Applications of Statistical Models that Detect Daily Changes Using Key Estimates from the American Community Survey Due to the U.S. Census Bureau Regional Office Restructure[1]

Lindsay McMillan, Robyn Sirkis
US Census Bureau, 4700 Silver Hill Road, Washington DC 20233

Abstract
In 2012, the U.S. Census Bureau shifted field operations from twelve to six regional offices (ROs). To monitor this shift in management, we built models for key variables for many different surveys to determine if there were significant changes in our estimates. Models were run on unedited Computer Assisted Personal Interview (CAPI) data on a daily basis. This paper discusses the models used for the American Community Survey (ACS). First, we will give a brief overview of the scope of this project and discuss the key variables, predictor variables, and how models were built using 2011 ACS unedited data. Second, we discuss the automated system used to generate graphs for each key variable. These graphs track the daily change on the coefficient for the management structure indicator variable from the models. Finally, we present the results of some of our models and discuss possible ways to improve/use these models in the future.

Key Words: Data Manipulation, Responsive Design, American Community Survey

1. Introduction

To reflect innovations in technology and current developments in survey methodology, the U.S. Census Bureau proposed a realignment of its national field office structure to begin in 2012. The regional office (RO) structure has remained considerably unchanged since 1961. The goal of the realignment was to minimize the cost of survey operations, increase responsiveness, and improve the quality of the surveys conducted by the U.S. Census Bureau. The changes resulted in the permanent closing of six of the organization's twelve ROs (U.S. Census Bureau, 2011).

As part of this restructuring, one of the tasks included creating a system to assess whether changes in the RO management structure had an affect on key estimates for demographic household surveys. The responses for the key variables were monitored daily from January through December 2012. In this paper, the focus is placed on the American Community Survey (ACS). Statistical models were constructed in which one of the covariates indicated whether the interview was conducted under the new or old

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

management structure. Graphs were generated to display the change in the coefficient of the management structure indicator variable over time for each ACS key statistic. A significant effect occurred if the coefficient for the key variable was out of the confidence bounds in the graph.

2. Background

2.1 Regional Office Realignment

The realignment of the national field office structure involved updating the number, size, geographic coverage and staffing of the ROs. Throughout 2012, the closing ROs in Boston, Charlotte, Dallas, Detroit, Kansas City, and Seattle slowly transitioned their field operations to the remaining ROs in Atlanta, Chicago, Denver, Los Angeles, New York, and Philadelphia. The transition occurred in seven waves (January, April, June, August, September, October, and November). The plan was to have minimal or no change in the responsibilities of the interviewers collecting the data in the field. However, the change had an impact on the supervisory structure. Within each of the six new ROs, there are eight Survey Statisticians in the Field (SSF) representing all Census Bureau demographic surveys in a given area. The SSFs each manage approximately twelve Field Supervisors (FS), and Field Representatives are supervised by FSs. Also, more of the supervisory staff work out of their homes. The new design has the potential to provide improved management information systems, maintain higher data quality and increased efficiency at lower costs.

2.2 American Community Survey Sample Design

The ACS is an ongoing survey that provides annual, three-year, and five year-estimates on demographic, socio-economic, and housing unit topics to help determine how federal and state funds are distributed each year.

The annual target sample size for the ACS is approximately 3.54 million housing units (Asiala, 2013). The ACS includes twelve independent monthly samples. Data collection for each independent sample is part of a three month panel. A questionnaire is mailed to the sample address in the first month of the panel. The second month includes a follow up with Computer-Assisted Telephone Interviewing (CATI) when a valid telephone number is available. The sample address may be selected for Computer-Assisted Personal Interviewing (CAPI) in the third month if the household refused to participate in the mail or CATI phase, or the address did not have a valid mailing address or telephone number (U.S. Census Bureau, 2009). Only data collected through the CAPI mode was used in the statistical models, to detect daily changes in ACS key estimates.

3. Methodology

3.1 Key Variables

The primary objective was to develop a model to assess whether changes in the RO management structure had an affect on ACS key estimates. Eleven key statistics were chosen by the sponsors to cover the social, demographic, economic, and housing characteristics of the ACS. Table 1 shows the key statistics and the response of interest used in the statistical models.

Table 1: Key Variables and Responses of Interest.		
Characteristics	**Key Variables**	**Response**
Housing	Building Description	Building with 2 Apartments
	Tenure	Rent
	Monthly Mortgage Amount	DK or Refused
Demographic	Year of Birth	DK or Refused
	Race	Some Other Race
Social	U.S. Citizenship	No
	Speak Another Language at Home Other than English	Yes
	Marital Status	Never Married
	Did you Live in this Residence 1 Year Ago?	No
Economic	Worked Last Week	Yes
	Wages	DK or Refused

3.2 Data Used for Model Building

The covariates used in the statistical models were pooled from several sources. The two major sources of covariates included November and December 2011 ACS unedited CAPI data, and tract-level demographic and socioeconomic proportions obtained from the 2000 Decennial Census planning database. Several covariates were considered for the household level statistical models, such as the number of rooms in the household, the year the household was built, the type of building, housing unit tenure (ex. owned, rented), whether there was a business on the property, number of household members, percent of population below poverty level, and percent of population age 65+. Some of the person level statistical model potential covariates included age, sex, race, education, income, percent Hispanic, and percent of population age 0-17.

Only cases completed by the CAPI data collection method were used in the statistical models. Throughout the production phase, models used cumulative 2012 ACS unedited CAPI data and historical CAPI data from November and December 2011 to assess whether there were changes in the key variables due to the RO realignment.

3.3 Model Building

Logistic regression models were constructed to determine if there was a significant effect on the key variables of interest due, to the change in RO field structure. All models take on the general form:

$$logit(E(y_i|z_i,\tilde{c}_i)) = \beta_0 + \beta_z z_i + \tilde{\beta}_i \tilde{c}_i$$

where z_i = 1 if the case was collected in an SSF area under the new management and $z_i = 0$ otherwise; and $\tilde{c}_i$ is a vector of covariates

All of the key variables were coded as binary (0 or 1) where 1 represents the response of interest. Covariates were not included in the statistical model if it was a variation of the key variable. Only main effects were considered in the analysis. Issues relating to sample

size, missing values, and collinearity among predictor variables were taken into account when choosing an appropriate statistical model. Furthermore, missing values were placed into their own category and not used in the models.

To account for seasonality, monthly means for each key variable were included in the statistical models with the exception of the models involving *don't know and refusal* values, since the values were imputed and not available. These "lagged" means were calculated from 2010 ACS edited data for each SSF area by month. Base weights were included in the final statistical models to account for the sample design.

In addition to the ACS data, several variables that dealt with the RO restructure were incorporated into all the statistical models. These are described below.

1. *Current_wave* takes on the values from 1 to 7, indicating the seven phases of the realignment.
2. *Wave_RO_change* is a geography variable that identifies, for each tract, when the management change will take place.
3. *Z-variable* is an indicator variable of whether or not the case was completed under the new management structure.

The z-variable was included in the statistical models in both the testing and production stages. During the testing stage, the z-variable was included to simulate the management structure change. The testing stage occurred in 2011, while the actual RO structure change occurred in 2012. The goal was to choose a model that was both highly significant overall and had a z-variable that was statistically insignificant, small in absolute value, or both. Throughout the production stage, the z-variable indicated whether the case was conducted under the new or old management structure. The purpose of constructing the statistical model was to utilize the coefficient of the z-variable to determine if there was an affect on the key variable due to the change in the RO structure.

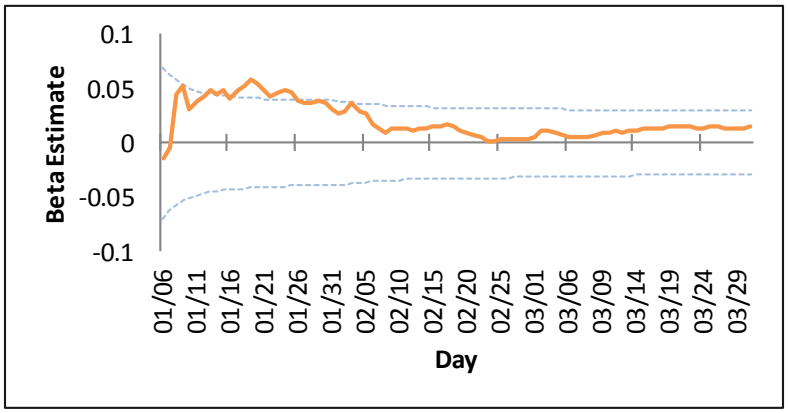
3.4 System and Graphs

An automated system was developed to upload data, run the models, and produce graphs showing the changes in the coefficient of the management structure over time for each key variable. Prior to producing the graphical representations, standard errors were calculated based on the z-variable for each key variable. In addition, validation checks were conducted on the key variables and covariates to ensure the values were within valid ranges. Once the statistical models were produced each day, the Oracle Application Express (Oracle APEX) web application was utilized to display the graphs. This developmental tool was also used to produce summary tables showing the beta coefficient, upper bound, lower bound, and an indicator on whether the z-variable was out of bounds. Each day, coefficients were computed in the models using cumulative data collected in 2012, including historical data from November and December 2011. Points were then added to graphs indicating the new value of the z-variable coefficient (Beta-z) in relation to its upper and lower bounds (95% confidence interval) under the hypothesis that the Beta-z value was equal to zero.

Figure 1 shows the Beta-z value from January 1 to March 30 for *Did you Live in this Residence 1 Year Ago (no)*. The solid line represents the Beta-z estimate for the key variable. The dashed lines represent the upper and lower 95% confidence bounds. A significant beta value occurred when the beta estimate was above or below the confidence bounds. As expected, the Beta-z was significant during the beginning of data

collection but as more cases were collected under the new management structure, it moved within its confidence bounds. Each day, the eleven ACS key variable graphs were viewed to determine if there were changes in the new field structure. Notifications were sent out to management to alert them if there were significant effects in the ACS key variables.

Figure 1: *Did you Live in this Residence 1 Year Ago (no) Beta-z.*[2]



4. Results

4.1 Statistical Model Fit Results

The predictor variables and fit statistics varied across the statistical models. Some of the statistical models had R-squared values between 0.2 and 0.45, while other models did not fit the data as well. R-squared values, for all models, remained fairly consistent throughout the year. The c-statistics for the models ranged between 0.56 and 0.93 with eight out of the eleven models having a value higher than 0.70. The statistical models are described for the *Housing Unit Tenure (rent)*, *Race* (*some other race)*, and *Wages (don't know or refused)*.

4.2 Graphs Demonstration

The majority of the ACS variables monitored did not have an effect due to the field management realignment. The Beta-z for some of the key variables was regularly outside of the confidence bounds, but the difference was fairly small. On any given day, it was expected that a few graphs would have significant Beta-z values, just by chance (approximately 5% on any given day). Some values were significant but steady due to model fitting difficulties. The Beta-z for a variable being monitored may be significant on a given day, especially when starting to collect data with z = 1 at the beginning of January, but then it slowly trends back toward zero. It trends slowly because these models are cumulative, in that the current day's data used in the model is the same as the previous days, with one more day's worth of collection. The graphs for the key variables *Housing Unit Tenure (rent)*, *Race* (*some other race)*, and *Wages (don't know or refused)* are described in this section.

[2] The data for all remaining figures and tables in the paper is from the American Community Survey

Housing Unit Tenure (rent)

The key variable, *Housing Unit Tenure (rent)* was one of the best fitting statistical models. Table 2 contains the covariates used in this model separated by categorical and continuous variables. Variables marked with an asterisk are common to all models.

Table 2: Covariates used in the model for *Housing Unit Tenure (rent)*.

Variable Type	Variable
Categorical	Building description
	Year built
	Wave_RO_Change*
	Current_wave*
Continuous	z-variable*
	Number of rooms
	Monthly mean for renters
	Percent of population below poverty
	Base weight

Figure 2 is the graphical display for *Housing Unit Tenure (rent)* from January 1 to December 31, 2012. The Beta-z value remained within the confidence limits for the entire year, indicating there was no effect due to the realignment of the ROs. The confidence limits were wider at the beginning of the time period; this is likely due to the small number of cases under the new management structure.

Figure 2: *Housing Unit Tenure (rent) Beta-z.*

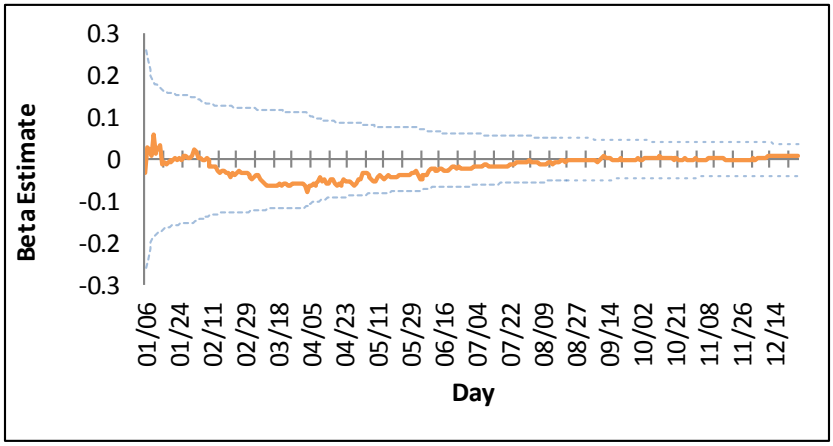


Table 3 contains some overall model evaluations to better assess the model fit. The data presented below included data from January 1 through December 31, 2012. All three statistical tests (Likelihood Ratio, Score, and Wald) indicated that the logistic regression model was more effective than an intercept only model. The R-squared value was 0.38 and the c-statistic was 0.87. It is also worth noting that the association statistics showed that the model for *Housing Unit Tenure* was indeed assigning higher probabilities to those who were renters.

Table 3: Final association statistics for *Housing Unit Tenure (rent).*			
Test	χ^2	df	**p-value**
Likelihood Ratio Test	112309.05	38	<.0001
Score Test	93278.76	38	<.0001
Wald Test	59193.48	38	<.0001
Association statistics: R-squared = 0.3836, Kendall's Tau-a = 0.365, Gamma = 0.731, Somers' D = 0.730, c-statistic = 0.865			

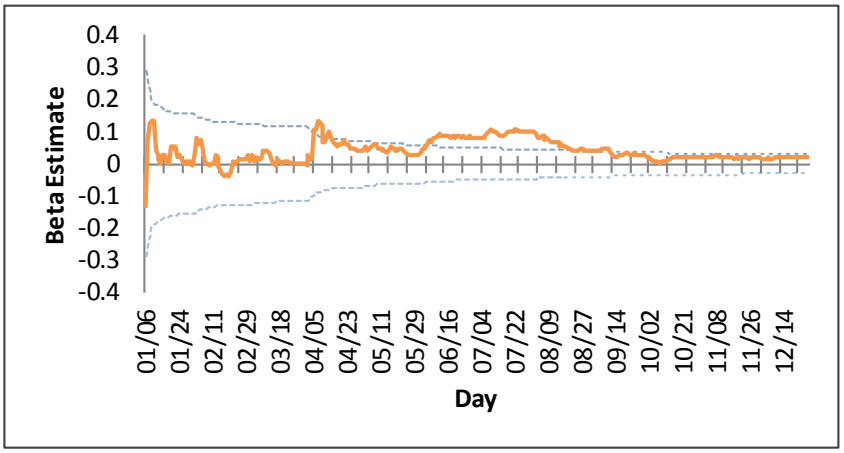
Race (Some Other Race)

Covariates for the key variable, *Race (some other race)*, are presented in Table 4. As in Table 2, variables are identified as continuous and categorical, with variables common to all models marked by an asterisk.

Table 4: Covariates used in the model, *Race (some other race).*	
Variable Type	**Variable**
Categorical	Speaks another language at home
	Educational attainment
	Citizenship
	Wave_RO_Change*
	Current_Wave*
Continuous	z-variable*
	Monthly mean for some other race
	Percent Hispanic
	Percent White
	Baseweight

During the first wave of data collection the Beta-z variable for Race (*some other race)* was somewhat unstable, but always close to the confidence limits. This was expected from this model as "some other race" was not selected during data collection as often as the other choices. Figure 3 shows that as wave 2 began, there was a sharp increase in the Beta-z variable causing it to become significant before stabilizing. Additionally, in wave 3 the Beta-z moved outside the confidence bounds, where it remained until late in wave 4. There was probably not an effect due to the change in management even though the Beta-z was significant during this time. It is plausible that the transitioned areas entered into the statistical model at the start of the wave had demographic characteristics different from the transitioned areas in previous waves. This would cause the Beta-z to move out of bounds. Eventually, the Beta-z became more stable later in the wave.

Figure 3: *Race (some other race) Beta-z.*



Final model evaluations for *Race (some other race)* are displayed in Table 5. As with the *Housing Unit Tenure* model, the three statistical tests indicated that the model chosen was better than an intercept-only model. The R-squared value for this model was 0.16 with a c-statistic of 0.87. The model seemed to fit the data considering the R-squared value together with the other association statistics.

Table 5: Association statistics for *Race* (*some other race*).

Test	χ^2	df	**p-value**
Likelihood Ratio Test	144800.88	52	<.0001
Score Test	162766.32	52	<.0001
Wald Test	100373.05	52	<.0001
Association statistics: R-squared = 0.1591, Kendall's Tau-a = 0.131, Gamma = 0.732, Somers' D = 0.729, c-statistic = 0.865			

Wages (Don't Know or Refused)

Table 6 contains the covariates used in the model for *Wages (don't know or refused)*. Variables marked with an asterisk are common to all models.

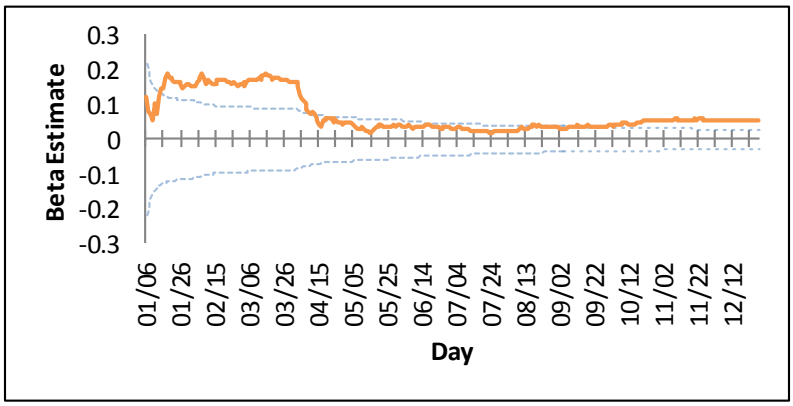
Table 6: Covariates for *Wages* (*don't know or refused*).

Variable Type	**Variable**
Categorical	Sex
	Wave_RO_Change*
	Current_Wave*
Continuous	z-variable
	Percent White
	Age
	Age Squared
	Baseweight

During the first wave of data collection for *Wages*, the Beta-z was significant the majority of the time. Figure 4 shows as more data was collected the Beta-z improved,

moving within the confidence intervals during wave 2 where it remained until the middle of wave 5. Even though the Beta-Z was out of bounds again in wave 5 through the end of the data collection, it was unclear if this was due to a change in management structure. It is possible that respondents living in areas transitioning into the model in waves 5 through 7 answered the *Wage* question more often than respondents in the earlier waves. More investigation into the causes would need to be conducted to verify this conclusion.

Figure 4: *Wages (Don't Know or Refused).*



Some final evaluation models are provided in Table 7. The three overall model evaluation tests (Likelihood Ratio, Score, and Wald) indicated the logistic model chosen was more effective than using an intercept only model. However, the association statistics for the *Wages* model implied the model was not assigning higher probabilities to those who answered the wage question with a *don't know or refused* response. In other words it appeared that although the model was significant it was not efficient in predicting the event outcome. T

Table 7: Association statistics for *Wages (don't know or refused).*

Test	χ^2	df	**p-value**
Likelihood Ratio Test	2947.44	21	<.0001
Score Test	2980.23	21	<.0001
Wald Test	2914.96	21	<.0001
Association statistics: R-squared = 0.0096, Kendall's Tau-a = 0.050, Gamma = 0.122, Somers' D = 0.121, c-statistic = 0.560			

5. Limitations

To give proper interpretation to the results, it is important to keep in mind limitations in the development of the evaluation methodology. The ACS is a large survey with a vast number of possible covariates to choose from when constructing the statistical models. Also the ACS data and estimates are subject to sampling and nonsamping error. Due to operational constraints, the team used variables that were highly correlated with the key variables. Then the 'forward selection' method was used to choose the covariates. It is possible adding or using different covariates could improve the models. Due to time constraints missing values for covariates were placed into a separate category, leaving the opportunity to develop an imputation procedure to better handle missing values.

Due to technological constraints, only historical CAPI data for November and December were used for model building and testing stages. It is conceivable that incorporating more data into the development of the statistical models could have resulted in models with better fit, especially for key variables that were considered rare events.

Since the change to the management structure did not start until January 1, 2012, it was important to incorporate certain variables, such as the z-variable and current_wave, into the statistical models that would not be available until the production stage. To build these variables into the statistical model, the management change was simulated during the testing stage using the historical data.

6. Future Research

Additional research includes using a different set of covariates in the statistical model, in an attempt to improve the fit statistics. Similarly, adding additional data to the models, such as CATI data, may also help reduce some of the day-to-day variability of the models and allow for a better fit.

Some of the covariates, such as educational attainment and year built contained many different response levels, some of which were very small. There is potential to resolve this issue by collapsing these extreme cells and replicating the models. In addition to collapsing the extreme cells, developing different methodology, such as imputation, for handling cases with missing values may also affect the statistical models.

Furthermore, some of our models (e.g., *some other race*) tended to have a significant Beta-z at the beginning of a wave but evened out by the middle/end of the wave. The areas that were transitioned into the model at the beginning of a wave may have different characteristics than those already in the model, causing the Beta-z to become significant for a short period of time. Exploring this hypothesis could lead to better model fitting and reducing day-to-day variability.

Finally, the team could explore different types of models, such as mixed models or multilevel models and determine if the key estimates were affected due to the change in the management structure.

7. Conclusion

In order to adapt to the new advances in technology and survey methodology it was necessary for the U.S. Census Bureau to make changes to the way data collection is conducted. One of the major changes was reducing the number of ROs from twelve to six in an effort to reduce the overall cost, and to improve the quality of the data. To monitor this transitional period the Data Monitoring Team developed a system to monitor key variables for different surveys. Generally, it was expected that a few graphs would have significant Beta-z values, just by chance on any given day. Some values were significant but steady due to model fitting difficulties. Therefore, it can be assumed that overall there is not an effect on the key estimates due to the RO realignment.

It was demonstrated in this paper the methodology used to develop the statistical models for the ACS key variables, and the system that was put in place to monitor the realignment in real time. Additionally, three models were explained in detail and were

chosen to represent the variety in results among the eleven key variables. Finally, future research was presented that could aid in future similar projects.

Acknowledgements

We would like to thank Reid Rottach and Carrie Lynch for overseeing the Data Monitoring Project and helping us meet our deadlines. We would also like to thank and congratulate the rest of the members of the survey modeling team for completing such a successful project. Finally we would like to thank Andre Harper and the rest of the IT specialists who developed the APEX monitoring system.

References

Asiala, M. (2013).” Topics on American Community Survey”, California Regional Affiliate Data Center Meeting. Retrieved from http://www.dof.ca.gov/research/demographic/state_census_data_center/meetings/documents/CASDC_AnnualMtg2012_Asiala-ACSUpdate.pdf

U.S. Census Bureau (2009). “Data Collection and Capture for Housing Units”. Retrieved from http://www.census.gov/acs/www/Downloads/survey_methodology/acs_design_methodology_ch07.pdf.

U.S. Census Bureau (2011). “U.S. Census Bureau Announces Field Management Reforms to Reduce Costs and Enhance Data Quality”. Retrieved from http://www.census.gov/newsroom/releases/archives/miscellaneous/realignment.html.