse spatial generalized linear mixed model (SGLMM) introduced by Hughes and Haran (2013) to reduce the dimension of the problem. This paper focuses specifically on the analysis of New York County to demonstrate the methodology, but future work could expand to larger geographies.

The rest of the paper proceeds as follows. Section 2 introduces the baseline zero-inflated negative binomial model based on previous modeling work, as well as a non-spatial Poisson GLMM. Section 3 describes spatial modeling methods used in the CAR model and the sparse SGLMM. Section 4 describes the data used for this analysis as well as the measures of model evaluation. The resulting modeling evaluations for New York County are listed and explained in Section 5. Section 6 gives our conclusions and potential future work.

2. Baseline Models

Previous work has evaluated various types of regression models on the 2009 Address Canvassing data, including Poisson, negative binomial, logistic, zero-inflated Poisson and zero-inflated negative binomial regression. We focus on zero-inflated negative binomial regression as the baseline model to which spatial modeling will be compared. We also present a Poisson generalized linear mixed model as its structure is similar to the sparse SGLMM, except that its random effects are not spatially dependent. In this paper, we will focus on the number of missed housing units as the response variable; we will refer to these as "adds", as Address Canvassing operations add these housing units to the MAF.

2.1 Zero-Inflated Negative Binomial Model

Poisson regression is the standard model for count data in the Generalized Linear Model framework (McCullagh and Nelder, 1989). Poisson regression often suffers from overdispersion in real data analysis, where variance of the outcome is much larger than the mean. The negative binomial (NB) distribution is used as a more flexible counterpart (Morel and Neerchal, 2012, Chapter 6).

The density of the classical NB distribution can be defined as

$$\Pr(Y = y) = \binom{y + r - 1}{y} p^y (1 - p)^r, \quad \text{for } y = 0, 1, 2, \ldots.$$

Here the random variable Y represents a number of failures observed until r successes are obtained, in a series of independent Bernoulli trials each having success probability p. The

form of NB which is commonly used to model counts is obtained by taking $r = 1/\kappa$ and $p = \kappa\mu/(1+\kappa\mu)$, so that $\mu = \mathrm{E}(Y)$ parameterizes the mean and $\kappa > 0$ is a real-valued dispersion parameter. The variance depends on both μ and κ, with $\mathrm{Var}(Y) = \mu(1+\kappa\mu)$. This form of the NB distribution can be written

$$\Pr(Y = y) = \frac{\Gamma(y + 1/\kappa)}{\Gamma(y+1)\Gamma(1/\kappa)} \left[\frac{\kappa\mu}{1+\kappa\mu}\right]^y \left[\frac{1}{1+\kappa\mu}\right]^{1/\kappa}, \quad \text{for } y = 0, 1, 2, \ldots \tag{1}$$

where $\Gamma(\cdot)$ is the Gamma function. The NB distribution can also be derived as a continuous mixture of Poisson distributions mixed by a gamma distribution,

$$\begin{aligned} Y \mid \lambda &\sim \text{Poisson}(\lambda) \\ \lambda &\sim \text{Gamma}(\alpha, \beta), \end{aligned}$$

taking $\mu = \alpha\beta$ and $\kappa = 1/\alpha$, and using the Gamma density $f(\lambda \mid \alpha, \beta) = \frac{\lambda^{\alpha-1}e^{-\lambda/\beta}}{\Gamma(\alpha)\beta^\alpha}$ (Morel and Neerchal, 2012, Chapter 6).

In a typical count regression setting, we observe data $\{(y_i, \boldsymbol{x}_i) : i = 1, \ldots n\}$ with y_i a count-valued outcome and $\boldsymbol{x}_i$ a k-dimensional covariate. Optionally, we may designate a covariate O_i, for $i = 1, \ldots, n$, as an offset to scale the count rate for interpretability. Model (1) can adapted to the regression setting by assuming that

$$Y_i \sim \text{NB}(\mu_i, \kappa), \quad g(\mu_i) = \boldsymbol{x}_i^T\beta + O_i.$$

The link function g is typically taken as the natural log function $g(x) = \log(x)$, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)$ is a vector of regression coefficients.

Zero-inflated modeling is used to account for the presence of excess zeros in observed y_i. Excess zeros often appear in area-based count data when populations are very small. Zero-inflated models assume that zero counts are caused by a mixture of two processes: a count distribution which naturally produces zeros on occasion, and a point mass at zero which always produces zeros. A Bernoulli draw selects between the point mass with probability φ and the count distribution with probability $1 - \varphi$ when producing a random draw. The zero-inflated negative binomial (ZINB) model uses NB as the count distribution, and is considered for data with both excess zeros and overdispersion. The density of ZINB is given by

$$\begin{aligned} \Pr(Y = 0) &= \varphi + (1-\varphi)\left[\frac{1}{1+\kappa\mu}\right]^{1/\kappa} \\ \Pr(Y = y) &= (1-\varphi)\frac{\Gamma(y+1/\kappa)}{\Gamma(y+1)\Gamma(1/\kappa)}\left[\frac{\kappa\mu}{1+\kappa\mu}\right]^y \left[\frac{1}{1+\kappa\mu}\right]^{1/\kappa}, \quad \text{for } y = 1, 2, \ldots. \end{aligned}$$

ZINB regression is considered in this paper so that the proposed spatial models can be compared to models from previous work (Raim and Gargano, 2015; Young et al., 2016).

2.2 Poisson Generalized Linear Mixed Model

Another way to extend the Poisson distribution to model extra variation is by the addition of random effects. A GLM with random effects is referred to as a Generalized Linear Mixed Model (GLMM). While ZINB is relevant for its role in previous non-spatial modeling attempts, Poisson GLMMs are more closely related to our spatial model and serve as a more direct assessment of the impact of spatial modeling. GLMMs are an extension of linear

mixed models (LMMs); LMMs are formulated for the setting of additive errors, usually in the form

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon},$$

where $\boldsymbol{y}$ is the response variable, $\boldsymbol{X}$ is a design matrix of covariates, $\boldsymbol{\beta}$ is a vector of fixed effects, $\boldsymbol{Z}$ is a design matrix for random effects, $\boldsymbol{\gamma}$ is a vector of random effects, and $\boldsymbol{\epsilon}$ is a vector of residuals. We assume $\boldsymbol{\gamma} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{G})$ where $\boldsymbol{G}$ is the covariance matrix of the random effects and $\boldsymbol{\epsilon} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{R})$ where $\boldsymbol{R}$ is the covariance matrix for the residuals. In a Bayesian setting (Gelman et al., 2003), we make further assumptions on prior distributions of parameters, such as $\boldsymbol{\beta} \sim \mathrm{N}(\boldsymbol{a}, \boldsymbol{B})$, so that $\boldsymbol{a}$ is the prior mean for the fixed effects and is the $\boldsymbol{B}$ prior covariance matrix.

In the case of count data with a Poisson response, the GLMM is usually formulated as

$$\begin{aligned}
\boldsymbol{y} &\sim \mathrm{Poisson}(\boldsymbol{\lambda}) \\
g(\boldsymbol{\lambda}) &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{O} \\
\boldsymbol{\gamma} &\sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{G})
\end{aligned}$$

with offset $\boldsymbol{O}$, whose likelihood is expressed as an integral to obtain the marginal distribution of $\boldsymbol{y}$. For the Bayesian setting, we use Markov chain Monte Carlo (MCMC) methods available in the `R` package `MCMCglmm` (Hadfield, 2010). For our application, we specifically assume the random intercept model

$$\begin{aligned}
y_i &\sim \mathrm{Poisson}(\lambda_i) \\
g(\lambda_i) &= \boldsymbol{x}^T\boldsymbol{\beta} + \boldsymbol{Z}\gamma_i + O_i \\
\gamma_i &\sim \mathrm{N}(0, \tau^2)
\end{aligned}$$

with $\boldsymbol{Z} = \boldsymbol{1}$, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)$, $\boldsymbol{\beta} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{C})$ with $\boldsymbol{C}$ taken to be a diagonal matrix with large values, and $\tau^2 \sim \mathrm{InverseGamma}(a, b)$ for shape and scale parameters a and b, respectively.

3. Spatial Models

3.1 Conditional Autoregressive Model

In data collected over a range of geographic areas, such as the 2009 Address Canvassing outcomes, we might suspect that nearby areal units tend to behave more similarly than units which are further away. However, explicitly modeling the nature of this dependence can be difficult. Conditional autoregressive (CAR) models are GLMMs which model spatial dependence by assuming certain covariance structures on the distribution of the random effects. CAR models allow the distribution of a particular random effect to be written conditionally on the others, while ensuring a valid joint distribution exists for the random effects together. CAR models formulate the distribution of each random effect conditionally on the others. The `R` package `CARBayes` (Lee, 2013) supports several types of CAR models with Normal, Binomial, and Poisson responses.

The CAR model, as specified in Lee (2013), assumes responses $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ are linked to n non-overlapping areal units $S = \{S_1, \ldots, S_n\}$ for a given study region S. The spatial pattern in the response is modeled by a vector of random effects $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)$, which capture spatial autocorrelation in $\boldsymbol{Y}$ not accounted for by the fixed effects.

For a given areal unit S_k, the vector of covariates is $\boldsymbol{x}_k^T = (x_{k1}, \ldots, x_{kp})$. The model that `CARBayes` implements is a generalized linear mixed model (similar to what was

shown in the previous section) for spatial areal unit data, which is given by

$$Y_k \mid \mu_k \sim f(y_k \mid \mu_k, \nu^2) \quad \text{for } k = 1, \ldots, n,$$
$$g(\mu_k) = \boldsymbol{x}_k^T \boldsymbol{\beta} + \phi_k + O_k.$$

The responses Y_k in this example come from a Poisson distribution and the vector of regression parameters are denoted as $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$. The expected values μ_k of the responses are related to the linear predictor by the link function $g(\cdot)$.

When there is likely to be residual spatial autocorrelation, one of the global or local CAR priors is required, which will incorporate information on the spatial adjacency into the Bayesian hierarchical model. We present the intrinsic prior (Besag et al., 1991) here, which is one of the simpler CAR prior specifications that is available in `CARBayes`. A common characteristic of CAR priors is that they can be written as a set of n univariate full conditional distributions $f(\phi_k \mid \boldsymbol{\phi}_{-k})$ for $k = 1, \ldots, n$, where $\boldsymbol{\phi}_{-k} = (\phi_1, \ldots, \phi_{k-1}, \phi_{k+1}, \ldots, \phi_n)$. The intrinsic CAR prior is

$$\phi_k \mid \boldsymbol{\phi}_{-k}, \boldsymbol{W}, \tau^2 \sim \text{N}\left(\frac{\sum_{i=1}^n w_{ki}\phi_i}{\sum_{i=1}^n w_{ki}}, \frac{\tau^2}{\sum_{i=1}^n w_{ki}}\right)$$
$$\tau^2 \sim \text{InverseGamma}(a, b).$$

You can think about the random effect ϕ_k as a draw from a normal distribution with the mean being the average of the random effects of the neighboring units. The adjacency matrix $\boldsymbol{W}$ determines the spatial autocorrelation of the random effects. The form of this non-negative symmetric $n \times n$ matrix depends on the selected CAR model. For our choice of intrinsic CAR, $\boldsymbol{W}$ is specified as a binary matrix based on geographic contiguity, with $w_{kj} = 1$ if two distinct blocks ($k \neq j$) share a common border and zero otherwise. The joint distribution of $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)$ is multivariate normal with $\boldsymbol{\phi} \sim \text{N}(\boldsymbol{0}, \tau^2(\boldsymbol{I} - \boldsymbol{B})^{-1}\boldsymbol{D})$ where $\boldsymbol{D} = \text{Diag}(1/w_{1+}, \ldots, 1/w_{n+})$, with $w_{k+} = \sum_{j=1}^n w_{kj}$, and

$$\boldsymbol{B} = \boldsymbol{D}\boldsymbol{W} = \begin{pmatrix} w_{11}/w_{1+} & \ldots & w_{1n}/w_{1+} \\ \vdots & \ddots & \vdots \\ w_{n1}/w_{n+} & \ldots & w_{nn}/w_{n+} \end{pmatrix}.$$

3.2 Sparse Spatial Generalized Linear Mixed Model

While it is feasible to model all blocks within a particular county jointly for most counties, the adjacency matrix for a large geography with many blocks becomes too large to construct explicitly. For example, the state of Pennsylvania has over 450,000 blocks which would suggest construction of a 450,000 $\times$ 450,000 adjacency matrix. Most cells in the adjacency matrix contain zeros since most blocks do not border one another. One alternative is to instead construct a sparse representation of the matrix using a list structure, which can be obtained using the `poly2nb()` function within the `spdep R` package. This list structure retains information on which blocks are connected without saving the zero cells. It is possible to modify the `CARBayes` package to use list structure; however, Bayesian hierarchical modeling with MCMC methods may still be too computationally intensive because of the large number of random effects. Model diagnostics may be improved with more covariates and longer MCMC chains, but these modifications further increase the time required to fit the model even once. Another issue with CAR modeling is spatial confounding. Introducing spatial random effects into the model can inflate the variance of the posterior distribution of $\boldsymbol{\beta}$, potentially distorting estimates of coefficients and changing their interpretation.

Hughes and Haran (2013) recently proposed a spatial generalized linear mixed model (SGLMM) for areal data that reduces spatial confounding and speeds computation. As with CAR, an SGLMM is a hierarchical model that induces spatial dependence through a latent Gaussian Markov random field. An SGLMM additionally reduces the dimension of the random effects; this reduction can potentially be dramatic with minimal loss to the quality of the model fit. SGLMMs can be fit with the function `sparse.sglmm()` in the `R` package `ngspatial` (Hughes, 2014), which uses a sparse reparameterization of the adjacency matrix.

Hughes and Haran reparameterize the areal SGLMM by considering the Moran operator $\boldsymbol{P}^{\perp}\boldsymbol{W}\boldsymbol{P}^{\perp}$, which is contained in the numerator of a generalized form of the Moran's I statistic

$$I_X(\boldsymbol{W}) = \frac{n}{\mathbf{1}^T\boldsymbol{W}\mathbf{1}} \frac{\boldsymbol{Y}^T\boldsymbol{P}^{\perp}\boldsymbol{W}\boldsymbol{P}^{\perp}\boldsymbol{Y}}{\boldsymbol{Y}^T\boldsymbol{P}^{\perp}\boldsymbol{Y}}.$$

where $\boldsymbol{P}^{\perp} = \boldsymbol{I} - \boldsymbol{P}$ and $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ is the projection matrix onto the column space of $\boldsymbol{X}$. The Moran operator captures the intrinsic geometry of the models and the information orthogonal to the $\boldsymbol{X}$ covariates.

Magnitudes of eigenvalues of the Moran operator measure varying degrees of spatial dependence: larger positive eigenvalues correspond to stronger "attractive" spatial dependence while negative eigenvalues correspond to stronger "repulsive" spatial dependence. The eigenvectors associated with a given eigenvalue represent patterns of spatial clustering.

The dimension of the spatial effects is reduced to $q << n$ by selecting q eigenvalues of the Moran operator, and constructing an $n \times q$ matrix $\boldsymbol{M}$ from the corresponding eigenvectors. In general, the choice of q and the selection of eigenvectors is a model selection problem. For the application in this paper, we selected the eigenvectors corresponding to the $q = 50$ largest positive eigenvalues. Once $\boldsymbol{M}$ is constructed, the regression model becomes

$$g\{E(Z_i \mid \boldsymbol{\beta}, \boldsymbol{\phi})\} = \boldsymbol{x}_i^T\boldsymbol{\beta} + \boldsymbol{m}_i^T\boldsymbol{\phi},$$

where $\boldsymbol{m}_i^T$, $i = 1, \ldots, n$, are the rows of $\boldsymbol{M}$. The prior density for the random effects is

$$p(\boldsymbol{\phi} \mid \tau) \propto \tau^{q/2} \exp\left(-\frac{\tau}{2}\boldsymbol{\phi}^T\boldsymbol{R}\boldsymbol{\phi}\right),$$

where $\boldsymbol{R} = \boldsymbol{M}^T\boldsymbol{Q}\boldsymbol{M}$ is a precision matrix with $\boldsymbol{Q} = \mathrm{Diag}(\boldsymbol{W}\mathbf{1}) - \boldsymbol{W}$.

4. Application to Address Canvassing Data

4.1 Databases

The ZINB, Poisson GLMM, and Poisson SGLMM models will now be applied to the Address Canvassing data. The three models are fit to data from the 2009 Address Canvassing operation and then validated using data from the 2014 MMVT. Our objective is to determine whether spatial modeling can improve the identification of blocks which contain MAF coverage errors. This study focuses specifically on adds, and does not make use of other outcomes from address canvassing.

The dependent variable y_i in the 2009 model is the count of newly added housing units on the ith block obtained from Address Canvassing. The offset is taken to be `logdeplist`, the log of number of housing units prior to canvassing. Because 2009 was the only year in which a full-scale Address Canvassing was conducted, 2009 is the only year for which we

will have add counts on a majority of blocks. For the 2013 data, the dependent variable is only available for the sample of 10,000 blocks from the 2014 MMVT test.

The covariates used in this work are a subset of those obtained from an extensive variable selection carried out in Raim and Gargano (2015). Here we consider variables from two of the previously studied data sources: the 2009 Address Canvassing database (AdCan DB) and the 2000 Planning Database (PDB). Some variables were removed after noticing poor diagnostics when performing initial fits of the spatial model. In addition to taking the same transformations as in the previous work (mostly involving taking the natural log), all variables were standardized to have mean zero and variance one; this was done to alleviate convergence issues which we believe are influenced by irregular distributions of these variables.

The 2009 AdCan DB was prepared by the Census Bureau to evaluate statistical models for Address Canvassing outcomes (Tomaszewski, 2014). It contains a block-level database with variables that have been summed from address-level data. A number of candidate predictor variables are available which were recorded prior the Address Canvassing operation; their sources included previous MAF extracts and United States Postal Service Delivery Sequence Files. The PDB contains variables related to nonresponse in the census (Bruce and Robinson, 2007). Note that later PDBs are available, but the 2000 PDB was the most recent one prior to 2009 Address Canvassing. Data in the 2000 PDB was recorded on 2000 tabulation geography; we interpolated it to 2010 geography using the method described in Raim and Gargano (2015) so that it could be linked to the AdCan DB. To compute updated predictors for the MMVT, the 2013 database of independent variables were used in conjunction with the 2015 PDB, which included 5-year estimates from the 2009-2013 American Community Survey (ACS). Some variables from the 2015 PDB were summed together to reproduce variables from the 2000 PDB. The same predictors were used across all model types for consistency. Table 1 gives the list of predictors.

We fit the baseline ZINB model, Poisson GLMM, and sparse SGLMM models to Address Canvassing data in New York County, NY, which contains 3,950 blocks. Two blocks were removed from the analysis because they are islands with no neighbors, and therefore could not be analyzed by our spatial models. Predictions were applied to the 2014 MMVT sample using the 2013 covariates. New York County contained only 64 blocks in the 2014 MMVT sample, the largest proportion of any county in the US yet only a small number of the blocks in the county.

4.2 Model Evaluations

We evaluate the models based on the following metrics:

1. Significance of the covariates.[2]
2. Prediction errors.
3. Correlations between predicted and observed values.
4. Significance of Moran's I values of the raw residuals.
5. Number of adds captured based on predictions for given canvassing thresholds.

For 2009 model evaluation we use 5-fold cross validation and split the blocks in the county into five pieces, created by sorting the blocks and labeling them 1 through 5 systematically. Setting aside the first piece as a test set, the model is fit to the remaining four pieces, and predictions are computed for the test set. This is repeated four more times, taking pieces $2, \ldots, 5$ as the test set and obtaining cross validated predictions for all blocks.

[2]While prediction is of primary interest, knowing which variables are truly important in explaining adds is of general interest in interpreting results.

Table 1: List of predictors used in count regression models.

teaMOM	Mailout/Mailback or Military type of enumeration area.
log_dpreac_ac_mafsrc2_Sum	Number of housing units in block whose original MAF source is the 1990 Address Control File (ACF)
log_dpreac_ac_mafsrc28_Sum	Number of housing units in block whose original MAF source is the 2000 Special Place/Group Quarters (SP/G) Enumeration.
log_dpreac_ac_delptypeBk_Sum	Number of housing units in block whose Postal Delivery Point Types are blank.
log_dpreac_ac_delptypeX_Sum	Number of housing units in block whose Postal Delivery Point Types are provided by the Delivery Sequence File (DSF).
log_dpreac_ac_business_Sum	Number of housing units in block with a business-related Postal Delivery Point Type.
tract_a9_count_sum	Housing unit count in block prior to Address Canvassing.
block_hu_density	Housing unit density.
log_dpreac_a9_adcanaf0_Sum	Number of housing units in block not valid for Address Canvassing.
log_dpreac_a9_adcanaf3_Sum	Number of housing units in block valid for Address Canvassing that are post-census DSF adds.
log_landmeters2_sq	Land area of block.
log_devel3_pct	Percent of block highly developed, from the National Land Cover Database (NLCD).
devel1_pct	Percent of block with low development, from the NLCD.
wetlands2_pct	Percent of block covered by emergent herbaceous wetlands, from the NLCD.
log_pct_crowd_occp_u	Percent of American Community Survey (ACS) occupied housing units with greater than 1.5 persons per room within the tract.
pct_occp_hu_moved_2010	Percent of ACS occupied housing units within the tract where the householder moved into the current unit in the year 2000 or later (for the 2009 model) or 2010 or later (for the 2014 Model).
pct_aian_zero	No people on the ACS indicating ‘American Indian or Alaska Native’ within the tract.
log_pct_pop_0_17	Percent within tract that were age 0 to 17 on the ACS.
pct_pub_asst_inc_2010	The percentage of ACS occupied housing units within tract that receive public assistance income.

Note that in the Poisson GLMM, a block excluded from fitting does not have posterior information for its random effect; therefore, only fixed effects are used in its prediction when performing cross validation. For the SGLMM, the random effects enter the model through the spatial structure $\boldsymbol{M}$, and we are able to use them in cross validated predictions by dropping rows of the overall $\boldsymbol{M}$ corresponding to the test set. For the 2014 model evaluations, we use parameter estimates from the full (without cross validation) run of 2009 models with updated 2013 covariates. For the spatial model, this translates to having the same spatial structure but an updated $\boldsymbol{M}$ matrix.

Moran's I is a measure of spatial autocorrelation with the standard form

$$I = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(X_i - \bar{X})(X_j - \bar{X})}{\sum_i (X_i - \bar{X})^2}, \tag{2}$$

where N is the number of blocks, w_{ij} is the adjacency between block i and block j (=1 if the blocks are adjacent, 0 otherwise), and X is the variable of interest. Here we focus on the Moran's I of the residuals from a model to see if there is spatial autocorrelation in the error term which has not been accounted for by the regression. Expression (2) is the global Moran's I value, which is the sum of the local Moran's I, I_i, that can be calculated for each block.

5. Results

ZINB, Poisson GLMM and sparse SGLMM models were fit to New York County. Trace plots and acceptance rates were monitored for each model to ensure there were no obvious failures of MCMC to converge to the posterior distribution — these diagnostics are not shown due to space restrictions. Table 2 displays the significance of predictors in each of the three example models. The SGLMM results in many more significant predictors than ZINB or Poisson GLMM; we suspect that this is due to the SGLMM's ability to reduce spatial confounding between predictors and spatial effects.

Table 3 evaluates the 5-fold cross validated models fitted to the 2009 Address Canvassing data. Prediction errors are measured by the sum of squared prediction error (SSPE) $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ and the sum of absolute prediction error (SAPE) $\sum_{i=1}^{n}|y_i - \hat{y}_i|$. The correlation (CORR) between the observed values $y_1, \ldots, y_n$ and the predicted values $\hat{y}_1, \ldots, \hat{y}_n$ provides another summary of prediction accuracy. The p-value for the global Moran's I test, computed from the raw residuals $y_i - \hat{y}_i$, measures the overall spatial dependence not captured by the model. The number of blocks with a significant local Moran's I statistic, also computed on the raw residuals, summarizes uncaptured spatial dependence at a more granular level.

The SGLMM attains the smallest SSPE while Poisson GLMM gives the smallest SAPE. The spatial model also has the largest correlation between predicted add counts and observed add counts on the 3,948 blocks. Although we do not see an improvement in the local Moran's I with the SGLMM, the p-value from global Moran's I indicates that the spatial model has accounted for the spatial autocorrelation which is present in the residuals from Poisson GLMM and ZINB. Overall, we see a slight improvement in predictions from the spatial model compared to the non-spatial models.

Table 4 presents evaluations for the 2014 MMVT add counts. Note that we are unable to calculate the Moran's I for these blocks as they are a noncontiguous sample of the full county. Similar to the results in Table 3, we can once again see an improvement in the sum of squared prediction error as well as the correlation between the predicted and observed add counts, albeit slight.

Table 2: The significance of predictors in each selected model; significant variables are designated with an X.

Variable	ZINB	Poisson GLMM	Sparse SGLMM
teaMOM	X	X	X
log_dpreac_ac_mafsrc2_Sum	X*	X	X
log_dpreac_ac_delptypeBk_Sum	X*	X	X
log_landmeters2_sq	X	X	X
log_dpreac_ac_delptypeX_Sum	X*	X	X
log_dpreac_a9_adcanaf0_Sum		X	X
log_dpreac_ac_mafsrc28_Sum	X		X
tract_a9_count_sum		X	X
log_dpreac_ac_business_Sum			X
log_dpreac_a9_adcanaf3_Sum	X*	X	X
pct_occp_hu_moved_2010	X	X	X
log_devel3_pct			X
log_pct_crowd_occp_u	X	X	X
pct_aian_zero	X*	X	X
devel1_pct			X
log_pct_pop_0_17	X*		X
pct_pub_asst_inc_2010	X		X
wetlands2_pct	*		X
block_hu_density	*		X
logdeplist*log_pct_crowd_occp_u	X	X	X
logdeplist*pct_aian_zero	*	X	X
log_dpreac_ac_business_Sum	*		
pct_occp_hu_moved_2010	X*		X

*Significant in zero component

Table 3: 2009 Address Canvassing results for 5-fold cross validated models.

	SSPE	SAPE	CORR	Moran's I p-value	Local Moran's I # Significant
ZINB	2,680,787	27,284.43	0.18	0.0558	44
Poisson GLMM	2,544,028	18,913.54	0.15	0.0001	29
SGLMM (M=50)	2,528,335	26,973.32	0.28	0.9998	36

Table 4: 2014 MMVT modeling results.

	SSPE	SAPE	CORR
ZINB	50,744	808.84	0.049
Poisson GLMM	25,900	502.66	0.130
SGLMM ($M = 50$)	23,397	689.63	0.251

Plots of predicted and observed add counts for the predictions based on the 2009 data and on the 2014 data are shown in Figure 1. A perfect prediction would have the points (representing the blocks) fitted to the red dashed line. We can see that this is not the case for any of the six plots. While the numbers from the results in Tables 3 and 4 show slight improvement, visually we do not see this improvement.

Maps of the predicted values were created to evaluate the models from a geographic standpoint. Figure 2 shows a map created in SAS JMP which displays predicted adds for each of the fitted models, as well as adds observed from Address Canvassing in 2009. Here we see overestimation in both the ZINB model and sparse SGLMM, and underestimation in the Poisson GLMM. There are particular areas in the sparse SGLMM (highlighted with circles) where the spatial model is more clearly identifying where there are actually no adds observed compared to the ZINB model, while also still capturing the blocks with a large number of adds that are not captured in the Poisson GLMM.

Figure 3 is based on the predictions for the 64 blocks in New York County that were in the 2014 MMVT data. This plot gives the estimated weighted[3] add count for a given cumulative number of blocks, where those blocks are sorted by descending predicted add counts. The blue line indicates how many adds would be captured for a given number of blocks if we knew exactly what was to be observed in the field; a desirable model would be one that approaches this line. For example, assume that our canvassing threshold is 8 blocks. The vertical dotted line at 8 indicates the (weighted) number of true adds captured when selecting blocks with the largest 8 predicted add counts for a given model. At this mark, the SGLMM is performing the best. However, if our canvassing threshold is 24 blocks (focusing on the dotted line at 24), we see that the order has changed and SGLMM is performing worse than ZINB and Poisson GLMM. Therefore, the best model depends on the threshold, and no one model outperforms the other based on this metric.

6. Conclusions

We have compared three count regression models to predict adds in the full-scale 2009 Address Canvassing operation and to produce extrapolated predictions for the later 2014 MMVT operation. The ZINB model used in previous work is an overdispersed Poisson model with the ability to handle excess zeros in the data. The recently proposed SGLMM accounts for spatial dependence through its random effects structure, and is designed to avoid spatial confounding and avoid some computational burden via dimension reduction. Poisson GLMM was also included as a non-spatial counterpart of SGLMM.

The set of significant predictors was noticeably larger under SGLMM than under ZINB or Poisson GLMM, which suggests spatial dependence was masking the effect of some predictors in the non-spatial models. SGLMM was also effective at accounting for spatial autocorrelation, as shown in the global Moran's I test of the residuals. However, the ability to predict adds in New York County was only slightly improved with the SGLMM. Our experience suggests that, while spatial random effects models are helpful in accounting for trends that depend on space and are difficult to model explicitly, the fixed effects part of the model determines its prediction ability.

Therefore, while further improvements to predictions may be possible through more sophisticated modeling, major improvements will likely require the acquisition of stronger predictors. This has proven to be an elusive problem to date, as the true cause of adds in a particular block may depend on a set of circumstances specific to that block.

Our study of New York County has revealed some opportunities for future work. In spite of the dimension reduction, the computational burden of SGLMM becomes large as

[3]Sampling weights are used to weight the sample of blocks up to represent the entire county.

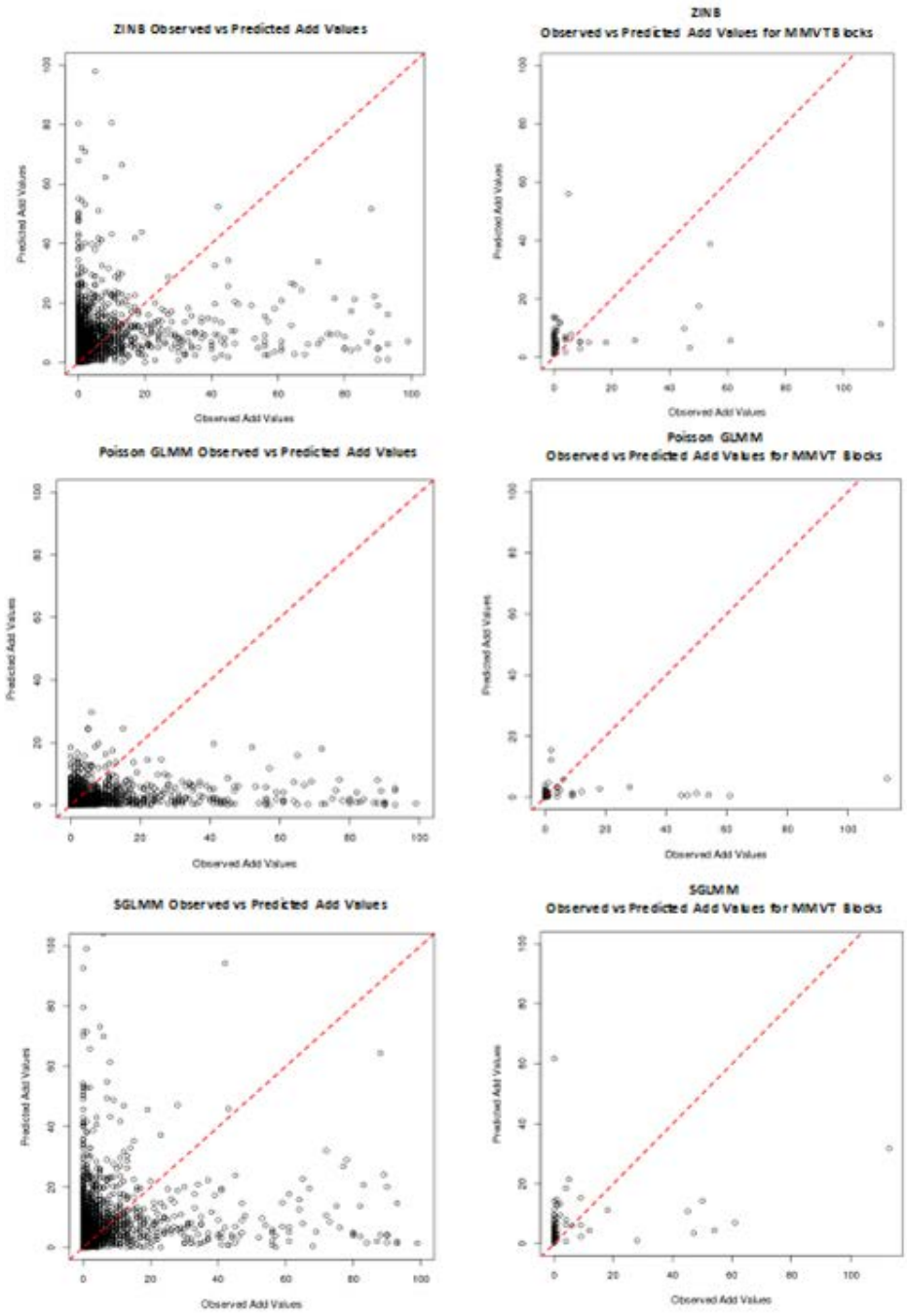

Figure 1: Predicted vs. observed add counts from the three models. Plots in first column display 5-fold cross validated predictions and observed counts for 2009 Address Canvassing. Right column shows 2014 MMVT predictions and observations.

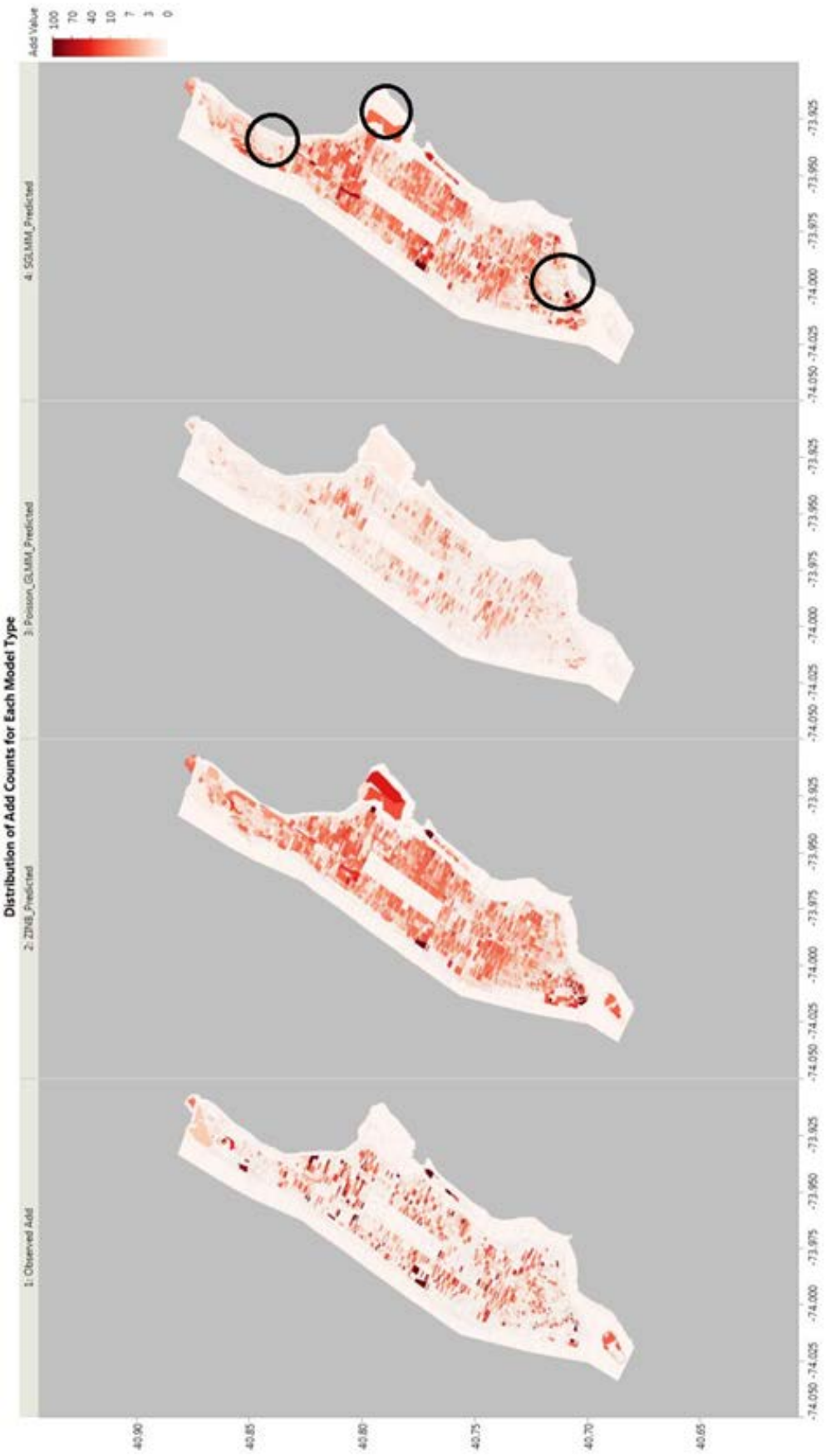

Figure 2: Map of 2009 adds in NYC: observed (left), predicted from ZINB (center-left), predicted from Poisson GLMM (center-right), and SGLMM (right).

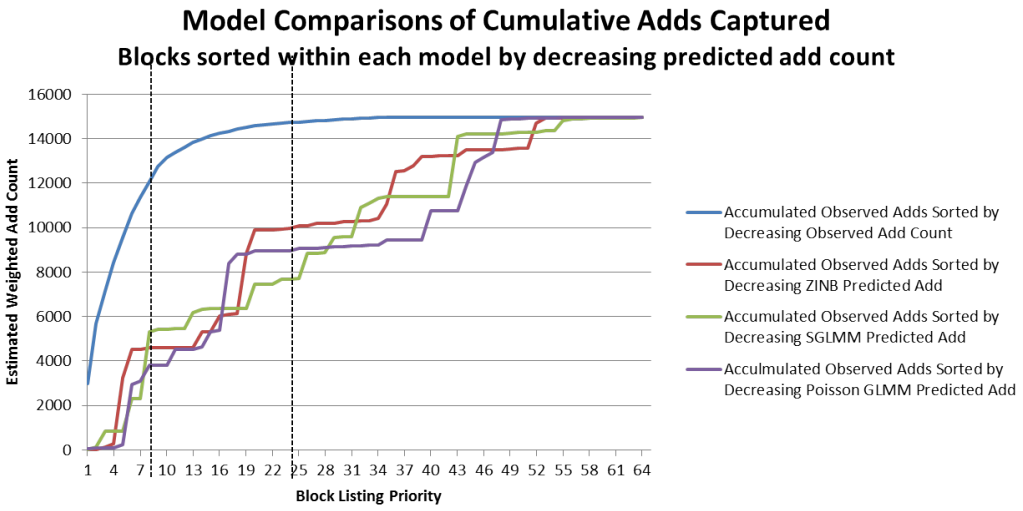

Figure 3: Plot of model comparison of cumulative adds captured

the number of areal units increase, and it would be of interest to scale it to large geographies. In this paper we have considered only adds, but several other outcomes of interest were recorded in Address Canvassing. A multivariate model of these outcomes might outperform our univariate model. Recently Bradley et al. (2015) have proposed a large scale multivariate spatio-temporal model which may help to address some of these issues.

If models with sufficiently accurate predictions can be developed, there are also questions about how they could be used operationally. It has already been determined for the 2020 Census that models alone will not be used to guide an in-field canvassing workload (Decennial Census Management Division, 2015, Section 2.3). However, models could be used as an additional input to inform planned canvassing activities. For example, an alert could be raised if an area is at high risk of having large coverage error based on past canvassing experience.

Acknowledgements

Thanks to Scott Holan from the University of Missouri for his guidance and advice on this problem, and to Timothy Kennel and Laura Ferreira from the U.S. Census Bureau for reviewing our manuscript.

References

Address List Operations Implementation Team. 2010 Census Address Canvassing Operational Assessment. In *2010 Census Program for Evaluation and Experiments*. 2012.

Julian Besag, Jeremy York, and Annie Mollié. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1): 1–20, 1991. ISSN 1572-9052. doi: 10.1007/BF00116466.

John L. Boies, Kevin M. Shaw, and Jonathan P. Holland. Address Canvassing targeting and cost reduction evaluation report. In *2010 Census Program for Evaluations and Experiments*. 2012.

Jonathan R. Bradley, Scott H. Holan, and Christopher K. Wikle. Multivariate spatio-temporal models for high-dimensional areal data with application to longitudinal employer-household dynamics. *Annals of Applied Statistics*, 9(4):1761–1791, 2015.

Antonio Bruce and J. Gregory Robinson. Tract level planning database with Census 2000 data. U.S. Government Printing Office, 2007.

Decennial Census Management Division. 2020 Census Detailed Operational Plan for the Address Canvassing Operation, 2015.

Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd edition, 2003.

Jarrod Hadfield. MCMC methods for multi-response generalized linear mixed models: The mcmcglmm r package. *Journal of Statistical Software*, 33(1):1–22, 2010. ISSN 1548-7660. doi: 10.18637/jss.v033.i02.

John Hughes. ngspatial: A package for fitting the centered autologistic and sparse spatial generalized linear mixed models for areal data. *The R Journal*, 6(2):81–95, 2014.

John Hughes and Murali Haran. Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(1):139–159, 2013. doi: 10.1111/j.1467-9868.2012.01041.x.

Duncan Lee. CARBayes: An R package for Bayesian spatial modeling with conditional autoregressive priors. *Journal of Statistical Software*, 55(1):1–24, 2013. ISSN 1548-7660. doi: 10.18637/jss.v055.i13.

Peter McCullagh and John A. Nelder. *Generalized Linear Models*. CRC press, 2nd edition, 1989.

Jorge G. Morel and Nagaraj K. Neerchal. *Overdispersion Models in SAS*. SAS Institute, 2012.

Andrew M. Raim and Marissa N. Gargano. Selection of predictors to model coverage errors in the Master Address File. Research Report Series: Statistics #2015-04, Center for Statistical Research and Methodology, U.S. Census Bureau, 2015.

Christine Gibson Tomaszewski. 2009 Targeted Address Canvassing User Documentation. In *DSSD 2020 Decennial Census Research and Testing Memorandum Series*. 2014.

U.S. Census Bureau. 2015 Address Validation Test. In *2020 Census Research and Testing*. 2016.

Derek S. Young, Andrew M. Raim, and Nancy R. Johnson. Zero-inflated modelling for characterizing coverage errors of extracts from the US Census Bureau's Master Address File. *Journal of the Royal Statistical Society: Series A*, 2016. doi: 10.1111/rssa.12183.