An Evaluation of Housing Unit and Block Cluster Effects on Small Area Census Coverage Variability in the 2006 Census Test

Donald Malec[1]

[1]U.S. Census Bureau

Disclaimer: This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed on statistical, methodological, technical, or operational issues are those of the author and not necessarily those of the U.S. Census Bureau.

Acknowledgments: The author would like to thank Aaron Gilary for initial data analysis and for data support and Doug Olson for detailed explanations of the 2006 census test and for providing auxiliary census data needed for the project. Useful comments and discussions provided by Doug Olson, Tom Mule, Harland Shoemaker and other members of the 2010 Census Coverage measurement team are also appreciated.

Abstract

Models are developed for the correlation effect of block clusters and housing units on the capture/recapture model of Census coverage measurement using data from 2006 Census test. The purpose of understanding this possible correlation is two-fold: 1) as noted by Malec and Maples (2005) a model for within small area variability is needed because design based estimates may be imprecise. Also, using 2000 census coverage information, Keller (2008) has shown evidence of both variable census capture rates and correct enumeration rates between block cluster and 2) the heterogeneity effects of between correlation due to block clusters and housing units on estimates of coverage have not been thoroughly evaluated. Random effects and the choice of a random effects model will be evaluated as a tool for measuring variability and correlation at these small levels.

Key Words: Unit level model

1. Introduction

The use of logistic models with person-level covariates is currently being planned for national-level estimates of the 2010 U.S. Census coverage (Mule 2010). One reasonable way to develop small area models for census coverage is to modify the national-level model to include small area effects. This was the approach used by Malec and Maples (2008) when developing a small area method for the Census 2000 coverage. As with most small area estimation models, the model can be characterized by a set of explanatory variables, a specification of error within the small area and a specification of variability between small areas (Rao 2003, page 4). This paper is concerned only with specification of the variability within a small area. Future work may include development of other parts of a small area model for the 2010 Census coverage.

The specific concern here is that design-based estimates of within small area variance and covariance may be unstable so that the usual practice of regarding design-based estimates of variation in a model as fixed and known may adversely influence both estimates of the variability of a small area estimator as well as the amount of "borrowing strength." An early recognition of the problem that small areas or domains may have unstable variances and unstable means can be found in the Normal-theory work of Singh and Sedransk (1988). In that work, a statisical model was included for the unknown sampling varainces. In this work, a model for the variances is not specified. Instead, data analysis is used with the aim of modeling the underlying variability of the binary data. Randomization tests are proposed and used as a first step to study the effect of clustered responses on this variability.

The interest in looking more carefully at the within small area estimation for census coverage arose as follows. Malec and Maples (2008) proposed a binomial model using a logistic model with added Gaussian random effects to account for small area variability (e.g. Rao (2003), sec 5.6), with an adjusted sample size to model match status and the correct enumeration status within a small area. They used a separate model for each of these two types. The model for small area, k, and domain, i, took the following form:

$$m_{ki} \sim binomial(n_{ki}, p_{ki}),$$

where n_{ki} was the sample size adjusted by an estimate of the miss-specification effect (i.e. $n_{ki} = \hat{p}_{ki}(1 - \hat{p}_{ki})/\hat{V}_D(\hat{p}_{ki})$ where $\hat{p}_{ki}$ is the design-weighted rate and $\hat{V}_D$its design-based variance estimate). Direct estimates of the adjusted sample size appeared very unstable due to the underlying small samples they were based on. Ultimately, the effects were smoothed using the following model of adjustment factors:

$$\log(\hat{V}_D(\hat{p}_{ki})) - \log(\hat{p}_{ki}(1 - \hat{p}_{ki})/n'_{ki}) = a_{ki} + \log\{(1 + (\bar{b}_{ki} - 1)\rho_{ki})\} + e_{ki},$$

where b_{ki} is the primary sampling unit (PSU) size of observations in small area k, domain i, n'_{ki} is the actual sample size, a_{ki} and ρ_{ki} are unknown parameters to be estimated, in addition to the definitions above. The e_{ki} denotes Gaussian error with unknown variance. After the parameters were estimated, the model was used to smooth out the sample sizes as a way to adjust for the sample design. However, the approach of Malec and Maples is inadequate because it is based on the assumption that an adjusted binomial distribution adequately describes the within actual small area sampling distribution. In addition, it used estimates of parameters without accounting for their error. The goal of this paper is to develop a systematic, more defensible method to account for the distribution of the direct estimates of small area sample rates. The approach described here builds on the work in Malec and Maple, extending the model below the small area level. At the block cluster level, using 2000 census coverage information, Keller (2008) has already shown evidence of both variable census capture rates and correct enumeration rates. Here, both housing unit and block clusters are evaluated using 2006 Census Test data.

2. The 2006 Travis County Texas Test site

The data used to evaluate the within small area variability is from the 2006 Travis county Texas Test site. The sample consists of approximately two hundred census blocks and the Travis county Test site is used to represent one, single small area. For more details on the 2006 coverage measurement test see Shoemaker et al. (2006).

The basic components of the census 2010 coverage includes includes components similar to those used in previous census coverage estimates which use a capture/recapture/correct enumeration approach. The basic components are:

1. data-defined (dd) rate - based on the proportion of data-defined (non-imputed) census enumerations,
2. correct enumeration (ce) rate - based on the proportion of data-defined census enumerations that are correct and
3. the census capture (match) rate - based on the proportion of persons in the follow-up sample who are also in the census.

Note that, because the current plans for National-level coverage estimates entail estimating the data defined rate via a logistic model the same approach will be applied here.

3. Randomization Tests of Block Effects and Housing Unit Effects

Each of the three rates described in section 2 are based on binary outcomes and each can be modeled. A logistic model that includes fixed effects which represent standard demographic characteristics, an effect for being in block j, and an effect for being in housing unit k of block j is specified as follows:

$$E(\text{binary outcome}|\underline{x} \text{ in block } j, \text{ housing unit } k) = e^{\underline{x}'\underline{\beta}+\mu_j+\alpha_{jk}}/(1+e^{\underline{x}'\underline{\beta})+\mu_j+\alpha_{jk}})$$

Block effects and housing unit effects are tested in sequence. That is, the presence of a block effect is tested first, assuming no housing unit effects. Next, the best estimate of a block effect is substituted into the model, and the presence of a housing unit effect is tested. The following summarizes the test procedure:

step 0: Fit the covariate-only fixed effects model

Based on the model:
$E(\text{binary outcome}|\ell(\underline{x}), \mu) = e^{\underline{x}\underline{\beta})}/(1+e^{\underline{x}\underline{\beta}})$,
fit the largest fixed effects model that excludes block effects and housing unit effects. Start with the main effects model determined by Olson and Springer (2008). Add the design strata as additional effects[1] and as many interactions as are estimable. Overfitting the model, in this way, helps avoid the possibility that either block effects or housing unit effects only compensate for important covariates that were left out.

step 1: Include block fixed effects model into covariate only model

Use the estimate of $\underline{\beta}$, determined in step 0, above, to form the offset:$\ell(\underline{x}) = \underline{x}\hat{\beta}$. Then, obtain the maximum likelihood estimate (MLE) of the μ_js using the model:

[1] One of three strata had two additional weighted classes which were not used as additional factors in the model.

$$E(\text{binary outcome}|\underline{x}\text{in block } j, \text{housing unit } k) = e^{\ell(\underline{x})+\mu_j}/(1+e^{\ell(\underline{x})+\mu_j})$$

These estimates will be of poor quality due to small sample size. However, they will only be used as a group to form test statistics. Specifically, two statistics are formulated and used:

LargeMLE - the percent of MLEs that are either larger than 15 or smaller than -15 and

AbsMLE - the average of the absolute values of the the MLEs that are within ± 15

Large values for either of these statistics indicate the presence of block effects.

step 2: Test the null hypothesis of no block effect

Assume that all block effects μ_j, are zero. Use the covariate-only fixed effects model and the offset (from step 0) to generate a new sample. Estimate the MLEs of the block effects for each new sample (keeping the offset $\ell(\underline{x})$ fixed), and compare the distribution of the two test statistics. Based on 10,000 samples from the distribution under the null hypotheses, the proportion of test statistics that are more extreme than the observed statistics (using the actual data) is as follows:

	LargeMLE		AbsMLE	
	observed	p-value	observed	p-value
census capture (match) rate	21	$< .0001$	.98	$< .0001$
correct enumeration (ce) rate	22	$< .0001$	.67	$< .0001$
data-defined (dd) rate	92	$< .0001$	1.19	$< .0001$

This indicates a significant block effect for all three rates.

step 3: Continuation: testing for housing unit effects

First, substitute the MLE's of the block effects (using ± 15 for the MLEs that are out of range) as part of the offset. Then, as with block effects, form two test statistics: one the percentage of housing unit MLEs outside of ± 15, and the other the average absolute value of housing unit MLEs that fall between ± 15.

step 4: Test the null hypothesis of no housing unit effects

As with blocks, generate new samples with zero housing-unit effects and compare resulting housing unit effect MLEs. In this case, an offset which includes the estimated block effect MLEs is used.

	LargeMLE		AbsMLE	
	observed	p-value	observed	p-value
match rate	88.8	$< .0001$	1.20	$< .0001$
ce rate	85.7	$< .0001$	95.5	$< .0001$
dd rate	71.1	≈ 1	.003	≈ 1

Housing unit effects for both capture rate and correct enumeration rate also appear to be significant indicating a need to include a housing unit in the model. The data-defined housing unit effects were not significant. Note that most of the block-level MLEs were out of range for the data-defined model, leaving few housing units that did not not have extreme offsets left to model.

4. Randomization tests of normality of block effects and housing unit effects

Based on the results of the randomization tests which included no block effect and no housing unit effect, this study determines that both block effects and housing unit effects should be included in the model. However, since sample sizes are too small to include block effects and housing unit effects as fixed effects, the shrinkage estimator technique of modeling them as random effects with mean 0 will be used. Since there is no reason to assume that these random effects are Normally distributed, randomization tests will be used to evaluate this assumption.

The procedure used is the following.

For block effects, if the block effects are normally distributed, the distribution of the MLEs will reflect it. Use the MLEs for the fixed block effect block model determined in section 3 to form test statistics of this new hypothesis of normality. Specifically, the block level effects will be assumed to be normally distributed with a single, unknown variance. New samples based on the normal assumption with estimated variance component are generated and the resulting distribution of MLEs are compared to the distribution of observed MLEs. The following steps detail the procedure:

Step 0: Use the block MLEs obtained in section 3 and determine the following features of their distribution: 1) the percentage of MLE's that are less than or equal to -15, the percentage of MLEs that were greater than or equal to 15 and, of the MLEs that were between ± 15: their minimum value, their first quartile, their median, their third quartile and their maximum. These six statistics, obtained from the observed data will be used as test statistics.

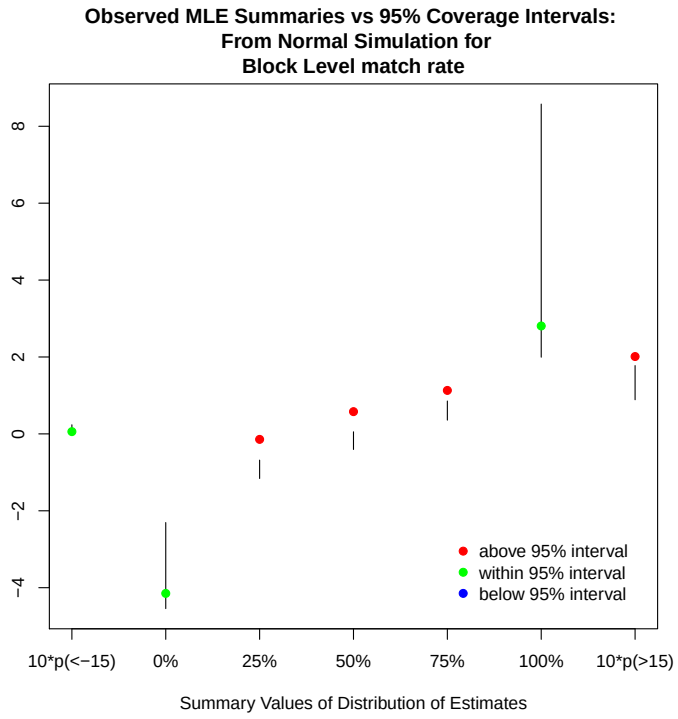
Step 1: Obtain the restricted maximum likelihood estimate (RMLE) of the variance of the assumed normally distributed random block effect distribution using "glmer" from the "lme4" package of the R-System (2005).

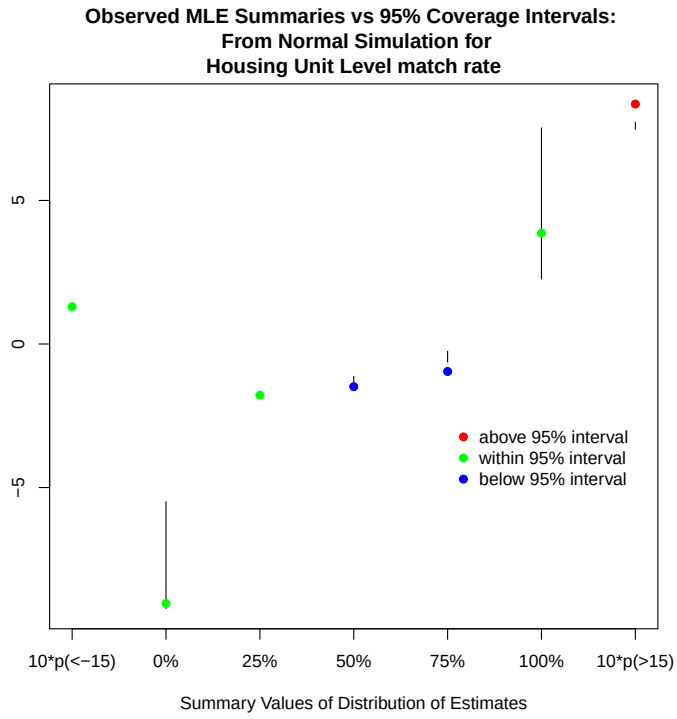
Step 2: Generate new data sets by first generating new block effects based on a normal distribution with a variance component equal to the RMLE and then generating binary data from the logistic model with combined offset and block-level random effect.

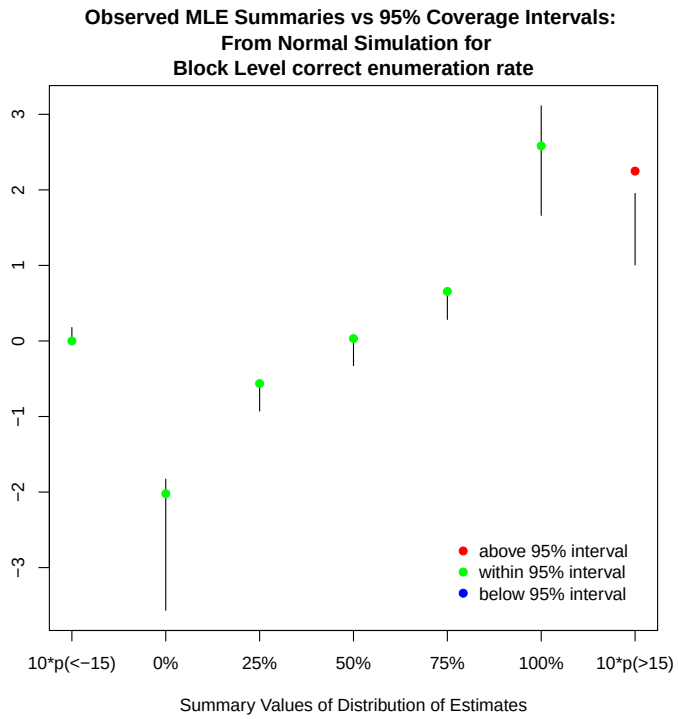
Step 3: Obtain the Block level MLE under the fixed block effect model and obtain the distributional summaries of these new MLEs (with the artificial, normally distributed random effects inserted into the model).

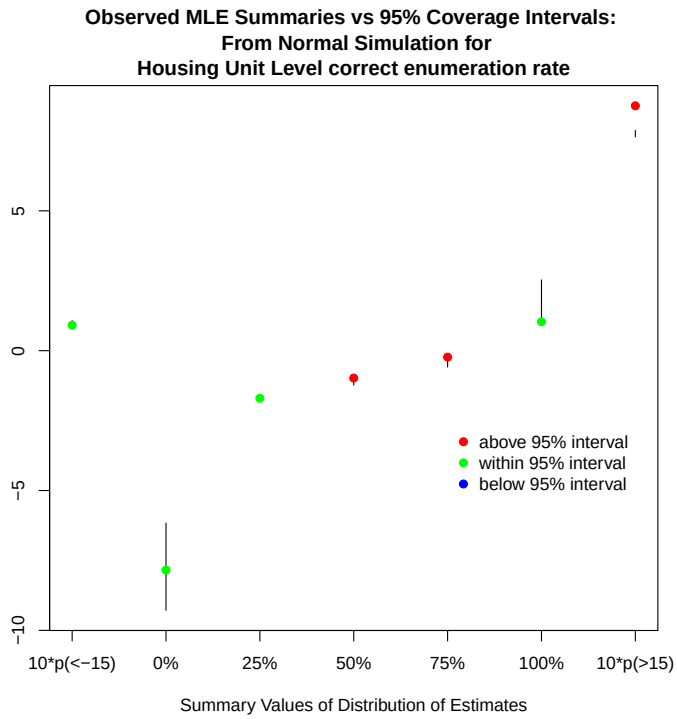
Step 4: Compare the observed MLEs against those based on the normal assumption. If the normal assumption is justifiable, the resulting MLEs based on the normal assumption should have a distribution close to the observed MLEs.

The following plots show the results for each distribution summary. Each vertical line represents the 95% probability interval based on the generated data. As shown in the graphs below, the observed data does not follow the hypothesized data very well, bringing the Normality assumption into question. A similar approach with similar conclusions (details not included) was taken for the housing unit effects after using the offsets that include block effects, as in section 3. The data-defined rates were not evaluated.


Observed MLE Summaries vs 95% Coverage Intervals:
From Normal Simulation for
Block Level match rate
above 95% interval
within 95% interval
below 95% interval
10*p(<−15)
0%
25%
50%
75%
100%
10*p(>15)
Summary Values of Distribution of Estimates


Observed MLE Summaries vs 95% Coverage Intervals:
From Normal Simulation for
Housing Unit Level match rate
above 95% interval
within 95% interval
below 95% interval
10*p(<−15)
0%
25%
50%
75%
100%
10*p(>15)
Summary Values of Distribution of Estimates


Observed MLE Summaries vs 95% Coverage Intervals:
From Normal Simulation for
Block Level correct enumeration rate
above 95% interval
within 95% interval
below 95% interval
10*p(<−15)
0%
25%
50%
75%
100%
10*p(>15)
Summary Values of Distribution of Estimates


Observed MLE Summaries vs 95% Coverage Intervals:
From Normal Simulation for
Housing Unit Level correct enumeration rate
above 95% interval
within 95% interval
below 95% interval
10*p(<−15)
0%
25%
50%
75%
100%
10*p(>15)
Summary Values of Distribution of Estimates

5. Use of estimates of sampling frame capture rate for model evaluation

Although more difficult to define and estimate in the 2010 Census coverage evaluation, the capture rate in the coverage sample frame (ce-match) can be estimated in the 2006 Test site (due to the use of the same housing unit address list for both the Census and coverage survey). Randomization tests of no block effects, no housing unit effects, Normal Block random effects, and Normal Housing unit random effects can all be performed in the same manner as before.

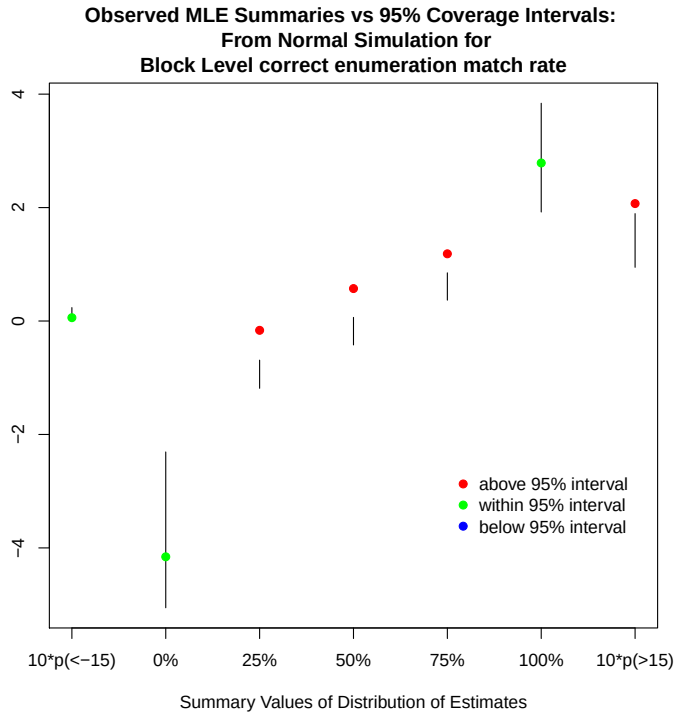
Testing of no block effect:

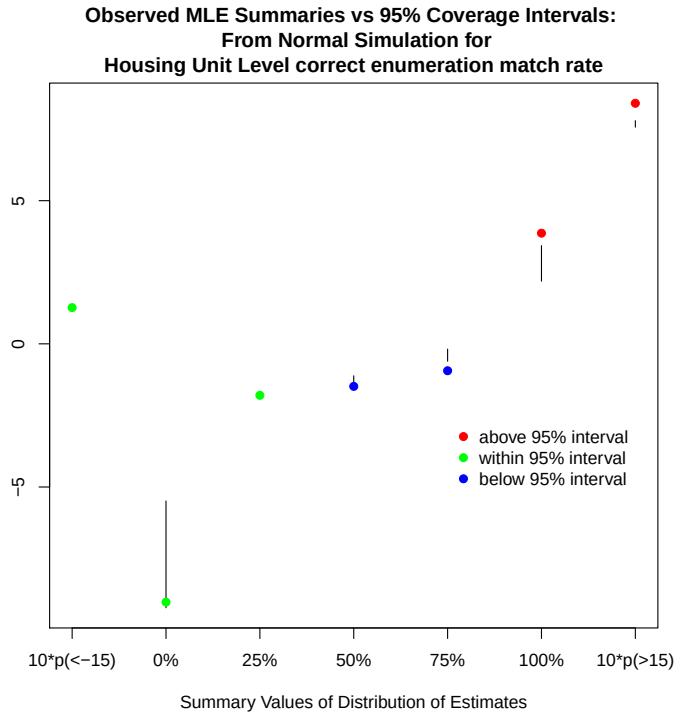
	LargeMLE		AbsMLE	
	observed	p-value	observed	p-value
ce-match	21	$< .0001$	.99	$< .0001$

Testing no housing unit effect:

	LargeMLE		AbsMLE	
	observed	p-value	observed	p-value
ce-match	88.7	$< .0001$	1.14	$< .0001$

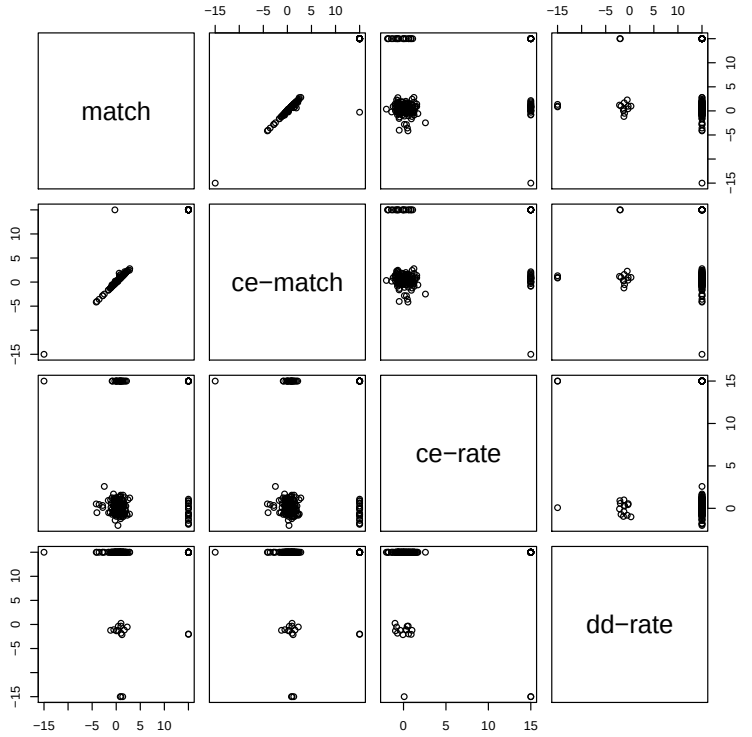
Testing the normality of block effects and housing unit effects, respectively:





6. Sample correlation between estimated components

Including random effects for each component still leaves the unanswered issue of whether or not they are independent of each other. The following figure plots the estimates of block effects for the four types of outcomes.



Most appear uncorrelated, but the two coverage rates (the census capture rate (match) and the followup-survey capture rate (ce-match)) are highly correlated. Although not substantiated here, this correlation between coverage rates suggests a hypothesis that the underlying block level capture rates may also be correlated, which could cause heterogeneity bias in the estimates if not accounted for.

7. Summary and Conclusion

In summary, the existence of block effects and housing unit effects were evaluated using randomization tests with fixed effect MLEs as test statistics. Both block and housing unit effects were found to be significantly different from zero. Additionally, the Normality assumption of the census capture rate, the census correct enumeration rate, and the capture rate of the coverage survey frame does not appear warranted. The correlation between effects was briefly evaluated using scatterplots, and it was noted that the sample estimates of the two capture rates appeared highly correlated while the others did not. Although this is not conclusive proof that the underlying capture rates are correlated, it does show that further evaluations are merited, since heterogeneous capture can result in bias.

In conclusion, a fully satisfactory unit level model requires more work. So far, a need for non-normally distributed random effects and possibly a need for correlation between certain random effects is indicated. The benefit of additional random effects (such as interactions with block level or housing unit effects) has not

been tested yet. In a different but related census coverage context in the 2000 Census, the use of additional housing unit effects was been reported by Olson (2009).

Perhaps even more problematic is the modeling of shifts in the small area due to time differences between census and coverage survey and to differences due to housing unit mapping. Although the population at the national level can be corrected to account for births and deaths between the census and coverage survey and for movers in and out of the country, the effects of movers as well as the effects of geographic miss-classification at the small area level will be hard to evaluate and will need to be based on a possibly incomplete data model.

REFERENCES

Keller, Andrew (2008), "Assessing synthetic error via Markov chain Monte Carlo techniques ASA Proceedings of the Joint Statistical Meetings," American Statistical Association (Alexandria, VA), 2017-2024.

Malec, Donald and Maples, Jerry (2008), "Small area random effects models for capture/recapture methods with applications to estimating coverage error in the U.S. Decennial Census," Statistics in Medicine, 27, 4038-4056.

Mule, Vincent Thomas , Jr. (2010), "U.S. Census Coverage Measurement Survey Plans," ASA Proceedings of the Joint Statistical Meetings, American Statistical Association (Alexandria, VA).

Olson, Douglas and Michael Springer, "2006 Census Coverage Measurement: Net Error Modeling Pseudo-Estimates," DSSD 2010 CENSUS COVERAGE MEASUREMENT MEMORANDUM SERIES #2006-E-09; July 22, 2008.

Olson, Douglas (2009),"A Three-Phase Model of Census Capture," ASA Proceedings of the Joint Statistical Meetings, 5107-5116, American Statistical Association (Alexandria, VA).

Rao, J.N.K. (2003) Small Area Estimation, John Wiley and Sons.

R Development Core Team (2005), R: A language and environment for statistical computing, reference index version 2.9.2. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Shoemaker, Harlan H. Jr, Andrew Keller and Tamara Adams, "2006 Census Coverage Measurement: Design of the Coverage Measurement Program for the 2006 Census Test," DSSD 2010 CENSUS COVERAGE MEASUREMENT MEMORANDUM SERIES #2006-B-02; June 28, 2006.

Singh, Bahadur and J. Sedransk (1988), "Variance Estimation in Stratified Sampling When Strata Sample Sizes Are Small," Sankhya: The Indian Journal of Statistics, Series B, Vol. 50, No. 3 (Dec., 1988), pp. 382-393.