Smart Meter Data Quality Insights

Juan Shishido, EnerNOC Utility Solutions

ABSTRACT

This paper focuses on the ways in which we can better understand residential smart meter data. There is uncertainty, on the minds of many stakeholders, regarding how smart meters collect data and the quality with which those data are collected. Issues covered are missing intervals, zeros, and spikes. We present data from a recent smart grid project, examining data at the 15-minute interval level, and discuss several systematic patterns which indicate inconsistencies and erroneous usage data.

Introduction

Smart meters provide more reliable and higher temporal-resolution data. In addition to more accurate billing, which benefits both the end use customer and the utility, smart meters enable a myriad of new possibilities in energy efficiency, demand response, and evaluation. Furthermore, many smart meter enabled customers are able to see real time usage through energy monitoring tools and web portals, which provide quicker feedback and allow customers to make better consumption decisions. As utility smart meter rollouts continue, researchers will have abundant amounts of data to process and energy monitoring tool vendors will have additional opportunities to convert that data into usable and actionable information for customers. As of the end of 2011, Pacific Gas and Electric (PG&E) had installed almost 9 million gas and electric meters; as of May 2012, Southern California Edison (SCE) had installed 4.3 million smart meters. With this expansion, and the questions surrounding smart meter accuracy and effectiveness, utilities will have an obligation to assuage customer concerns regarding these devices.

As has been well documented, especially in California, customers have not been as willing to accept smart meters and their purported benefits as utilities had hoped. A recent Pike Research survey found that the major cause of trepidation, for almost 60 percent of respondents, had to do with worries about increases in electricity bills. In 2009, Google advocated to the California Public Utilities Commission (CPUC) that smart meter data be available to consumers free of charge and in standard formats, as a result of their belief that usage data and intelligible information could reduce energy consumption by as much as 15%. Similarly, a 2010 report released by the American Council for an Energy-Efficient Economy (ACEEE) found that, if customers are given context, suggestions, and encouragement, household energy savings from smart meter interval data could potentially be as high as 12 percent. Large Commercial and Industrial customers have had access to their interval data in California, upon request, for several years, and often use that data to inform the operation and management decisions of their plants and facilities. Smart meter data is also playing an increasingly important role in evaluations of both energy efficiency and demand response programs as the data become ubiquitous. For example, smart meter interval data enables utilities to evaluate savings associated with demand response events, providing customers prompt feedback on their curtailment efforts. Interval data is also invaluable in the estimation of baseline energy consumption models to track savings from

energy efficiency activities. It is particularly important when evaluating savings from programs with both technological and behavioral components, such as Strategic Energy Management (SEM).

In all of the cases described above, whether being used by an evaluator, an end use customer, or the utility, the integrity of the interval data is paramount because it is being used to inform consumption decisions and measure the success of programs. However, as many in the industry are discovering, it is often more challenging than expected to establish the necessary protocols to access and maintain high quality interval data, especially in the first few years after implementation. Through our involvement in a recent smart grid program evaluation, we were exposed to a significant amount of interval data being extracted from a recently completed smart meter deployment. The following paragraphs outline our data quality findings on duplicates, intervals whose values were zero, and spikes, using data from that evaluation.

Analysis

Our data, which were 15-minute interval data, were structured as shown in Table 1. Each customer, or household, distinguished by an account identification number, had a record for each day of available usage data as well as information regarding the meter from which that data came. On average, each customer had 224 records. Ordered columns identified each of the 96 15-minute intervals. We also had information on move-in and move-out dates as well as information on household demographics.

Table 1

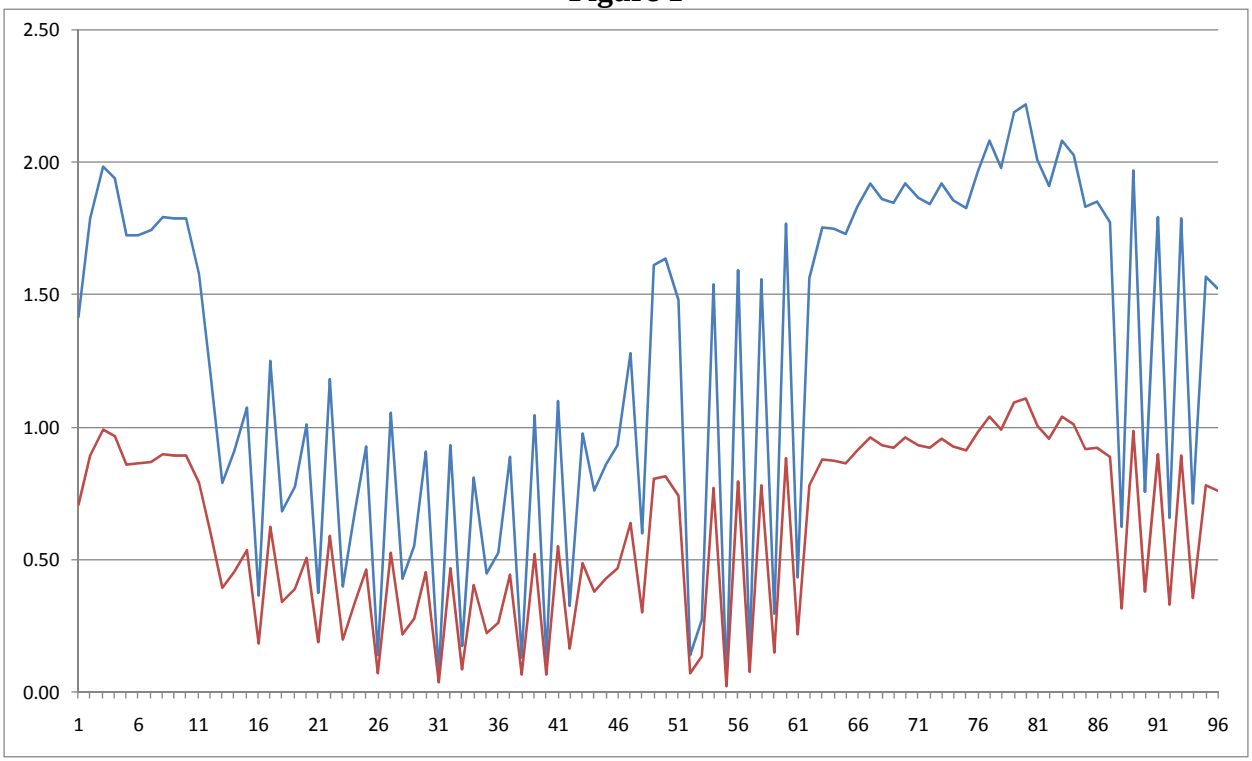
accountid	meterid	installdate	date	[kw1-kw96]	move-in	move-out	[demographics]

Source: Juan Shishido 2012

Duplicates

The first issue we encountered was duplicate records. Usually, duplicates are entirely identical, which means that keeping one and deleting all others is appropriate. In addition to these types, however, we found non-identical duplicates, or customer records with differing consumption values for the same date. Some of these duplicates resulted from differences in resolution, or the number of significant digits displayed. While most records listed usage to a thousandth of a kilowatt, some listed these values at a ten-thousandth of a kilowatt. Once corrected, these duplicates became identical and records were then removed from the analysis. When we analyzed the remaining duplicates, which notably accounted for only 1.05% of all records, we found that, in many cases, consumption values in one record were sometimes just multiples of those in another. This is illustrated in Figure 1, below. These lines have the same shape, but the data points for the blue line are always two times larger than those of the red line. These findings suggest that a meter multiplier was applied to usage in one record but not in the other(s). The unaccounted for duplicates—those for which we had no solution or explanation—simply differed across certain intervals and not others, but without any discernible pattern. As a result of not being able to determine the appropriate record, all remaining duplicates were excluded from the subsequent analysis.

Figure 1



Source: Juan Shishido 2012

A distinct duplicates scenario, not dealt with above, occurred as a result of a meter change out—what we call the meter swap date. In cases where a change out occurred within our observation date range, we often had two records for the same day, with one associated with the old meter and the other associated with the new meter. Table 2 illustrates this below. In this example, the new meter, meter Z, went online at 12:00 PM on July 25, 2010. To handle this, we simply collapsed the records, adding the kW values across rows. This was appropriate because there were no cases of data overlap. That is, both meters were not able to collect consumption data at the same time.

Table 2

accountid	meterid	installdate	date	kw1	kw2	…	kw48	kw49	…	kw95	kw96
X	Y	5/10/1999	7/25/2010	0.261	0.292	…	0.294	0	…	0	0
X	Z	7/25/2010	7/252010	0	0	…	0	0.287	…	0.616	0.596

Source: Juan Shishido 2012

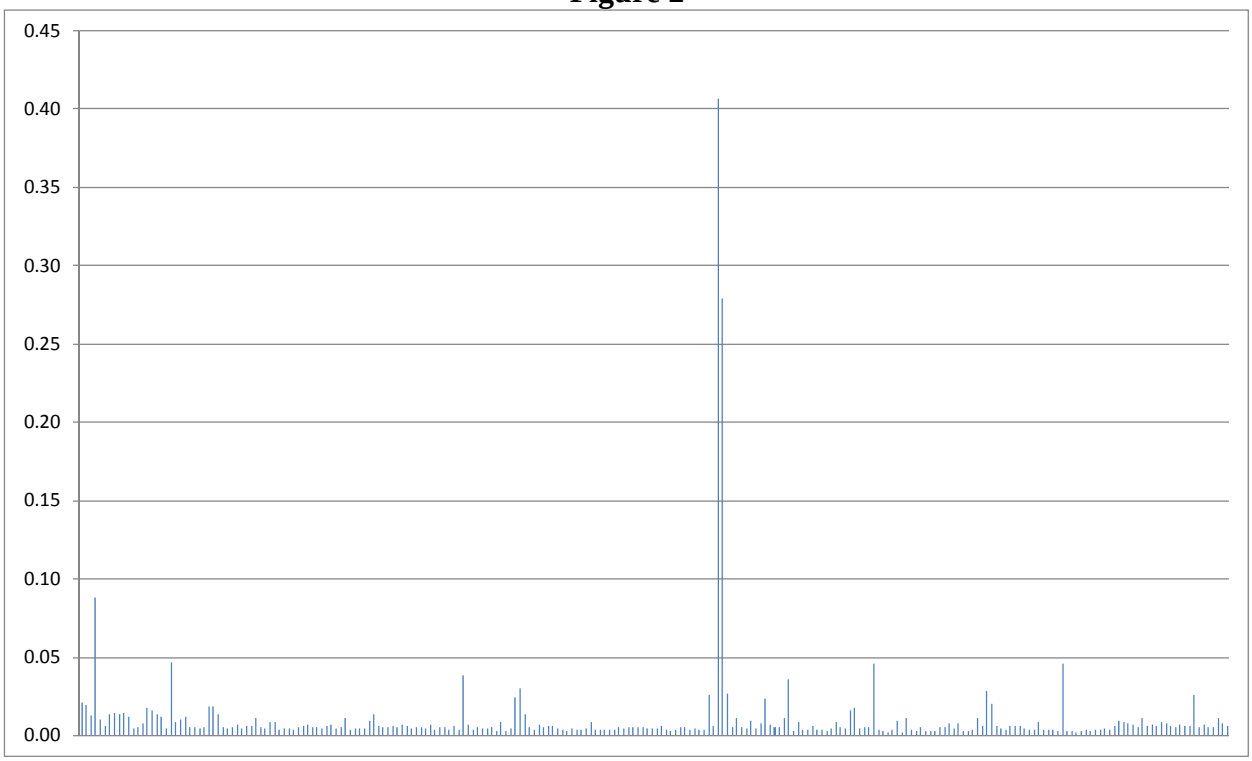
Zeros

The data quality issue we spent the most time analyzing was no usage, or zero-valued intervals (sometimes referred to as “zero records”). With the proliferation of consumer electronics, especially ones that cannot be switched off completely, one would expect residential customers to never have no usage. This is because of the phenomenon known as standby power, which is, according to the Lawrence Berkeley National Laboratory, “power consumed by power supplies, the circuits and sensors needed to receive a remote signal, soft keypads, and displays including miscellaneous LED status lights.” Although no one knows for sure, it is generally

agreed upon that standby power makes up 5-10% of residential electricity use. A recent New York Times article, for example, noted that, in a year, a high-definition digital video recorder and a high-definition cable box use, on average, "about 10 percent more than a 21-cubic-foot energy-efficient refrigerator." Given this information, our expectation was to always see usage values greater than zero. However, this is not what we found.

We first analyzed zeros based their occurrence by date. Specifically, for each date, we calculated the percentage of records with at least one zero value. Our findings are shown below in Figure 2. Zeros could have been random events or could have been associated with only certain customers, in which case the distribution would have been flat, assuming that those customers have a consistent number of records with zeros. This chart, however, suggests that exogenous factors may have been responsible for the surge in zeros. Across all dates, 1.04% of all records have at least one zero interval value. While an overwhelming number of data points in Figure 2 are within less than a tenth of a standard deviation of that value, we see values as high as 40.65%. Further investigation revealed that a storm-caused-outage was responsible for this particular spike in zeros. Still, there are 8 days where the percentage of records with at least one zero exceeds 3%, or 0.20 standard deviations from the mean. Even though three of the eight days occur in the same month, there is no apparent pattern in when the higher-than-average percentage of zeros occur, especially given that the outage affected customers across two consecutive days.

Figure 2

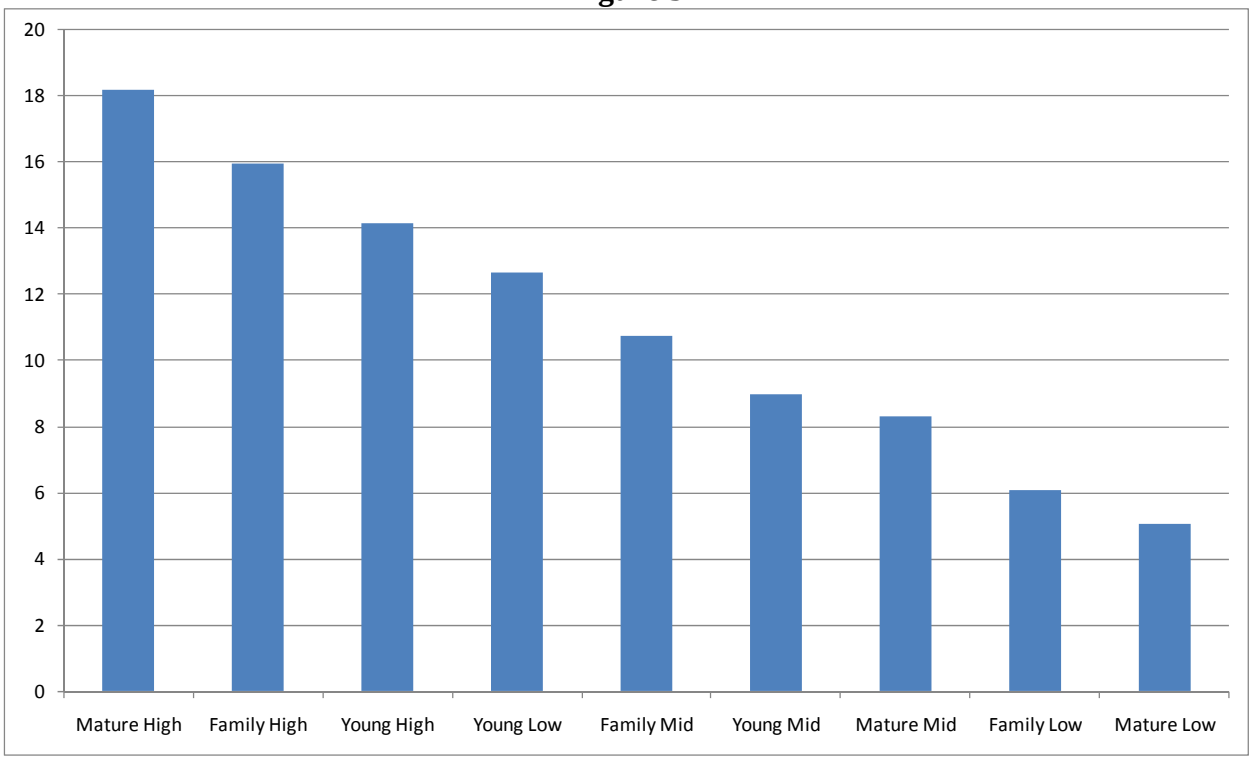


Source: Juan Shishido 2012

We expanded this analysis by parsing the observations by both the three age and three income segments. All observations—both with zeros and without—were distributed across the nine age/income segments as shown in Figure 3. In terms of age, the observations are distributed

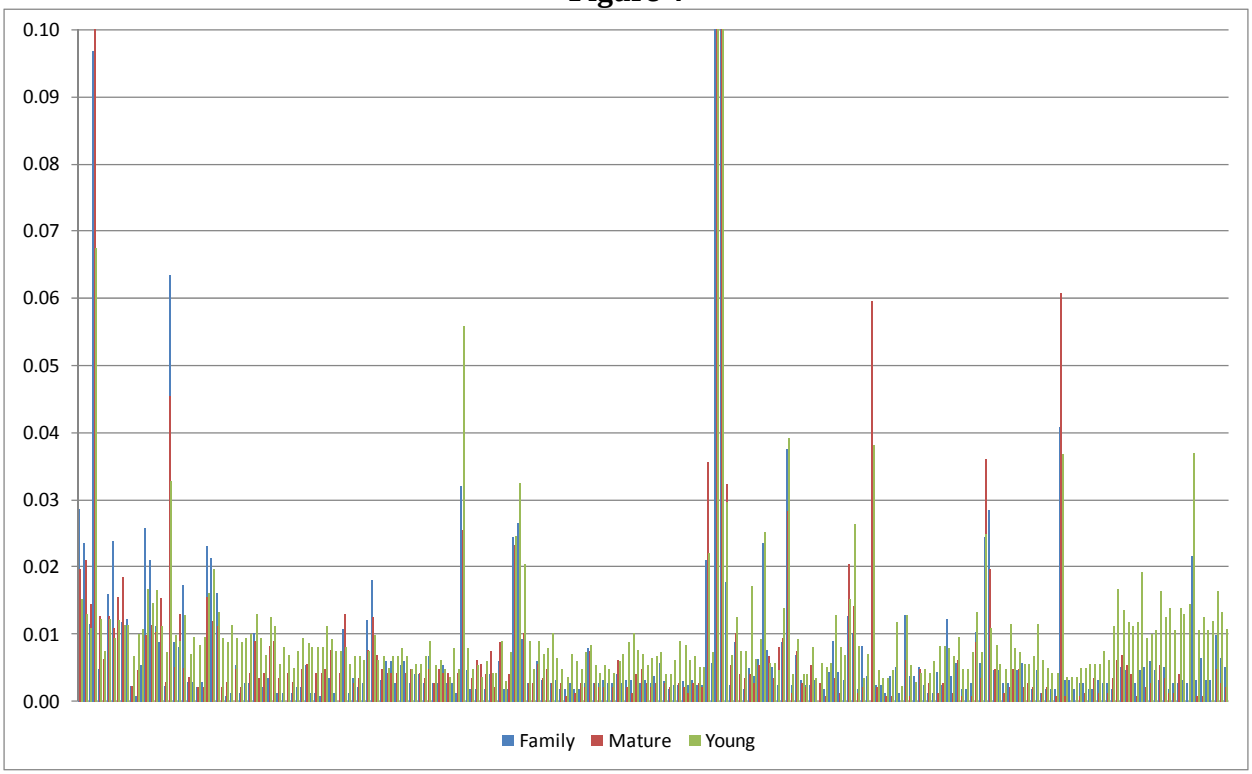
fairly evenly. However, 48% of all observations belong to high income households while only 28% and 24% belong to middle and low income households, respectively. Across the three age segments, the young have more records zeros than do the other two groups. This can be seen in Figure 4, which sets the y-axis to a maximum of 10% for easier viewing. Across income segments, shown below in Figure 5, the distribution of zeros is roughly similar, though the groups are more different from one another than those in the age chart. There are certain days where high income households have significantly more zeros than the two other groups. In the "normal" range, however, defined as days with less than two percent of records having zeros, low income households have many more zeros than the other two groups. The percentages of records with zeros, which are disproportionately associated with young households as well as with low income households, are summarized in Table 3 and Table 4 below.

Figure 3



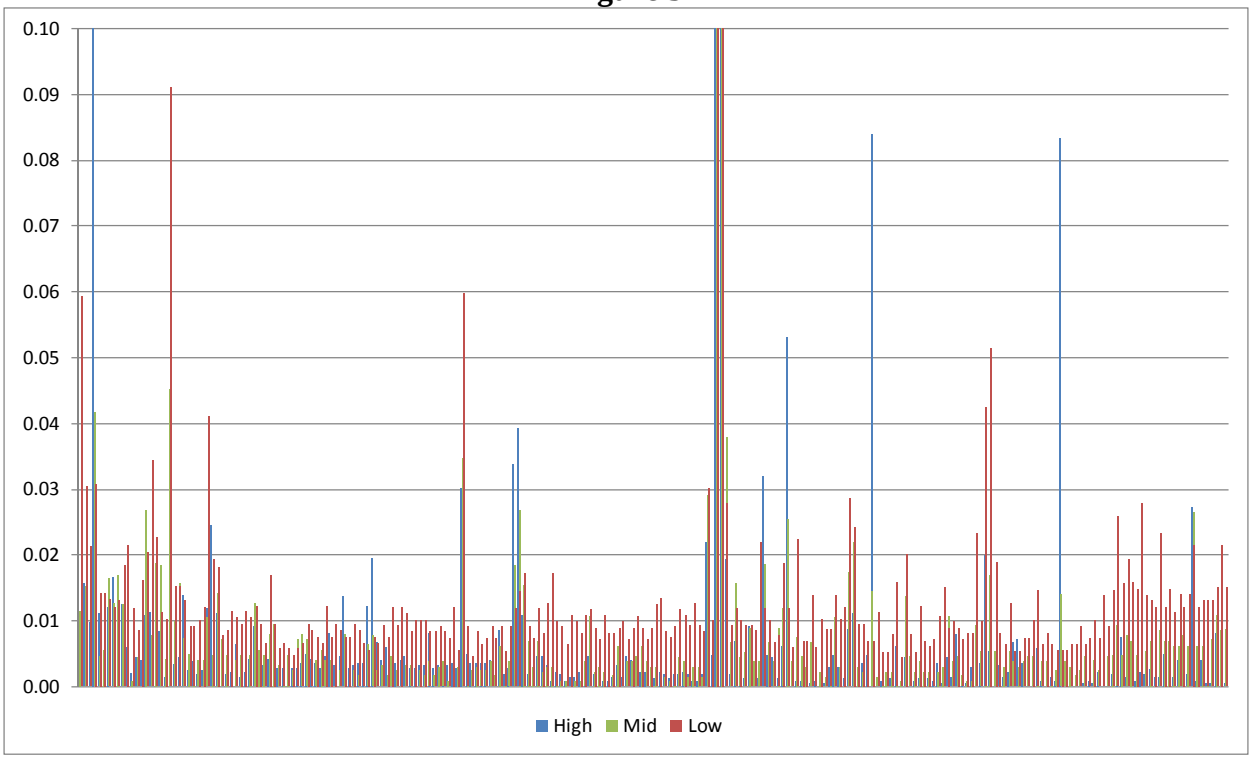
Source: Juan Shishido 2012

Figure 4



Source: Juan Shishido 2012

Figure 5



Source: Juan Shishido 2012

Table 3

Family	Mature	Young
0.87%	0.93%	1.30%

Source: Juan Shishido 2012

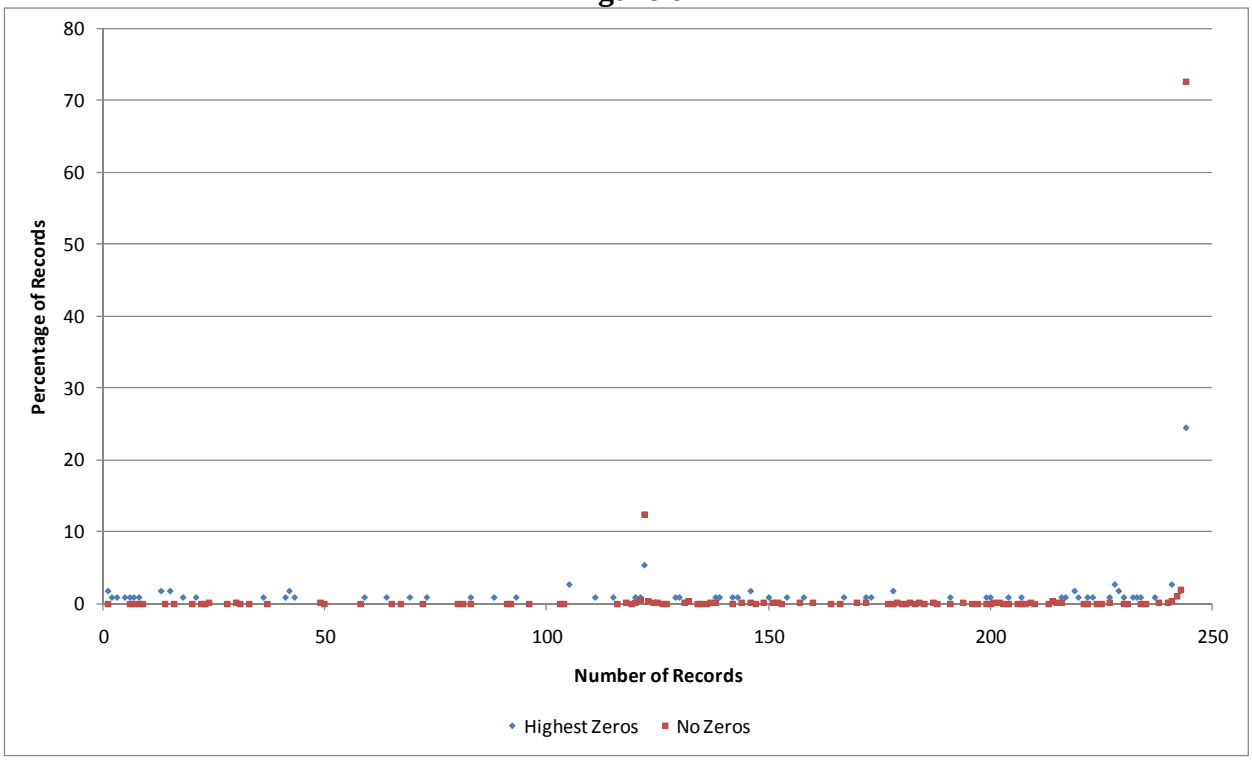
Table 4

High	Mid	Low
0.87%	0.91%	1.53%

Source: Juan Shishido 2012

Next, we examined how the zero records were distributed across customers. We established a cutoff point of 5%, which returned 110 accounts. These customers had the highest percentage of zero records. We contrasted this group with accounts having no zeros, of which there were 1,445. This comparison is shown in Figure 6, below. Overall, the no zeros group had almost three times as many customers, on a percentage basis, with a complete set of records as those in the highest zeros group. It follows, then, that the no zeros group had a smaller percentage of its customers in the other number of records categories. In others words, the highest zeros group tends to have, for whatever reason, incomplete data.

Figure 6



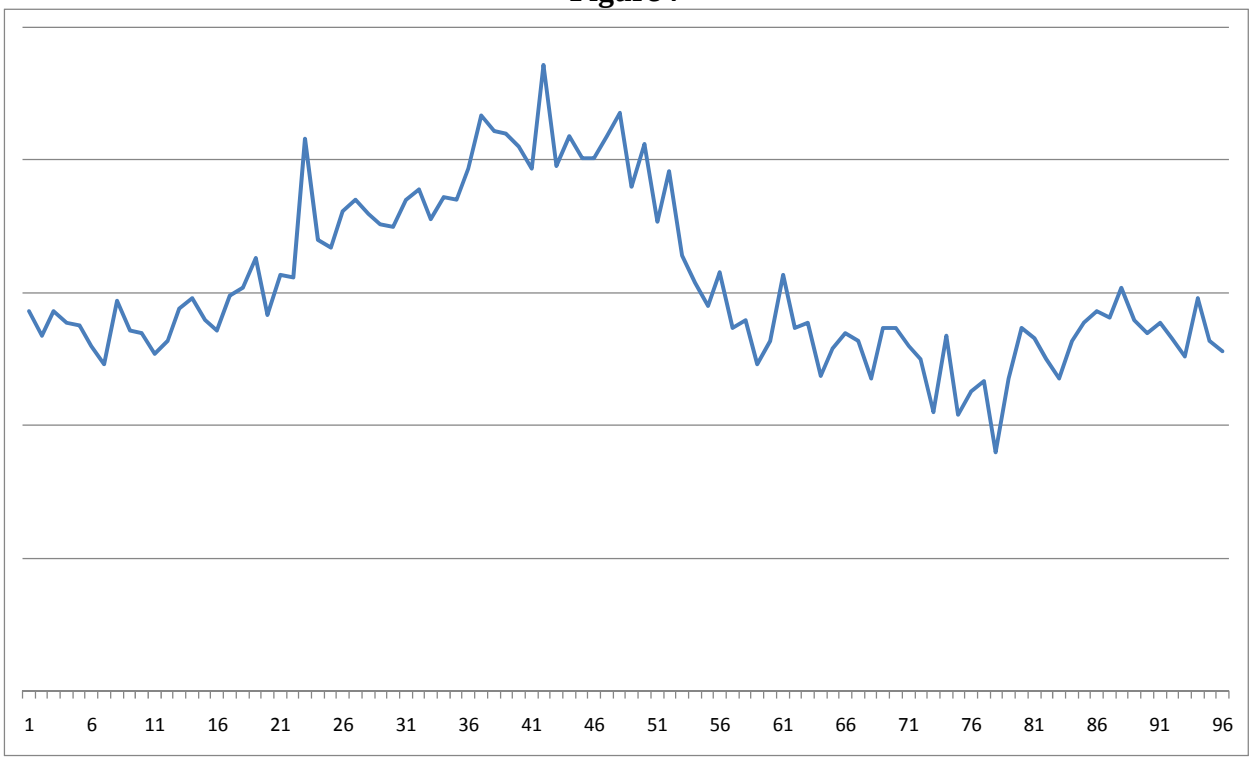
Source: Juan Shishido 2012

There are two ways in which we analyze the information above. First, we look at the reasons why these groups have the number of records they do and, second, we try to explain why certain customers fall into each of these group. In terms of demographics, in the highest zeros group, an overwhelming majority are either young, low-income or young, middle-income customers. These are the types of customers who tend to move more often, perhaps even out of

the study participation service territory, potentially explaining their lack of complete data. The no zeros group, on the other hand, is made up of mostly mature, high-income and young, high-income customers. Examining age and income separately, the highest zeros group is made up of mostly young or low-income households while the no zeros group, while evenly distributed across age segments, is more than 50% high-income. This leads us to the conclusion that the reason the no zeros group is made up of mostly high income households is that electricity use is positively correlated with income, due to larger homes and a higher prevalence of consumer electronics.

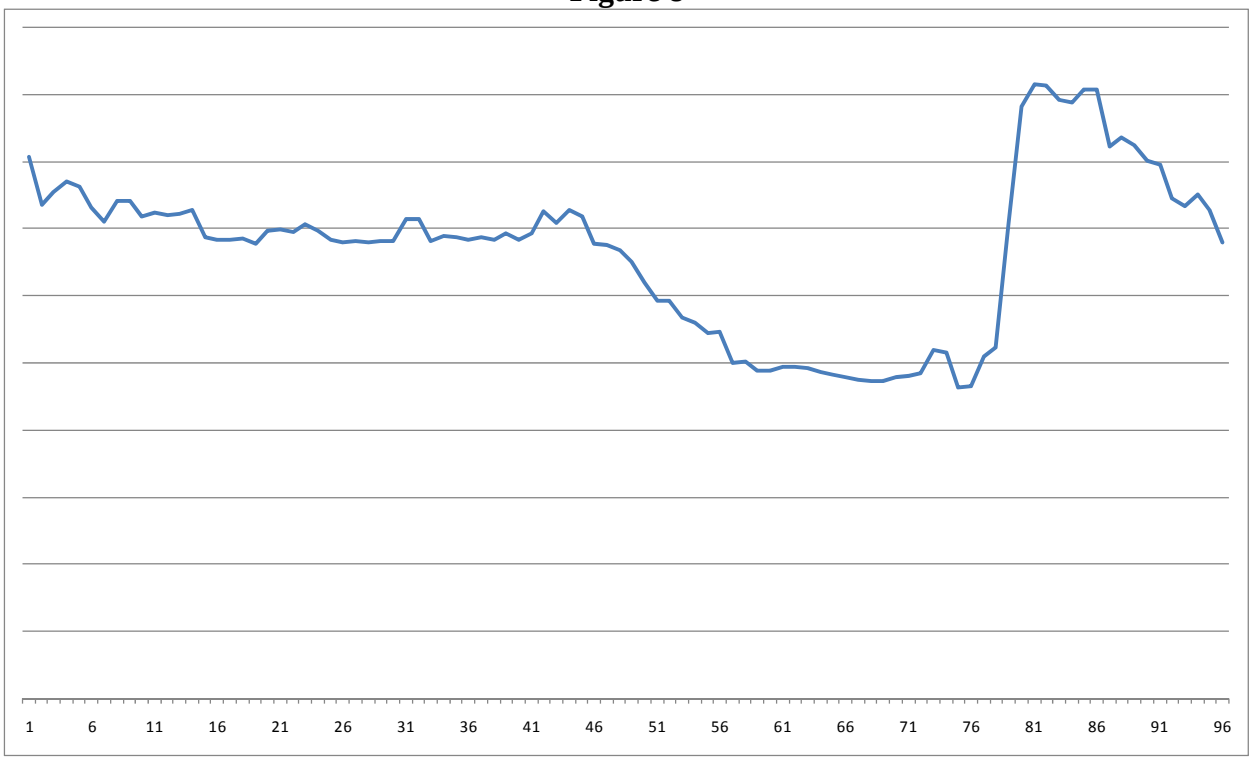
We were also interested in how the zeros were distributed across the hours of the day for those in the highest zeros group. As seen below in Figure 7, we find that there are more zeros during the first half of the day and less during the second half. It makes sense that there are less zeros during the evening hours, as that's when residential load is at its highest, and that there are more zeros during the morning and nighttime hours, when residential load is at its lowest. It's curious, however, that the most zeros tend to occur around mid-day, at a time that does not usually have the lowest consumption. It is possible that the higher-than-expected amount of zeros is due to meter change outs. We see evidence of this, where meter change outs resulted in all zeros on the day of the install as well as on subsequent days. In addition, meter change outs usually happened during the daylight hours. Across all observations, zeros are distributed as shown in Figure 8. Here, we see what we expect. The number of zeros is mostly flat for the first half of the day, then drops off during the on-peak time, and finally increases dramatically during the last 15% of the day, when people are usually sleeping. Our analysis suggests that it is still plausible to see zero-valued intervals, especially at the 15-minute level and given the demographics of those most affected.

Figure 7



Source: Juan Shishido 2012

Figure 8



Source: Juan Shishido 2012

Spikes

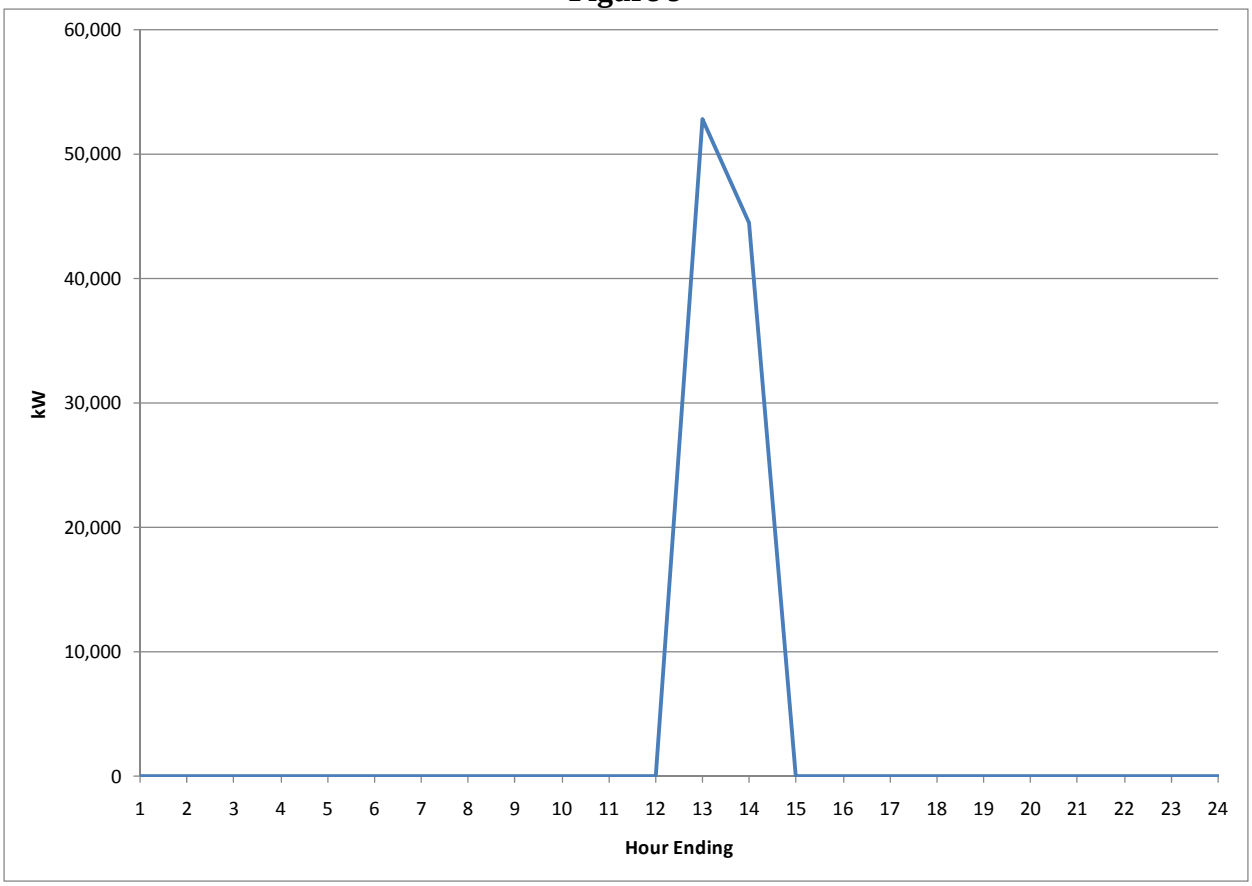
Finally, we explored whether there were cases of extremely high usage and/or spikes in the interval data. To do this, we first calculated average daily usage and its standard deviation. We then looked at how interval values changed from one interval to the next, and captured the highest percentage change per record. Using this distribution, we used the 95th percentile as a benchmark. We next isolated the records that were greater than or equal to the benchmark and where the average daily usage was greater than 0.25 standard deviations from the mean. This algorithm produced 27 records which we then scrutinized by graphing their load shapes and considering their mean and max daily usage. We found that five of the 27 records were significantly different than the others by every measure we looked at. Table 5 shows a comparison of the average mean and max usage values across both groups. Figure 9 is a representative load shape for the isolated five group. This distinctive obelisk shape makes it seem as if all usage values across the other hours of the day are zero. This isn't the case. When we edited these data points, using linear interpolation, we found that these load shapes actually looked like the one in Figure 10. While our algorithm identified twenty five additional records, their load shapes were what we expected and their mean usage values, while higher than average, were not extreme.

Table 5

Group	Mean Usage (kWh)	Max Usage (kW)
Isolated Five	4,774	41,995
Other Twenty Five	6	12

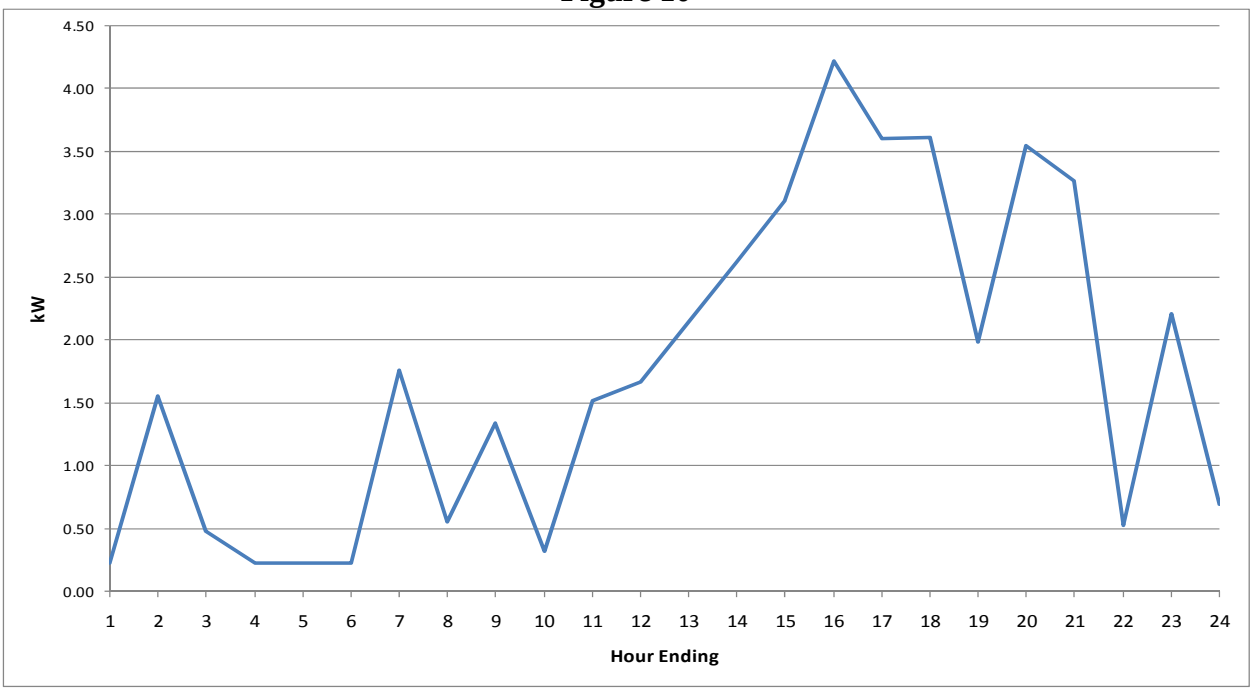
Source: Juan Shishido 2012

Figure 9



Source: Juan Shishido 2012

Figure 10



Source: Juan Shishido 2012

Conclusion

We've discussed our data quality findings on duplicates, intervals whose values were zero, and spikes. Any load researcher, data analyst, or evaluator is going to face these issues when working with smart meter interval data, even when the data have been through the Validation, Estimation, and Editing (VEE) process on the utility side. Careful examination often rewards the patient investigator, making their data more robust and their results more defensible. We have also considered some of the ways in which data quality can affect and promote energy efficiency, demand response, and evaluation activities. Smart meter interval data allows customers, through energy monitoring, to make more informed energy consumption decisions and find potential energy savings—savings which can influence them to purchase more energy efficient equipment. However, if customers aren't confident in the accuracy of smart meters, then they will be less likely to take advantage of the potential cost saving benefits they can provide. Program evaluation is often the best tool for gauging the efficacy of energy efficiency and demand response efforts. Here, data quality drives the results. These results, in turn, often inform future planning and program design endeavors, amplifying the importance of past findings. While the magnitude of all of these issues was certainly minimal, one can imagine how these invalid values may impact results. Still, there are reasons to be hopeful this will improve. Smart meters are still relatively new in the United States and as the technology and data collection strategies improve, so will the data.

References

Brown, Aaron & Weihl, Bill. 2011. *An update on Google Health and Google PowerMeter*. http://googleblog.blogspot.com/2011/06/update-on-google-health-and-google.html

Greener Buildings Staff. 2010. *Smart Meters Alone Won't Reduce Energy Use, Study Says*. http://www.greenbiz.com/news/2010/07/01/smart-meters-alone-wont-reduce-energy-use-study-says

Hessman, Kristy. 2012. *Smart Meters Get Chilly Consumer Reception.* http://www.earthtechling.com/2012/03/smart-meters-get-chilly-consumer-reception/

Lu, Edward. 2009. *Senate Committee on Energy and Natural Resources Hearing on Smart Grid.* http://www.google.com/powermeter/about/sgtestimony.html

[PG&E] Pacific Gas and Electric. 2012. *Smart Meters by the Numbers*. http://www.pge.com/myhome/customerservice/meter/smartmeter/deployment/

[SCE] Southern California Edison. 2012. *Meet the Meter & Get to Know the Grid*. http://www.sce.com/info/smartconnect/basics/smart-meters.htm

Acknowledgements

I would like to thank Richard Hart and Richard Milward for their guidance and suggestions. The analysis in this paper began as part of project work evaluating a smart grid program. Discussions with Craig Williamson enabled and encouraged me to expand upon and share my findings. I would like to extend a special thanks to Kelly Marrin who contributed greatly to helping me frame the findings and put them into context.