Commonwealth Edison's Anonymous Data Service: A Review and Recommendations

In 2011, the Illinois General Assembly passed the Energy Infrastructure Modernization Act (EIMA). This legislation was the impetus for grid modernization and advanced metering infrastructure (AMI) investments by the utilities. A key justification for the significant capital expenditures required to enable AMI deployment was the ability for customers to take advantage of smart grid capabilities[1], including access to better, more complete customer usage data. Accessible energy use data offers a variety of benefits, including the development of innovative energy programs, analysis to better understand and predict energy demand and savings potential, and the ability to calculate individual customer savings from behavioral changes or participation in programs.[2] Stakeholders, including the utilities, touted the new programs and services that would be developed when third parties such as researchers, service providers, and program implementers could identify customers' needs by accessing robust stores of their energy data.

However, it is important to find the right balance between data access and data privacy, so that customer information is protected. This means restricting third party access to personally-identifiable customer utility data in order to protect customers and preserve data privacy, while also enabling third parties to access anonymous data. Because customer privacy is a vital concern for the regulators and a requirement of the utilities[3], the Illinois Commerce Commission (ICC) initiated a proceeding to determine the parameters around sharing energy use data while ensuring its security and maintaining customer privacy. Between September 2013 and July 2014 parties deliberated over how to achieve the proper balance between data access and data privacy. Ultimately, the ICC ordered that the utilities provide data anonymously[4] based on the "15/15 Rule"[5], among other provisions in a standardized Data Protocol.

The 15/15 Rule is the key to maintaining customer privacy while providing third party access to energy use data. In Illinois, this rule states that anonymized data about individual customers can be provided for groups of at least 15 customers in the same 9-digit zip code (zip+4), as long as no single customer's load makes up more than 15 percent of the total for that group. If a zip+4 contains fewer than 15 customers, then the geographic area must be expanded to reach the 15 customer threshold, or those customers must be removed from the data.

ComEd's new anonymous data service enables third parties to access residential customer energy use data that is anonymous and complies with the 15/15 Rule. This service will help ensure that ComEd's customers can take advantage of the benefits of smart grid investment. This report will describe ComEd's anonymous data service, discuss the research questions it can help answer, and make recommendations for potential improvements to the service that could expand the benefits for customers, building owners, businesses, and others.

Description

In February 2017 ComEd launched a service for accessing anonymous customer energy use data. This service allows third parties to download anonymous data through ComEd's website, at https://www.comed.com/SmartEnergy/InnovationTechnology. These data include all residential customers with smart meters who can be made anonymous under the 15/15 Rule. A third party who purchases access to the service will have 35 days to download data using ShareFile[6]. The available data cover the previous 24 months on a rolling basis; when the

[1] 220 ILCS 5/16-108.6. http://ilga.gov/legislation/ilcs/fulltext.asp?DocName=022000050K16-108.6
[2] ACEEE, "Who Benefits from Statewide Data Access Guidelines?", in *Energy Use Data Access: A Getting-Started Guide for Regulators.* http://aceee.org/sector/state-policy/toolkit/data-access
[3] 220 ILCS 5/16-108.6 requires utilities to secure the privacy of the customer's personal information (name, address, phone number, and other personally identifying information)
[4] "Anonymous" is defined for the purposes derived in this proceeding as individual customer data with customer specific information removed.
[5] ICC Docket No. 13-0506, Final Order at 17. 28 January, 2014. https://www.icc.illinois.gov/downloads/public/edocket/367604.pdf
[6] ShareFile is a secure file sharing and transfer service operated by Citrix: https://www.sharefile.com/

service was launched, the oldest available data were from January 2015. Access for all data costs $1,300 (including access to 5-digit zip code data and 9-digit zip+4 data) and access for just the 5-digit zip code data costs $900. There is a 50 percent discount for educational institutions, grant-based researchers, and government agencies. There is also the option to pre-pay for access to new data after six months ($150, $75 with discount) or 12-months ($200, $100 with discount). ComEd will also consider special requests for data, and if a request is compliant with all applicable anonymous data protocols then additional data can be provided for a separate fee.

These data are stored in .csv files, with one file per zip code per month. The .csv files are compressed into .zip files for easier transfer. For example, the 5-digit zip code data for January 2015 are stored in 98 files, each one covering a single zip code that passed the 15/15 Rule. As the AMI rollout progresses, more zip codes will be added to the data service. Therefore, in December 2015 there are 268 .csv files covering the much larger set of zip codes that passed the 15/15 Rule. These files can be downloaded individually, or an entire month can be grouped together for a batch download of all zip codes or zip+4s available in that month. After downloading the compressed .zip files, they can be "unzipped" or extracted to access the original .csv file, and then imported into a variety of data analysis programs.

The data are structured as one row per customer per day, with columns for energy use in half-hour intervals along with other customer details. Figure 1 shows how these data look. The specific fields are:

- Zip code: either 5-digit or 9-digit (zip+4).
- Delivery Service Class and Delivery Service Name: these fields specify the name and associated code for the four delivery service classes: single-family, multifamily, single-family with electric space heat, and multifamily with electric space heat.
- Account Identifier: this is an anonymous identifier for each ComEd account. New identifiers are generated for each month of data, so that customer X in one month is not the same as customer X in the next month.
- Interval Reading Date: the date when these energy use readings occurred.
- Interval length, in minutes: this should be 1,800 in all cases, since these are half-hour intervals.
- Total Registered Energy: this is the total energy use by this customer on this date.
- Interval Energy Qty: there are 50 columns for kWh energy usage; one column for each half hour interval from the half hour ending at 12:30 am (interval 0030) through the end of the day. The additional two columns for the 25th hour are included to account for daylight saving time.

Figure 1. Data structure

	A	B	C	D	E	F	G	H
1	#ZIP_CODE	DELIVERY_SERVICE_CLASS	DELIVERY_SERVICE_NAME	ACCOUNT_IDENTIFIER	INTERVAL_READING_DATE	INTERVAL_LENGTH	TOTAL_REGISTERED_ENERGY	INTERVAL_HR0030_ENERGY_QTY
2	60005	C23	RESIDENTIAL SINGLE	1000610279627580000	1/1/2015	1800	64.3927	1.68
3	60005	C23	RESIDENTIAL SINGLE	1000610279627580000	1/2/2015	1800	59.9403	1.245
4	60005	C23	RESIDENTIAL SINGLE	1000610279627580000	1/3/2015	1800	49.9033	1.3025
5	60005	C23	RESIDENTIAL SINGLE	1000610279627580000	1/4/2015	1800	49.9962	0.8325
6	60005	C23	RESIDENTIAL SINGLE	1000610279627580000	1/5/2015	1800	62.6776	0.9875
7	60005	C23	RESIDENTIAL SINGLE	1000610279627580000	1/6/2015	1800	51.146	0.8325
8	60005	C23	RESIDENTIAL SINGLE	1000610279627580000	1/7/2015	1800	58.7143	0.87
9	60005	C23	RESIDENTIAL SINGLE	1000610279627580000	1/8/2015	1800	56.3065	0.9225
10	60005	C23	RESIDENTIAL SINGLE	1000610279627580000	1/9/2015	1800	53.2966	0.855
11	60005	C23	RESIDENTIAL SINGLE	1000610279627580000	1/10/2015	1800	56.5279	1.0925
12	60005	C23	RESIDENTIAL SINGLE	1000610279627580000	1/11/2015	1800	48.853	0.885
13	60005	C23	RESIDENTIAL SINGLE	1000610279627580000	1/12/2015	1800	49.4689	0.685
14	60005	C23	RESIDENTIAL SINGLE	1000610279627580000	1/13/2015	1800	49.0927	0.8125
15	60005	C23	RESIDENTIAL SINGLE	1000610279627580000	1/14/2015	1800	53.0464	0.7125
16	60005	C23	RESIDENTIAL SINGLE	1000610279627580000	1/15/2015	1800	48.1843	0.8375
17	60005	C23	RESIDENTIAL SINGLE	1000610279627580000	1/16/2015	1800	44.8339	0.7325
18	60005	C23	RESIDENTIAL SINGLE	1000610279627580000	1/17/2015	1800	52.1167	0.9675
19	60005	C23	RESIDENTIAL SINGLE	1000610279627580000	1/18/2015	1800	42.2414	0.66
20	60005	C23	RESIDENTIAL SINGLE	1000610279627580000	1/19/2015	1800	47.2766	0.7725

ANONYMOUS_DATA_201501_60005

These datasets are very large due to the granularity. For example, the data for 5-digit zip codes in all months of 2015 are stored in 2,159 files that use 198 gigabytes (GB) of disk space (34.9 GB when compressed), and they include data for more than 1.5 million ComEd customers with more than 20 billion observations of energy use in half-hour intervals. There are fewer customers in the 9-digit zip code data because there are fewer 9-digit zip codes that pass the 15/15 Rule for anonymous data; however, they are stored in substantially more files. Specifically, for the time period from January 2016 through February 2017 there are 2,954 files for the 5-digit zip code data and 314,248 files for the 9-digit zip code data. As the AMI rollout continues, the number of available files, customers, and observations will continue to grow. With such a

rich dataset, the anonymous data service can provide a deep understanding of how ComEd's residential customers use energy.

Table 1. Size of the dataset

	# of customers	# of observations of half-hour energy use	# of .csv files	
	5-digit zip codes	5-digit zip codes	5-digit zip codes	9-digit zip codes (zip+4)
January-December 2015	1,673,711 (in Dec)	20,075,556,883	2,159	data not available[7]

Potential Research Questions

There are a variety of research questions that can be explored with such a rich and granular dataset. Some of the largest value comes from the detailed information about residential customer load shapes across a given day or month. Analysis of customer load shapes from half-hour interval data provide a thorough description of the hourly, daily, and weekly patterns of residential energy use, as well as how those patterns vary in single-family versus multi-family households, households with or without electric space heat, and in different neighborhoods. For example, an analyst can track the timing of daily energy use peaks, identify and segment different types of customers by their daily usage patterns[8], and more. Because the tool generates fresh anonymous identifiers for each month of data, it is not possible to follow the same customers over time outside of a single month. However, these data do allow for analysis of aggregate and average patterns across months, such as seasonal fluctuation in average energy use for single-family versus multifamily customers, average energy use during the school year versus summer, and similar average or overall patterns. These are just a handful of examples of the many research questions that could be explored, and tracking these sorts of average patterns will help illuminate which household or building characteristics drive fluctuations in residential energy use. Understanding *why* customers use energy the way they do can help identify points of intervention for energy efficiency programs, behavioral changes that could help customers save money on their energy bills, and more.

An analysis of residential load shapes can also be expanded into related areas. For example, an energy supplier or consumer group could explore different rate structures and calculate average customer savings or losses under time of use rates, demand-based rates, and other rates. By understanding how many and what type of customers are likely to save money on a particular rate, this type of analysis can estimate the impact of a proposed new rate before it is enacted, and help determine which customers are more likely to succeed in demand response programs. Additionally, disaggregation algorithms could be applied to these data to estimate how much residential energy use is devoted to different purposes, such as heating and cooling, appliances, lighting, or other uses. By estimating how much residential energy use goes toward particular appliances or uses, disaggregation could be used to predict savings potential for an appliance replacement service or similar equipment-based energy efficiency program. Similarly, by estimating how much energy use is sensitive to temperature changes or extreme weather, these data could be used to predict how total residential energy demand might be affected by warmer summers, more frequent heat waves, or other climate change scenarios. This information could be used by states or municipalities developing climate resiliency plans, and energy demand predictions can help utilities plan future infrastructure needs as the climate continues to change.

Additionally, a broader set of research questions can be explored by linking these data to other data sources, such as census tracts or other geographic data. For example, linking zip+4 neighborhoods to census tracts would enable an analysis of typical energy use patterns in low-income neighborhoods, neighborhoods with mostly larger families versus neighborhoods with more small households, and similar neighborhood-level comparisons. Neighborhood-level analysis can help program implementers target their service to particular communities, or supplement any of the analyses described above.

[7] These data are not available because in 2015 the majority of 9-digit zip codes did not pass the 15/15 Rule.
[8] Sam Borgeson, "Load Shape Clustering of Residential Smart Meter Data," BECC 2016. http://beccconference.org/wp-content/uploads/2016/10/Borgeson_presentation.pdf

Limitations

Although there are a lot of research questions that can be investigated using these data, there are limitations. First, since the tool generates fresh anonymous identifiers for each month of data, it is not possible to follow the same customers over time outside of a single month. This significantly limits the available options for longitudinal analysis. For example, it would be useful to track variation in energy use across a full year, or to identify how many customers would save money on a different rate across all four seasons. In the past, Elevate Energy has conducted an analysis of the potential for savings on ComEd's Hourly Pricing program, a dynamic rate option for residential customers that Elevate Energy administers on behalf of ComEd. The analysis calculated how many customers would have saved money if they had been enrolled in ComEd's Hourly Pricing program, using anonymous data provided by ComEd that compiled with the 15/15 Rule to preserve data privacy.[9] However, because Hourly Pricing savings typically vary by season it is important to conduct this analysis with a full year of data for each customer, so it is not possible to replicate it using data from the anonymous data service.

A second type of limitation is that there are many data fields for customer characteristics that are not included, such as the rate they are currently on, enrollment in ComEd programs, or other details. For example, if tariff rate were included as a data field, then an analyst could track adoption of alternative retail electric suppliers in different neighborhoods. The tariff rate codes include information about delivery service classes along with rate information, and could provide additional information as a replacement for the delivery service class fields. Adding additional data fields would be useful for providing context for analysts to better understand the energy use patterns in the data.

Recommendations

ComEd has been a national leader in smart grid innovation, and this new service is no exception. Like any other new service the initial offering can be improved, and we suggest a few changes that could improve the functionality of the anonymous data service as well as the types of analysis that are possible with these data, while still preserving anonymity and data privacy. One simple improvement would be the option to download all data for a given month as a single csv. file, rather than one file per zip code per month. This small change would make it easier to process the data and import it into an analysis program, particularly for the zip+4 data that is currently stored in hundreds of thousands of compressed .csv files.

A more impactful change would be to offer anonymous identifiers that persist over time so that the same anonymous customers can be followed over at least a full year. This change would expand the types of analysis that are possible with these data, and greatly enhance the value of the service by expanding the options for longitudinal analysis. For instance:

- Seasonal and annual fluctuations in energy use patterns are important for fully understanding a household's load shape, possibilities for conservation, potential for saving money on different rates, and more.
- Many of the analysis possibilities described above could also be enhanced with more longitudinal data. For example, disaggregation algorithms could show how appliance-level energy use changes across seasons, and analyses of rates could identify how many customers might gain or lose under a given rate, in addition to the average effects.
- Finally, many analyses adjust for weather changes to track energy use patterns in a typical year, by applying weather normalization procedures. However, all weather normalization approaches currently available are somewhat unreliable when they are applied to a single month or even a single year of data.

Although there might be some concern about identifying a customer based on their energy use patterns, the varied nature of residential energy use provides a strong privacy protection. Statistically speaking, household energy use is very volatile and fluctuates based on a large number of household characteristics, temporal cycles, and ad-hoc changes that occur in a household over time. Even with multiple years of interval energy use data for the same anonymous customer, it would be essentially impossible to identify that customer without access to restricted personally-identifiable energy use data to match with the anonymous data. For these reasons we recommend that ComEd consider offering anonymous identifiers that persist for at least two years, ideally three to five years. This would preserve data privacy and significantly increase the value of the anonymous data service.

[9] Elevate Energy, "ComEd Hourly Pricing Performance vs. Fixed-Price Rate During 2013," http://www.elevateenergy.org/prod/httpdocs/wp/wp-content/uploads/HourlyvsFlatPrice-2015-09-25-FINAL.pdf and Elevate Energy, "ComEd Hourly Pricing Performance vs. Fixed Price Rate During 2014," http://www.elevateenergy.org/wp/wp-content/uploads/HourlyvsFlatPrice-2016-09-15-FINAL.pdf

About Elevate Energy

Elevate Energy is a mission-driven organization that designs and implements programs that help people do more with less energy. We conduct research to inform the energy industry and the programs we administer.

Acknowledgments

We would like to thank ComEd and the Citizens Utility Board for their assistance with this report.