Title: Detrimental Effects of Poor Data Quality: Power Outage Data from 2002-2015

Authors: Shawn Adderly[1], Todd Peterson[1], Tim Sullivan[2], and Mun Son[2]

1. Pacific Gas and Electric Company, San Francisco CA
2. University of Vermont, Burlington, VT

Data for Decisions

During the analysis of large data sets, one of the biggest challenges for statisticians, data scientists and anyone performing data analysis is obtaining high-quality data. Such data is most often presented in tabular form where each line item (such as an electrical power outage event) contains several or (several dozen) associated data fields (such as time of occurrence, duration, size, etc). Part of the analysis involves cross correlation of all of these variables. When a field is empty, has the wrong format or contains faulty data, the entire line item must be dropped. The demand for data-driven decisions increases daily, but data quality does not always keep pace. For example, the US electric utility industry is in the midst of a prolonged and concerted effort to leverage outage data to guide the investment strategy for improving power grid reliability through the use of smart grid assets.

The Smart Grid

A smart grid is the enhancement of the electric grid to incorporate sensors, feedback devices, monitoring, and control devices to enable efficient grid operation (Table 1). In the US, smart grid funding matched by government funding has nearly $10 billion (Figure 1), and the impact of this investment on outage events is being sought.

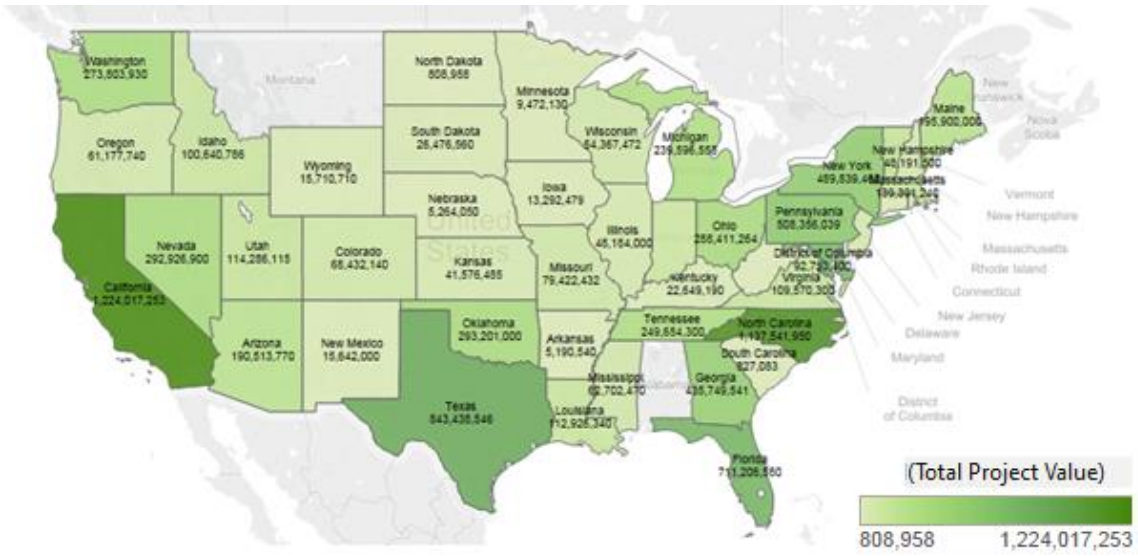

Figure 1: Smart Grid funding by State

Table 1: Smart Grid Technologies at a glance

Asset	Description
Advance Interruption Switch	Switches that can detect and clear faults more quickly or without traditional reclosing sequence
Advanced Metering Infrastructure	Electricity meters that use two-way communications to collect electricity usage and deliver information to the customer
Phasor Measurement Technology	Phasor technology enables the system operator to collect and analyze synchrophasor data
Load Monitoring	Technology that can measure and communicate line, feeder, and/or device-loading data via a communication network in real-time
Distribution Automation	Distribution Devices that can be used to perform automatic switching, reactive device coordination and other feeder operations/control

The data for electric disturbance events is recorded on form OE-417, maintained by the Office of Electric Delivery and Energy Reliability an organization inside the United States Department of Energy. US law requires this form to be submitted for any electrical disturbance greater than the reporting threshold of 300 MW. Understanding the causes of the outages, the number of customers impacted and the outage duration, enables policy makers and utilities to determine where funds should be invested and equipment

installed. Inaccurate, incomplete, missing or incorrect data require that such data be rejected from the dataset used for analysis. When data rejection does not occur on a random basis, it can skew the analysis, leading to bad decisions and, ultimately, to less effective application of resources.

Data Errors

We examine the outage data from the Department of Energy from 2002-2015 to find trends in the duration, causes, and number of customers impacted. Due to erroneous data in some fields, a substantial percentage, (61% to 88%) of the data could not be used for the analysis. Much of the remaining data had to be corrected or reformatted before analysis could be performed. Descriptions of errors encountered in the dataset are shown in Table 2. For example, in the row "Demand loss", we sometimes see the "N/A" inserted for the magnitude of the power loss. Supposing that we know for certain that the power outage event occurred, gaining meaningful insight without knowing the size of the outage event is difficult. In the fields recording time, inconsistent notation for AM and PM or using both 12 and 24 hour formats require all of the data to be reformatted by hand, which is very time consuming and also prone to generation of additional errors. Description of the causes of outages is inconsistent and uninformative.

Table 2. Problems observed in specific variables in the Electrical Outage Data.

Variable	**Meaning**	**Problem**
Date	Date event occurred	Date incorrect, multiple formats used
Time	Time event occurred	PM/AM reversed, 12 hr and 24 hr format
Restoration Date	Date power restored	Date incorrect, multiple formats used
Restoration Time	Time power restored	PM/AM reversed, 12 hr and 24 hr format
Respondent	Reporting Utility	Utility name inconsistent
Affected Area	Region covered by event	Inconsistent ways of naming area
NERC Region	US Reliability Power Regions	Misspelled NERC regions

Event Type	Cause of Outage	Inconsistent or incoherent description of event type
Demand loss	Magnitude of power loss	Useless entries: Unknown, --, O, NA
Customers Affected	Number of customers impacted by power loss event	Useless entries: Unknown, --, O, NA

Table 3. Examples of inconsistencies in the "Cause of Outage" column.

Cause of Outage	
Severe Weather - Winter Storm	Severe Weather - Snow/Ice
High Winds and Heavy Rains	Wind
Severe Thunderstorms with Strong Winds	Thunderstorms
Major Storm	Storm

Classification of the cause of outage events is inconsistent, some examples are shown in Table 3 and there is not much explanation between classifications such as storm and a major storm.

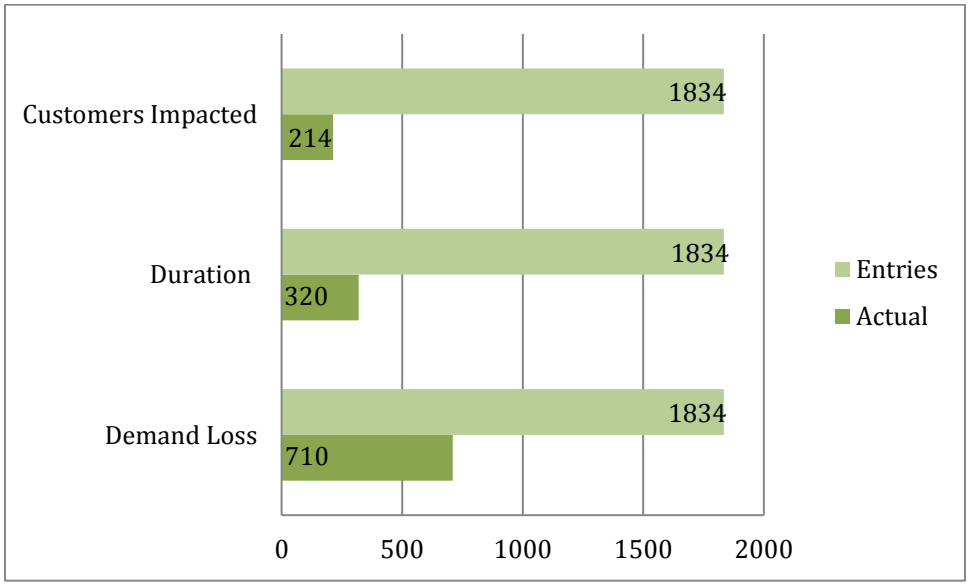

Figure 2: Graph of Available data vs. the number of entries. *

*Duration doesn't exist as a column in the original dataset but is calculated from the data and time the outage is reported and the date and time service is restored.

Large percentages of data are missing for each variable: Duration, Customer Impacted and Size of Outage as seen in in Figure 2.

Multiple Methods for Submission

Several examples of inconsistent information are presented in Tables 2 and 3. There are currently four ways to submit data to the DOE: online, email, phone and fax. While the new form standardizes a great amount of the data entry, several parts of the form still leave room for error. Submission to the DOE should only be allowed through online forms that would be rejected until the incorrect entries are corrected.

Missing Data

Missing data is the most serious problem. Poorly formatted data can be reformatted, but nothing can be done with missing data. One cause of missing data may be the need to file the form by a deadline following an outage. Or perhaps the person reporting does not know the required information and cannot obtain it quickly. In such instances, entries on drop-down menus such as "data not available until —" (where a date would be entered in the blank) would allow the form to be completed on time, but would generate a reminder asking for the missing data on the date indicated. We noticed that date errors from 2002-2015 were mostly in the 2011-2015 data, for onset and termination times of outage events. This is probably due to online submissions that require semi-validation.

Filling in Data

Some may argue that it is possible to run a predictive model and fill in the missing data, but as Figure 3 shows, predicting the size of an outage using the number of customers impacted is tricky. The predicted demand is on the top graph while the actual data is on the bottom graph. Industrial customers vary widely in size and power consumption. So a single industrial customer experiencing a large outage event does not fit an interpolation based on number of customers. Another problem is when two homes (say California and

Illinois) with the same square footage many not consume the same amount of power when LED bulbs and efficient appliances are in use in one and not the other. This may cause an analyst to question the data validity. For these and other reasons, use of interpolated data in the case of analyzing power outage events is neither advisable nor desirable.

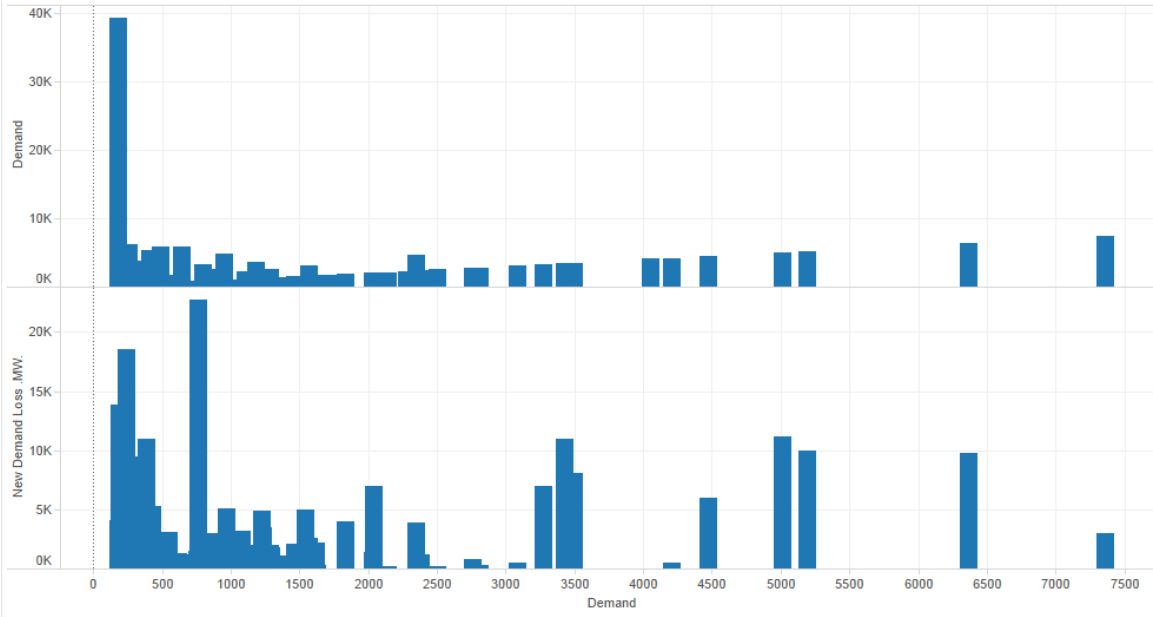

Figure 3: Raw Demand Data vs. Predicted Demand

Data Governance

The errors in the observed in the OE-417 dataset highlight the need for data governance in every industry from energy, to healthcare to the automotive industry. Data governance ensures that important data is managed with quality, integrity, security, and usability within the organization. In the case of the power data perhaps governance needs to happen inside the utility to make sure they provide the most complete and accurate information to the DOE in a timely manner. There is incentive on providing good information as the utility would use this same information to evaluate how to direct their

spending to reduce the frequency and duration of outage events. With effective data governance users of the data know the information can be trusted and if errors exist they can be corrected. This will allow researchers to have access to information that is accurate and be able to draw meaningful conclusions.

Closing Remarks

Using electric power outage data in the US, we have provided some examples of factors responsible for data loss and the resulting weakening of the associated analysis. We have also provided suggestions for the elimination of these factors to encourage improvements in future reporting. Although we have focused on electric power outage data in this article, we believe the comments apply to any dataset.

References:

1.) Batini, Carlo, and Monica Scannapieco. "Data Quality Dimensions." *Data and Information Quality*. Springer International Publishing, 2016. 21-51.
2.) Wende, Kristin. "A Model for Data Governance-Organising Accountabilities for Data Quality Management." (2007): 417-425.

About the authors:

Shawn Adderly is a Senior Data Scientist at Pacific Gas and Electric. He is part of the Electric Strategy and Asset Management group at PG&E, where he uses data to drive business decisions.

Todd Peterson is an energy economist at Pacific Gas and Electric. He is responsible for maintaining and developing gas market models and analyzing gas and electric market trends.

Tim Sullivan, PhD is a Lab Instructor at the University of Vermont.

Mun Son, PhD is a Professor of Statistics at the University of Vermont in Burlington.