Zero-Inflated Regression Modeling for Coverage Errors of the Master Address File

Derek Young and Andrew Raim (Presenter)

Center for Statistical Research & Methodology
U.S. Census Bureau

2014 Joint Statistical Meetings
Boston, MA, USA

Outline of Talk

1. Introduction to the MAF Error Model project.
2. Discussion of zero-inflated count modeling in (Young and Johnson, Submitted).
3. Some results using an updated database.

United States™ Census Bureau

MAF Background

- The **Master Address File (MAF)** is an inventory of all known living quarters in the U.S. and Puerto Rico.
- A MAF extract (MAFX) is used as a frame to support several household surveys (e.g. ACS, decennial census, and ongoing demographic surveys).
- The MAF is regularly updated by operations related to the decennial census: e.g. **canvassing**, Delivery Sequence File (DSF) from U.S. Postal Service
- Two types of coverage errors:
  1. **Undercoverage**: addresses missing from the MAF
  2. **Overcoverage**: addresses which should be removed from the MAF
- **Adds** fix undercoverage and **deletes** fix overcoverage

United States™ Census Bureau

MAF Error Model (MEM) Project

- Develop statistical models for the MAF that will produce estimates of coverage errors at the census block level.
- Help characterize the quality of a particular MAFX and lend insight to frame improvement.
  - Surveys and Census operations using a MAFX could quickly estimate coverage errors at different levels of geography with the current MEM.
- Toward the goal of Targeted Address Canvassing: select an optimal set of blocks to canvass in 2019 to support 2020 Census.

United States™ Census Bureau

Data for MEM

Housing Units (HUs) from 2010 Census Address Canvassing (AdCan).

Blocks with > 0 HUs.

Dependent variables: Counts from AdCan operation:

- Adds
- Deletes

Independent variables: Candidate predictors, such as:

- American indian/hawaiian homeland indicator
- Urban/rural area
- # small, # large multi-units
- Presence of LUCA (Local Update of Census Address)
- MAF source variables
- DSF coverage variables

HU-level counts are aggregated up to block level.

United States™ Census Bureau

Distributions of Adds and Deletes

Updated Database

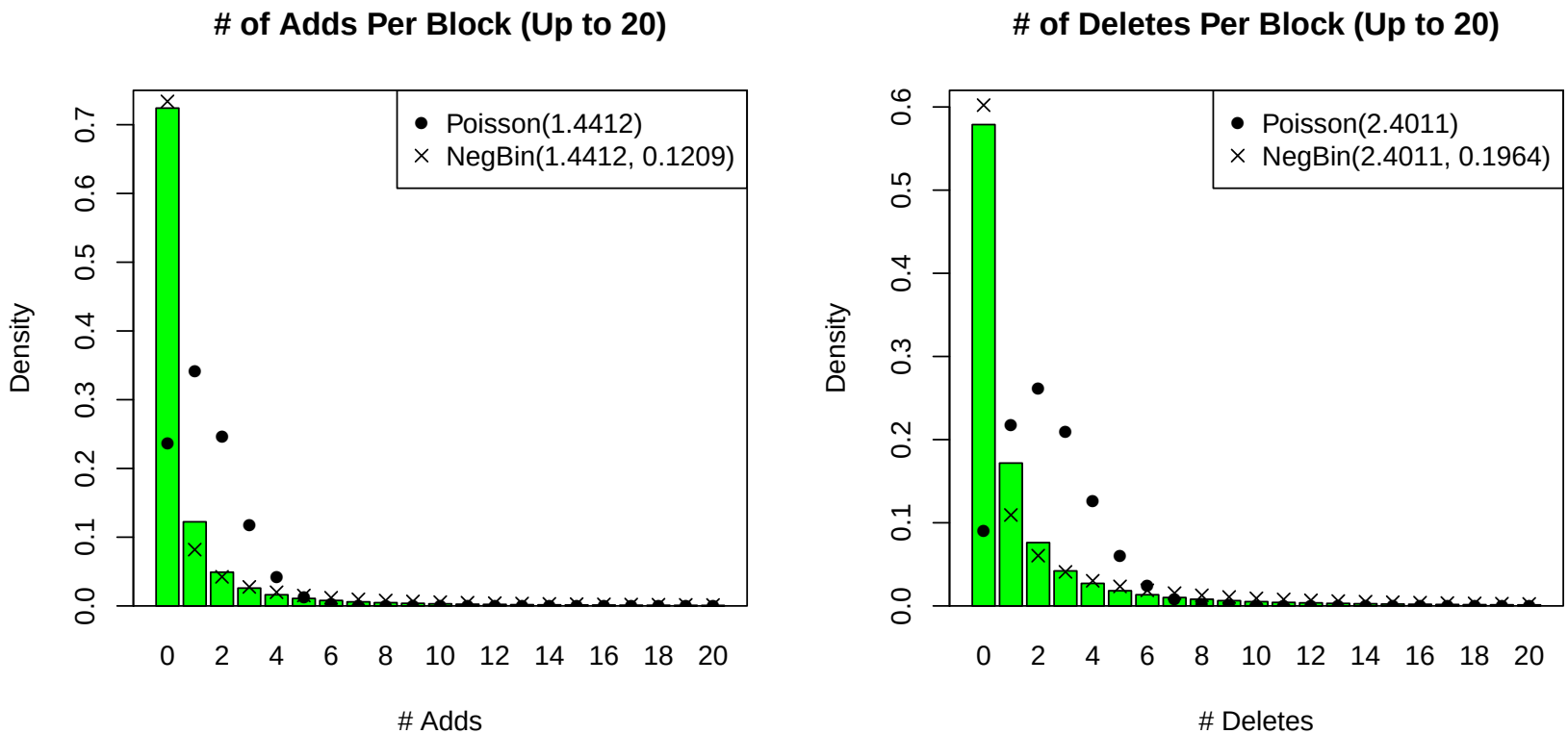


Figure: Distribution of blocks with 20 or fewer adds (left) and deletes (right).

Young and Johnson Methodology

ZI Regression Models for Counts

Investigated two zero-inflated (ZI) regression models for counts.

- **Zero-inflated Poisson (ZIP)** regression.
- **Zero-inflated negative binomial (ZINB)** regression.

Regression: To explain variability (distribution) of the observed # of adds/deletes through contributions of predictor variables.

Count model: Responses are (non-negative) counts.

- Ordinary least squares (OLS) regression is not quite appropriate.
- More granular at block level than logistic regression.

Zero-inflation: More zero-counts than expected under a traditional count regression model (e.g. Poisson or negative binomial).

United States™ Census Bureau

Young and Johnson Methodology

Initial Variable Screening

Omitted some candidate predictor variables because they had:

- Sparse counts;
- Were collected after AdCan; or
- Caused numerical issues when included.

Looked at each non-categorical predictor variables correlation with the number of adds and deletes.

- Variables omitted if correlation coefficient < 0.05.

Resulted in 51 candidate variables for the "adds" model and 58 candidate variables for the "deletes" model.

United States™ Census Bureau

Young and Johnson Methodology

Collinearity

- Collinearity (i.e. high correlation) between predictors can inflate standard errors, impact numerical accuracy, and cause other issues.
- Assessed collinearity at varying thresholds of the variance inflation factor (VIF).
- Systematically omitted variables from the model at various thresholds of the VIF (i.e. 100, 25, and 10) until all remaining variables had a VIF < 10.

United States™ Census Bureau

Young and Johnson Methodology

Variable Selection

- Using all remaining candidate predictors, did a limited variable selection investigation to identify the "best" subset of predictors.
- Classified predictors into 6 categories.
- Used the Bayesian Information Criterion (BIC) to assess inclusion/exclusion of each set of predictors.
  (Smaller BIC $\Longrightarrow$ "better" fit).

United States™ Census Bureau

Young and Johnson Methodology

Model Selection and Prediction Error

Using Vuong's test (Vuong, 1989), we found:

- Negative binomial fits better than Poisson.
- ZI models fits better than non-ZI models.

Prediction errors:

- Compared observed counts with counts predicted from ZI models.
- Calculated percentage of correct and incorrect predictions.
- Used 5-Fold cross-validation (CV) to check for overfitting.

United States™ Census Bureau

Young and Johnson Methodology

Some Variables Appearing in Both Models

- **Block-Level Categorical Variables**: urban/rural area; American Indian/Hawaiian Homeland
- **Census 2000 Variables**: enumeration status - not in Census; enumeration status - respondent return
- **Pre-AdCan Collection Block Variables**: Basic Street Addresses that are small multi-units
- **Pre-AdCan Delivery Point Variables**: business curbline; residential curbline
- **AdCan Filter Variables**: Flag percentage of records valid for AdCan delivery
- **All Other Variables**: percentage of seasonal records; percentage of vacant records

For ZINB: Out of 307 candidate predictors, 39 used in "adds" model and 38 used in "deletes" model.

United States™ Census Bureau

Young and Johnson Methodology

Informing Address Canvassing

- Good predictions can identify regions with many potential adds and/or deletes.
- Using the models, we can estimate coverage errors vs. canvassing effort (e.g. % of blocks canvassed).
- Receiver operating characteristic (ROC) curve:
  - True Positive Rate (TPR): $\frac{\#\text{ blocks marked for canvassing that need it}}{\#\text{ blocks that need canvassing}}$.
  - False Positive Rate (FPR): $\frac{\#\text{ blocks marked for canvassing that don't need it}}{\#\text{ blocks that don't need canvassing}}$.
- A limitation is the version of the data used to fit models. The further removed we are from the 2010 AdCan operation, the more difficult it becomes to accurately quantify block-level adds/deletes.

United States™ Census Bureau

Results from Updated Database

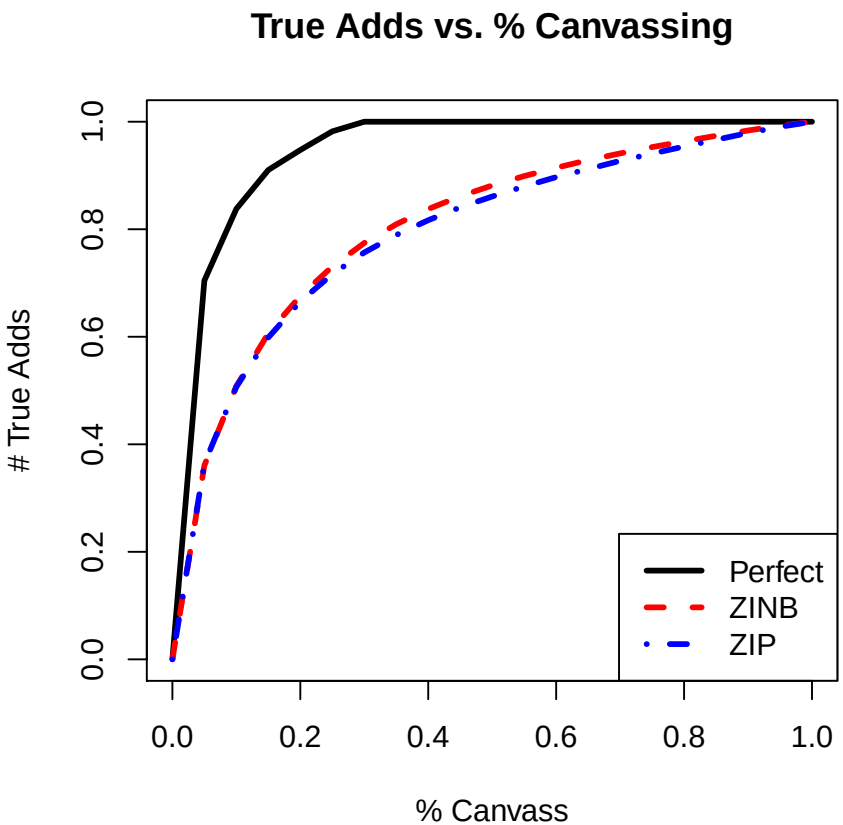
Refitting the Adds Model

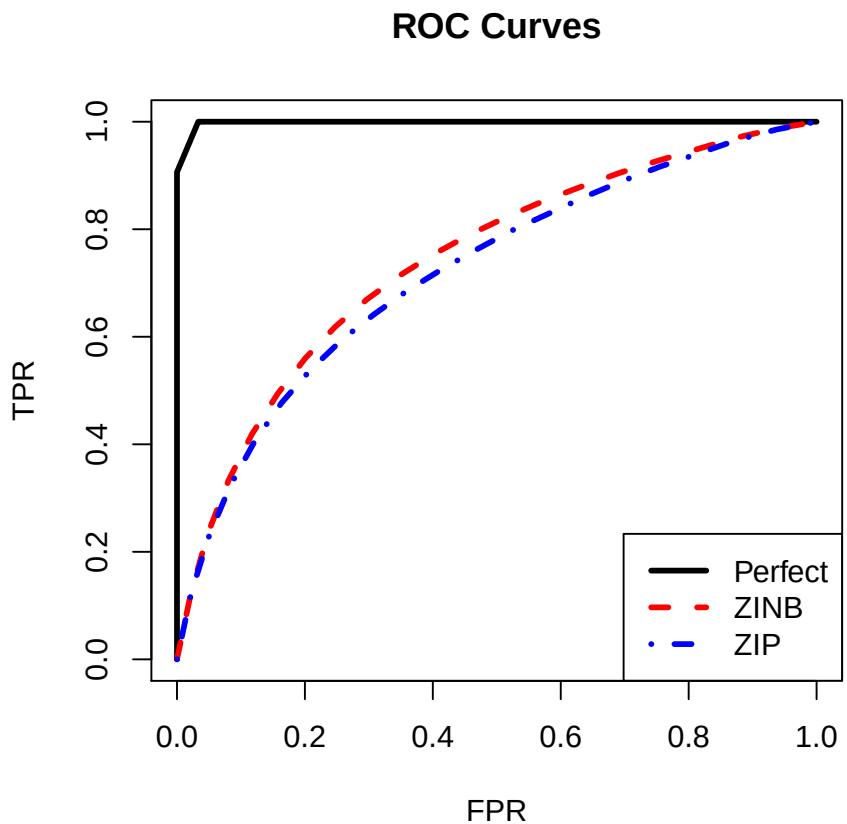
- Data contains 6,585,686 blocks with 144,617,812 HUs.
- 873 available predictors: One binary, one categorical (4 levels), rest are proportions.
- Removed 155 proportions which were constant over all blocks.
- Kept 167 of proportion variables, having $|\text{corr}| > 0.01$ to adds.
- After VIF analysis, kept 55 variables.
- Dropped variables with low significance (p-value > 0.01) in ZINB and ZIP models.
- Obtained models:
  - ZINB: 52 variables in NegBin regression, 1 (categorical) in ZI regression.
  - ZIP: 53 variables in Poisson regression, 2 (categorical + binary) in ZI regression.

United States™ Census Bureau

Results from Updated Database

Quality of Canvassing using Predictions and ROC





Results from Updated Database

Coverage Estimates

- Consider the following measure of MAF Undercoverage error

$$\text{Undercoverage} = \frac{\text{Total \# Adds}}{\text{Total \# HUs}},$$

	Actual	ZINB	ZIP
Undercoverage	0.066	0.085	0.054

- These quantities are simply for our assessment and are not based on any previous Census methodology, such as that in (Mule, 2008).
- "Actual" is computed from the database. It is not referencing published numbers and is used solely for comparison with models.

United States™ Census Bureau

Some Next Steps

Investigate new variables being added to database.

- Title 26 data, AdRecs, ...

Want flexibility to update models with new data sources.

- Spatial data, economic indicators, ...

Develop an approach for determining adds in zero-blocks.

- E.g. use distance measures between blocks with larger number of adds and the zero-blocks

Bivariate regression models for adds and deletes

- Diagonal inflation?

United States™ Census Bureau

References

C. Mazur and E. Wilson. Housing characteristics: 2010. In *2010 Census Briefs: C2010BR-07*. October 2011.

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

T. Mule. 2010 census coverage measurement methodology. In *DSSD 2010 Census Coverage Measurement Memorandum Series 2010-E-18*. October 2008.

Q. H. Vuong. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57:307–333, 1989.

Derek S. Young and Nancy R. Johnson. Zero-inflated modeling for characterizing coverage errors of extracts from the U.S. Census Bureau's Master Address File, Submitted.

United States™ Census Bureau

Contact Information

Andrew M. Raim
andrew.raim@census.gov

Thank you!

United States Census Bureau