Proceedings of Statistics Canada Symposium 2014
Beyond traditional survey taking: adapting to a changing world

A Big Data Pilot Project with Smart Meter Data (abridged version)

Lily Ma[1]

Abstract

What is big data? Can it replace and or supplement official surveys? What are some of the challenges associated with utilizing big data for official statistics? What are some of the possible solutions? Last fall, Statistics Canada invested in a Big Data Pilot project to answer some of these questions. This was the first business survey project of its kind. This paper will cover some of the lessons learned from the Big Data Pilot Project using Smart Meter Data.

Key Words: Big Data; Smart Meters; Official Statistics.

1. Introduction

Last fall, Statistics Canada demonstrated its enthusiasm for big data by funding a big data pilot project that originated from the Big Ideas Conference and The Innovation Channel. The two key objectives of the Big Data Pilot Project were: 1) to use smart meter data as an example of big data, to explore what is and isn't feasible, as well as the tools and skills required, and the potential benefits and pitfalls of utilizing data of that magnitude at Statistics Canada, and 2) to test the feasibility of replacing and/or supplementing Statistics Canada's residential electricity consumption survey data with smart meter data.

The objectives of this paper are as follows: first, to describe the project by providing an overview of the data, approach and methods used; next, to present the findings in terms of the six data quality dimensions defined by Statistics Canada; and finally, to make a number of recommendations based on the project's findings for moving forward with smart meter data (in particular) and big data (in general).

2. Project Summary

2.1 Smart Meters in Canada

Smart meters are electronic meters that enable automated collection of electricity consumption data of households and small businesses (IESO, 2014). According to the International Energy Agency, between year 2008 and 2012, cumulative smart meter deployment in the world increased by 500%, and global cumulative smart meter installation will grow from under 300 million meters in 2012 to 1 billion meters in 2018 (IEA, 2013). In Canada, smart meters are fully implemented across Ontario (ON) and British Columbia (BC). The remaining provinces are at various stages of either investigating or implementing smart meters.

2.2 Smart Meter Data as an Example of Big Data

Smart meter data is a good example of big data because not only does it satisfy the high volume and high velocity criteria of big data, its source is reliable and its format is semi-structured and workable with existing Statistics Canada tools. To give one a sense of scale, as of 2014, there are approximately 4.8 million smart meters installed in

---

[1] This is a condensed version of a longer paper. For the full version, contact the author: Lily Ma, Statistics Canada, Ottawa, ON, Canada, K1A 0T6 (lily.ma@statcan.gc.ca) This research was funded by Statistics Canada's Big Ideas Conference. The author would like to thank Michael Scrim and Yves DeGuire for their ongoing support, as well as André Bernard, Larry Mckeown, Russell Kowaluk, Karla Fox, Hélène Bérard, Jean Pignal, Jean-Pierre Simard and André Loranger for their comments on earlier versions of this paper.

ON, collecting data from almost every household and small business at a rate of over 3.456 billion data points per month (IESO, 2013).

2.3 Smart Meter Data and Statistics Canada Surveys

There are currently several surveys and statistical programs at Statistics Canada that either collect or utilize residential electricity consumption data or data related to residential electricity consumption; they include the Electricity Disposition - Quarterly Sector Survey, Electricity Supply and Disposition Annual Survey, Households and the Environment, Survey of Household Spending, Quarterly Household Final Consumption Expenditure, Detailed Household Final Consumption Expenditure, Consumer Price Index, Purchasing Power Parities, Inter-city Indexes of Price Differentials, Census and System of National Accounts programs. Each of these surveys and programs could potentially benefit from smart meter data, and each represents an opportunity to be explored.

The potential of smart meter data is that in the future, instead of surveying individual utilities and households, we could collect the data directly from the smart meter entity (once we have permission). This would reduce response burden, enhance the efficiency of data collection and potentially improve the accuracy, timeliness, coherence and relevance of the data.

2.4 Obtaining Datasets

Currently, only local distribution companies (LDCs) and their authorized agents have the ability to transmit or request information from the Meter Data Management and Repository system, which processes and stores smart meter data for a geographic location. Thus, in order to obtain smart meter data, we first needed to obtain data sharing agreements with LDCs.

After initiating contact and subsequently meeting with the Smart Meter Entity and two LDCs, two tailored data sharing agreements were reached after several months of negotiations. To maintain the anonymity of the two companies, they shall be referred to as LDC A and LDC B.

In early 2014, we received over 200 GB of smart meter data from the two LDCs. We were informed that the datasets were representative samples of their residential and small business customer base.[2] Each dataset is an unbalanced panel containing anonymous, randomly selected,[3] time-stamped, hourly consumption data in kWh at the household level between the years 2007 and 2013.

To further explore the analytical potential of smart meter data, we also obtained hourly weather data from Environment Canada's weather stations (Gov't of Canada, 2014), as well as hourly Time-Of-Use price data from the provincial Energy Board (Ontario Energy Board, 2014). All of these data were matched to the geographic locations and time spans of the smart meter data obtained.

2.5 Data Transfer

Since the size of the data files obtained exceeded Statistics Canada’s current electronic file transfer (EFT) system limits, they were transported via encrypted hard drives following RCMP-approved sensitive data transportation procedures.

2.6 Data Storage

Once at Statistics Canada, the data files were uploaded to the SAS Test Grid via Secure Shell and Secure File Transfer Protocol (SSH/SFTP). 1.2 TB of server space was obtained from Shared Services Canada and utilized after other datasets were merged and backed up.

[2] Customers below 50 kW per month of demand.
[3] As defined by each LDC.

2.7 Variables of Comparison

Given the study's budget and time constraints, we focused on comparing the smart meter data sample with data from the Electricity Disposition - Quarterly Sector (QERS) Survey, as it offered the most direct comparison of concepts (i.e., consumption) and frequency (i.e., quarterly). In addition, to further explore the potential analytical applications of smart meter data, we linked hourly electricity consumption data with hourly price data and hourly weather data. Our data and variables of comparison are as listed in Table 2.7-1.

Table 2.7-1
Datasets and variables of comparison.

Data	Variables	Source
QERS Survey	Electricity delivered to residential customers (mWh)	Utilities A & B
Smart Meter Sample	Anonymous Customer ID, Hourly Electricity Consumption (kWh), TimeStamp (DDMMYY:HH:MM:SS)	Utilities A & B
Weather	Temperature (C), TimeStamp (DD/MM/YYYY/HH:MM)	Environment Canada
Time-of-Use Price	Low-, Mid- and High-Peak Price (¢)	Provincial Energy Board
Created Variables	Year (YYYY), Month (MM), Date (DD), Time (HH), Off (0,1), Mid (0,1), On (0,1), Weekend (0, 1), Holiday(0, 1)	Created according to the local calendar

2.8 Period of Comparison

After the data were converted and loaded, frequencies were run on the number of smart meter readings by quarter. For LDC A, frequencies stabilized with an average quarterly difference of less than 1% point after the fourth quarter (Q4) of 2011. For LDC B, frequencies stabilized with an average quarterly difference of around 1% point between the first (Q1) and second quarter (Q2) of 2009. Thus, the reference periods used for the comparisons in this study begins with 2011 Q4 for LDC A and 2009 Q1 for LDC B.

2.9 Data Processing

Using these timeframes, we ran descriptive statistics at the household level to identify data issues such as duplicates, missing data, and significant outliers. Often, the significant outliers were just miscategorized data. After we examined the cause(s), similar outliers, as well as duplicates and missing data, were automatically detected and deleted via algorithms with SAS programs.

After the data were cleaned, hourly weather data and hourly price data were matched to hourly smart meter data by time of consumption. Descriptive statistics and graphs were then run on various sets of data to identify their relationships.

Finally, we aggregated hourly electricity consumption on a quarterly basis to match the quarterly electricity delivered to residential consumers data from the QERS survey, and conducted trend analysis and period-to-period change comparisons between the two datasets.

All of the data processing described above was completed via parallel processing on the SAS Test Grid at Statistics Canada. Grid computing can be thought of as a parallel processing architecture in which computer resources are shared across a network, enabling one or multiple users to fully utilize processors distributed on multiple machines, as well as fully utilize multiple processors on a single machine.[4] The SAS Test Grid is composed of 6 nodes, each node composed of 16 cores, with a total of 96 cores.[5] To give a sense of scale, a standard desktop computer has 2 cores. For our project, a program that would typically take hours to complete sequentially on a workstation, took less than 15 minutes via parallel processing on the SAS Test Grid.

[4] Consultation with Deguire, Yves, Nov 1, 2013.
[5] Consultation with SED team, April 1, 2014.

2.10 Dimensions of Comparison

The quality of the smart meter data was assessed in comparison with QERS survey data using Statistics Canada's six dimensions of quality: Accuracy, Relevance, Timeliness, Coherence, Interpretability, and Accessibility (Statistics Canada, 2014).

3. Project Findings

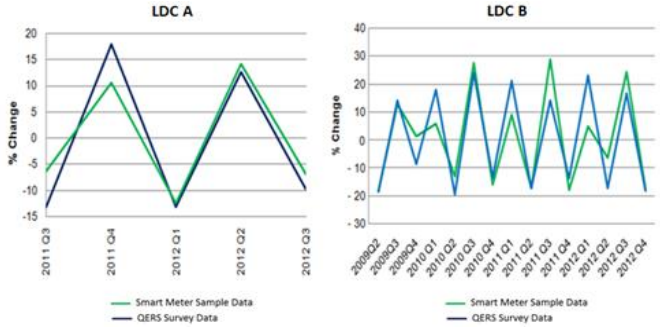
3.1 Quality Dimension 1: Accuracy

The *accuracy* of statistical information is the degree to which it correctly describes the phenomena it was designed to measure.

Since survey respondents with smart meters installed would base their survey responses on aggregated smart meter data, one can conclude *a priori* that smart meter data is at least as accurate, if not more accurate, than survey data. Smart meter data, if we were able to obtain all of it, would be free of respondent errors (e.g., reporting inaccurate numbers), non-response bias, data processing errors (e.g., data entry errors), and imputation errors. While there is the possibility of mechanical errors, whatever errors that could occur as a result of smart meters would also show up in the survey data since smart meters are what survey respondents rely on to obtain data to respond to surveys. In addition, it is in each utility's best interest to have accurate meter readings and all have regulated checks in place to keep such errors at a minimum. On the other hand, if it is the case that we could only obtain samples of smart meter data, a potential source of error would stem from the samples being non-representative. A possible solution would be to obtain the cooperation of utilities to ensure that the samples are representative (e.g., establish rules for data selection).

A posteriori, we conducted trend analysis and compared period-to-period changes between our smart meter sample data and the QERS survey data. We chose the QERS survey because it offered the most direct comparison in terms of concepts and frequency. While the totals from the samples are not directly comparable to the survey data, since the survey covers more households than the sample, because the sample is representative,[6] we were able to aggregated hourly electricity consumption for each quarter to match total electricity delivered to residential consumers' data from the survey. An alternative approach would be to do comparisons in per-household terms, but the survey doesn't contain household-level data or the number of households. Another possibility would be to scale up the smart meter based aggregate by the fraction of households sampled, but as yet we have only rough information about that, and we did not want to make assumptions about it. Therefore, we present the comparisons between the two datasets in terms of quarterly growth rates (see Figure 3.1-1).

Figure 3.1-1
Total Residential Consumption. % Quarterly Change.



When we compared the two datasets, what we found was that the period-to-period change trends generally match well, both in terms of direction and amplitude. The points where they did not match as well were when the survey data were imputed. Once we excluded the imputed survey data, the correlation between the two data sources improved, suggesting some inaccuracies caused by the current imputation method.

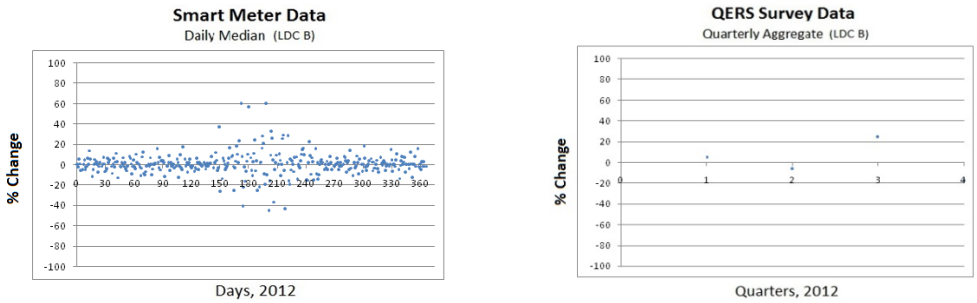
[6] As defined by the LDCs.

3.2 Quality Dimension 2: Relevance

Relevance reflects the degree to which a measure meets clients' need by shedding light on the issues that are important to them.

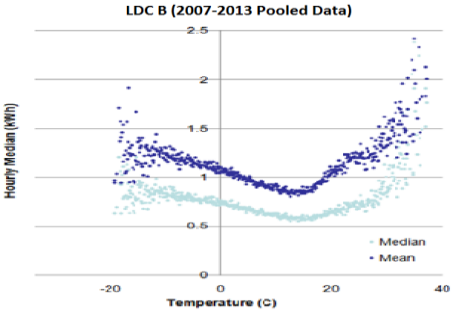
Because smart meter data is more granular and frequent than QERS survey data, it is able to reproduce all of the information that the survey can provide, and thus, any client need that can be met by the survey can be met by smart meter data. In addition, granular data is able to reveal high-frequency volatility in electricity consumption that the quarterly aggregates cannot (see Figure 3.2-1).

Figure 3.2-1
Frequency of smart meter data v survey data



Moreover, the more disaggregated smart meter data can broaden client uses. For example, smart meter data can help us identify the relationship between weather and electricity consumption by households (see Figure 3.2-2).

Figure 3.2-2
Weather Effects



This V shaped pattern in Figure 3.2-2 was found by linking hourly weather data with hourly consumption data, and it shows the extent to which most of the increases in energy consumption come from heating and cooling. This pattern is consistent with existing literature. Detailed information describing relationships between weather and electricity consumption could have policy implications, for example, on residential consumption behavior and social benchmarking.

Smart meter data can also help us identify the price elasticity of electricity demand. For example, Figure 3.2-3 illustrates that as price increases, hourly electricity consumption appears to decrease, suggesting that the Time-Of-Use pricing structure aimed at reducing consumption at peak hours may be working. This is a topic that could be further explored and would be of interest to policy makers as well as consumers.

Figure 3.2-3
Price Effects

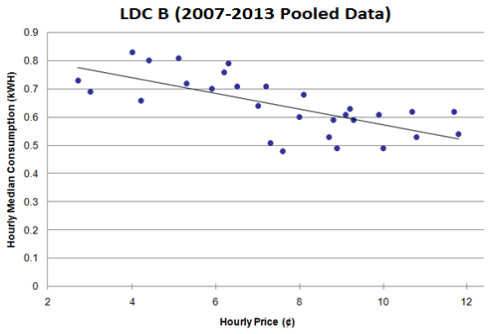
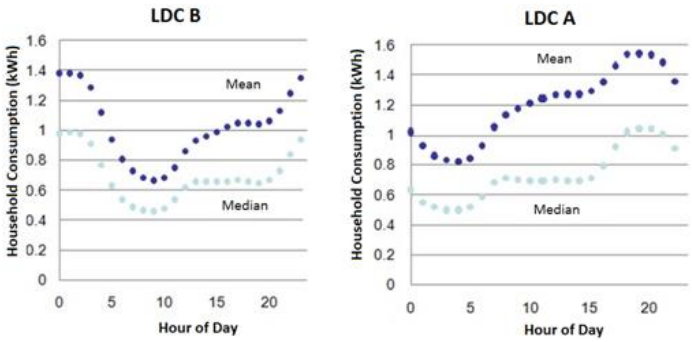


Figure 3.2-4
Geographic Differences



In this other example (see Figure 3.2-4), smart meter data could help us to identify geographic differences in consumption patterns. With only two utilities from the same province, we were able to discern differences in consumption patterns across geographic locations. This indicates that aggregate results could potentially hide significant regional differences that smart meter data could highlight, which would be of interest to regional regulators and other data users.

Finally, more granular smart meter data also offers us the opportunity, for example, to run cluster analysis or neural networks. The possibilities are vast.

3.3 Quality Dimension 3: Timeliness

Timeliness is defined as the delay between the reference period and the date on which information becomes available to users.

Since smart meter data are produced on an hourly basis, it offers a greater range of possible reference periods than existing survey data. However, like survey data, smart meter data is only as timely as the rate at which the data can be obtained, processed and disseminated. As discussed previously, obtaining smart meter data depends on factors like obtaining LDC permissions and having access to the right tools. Given the Smart Meter Entity and LDCs' day-to-day operation priorities, smart meter data cannot be extracted on demand for third party use. Moreover, even if they did give us permission to obtain the data we needed when we needed it, we would need to have the tools in place to transfer and process the data.

3.4 Quality Dimension 4: Interpretability

Interpretability reflects the availability of metadata and supplementary information necessary to interpret and utilize data appropriately.

While the metadata we received (e.g., data labels & descriptions) were accurate and sufficient, the availability of the supplementary information (e.g., sample size, sampling methods) depends on obtaining LDC cooperation.

3.5 Quality Dimension 5: Coherence

Coherence reflects the degree to which data can be brought together with other sources within a broad analytic framework and over time.

As we have demonstrated previously, the granular, malleable, frequent and transparent nature of smart meter data enable it to be linked with other data sources and fit into other frameworks with great ease. If in the future, we are able to obtain additional data like postal codes with smart meter data, then that information could be used to update survey frames. For example, if there are *X* number of smart meters in location *Y*, one could safely assume that there are at least *X* number of households in location *Y*. This information could benefit our census, for example.

3.6 Quality Dimension 6: Accessibility

Accessibility is defined as the ease with which clients can obtain and use pertinent data from a statistical agency.

We do not anticipate any significant differences between the accessibility of smart meter data and other sources of micro-level confidential data here at Statistics Canada, as they are all protected under the Statistics Act.

4. Recommendations

4.1 Smart Meter Data

Given that Canada is still in the early stages of implementing smart meters, to move forward with this initiative at Statistics Canada, relationships with utilities, smart meter entities and governing energy boards and SSC should be further cultivated to identify feasible and sustainable options in the long run. For instance, Information Technology teams from both sides could examine ways to expedite data transfer solutions. Public concerns about privacy should be addressed. Additionally, data usability studies with other surveys and programs should also be explored to maximize the potential benefits.

4.2 Big Data Acquisition

With the exception of analytical competitors[7] (e.g., Google, Opower), many big data sources are the by-product of organizations' machine-generated processes (e.g., business transactions, sensor readings, event logs), and thus depending on the analytical capacity and the business priorities of the organization, their capacity and willingness to extract data for outside use may vary. For example, while some organizations are capable of collecting, storing, cleaning, analyzing and extracting data with ease, others may be restricted in their capacity to manipulate data due to tight business schedules and availability of resources required to extract data for outside use. Some organizations may be hesitant to provide their proprietary data or data analysis services to Statistics Canada without compensation, while others may have strategic reasons for wanting to keep their proprietary data private. Thus, the time and effort that it takes for organizations to provide data to external parties should be addressed.

Ultimately, it is up to the organization to provide the data, and there must be incentives for organizations to provide their data to Statistics Canada. One possible approach for facilitating data acquisition is to emphasize to our partner organizations that they can benefit Canadians by providing their data. Another approach would be to emphasize that the Statistics Act strictly prohibits the public revelation of confidential information and enforces this prohibition with penalties up to and including imprisonment. These measures would protect the confidentiality of smart meter data just as they already protect the confidentiality of other sensitive information, e.g., census data. Additionally, analytical feedback could be offered as a form of compensation to the data providers. In any event, Statistics Canada will have to form partnerships with key stakeholders and create data sharing agreements tailored to specific data providers.

4.3 Data Transfer

While transporting data via encrypted hard drives and RCMP-approved procedures met our needs for this project, it is not the most efficient method of transporting sensitive data in the long run. Since Shared Services Canada (SSC), a new organization created to centralize federal IT infrastructure, must support any change to Statistics Canada's data storage and data processing capacity to accommodate big data, possible solutions include partnering up with SSC to investigate ways to increase capacity of current EFT service and/or to make the RCMP-approved procedures more accessible and expedient.

[7] Analytics is defined as the extensive use of data, statistical and quantitative analysis, explanatory and predictive models to drive decision and actions and an analytical competitor is defined as an organization whose competitive advantage is their analytical capabilities. Davenport,T. H., & Harris, J. G. (2007). p. 7.

4.4 Data Storage and Processing

We recommend that Statistics Canada partner with SSC to explore the possibility of scaling up and scaling out the grid (which would give a greater number of users access to greater processing power), and explore the possibility of leveraging Hadoop's distributed storage and processing architecture, and how some of the other tools available (e.g., RHadoop) can be leveraged for more complex techniques.[8]

4.5 Data Visualization

All of our data visualization was performed on summary statistics. We recommend that Statistics Canada investigate the possibility of implementing big data visual analytic tools.

5. Further Considerations

Obtaining the cooperation and partnership of external parties will be a key issue for official statistical agencies in moving forward with big data. Statistics Canada will need to address how it will elicit the cooperation of external organizations as well as how to proceed if cooperation (e.g., reaching a data sharing agreement) is not obtained. Additionally, some companies spend millions on their analytical infrastructure, and it may be more cost-effective and efficient for official statistical agencies to conduct data analysis on their system instead of duplicating datasets and systems.

If Statistics Canada is to pursue big data further, public perceptions of privacy will need to be addressed. For example, Statistics Canada can emphasize to the public that it has a long history of working with external data sources and that there is a robust system in place to protect confidential data. In cases in which big data sources are used to replace existing surveys, Statistics Canada can emphasize that the information was already being collected and big data methods simply increase accuracy and reduce response burden.

References

BC Hydro (2014), *Meter Choices*, available at: https://www.bchydro.com/energy-in-bc/projects/smart_metering_infrastructure_program/smart_meter_installation/installation_preparation/meter-choice.html (accessed June 1, 2014).

Davenport, T. H., & Harris, J. G. (2007), *Competing on analytics: The new science of winning.* Boston, MA: Havard Business School.

EMC^2 (2014), *Digital Universe*, available at: http://www.emc.com/leadership/programs/digital-universe.htm (accessed June 1, 2014).

Federation Of Canadian Municipalities (2013), *Wind Farm and Smart Grid Pilot Program*, available at: http://www.fcm.ca/home/awards/fcm-sustainable-communities-awards/2013-winners/2013-energy-projects-co-winner-2.htm (accessed June 1, 2014).

Fortis Alberta (2012), *Our Meters*, available at: http://www.fortisalberta.com/residential/customerservice/meters/Pages/FortisAlberta-Meters.aspx (accessed June 1, 2014).

Government of Canada (2014), *Historical Climate Data*, available at: http://climate.weather.gc.ca/ (accessed February 1, 2014).

Hydro Quebec (2014), *Project*, available at: http://meters.hydroquebec.com/questions-answers/project/meter-replacement-hydro-quebec (accessed June 1, 2014).

IDC (2011), *Top 10 Predictions*, available at: http://cdn.idc.com/research/Predictions12/Main/downloads/IDCTOP10Predictions2012.pdf (accessed June 1, 2014).

IESO (2014), *How Your Smart Meter Works*, available at: https://www.ieso.ca/imoweb/siteshared/smart_meter_information.asp?sid=ic (accessed June 1, 2014).

[8] Hadoop is an open source framework for distributed storage and processing of large sets of data on commodity hardware.

IESO (2013), *Ontario Smart Grid Progress Assessment: A Vignette*, available at: http://www.ieso.ca/documents/smart_grid/Smart_Grid_Progress_Assessment_Vignette.pdf (accessed June 1, 2014).

International Energy Agency (2013). Tracking Clean Energy Progress 2013. Paris: IEA.

Manitoba Hydro (2010), available at: http://www.hydro.mb.ca/regulatory_affairs/electric/gra_2010_2012/Appendix_24.pdf (accessed June 1, 2014).

Newfoundland and Labrador Hydro (July 2012), *A Report To The Board Of Commissioners Of Public Utilities*, Available at: http://www.pub.nf.ca/applications/NLH2013Capital/files/application/NLH2013Application-VolumeII-Report23.pdf (accessed June 1, 2014).

Nova Scotia Power (2014), *Nova Scotia Power Answered*, available at: http://tomorrowspower.ca/answer/106 (accessed June 1, 2014).

Ontario Energy Board (April 29, 2014), *Time-of-use (TOU) Prices*, available at: http://www.ontarioenergyboard.ca/OEB/Consumers/Electricity/Electricity+Prices (accessed June 1, 2014).

SaskEnergy (n.d.), *Advanced Metering Infrastructure*, available at: http://www.saskenergy.com/residential/AMI.asp (accessed June 1, 2014).

Smartmeters (February 13, 2013), *New Brunswick Opens Smart Grid Center*, available at: http://www.smartmeters.com/the-news/smart-grid-news/3844-new-brunswick-opens-smart-grid-center.html (accessed July 1, 2013).

Statistics Canada (April 10, 2014), *Defining Quality*, available at: http://www.statcan.gc.ca/pub/12-539-x/4147797-eng.htm#elements (accessed June 1, 2014).