Selection of Predictors to Model Coverage Errors in the Master Address File

Andrew M. Raim

Center for Statistical Research & Methodology
U.S. Census Bureau
andrew.raim@census.gov

2015 International Total Survey Error Conference
Baltimore, MD, U.S.A.

Joint work with Marissa N. Gargano (Center for Statistical Research & Methodology, U.S. Census Bureau)

Disclaimers

- This presentation is to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

- This work discussed in this presentation was a research effort conducted outside of the Master Address File Model Validation Test (MMVT) project. Namely, the Title 26 datasets used in the present work were not used in MMVT.

Overview

- To prepare the Master Address File (MAF) for the 2010 Decennial Census, the Census Bureau conducted the 2010 Address Canvassing (AdCan) operation.
- $\sim$111,000 field representatives (FRs) walked $\sim$6 million census blocks in the United States and Puerto Rico.
- AdCan provided a wealth of data on MAF coverage errors.
- If a valid address was missing from the MAF
  1. Indication of an **undercoverage** error.
  2. Address was added to the MAF. AdCan outcome: an **“add”**.
- If an invalid address was present on the MAF
  1. Indication of an **overcoverage** error.
  2. Address was removed to the MAF. AdCan outcome: a **“delete”**.
- The Census Bureau has been interested in using 2010 AdCan data to develop statistical models to study and predict MAF error.

Overview

- There are many factors from data collection which (we suspect) complicate the analysis. These include:
  1. Selection of housing units sent out in the dependent list.
  2. Variation between field representatives who collected the data.
  3. In-office processing to determine the final outcomes.

- Young et al. (2015) proposed count modeling for adds (or deletes) at the census block level, based on zero-inflated negative binomial (ZINB) and zero-inflated Poisson (ZIP) distributions.

- This work builds on the ZINB approach with a more exhaustive variable selection method. We consider main effects and two-way interactions selected from the main AdCan DB and six supplementary data sources.

2010 AdCan Database

From *Reengineered Address Canvassing Fact Sheet* by John Boies

Outcome for Housing Unit	Code	Count
Sent out for canvassing	--	144.9m
True Adds	A	6.7m
Matched / Reinstated Adds	R	4.2m
Deletes	D	15.8m
Moves (found in wrong collection block)	M	5.5m
Changes (Error found in address)	C	19.6m
Verify (Address was correct)	K	97.6m

Block Description	Blocks	HUs	A's	R's	D's
Sent out for AdCan	6.6m	144.8m	6.1m	3.5m	15.8m
Empty w/ AdCan outcomes	210k	1.3m	630k	630k	--
Empty w/ no AdCan outcomes	4.0m	--	--	--	--
Water only	550k	--	--	--	--
100% Public Land	520k	1.4m	210k	64k	310k
Total	11.2m	145.1m	6.7m	4.1m	15.8m

2,138 total variables in main database.

- Almost all are counts/means of HUs that meet some criteria in a block.
- 305 have six versions corresponding to six filters: ac, a9, gc, nc, n9, ug.
- Also, urban vs. rural, TEA, land area, water area.

Modeling universe contains 6,539,119 blocks. Of those, training set obtained by sampling 100,000 blocks.

Supplemental Data Sources

- **2000 Planning Database (PDB)**: contains variables correlated with mail nonresponse (Bruce and Robinson, 2004).
- **Land use data** provided by the Geography Division at the Census Bureau (GEO): Contains percentages of geographical features on each block, provided by the National Land Cover Database (Homer et al., 2007).
- **DSF stability Index** provided by GEO: Block level measure of stability for coverage of housing units by the USPS Delivery Sequence File.
- **2007–2008 Longitudinal Employer-Household Dynamics (LEHD) Origin-Destination Employment Statistics (LODES) data**: residence and workplace characteristics for the workforce.
- **2005–2008 RealtyTrac data**: data on foreclosured homes.
- **IRS 1040 data**: estimates of IRS 1040 returns that had no block ID, no MAFID, and both no block ID and no MAFID.

Notes on Candidate Predictors

- Our main interest is in a predictive model. Therefore, we only consider predictors which would have been available before AdCan.

- Special gotcha: should not use geocoding (attributing data to blocks) which would not have been available before AdCan.

- Our fundamental hypothesis is: MAF error $=$ change $+$ hard-to-detect.

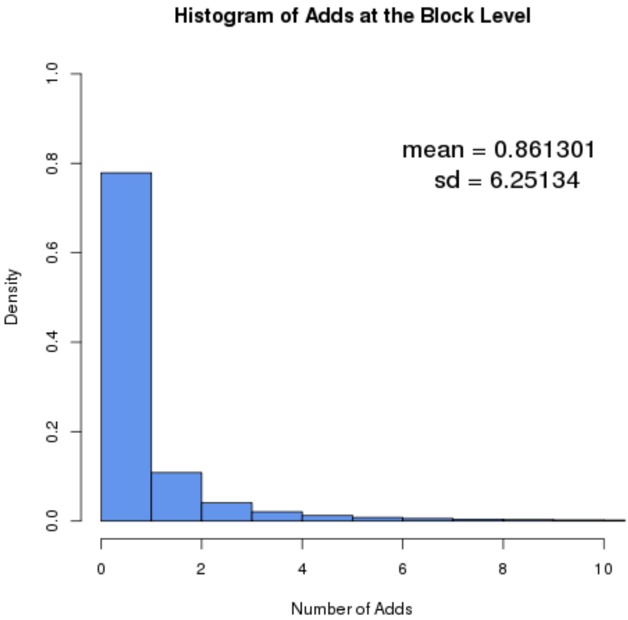
ZINB Regression

- ZINB is commonly used to model count data with many zeros that cannot be explained only by a count distribution (Hilbe, 2011).
- We consider $Y \sim \text{ZINB}(\mu, \kappa, \pi)$ with density

$$f(y \mid \mu, \kappa, \pi) = \pi 1_{\{0\}}(y) + (1-\pi) \frac{\Gamma(y + 1/\kappa)}{\Gamma(y+1)\Gamma(1/\kappa)} \frac{(\kappa\mu)^y}{(1+\kappa\mu)^{y+1/\kappa}},$$

  where $y \in \{0, 1, 2, \ldots\}$, $\mu > 0$, $\kappa > 0$, and $\pi \in (0, 1)$.
- Can show $\text{E}(Y) = (1-\pi)\mu$ and $\text{Var}(Y) = (1-\pi)\mu\{1 + \mu(\kappa + \pi)\}$.
- Special cases:
  1. When $\pi \to 0$, ZINB becomes Negative Binomial.
  2. When $\kappa \to 0$, ZINB becomes Zero-Inflated Poisson.
  3. When $\pi \to 0$ and $\kappa \to 0$, ZINB becomes Poisson.
- Given predictors $\boldsymbol{x} = (x_1, \ldots, x_{d_1})$ and $\boldsymbol{w} = (w_1, \ldots, w_{d_2})$, we consider ZINB regression by linking $\log(\mu)$ to $\beta_1 x_1 + \cdots + \beta_{d_1} x_{d_1}$ and $\text{logit}(\pi)$ to $\gamma_1 w_1 + \cdots + \gamma_{d_2} w_{d_2}$.
- Model for AdCan Data: $Y_i \overset{\text{ind}}{\sim} \text{ZINB}[\exp(\boldsymbol{x}_i^T \boldsymbol{\beta}), \kappa, \text{logit}^{-1}(\boldsymbol{w}_i^T \boldsymbol{\gamma})]$.

Exploratory Analysis



Variable Selection

- Exhaustive variable selection by manually sequencing forward and backward selection steps.

- Select two components of the model individually.
    1. Use negative binomial regression with response y_i to select predictors for **count component** $(\log \mu)$.
    2. Use logistic regression with response $I(y_i = 0)$ to select predictors for **zero-inflated component** $(\operatorname{logit} \pi)$.

- For each of the two components, select in three phases:
    1. From the **2010 AdCan database**.
    2. From the **six supplementary data sources**.
    3. From all **two-way interactions** between main effects in the model.

- Consider models by several criteria:
    1. Likelihood based: log-likelihood, Akaike's Information Criterion (AIC), Bayesian Information Criterion (BIC).
    2. Prediction based: $\text{SSPE} = \sum_i (y_i - \hat{y}_i)^2$ and $\text{APE} = \sum_i |y_i - \hat{y}_i|$.

Variable Selection

- **Add Step**
  1. Specify initial predictors $\boldsymbol{x}$ and candidate predictors $\boldsymbol{x}^* = (x_1^*, \ldots, x_q^*)$.
  2. Fit q models using $(\boldsymbol{x}, x_1^*), \ldots, (\boldsymbol{x}, x_q^*)$ respectively, and record fit statistics for each.
  3. Compare the q models with the initial model using fit statistics.
  4. Add the "best" candidate predictor or keep initial model.
- **Drop Step**
  1. Specify initial predictors $\boldsymbol{x} = (x_1, \ldots, x_p)$.
  2. Fit p models using $\boldsymbol{x}_{(-1)}, \ldots, \boldsymbol{x}_{(-p)}$ respectively, and record fit statistics for each.
  3. Compare the p models with the initial model using fit statistics.
  4. Drop the "least useful" predictor or keep initial model.
- Perform a sequence of Add and Drop steps until no substantial improvement can be made.
- Restrict to the training set to protect against overfitting.
- Check Generalized Variance Inflation Factor (GVIF) from Fox and Monette (1992) to protect against multicollinearity.

Variable Selection

Bernoulli Regression for Zero-Inflated Component

	AdCan DB Steps	AIC	SSPE
0	Initial	87,663.22	13,846.82
1	Drop log_acs_hu_ratio	87,661.52	13,846.78
2	Drop log_gc_sum	87,660.09	13,846.44
3	Drop log_business_sum	87,659.91	13,845.82
4	Add log_mafsrc1_sum	87,397.60	13,806.06
5	Add log_compcity1_sum	87,328.73	13,793.74
	Supplemental Steps	AIC	SSPE
1	Add log_forest*_pct	86,843.09	13,702.57
2	Add log_irs1040ng	86,333.96	13,613.40
3	Add log_pct_crowd_occp_u	86,032.09	13,565.01
4	Add log_crops_pct	85,830.21	13,527.67
5	Add log_dsf_si_spr09	85,669.60	13,503.96
6	Add log_shrub_pct	85,544.47	13,481.73
7	Add log_devel*_pct	85,457.89	13,466.32
8	Add stability_index	85,381.30	13,454.58
9	Add hu_block2tract_ratio	85,330.91	13,445.67
10	Add log_pct_pop_0_17	85,282.20	13,436.57
11	Add log_irs1040nb	85,146.95	13,409.99

Variable/Group Definitions

- log_devel*_pct: log_devel0_pct, log_devel1_pct, log_devel2_pct, log_devel3_pct
- log_forest*_pct: log_forest1_pct, log_forest2_pct, log_forest3_pct

Variable Selection

Bernoulli Regression for Zero-Inflated Component

	Supplemental Steps	AIC	SSPE
12	Add log_irs1040nm	84,996.35	13,386.86
13	Add log_htc	84,920.42	13,374.92
14	Add log_pct_mlt_u_10p_str	84,875.15	13,368.77
15	Add log_pct_not_single_u_strc	84,827.39	13,360.71
16	Add log_pct_black	84,785.28	13,352.51
17	Drop log_hu_density_ratio	84,783.29	13,352.50
	Interaction Steps	AIC	SSPE
1	Add I1	84,516.08	13,305.19
2	Add I2	84,364.66	13,276.19
3	Add I3	84,217.50	13,244.43
4	Add I4	84,095.73	13,226.09
5	Add I5	83,973.47	13,202.67
6	Add I6	83,898.05	13,190.87
7	Drop urbanZERO	83,898.05	13,190.87
8	Drop teaUER	83,902.17	13,191.83

Variable/Group Definitions

- I1: log_compcity1_sum:log_devel1_pct
- I2: log_dep_list:log_dsf_si_spr09
- I3: log_landmeters2:log_dsf_si_spr09
- I4: log_delptypeBk_sum:log_dsf_si_spr09
- I5: log_dsf_si_spr09:log_irs1040nm
- I6: log_devel2_pct:log_irs1040nb

Final Drop1

Bernoulli Regression for Zero-Inflated Component

Drop	AIC	SSPE	GVIF
<FULL MODEL>	83,902.17	13,191.83	--
log_landmeters2	83,900.30	13,191.93	8.30
log_irs1040nm	83,900.33	13,191.78	2.52
log_compcity1_sum	83,908.40	13,192.75	15.29
hu_block2tract_ratio	83,910.27	13,194.03	2.73
hasSeasonalY	83,915.18	13,194.58	1.05
log_unitstat1_sum	83,917.88	13,193.28	19.97
teaMOM	83,935.48	13,197.91	1.73
log_dep_list	83,937.69	13,195.97	17.12
...	...	...	...
log_crops_pct	84,068.26	13,224.24	1.88
I5	84,073.43	13,225.13	26.48
log_irs1040nb	84,078.61	13,223.37	2.07
stability_index	84,084.10	13,223.06	2.74
I6	84,159.79	13,241.89	6.40
log_pct_crowd_occp_u	84,173.78	13,237.98	1.99
log_irs1040ng	84,188.03	13,246.17	1.89
log_dsf_si_spr09	84,586.77	13,312.72	56.23
log_delptypeBk_sum	84,722.85	13,323.03	19.97

(34 variables were selected)

Variable Selection

Negative Binomial Regression Count Component

	AdCan DB Steps	AIC	SSPE
0	Initial	177,405.4	2,241,029
1	Add log_mafsrc2_sum	177,403.8	2,212,585
	Supplemental Steps	AIC	SSPE
1	Add stability_index	176,023.2	2,322,641
2	Add log_irs1040ng	175,476.9	2,212,165
3	Add log_irs1040nb	175,000.5	2,187,939
4	Add log_devel*_pct	174,781.9	2,151,479
5	Add log_crops_pct	174,461.2	2,146,343
6	Add log_pct_crowd_occp_u	174,254.6	2,131,212
7	Add log_pct_pop_0_17	174,123.6	2,131,989
8	Add log_pct_not_single_u_strc	173,944.1	2,122,462
9	Add log_forest*_pct	173,859.5	2,108,162
10	Add log_dsf_si_spr00	173,724.2	2,124,208
11	Add log_shrub_pct	173,626.2	2,123,697
12	Add log_dsf_si_spr09	173,467.5	2,180,394
13	Add pct_unemploy_zero	173,387.2	2,167,083

Variable/Group Definitions

- log_devel*_pct: log_devel0_pct, log_devel1_pct, log_devel2_pct, log_devel3_pct
- log_forest*_pct: log_forest1_pct, log_forest2_pct, log_forest3_pct

Variable Selection

Negative Binomial Regression Count Component

	Supplemental Steps	AIC	SSPE
14	Add log_pct_li_hh_indo_europe	173,288.6	2,166,127
15	Add log_irs1040nm	173,201.4	2,165,367
16	Add log_pct_mlt_u_2p_strc	173,122.0	2,173,598
17	Add realtrac_*_2007	173,027.2	2,189,715
18	Add log_pct_api	172,969.8	2,193,304
19	Add uni_dist*	172,919.6	2,198,854
20	Drop log_acs_hu_ratio	172,918.1	2,199,715
21	Drop uni_dist3	172,916.6	2,200,831
22	Drop urbanZERO	172,915.8	2,201,206
23	Drop realtrac_6_10_2007	172,915.2	2,201,698
24	Drop uni_dist5	172,914.8	2,201,086
25	Drop uni_dist1	172,916.3	2,201,842
26	Drop uni_dist4	172,920.0	2,200,246
	Interaction Steps	**AIC**	**SSPE**
1	Add I1	172,501.6	2,208,588
2	Add I2	172,322.2	2,204,527
3	Add I3	172,150.1	2,195,509
4	Add I4	172,031.7	2,116,908

Variable/Group Definitions

- realtrac_*_2007: realtrac_1_5_2007, realtrac_6_10_2007, realtrac_11plus_2007
- uni_dist*: uni_dist0, uni_dist1, uni_dist2, uni_dist3, uni_dist4, uni_dist5
- I1: log_dep_list:log_devel1_pct
- I2: log_landmeters2:log_dsf_si_spr00
- I3: log_unitstat1_sum:log_hu_density_ratio
- I4: log_eds_res_sum:stability_index

Final Drop1

Negative Binomial Regression for Count Component

Drop	AIC	SSPE	GVIF
<FULL MODEL>	172,031.7	2,116,908	--
log_landmeters2	172,039.5	2,109,846	9.29
log_business_sum	172,040.1	2,111,715	2.26
Intercept	172,054.5	2,120,100	--
teaUER	172,057.9	2,117,310	1.11
log_gc_sum	172,059.1	2,117,925	31.04
teaMOM	172,068.8	2,115,589	1.86
uni_dist*	172,071.9	2,107,489	1.11
...	...	...	...
I3	172,304.0	2,117,005	4.97
log_devel*_pct	172,315.1	2,128,059	19.75
log_pct_pop_0_17	172,350.5	2,124,477	9.59
log_hu_density_ratio	172,362.5	2,089,070	8.33
I4	172,407.5	2,128,475	8.16
stability_index	172,467.1	2,147,184	2.90
log_irs1040ng	172,475.9	2,122,896	1.93
has_delptypeBk	172,593.4	2,159,779	1.53

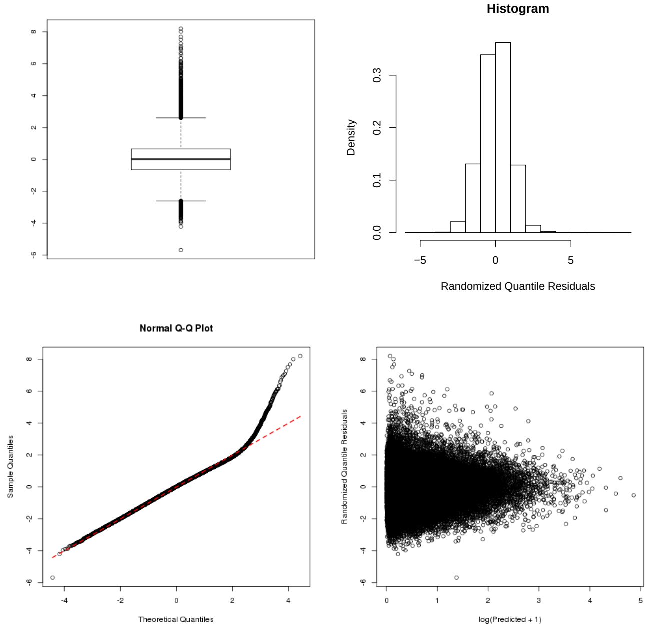
(37 variables were selected)

Resulting ZINB Model

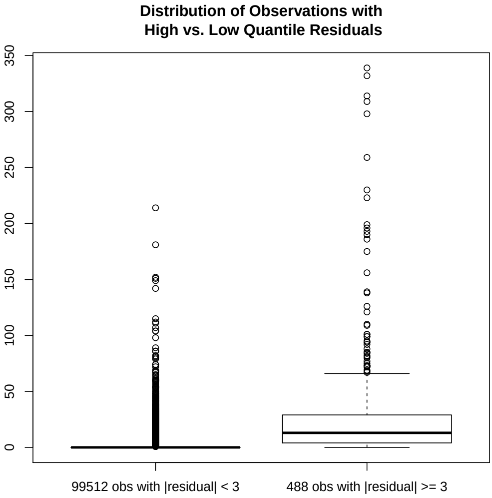
Count Coefficient	Estimate	SE	95% CI Lo	95% CI Hi
Intercept	0.6101	0.0980	0.4180	0.8022
log_dep_list	-0.6519	0.0369	-0.7243	-0.5796
log_landmeters2	-0.0226	0.0113	-0.0448	-4e-04
...	...	...	...	...
log_landmeters2:log_dsf_si_spr00	0.0480	0.0033	0.0415	0.0545
log_unitstat1_sum:log_hu_density_ratio	0.0670	0.0048	0.0576	0.0765
log_eds_res_sum:stability_index	0.2609	0.0401	0.1823	0.3394
ZI Coefficient	**Estimate**	**SE**	**95% CI Lo**	**95% CI Hi**
Intercept	0.0221	0.2162	-0.4016	0.4459
log_dep_list	-0.1813	0.0463	-0.2721	-0.0904
log_landmeters2	0.0888	0.0286	0.0327	0.1450
...	...	...	...	...
log_dsf_si_spr09:log_delptypeBk_sum	0.1493	0.0239	0.1024	0.1962
log_dsf_si_spr09:log_irs1040nm	-0.1092	0.0131	-0.1350	-0.0835
log_irs1040nb:log_devel2_pct	0.0384	0.0052	0.0282	0.0486
Dispersion	1.9918	0.0328	1.9276	2.0560

		ZINB	NegBin	Poisson
Training Set	LogLik	-83,113	-85,971	-152,561
	AIC	166,393	172,032	305,210
	BIC	167,192	172,460	305,629
Universe	SSPE	235,779,143	240,267,978	232,626,457
	MSPE	36.0567	36.7432	35.5746
	APE	6,897,446	7,054,974	6,900,898
	MAPE	1.0548	1.0789	1.0553

Randomized Quantile Residuals (Dunn and Smyth, 1996) for training set.



Resulting ZINB Model



Conclusions

- After exhaustive variable selection, some of the add activity from 2010 AdCan is not well-explained by our model.
- Full details are being assembled into a report (Raim and Gargano, 2015).
- A more automated method of variable selection would be desirable before considering other variables and data sources.
- Other models can be considered to handle extra variation in the absence of stronger predictors:
    1. Finite mixtures of regressions, and related distributions.
    2. Models for spatial dependence.

References I

Antonio Bruce and J. Gregory Robinson. Tract level planning database with census 2000 data. 2004.

Peter K. Dunn and Gordon K. Smyth. Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, 5(3):236–244, 1996.

John Fox and Georges Monette. Generalized collinearity diagnostics. *Journal of the American Statistical Association*, 87(417):178–183, 1992.

Joseph M. Hilbe. *Negative Binomial Regression*. Cambridge University Press, 2 edition, 2011.

Collin Homer, Jon Dewitz, Joyce Fry, Michael Coan, Nazmul Hossian, Charles Larson, Nate Herold, Alexa McKerrow, J. Nick VanDriel, and James Wickham. Completion of the 2001 National Land Cover Database for the conterminous United States. *Photogrammetric Engineering and Remote Sensing*, 73(4):337 – 341, 2007.

Andrew M. Raim and Marissa N. Gargano. Selection of predictors to model coverage errors in the Master Address File, In Progress, 2015.

Derek S. Young, Andrew M. Raim, and Nancy R. Johnson. Zero-inflated modeling for characterizing coverage errors of extracts from the U.S. Census Bureau's Master Address File, Submitted, 2015.