Using Statistical Process Control Techniques in the American Community Survey

Steven P. Hefter, Erica Marquette
U.S. Census Bureau
Washington, DC 20233

Abstract
Each year the American Community Survey collects data roughly two million housing units (HUs), 4.5 million people in the household population, and 150,000 people in group quarters facilities. We have begun developing automated statistical process control methods to uncover potential errors in the data. Several methodologies are being used to investigate responses from all three data collection modes (mail, Computer Assisted Telephone Interview (CATI), and Computer Assisted Personal Interview (CAPI)) using traditional Shewhart charts. For mail, we compare the individual responses of each question at various levels of geography. For CATI, we use similar methods, but also compare data for each telephone center. For the CAPI data collection mode, we concentrate our initial efforts on Field Representative (FR) item missing data rates and compare each FR to all FRs within clusters of counties. Our paper presents the details of the methodology, and several examples and results, as well as discussion of the inherent challenges and obstacles faced when applying traditional process control methods to a large-scale, multi-mode, demographic survey.

Key Words: American Community Survey, statistical process control

1. Introduction

1.1 What is the American Community Survey?

The American Community Survey (ACS) is a national rolling monthly household survey. The data collected from the ACS sample is used to produce estimates of the demographic, housing, and socio-economic characteristics of the U.S. population and replaces the estimates traditionally produced from the Decennial Census long form. Beginning with the 2010 Census, only the basic demographic items were asked on the Census questionnaire. The long form census questionnaire has been discontinued and the ACS has taken its place. Examples of the kind of data collected by the ACS include income, educational attainment, year structure was built, tenure, etc. We collect data from a sample of housing unit addresses as well as from a sample of people in Group Quarters (GQs). GQs are non-traditional living arrangements such as college dorms, prisons, nursing homes, etc. This paper focuses only on the HU address sample, though our intent is to include the GQ component in our work in the future.

We release three sets of period estimates each year: 1-year estimates for areas with a population of 65,000 or more; 3-year estimates for areas with a population of 20,000 or more; and 5-year estimates for all geographic areas down to the tabulation block group level. Note that the population threshold used is the most current estimate of total

---

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

population from the Population Estimates program – the official estimates of population released each year by the Census Bureau.

1.2. Sample Selection and Data Collection

Each year we select a sample of addresses from the Census Bureau's Master Address File. We use several sampling rates and assign every block to one sampling strata. We do this based on our estimate of occupied housing units for various geographic levels. In general, blocks in the least dense areas are sampled at our higher rates. Areas of higher density (in terms of occupied housing units) are sampled at lower rates. From 2005 to 2010 the sample size was approximately 2.9 million addresses per year. Beginning in June of 2011 the sample was increased to 3.54 million addresses per year.

Each sample address is randomly assigned to one of the twelve calendar months or panels. In month one, all units determined to have good mailing addresses (approximately 95 percent of the sample) are mailed out. Currently, all addresses that are mailed a questionnaire and that do not respond by the middle of the month are mailed a second questionnaire. In month two, all mailable addresses that did not respond to either the first or second mailing in month one, and for which we have a telephone number are sent to Computer Assisted Telephone Interviewing (CATI). In month three, all addresses that remain non-respondents from mail and CATI, and all unmailable addresses, are sampled for Computer Assisted Personal Interviewing (CAPI). We also use differential sampling rates for CAPI which are based on historical tract level mail and CATI response rates. Tracts with higher response rates are sampled for CAPI at the lower rates and, conversely, tracts with the poorest historical response rates are sampled for CAPI at the highest rates. See Figure 1 for an illustration of the ACS data collection strategy through time.

Figure 1. ACS Data Collection Strategy – Sequential Modes of Data Collection

Sample Panel	Calendar Month				
	Jan 2010	**Feb 2010**	**Mar 2010**	**Apr 2010**	**May 2010**
Dec 2009	Phone	Personal Visit			
Jan 2010	Mail	Phone	Personal Visit		
Feb 2010		Mail	Phone	Personal Visit	
March 2010			Mail	Phone	Personal Visit

2. Statistical Process Control and the ACS

2.1 What is Statistical Process Control?

There are many definitions for statistical process control (SPC) all of which share the same central theme. Pioneered by Walter A. Shewhart in the 1920s and later by W. Edwards Deming, SPC is the application of statistical methods and procedures (such as control charts) to analyze the inherent variability of a process or its outputs to achieve and maintain a state of statistical control, and to improve the process capability. SPC methods have traditionally been applied to manufacturing processes with relatively stable variation in each process over time. There are two types of variation to consider when using SPC techniques: 1) controlled variation that is natural to the process (common causes of variation), and 2) uncontrolled variation that is not present in the process at all times (special causes of variation). While the first type is inherently interesting and potentially useful in understanding the patterns of ACS respondents, it is the second type of variation that we are most concerned with. In the manufacturing sector, SPC has been used to identify processes that should produce a product with known, specified parameters, such as washers with known diameter, center hole diameter, and thickness. SPC techniques, in particular, control charting has been shown to be effective in identifying process shifts which can lead to corrective action. W. Edwards Deming (1986) stated the following,

> "The control chart sends statistical signals, which detect existence of a special cause (usually specific to some worker or group or to some special fleeting circumstance), or tell us that the observed variation should be ascribed to common causes, chance variation attributable to the system." (p. 319).

2.2 Why Use SPC Methods with the ACS?

2.2.1 Data Quality

Over the past several years, the ACS has grown substantially in both size and scope. The ACS began as a very small scale test survey in the late 1990s, and has become the largest demographic household survey conducted by the U.S. Census Bureau with a current annual sample size of approximately 3.54 million addresses. Each year the ACS produces and publishes over 11 billion estimates (Census Bureau, 2010). Over the past several years, a number of errors have been found in our estimates. On the ACS web page containing our errata, there are currently 70 notes documenting errors in our data that have been discovered either by Census Bureau staff or by data users (Census Bureau, 2010). Some of these errors have been caused by internal Headquarters data processing mistakes, while others have resulted from field representatives (FRs) misunderstanding procedures or training and introducing non-sampling error during the collection of the data. No matter the source of these errors, our hope is that SPC methods can identify potential errors in the ACS data early in the process, before the annual processing of the raw response data begins, and maybe in time to take corrective action during the year.

Figure 1 shows that the sample assigned to the January 2010 panel is finished with data collection activities by the end of March 2010. This data can sit (after minimal post-processing) for up to a year before the pre-weighting and edit/input steps begin in March of 2011 for the 2010 tabulation year – which includes all of the data collected during the 2010 calendar year. We feel that we may be able to take corrective action in near real-time if an automated system is running on the ACS response data monthly.

We believe that the use of SPC techniques will allow us to identify errors introduced through the actual data collection and errors introduced through processing changes. Many computer programs are written and maintained that process the data and there are many "risk points" along the way. The ACS is currently undergoing a large-scale process improvement effort to better document the processing and to identify those areas that may be lacking in either documentation of the process itself or in the dependencies across many linked processes. By reducing or eliminating processing changes made in isolation, this effort will undoubtedly lead to fewer errors in our data.

In addition, we hope that through our research we can verify that changes introduced into the survey, such as new questions, format changes, instruction changes, or processing changes do not cause unexpected results in the response data. These are special causes that we are aware of and can adjust for in our models. Moreover, we usually have an idea of what effect a given change will have on the data since our experimental design program tests the majority of changes before being implemented in production.

2.2.2 Repeated Measures

SPC methods require repeated measures over a relatively long period of time in order to measure the process average. The ACS has collected data from a sample of addresses in every county in the U.S. and in Puerto Rico every month since January 2005 with very few significant changes to the questionnaire. This gives the ACS program a unique opportunity to develop quality control methods that take advantage of the amount of data we have collected. Between January 2005 and December 2010, we interviewed roughly 11.5 million households and approximately 27 million people. We have amassed an unprecedented time-series of data points from survey response data, not only at the national level, but also at the county, and in some cases, tract level. We collect and are testing all 21 housing and 48 person questions on the mail questionnaire with a total of 685 unique values. We also include the value of missing or null as a valid response value to be tested. There are roughly the same number of variables and responses in the CATI and CAPI data. Some variables are seasonal or are trending and we want to remove these signals (known causes of variation) and test the residuals against limits.

2.3 Challenges

2.3.1 The Methodology – Data Interaction

One of the biggest challenges in developing an SPC system to run on the ACS response data is determining how to monitor data that is, by the very nature of the ACS, changing over time. Traditional SPC charts monitor fairly well understood processes and identify units that are statistically significantly different from the process average. Fluctuations in survey response data can be attributed to a number of different causes, most of which are likely not errors in the data. Changing economic conditions, natural disasters, attitudes, politics, etc., can (and do) play a part in the observed responses to the ACS. The goal of our research is to identify the most egregious outliers and investigate these as potential errors.

2.3.2 Analysis

Another challenge of developing the ACS SPC system is finding the most efficient ways to get the appropriate data to the subject matter experts and analysts to adjudicate outliers identified by the system. A fair amount of work went into developing an internal Intranet application that will allow the analysts to log in from their desktop computers and review the output from the SPC system. We have contracted with a private company to help us develop our SPC system. In section 3 we discuss the methodological work they have developed with the Census Bureau, building on the work that we have done to date in house. In addition to their work on the actual SPC testing, the contractor is also developing a prototype of an output delivery system using SAS® Internet software. This software is designed such that the analysts and subject matter experts who are charged with researching potential problems identified by the system have easy access to the data from their work areas.

3. Methodology

3.1 Time Window

In developing the SPC methodologies for our specific application, one of the first questions we had to answer was: *What is the appropriate length of time to use to calculate the process average for each response value?* This question illustrates why the ACS requires a new way of thinking about survey data. Just as data users do, we had to balance our need for reliable estimates with our desire for the most up-to-date data. The ACS program continues to spend considerable time and effort informing and educating data users so that they may make these same types of decisions when deciding whether 1-, 3-, or 5-year estimates best suit their needs. For our use, and because the ACS questionnaire has been relatively stable since 2008, we decided to use all data collected in the two previous years and the current year, up to the month of interest to construct our process averages. In this way we believe we have sufficient sample to construct reliable upper and lower bounds while using relatively recent data for our process averages.

3.2 Mail Response Data

3.2.1 Input Data

To focus our efforts, we are currently reviewing only the response data from the U.S. English questionnaire and the CATI data. Preliminary work has been done to review the CAPI data, but this paper will not discuss this in detail. Our intent is to include data from all form types in all languages. The input files we are using are the daily keying files. These files are created using the unedited batch files from the National Processing Center (NPC), created from the questionnaires received each month. The forms for each month are accepted and processed at NPC until a set closeout date. These data are used to create the Failed Edit Follow-Up (FEFU) workload. The FEFU operation contacts respondents whose questionnaire triggers one or more data quality checks, such as listing more than five people on the roster. Thus, we do not include any changes resulting from FEFU. We review the data by sample panel (month) and not the NPC check-in date.

3.2.2 Statistical Testing of Response Distributions

For each question, the proportion of each legal value is reviewed including the "non-response" or missing value. For example, a yes/no question has three values: the "yes" response, the "no" response, and the "non-response" or "missing" value. Only the responses of those in the eligible universe are reviewed for all of the values. The eligible universe for each question has been determined, primarily based on the edit universes. For example, the marriage questions are only asked of respondents who are at least 15 years old, therefore, only those respondents are included in these calculations. Due to the large amount of data that could be output (because of the length of the questionnaire and all the possible values), we decided to create a conservative filter to identify variable values considered to be the "worst" offenders and create SPC charts only for these variable values.

3.2.2.1 Z-Score Filters

Our filter includes three individual tests that compare each proportion from the month of interest (the current month of analysis) to the proportions from: 1) the same month from the previous year; 2) the entire previous calendar year; and 3) the previous month. Each test is a basic significance test of the difference of two proportions. A test fails if the z-score is greater than 2.5 or less then -2.5 (which is the level of alpha equal to 0.01). Each variable value (or range of variable values) is tested individually. The proportions, as well as the standard errors, are based on unweighted data. The following formulas were used to calculate the z-score for each test:

$$z_t = \frac{\hat{p}_{1t} - \hat{p}_2}{SE_t} \quad \text{where:} \quad SE_t = \sqrt{\hat{p}_t(1-\hat{p}_t)\left(\frac{1}{n_{1t}} + \frac{1}{n_2}\right)} \quad \text{and}$$

$$\hat{p}_t = \frac{\hat{p}_{1t}n_{1t} + \hat{p}_2 n_2}{n_{1t} + n_2}$$

where

t=1, 2, 3	represents each test, respectively,
z_t	is the z-score for the t[th] test,
$\hat{p}_{1t}$	is the sample proportion of the month(s) being tested (excluding the month of interest),
n_{1t}	is the sample size of the month(s) being tested (excluding the month of interest),
$\hat{p}_2$	is the sample proportion of the month of interest,
n_2	is the sample size for the month of interest,
SE_t	is the standard error based on the month of interest and the other month(s) being tested.
$\hat{p}_t$	is the overall process proportion (from January 2008 to the month of interest).

Note that a negative z-score means that the proportion of the month of interest is higher than that of the other month(s) being tested. A positive z-score means that the proportion of the month of interest is lower than that of the other month(s) being tested. The farther away from zero the z-score is, the bigger the difference between the two proportions. The variable value makes it through the filter if all three tests fail.

We are also working with the contractor to incorporate more traditional SPC methods into our system, in particular, the Western Electric SPC rules. In general, these rules look for particular patterns in consecutive measurements.

3.2.2.2 Seasonal Adjustment

We observe that many of these variable values have a nonrandom pattern brought about by predictable changes in the population and the nature of the question. For example, one of the ACS questions ask “LAST MONTH, what was the cost of electricity for this house, apartment, or mobile home?” At the national level, this question yields seasonal responses, as households tend to pay more for electricity in the summer months. Known patterns may actually confound real problems in the data. Therefore, a seasonal or trending adjustment must be made to the data before sending it through a process control system (Alwan and Roberts, 1988).

Specific variables were selected for this treatment based on past knowledge. For each of these variables, we select a model using time series techniques and obtain the predicted and residual values. We then send all three sets of values through the rest of the system.

3.2.2.3 Shewhart Charts

We create Shewhart charts for each of the variable values that make it through the filter. The average is based on all of the data in the chart (see formula for $\bar{p}_l$ where n_{il} is the sample size for each month, *i*, for each variable value, and p_{il} is the percent of responses for the specific value for that particular month). N represents the number of panels that are being passed through the formula and l represents the fact that each variable value has a different proportion. The control limits are based on $\bar{p}_l$ and the sample size for that particular month (k = 3 = the sigma limit for the charts). Figure 2 shows an example a Shewhart graph with the lines labeled.

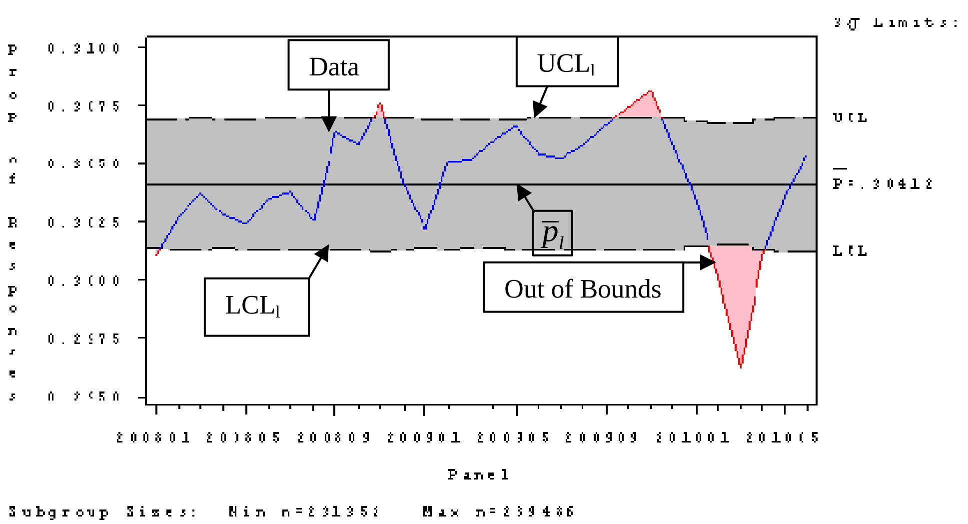
$$\bar{p}_l = \frac{n_{1l}p_{1l} + \ldots + n_{il}p_{il} + \ldots + n_{Nl}p_{Nl}}{n_{1l} + \ldots + n_{il} + \ldots + n_{Nl}}$$

$$LCL_l = \max(\bar{p}_l - k\sqrt{\bar{p}_l(1-\bar{p}_l)/n_{il}}),0$$

$$UCL_l = \min(\bar{p}_l + k\sqrt{\bar{p}_l(1-\bar{p}_l)/n_{il}}),1$$

where, LCL and UCL stand for lower control limit and upper control limit, respectively.

Figure 2. Shewhart Chart Example



We also perform this check at the Census segmentation group and state levels. The segmentation groups were created to try to group tracts that were similar based on the following characteristics: homeowner status, past response rates, income, and ethnicity.

We note that there are a few limitations to these methods. The ACS sample size is a lot larger than those typically monitored with SPC. The results are that there is an extremely small scale for charts and the control limits are very narrow. We note that this results in relatively small variances and more false signals in the system. Very often we ask ourselves the question, “How much of a difference is actually significant?” and remain cautious when analyzing the charts.

3.3 CATI Response Data

The input data we use for this analysis are from files provided by the Technologies Management Office (TMO) at the Census Bureau. These files are created from the output of all the housing unit automated instruments. These data sets are delivered after the each panel of data collection is closed out.

The CATI uses specific skip patterns based on the answer or answers to certain questions. When we calculate the variable values, as with mail, these skip patterns are taken into account.

We analyze the CATI data similarly to the mail data. We currently only look at the distribution of responses, but the data is calculated the same way. Currently we analyze the data for each telephone center.

3.4 Additional Data Checks

3.4.1 CAPI Response Data

We have begun to review data from the CAPI data collection mode. This data is from files provided by the ACS Data Capture File (DCF). This file is updated monthly and is a collection of files output from the various data collection entities. As with the mail and CATI data, the universes for CAPI are defined based on the skip pattern. Currently we are only analyzing the non-response rates for seven variables.

Shewhart charts, like in the previous sections, are created not only for previously mentioned geographies, but also for the Regional Offices (ROs) and calculated clusters. Clusters are counties, within the same RO, grouped by similar socio-economic characteristics. Currently, we group counties by the following variables using January 2008 through December 2009 data: median household income, percent rural, and overall CAPI unit nonresponse rate. These clusters are currently being held constant over time and are collapsed so that at least five field representatives (FRs) are in each cluster. A FR may fall into multiple clusters if he or she works multiple counties that fall into different clusters. Charts are then created for those FRs who are out of control (beyond the three sigma limit) for a particular variable’s non-response in the month of interest. This work is still in development and results are not yet available for distribution.

3.4.2 Multiple Responses in Mail

Another mail check we perform involves multiple checkboxes. The data capture system allows for the recording of items where respondents marked multiple checkboxes in the response to a question where, often, only one mark was expected. This rate is defined as the number of times multiple checkboxes were marked for an item as a percentage of the total number of times that the item had a response (multiple or single response). We are interested to see if an unexpected variable shows up or if one question shows sudden jumps in the number of multiple responses. Two lists are then created, one that includes items where multiple responses are valid and the second does not include these variables.

3.4.3 Mail Completeness

We are also reviewing the percent of each section of the U.S. English mail questionnaire that was deemed “complete.” We are interested to see if the overall questionnaire completeness had similar patterns to the non-response patterns of the individual items in the Shewhart charts. We define this rate as the total number of completed questions a household or person is eligible to respond to divided by the total number of questions a household or person is eligible to respond to for each section. A mail form is considered non-blank if the cover contains a phone number or if there is at least one data-defined person on the form. Currently, we are looking at this by section of the form (front page, basic person questions, housing questions, and detailed person questions) and within Census Segmentation Groups.

4. Results

We have chosen not to include the results of our checks on the multiple check boxes marked or the completeness rates because we did not observe any unusual results when analyzing these data.

4.1 Mail Response Data

In this section, we discuss the results of the monthly mail run on the January 2011 data. We only cover selected results due to the amount of output.

4.1.1 Distribution of Responses

4.1.1.1 Z-Score Filters

Table 1 lists the ten largest housing variable values that passed through the filter, sorted by the absolute values of the z-score based on comparisons of the month of interest to that same month of the previous year. Overall, 40 of the 292 housing variable values failed all three tests. Table 2 lists the same information for the population variables. Overall 60 population variable values failed all three tests out of 393 total population variable values.

Table 1: January 2011 - Ten Largest Z-scores for Housing Variables

Variable Values		Z-Scores		
Variable Name	**Variable Value Definition**	**Previous Year, Same Month**	**Entire Previous Year**	**Previous Month**
fs2	Did not receive food stamps	11.67	7.48	5.75
mrgx1	Mortgage, deed of trust, etc.	11.34	9.27	2.70
fs1	Received food stamps	-11.02	-12.64	-7.72
acr1	Less than 1 acre	8.51	10.78	5.36
acr3	10+ acres	-8.29	-9.55	-5.64
mrg11	Monthly mortgage of 1-999	-8.23	-10.23	-6.39
wat13	Water cost $1,000 or more	-8.11	-9.51	-6.37
ybl2	Building Built 1990 to 1999	7.74	8.04	5.36
ten1	Tenure: owned with a mortgage	7.71	6.44	4.90
hfl2	Type of fuel is bottled, tank, or LP gas	-7.53	-9.68	-7.57

Table 2: January 2011 - Ten Largest Z-scores for Population Variables

Variable Values		Z-Scores		
Variable Name	**Variable Definition**	**Previous Year, Same Month**	**Entire Previous Year**	**Previous Month**
hins4_2	No, Health Insurance from MEDICAID	16.02	8.08	2.78
int7	Yes, interest earned and amount written	13.78	14.57	4.98
nwla1	Yes, on layoff	13.72	3.43	-3.85
int1	Interest earned between $1 and $9999	13.29	14.48	3.88
wkl1	last worked within the past 12 months	13.16	-6.66	-10.99
wagnr	No response to Wages	-12.07	-8.50	-3.74
hins1_1	Yes, Health Insurance from current employer	11.41	7.15	5.59
age5	Age is 65 or greater	-11.14	-6.59	5.22
tinr	No response to Total Income	-10.85	-7.18	-3.41
marhynr	No response to Year Last Married	-10.70	-4.82	-4.65

4.1.1.2 Seasonal Adjustment

Figure 3 shows the chart for those who responded that they paid between $1 and $99 for electricity cost last month. As you can see, it has a very seasonal pattern. Using the proc esm procedures in SAS®, the data is seasonally adjusted and the residuals are output. Figure 4 charts the residuals in an individual measurements and moving ranges chart from the Shewhart charts. As is shown, the seasonality has been removed and the process is mostly in control.

Figure 3. Shewhart Chart December 2010 – Electricity Cost $1 - $99

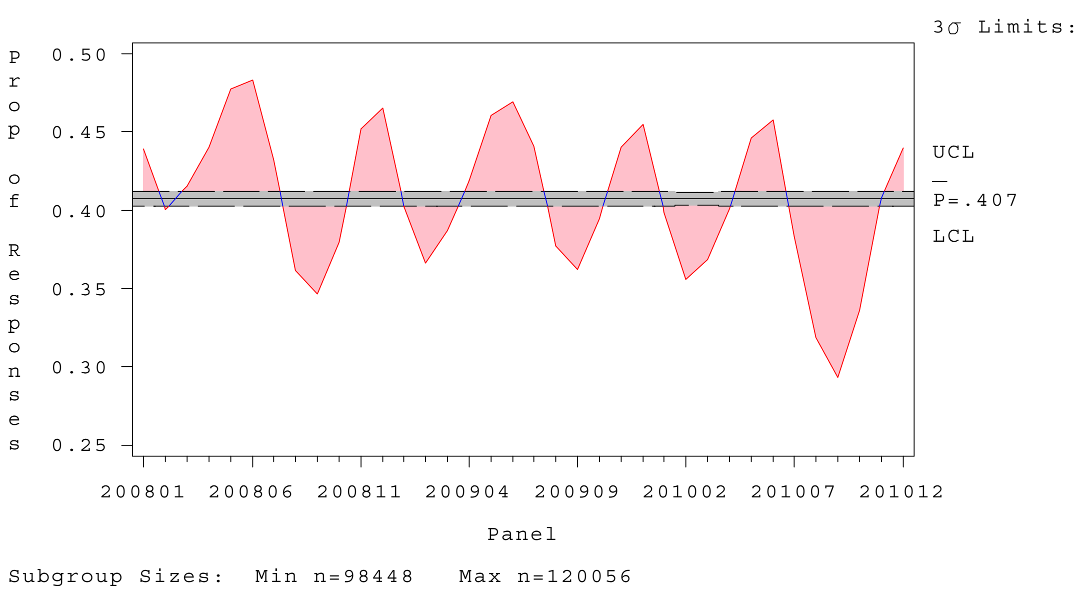
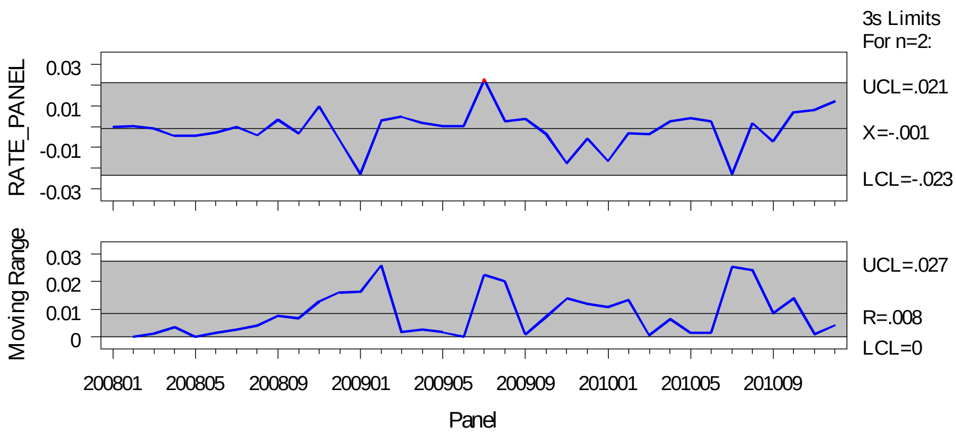


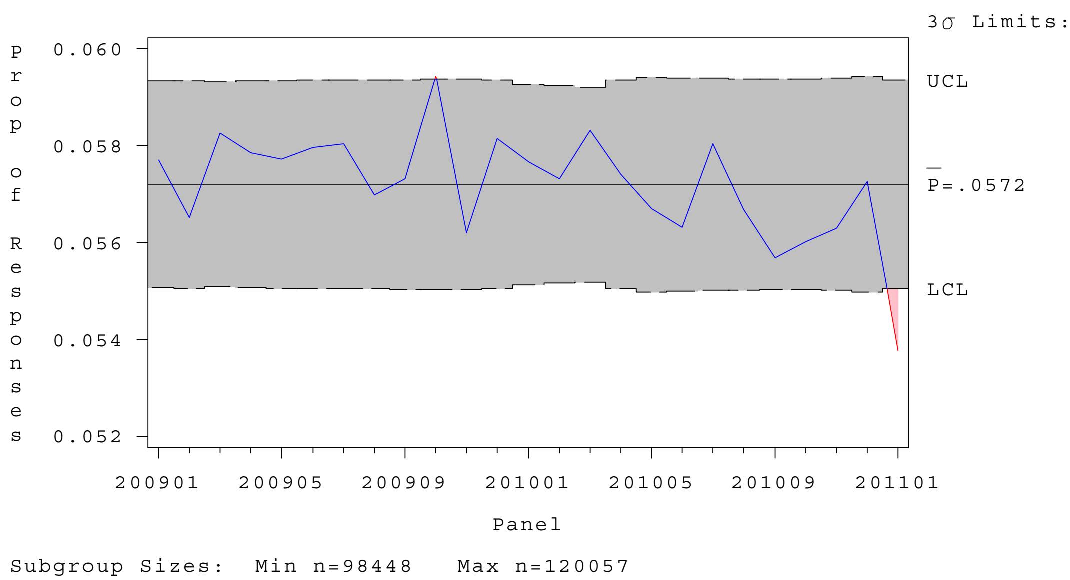
Figure 4. Shewhart Chart December 2010 – Residuals of Modeled Electricity Cost $1 - $99



4.1.1.3 Shewhart Charts

We produce a Shewhart chart for all of the variables that fail all three z-scores. Figure 4 shows the graph for all the respondents who answered “Which best describes this building?” with “A one-family house attached to one or more houses.” We note that even though the data goes beyond the three-sigma limits in January 2011, the difference is only less than a percent, as is the entire range of the chart. Figure 5 shows the same question, but only the response is “A building with 2 apartments.” This graph goes “out of control” four times. We see that the difference in January 2011 are opposite between the two graphs. This could possibly show a change in the population. We are currently investigating this issue.

Table 5. January 2011 Shewhart Chart, One-Family House Attached



4.2 CATI Response Data

When analyzing the CATI data, we create the p-charts for each variable value for each telephone center. These centers have different properties, including being in different time zones and interviewers with different language skills. This means that they have different processes. Figures 8, 9, and 10 show the same variable value (those who respond Asian to the Race Question) for each telephone center. As you can see, they have three different patterns and the process average is different. This shows that looking at any of the data at a higher level does not make sense.

Figure 8. Shewhart Chart January 2011 – Asian, Hagerstown

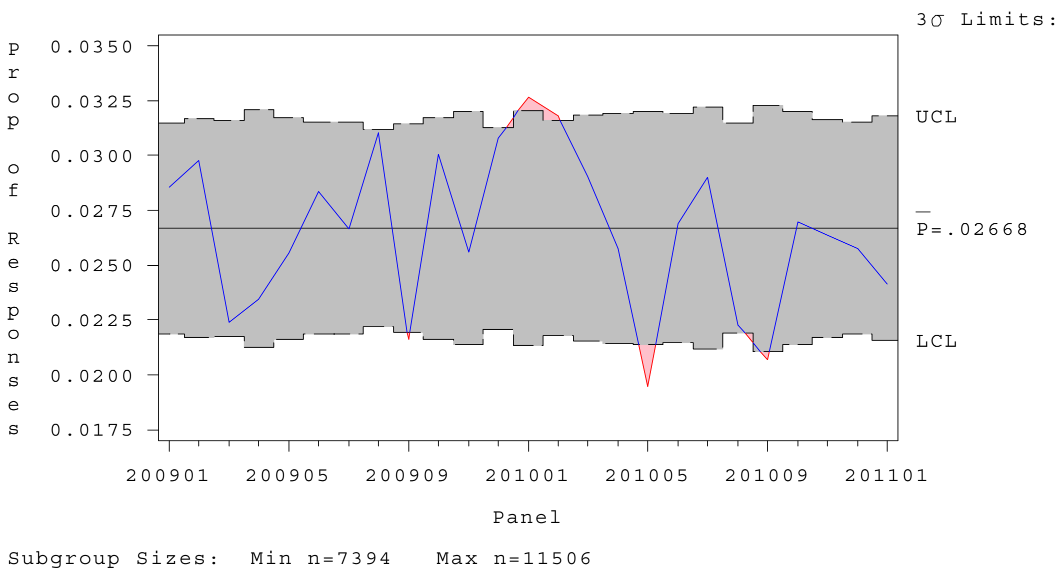


Figure 9. Shewhart Chart January 2011 – Asian, Jeffersonville

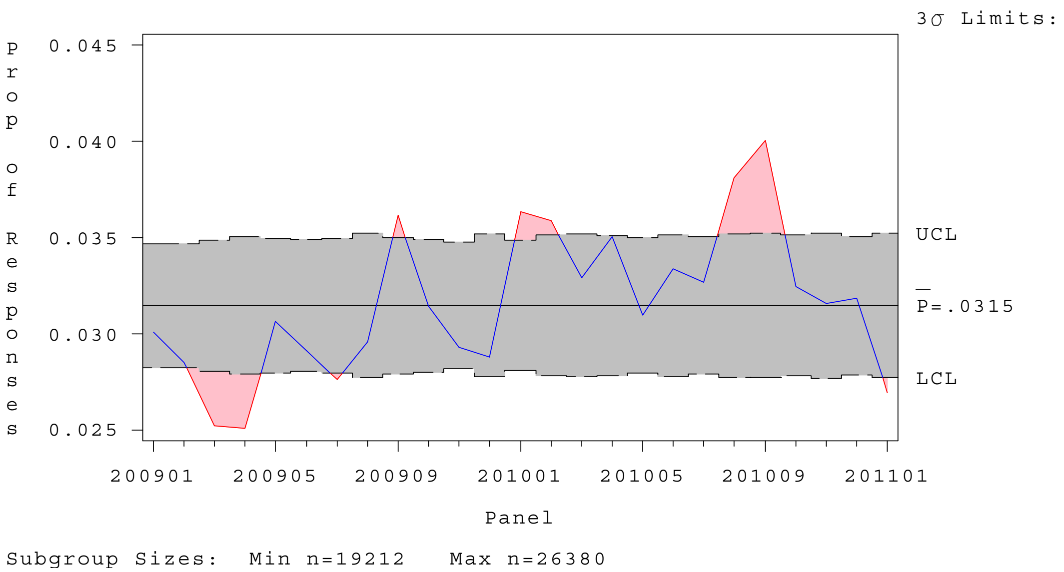
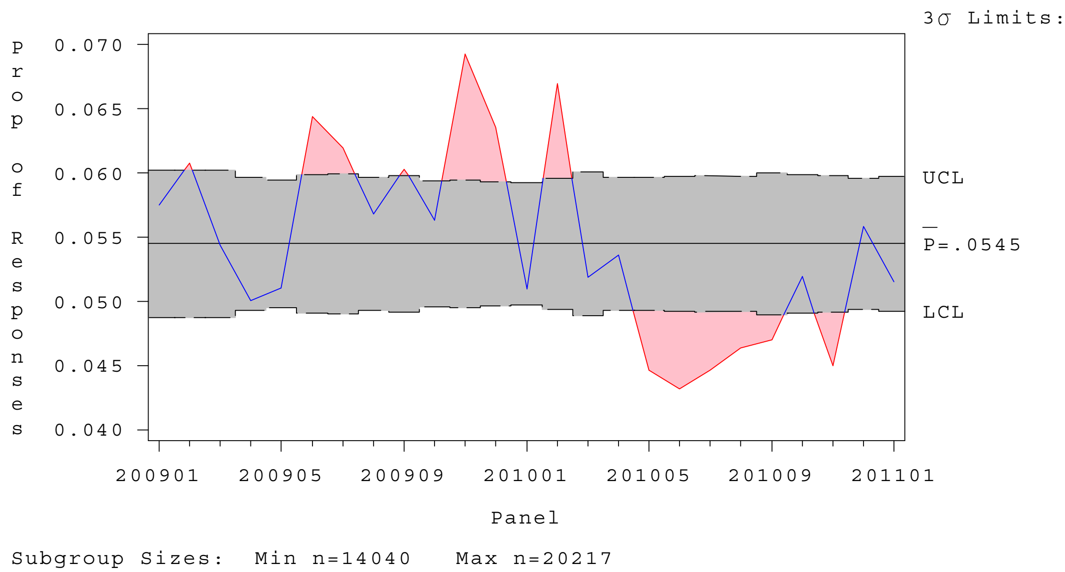


Figure 10. Shewhart Chart January 2010 – Asian, Tucson



5. Conclusions and Future Work

Our initial work using SPC methods with ACS response data shows a great deal of promise. Our preliminary results have shown that while some response values appear to be errors, changes in methodology and in the population may be causing the observed differences. These methodologies must be used in conjunction with knowledge of the variables and expected changes in the population. Anomalies need to be explored by the appropriate subject matter experts.

Our next steps will be to include the methodology to review the response data from the housing unit CAPI data. In addition, we will be developing similar methods for reviewing the response data from the Group Quarters data collection efforts. We also continue to develop an Intranet based review system and eventually hope to include metrics on important paradata items such as response rates, delete rates, interview length etc.

References

Alwan, Layth C. and Harry V. Roberts (1988), "Time-Series Modeling for Statistical Process Control" *Journal of Business & Economic Statistics*, 6, 1, 87-95

Census Bureau, a, http://www.census.gov/newsroom/releases/archives/american_community_survey_acs/cb10-cn90.html

Census Bureau, b, 2011, http://www.census.gov/acs/www/data_documentation/errata/

Deming, W.E. (1986), *Out of the Crisis*, Cambridge, MA: MIT Press.