EPRI | ELECTRIC POWER RESEARCH INSTITUTE

# Advanced Analytics for Smart Meter Data

**3002011236**

# Advanced Analytics for Smart Meter Data

**3002011236**

Technical Update, March 2018

EPRI Project Manager

D. Dorr

ELECTRIC POWER RESEARCH INSTITUTE
3420 Hillview Avenue, Palo Alto, California 94304-1338 ▪ PO Box 10412, Palo Alto, California 94303-0813 ▪ USA
800.313.3774 ▪ 650.855.2121 ▪ askepri@epri.com ▪ www.epri.com

## DISCLAIMER OF WARRANTIES AND LIMITATION OF LIABILITIES



THE FOLLOWING ORGANIZATION, UNDER CONTRACT TO EPRI, PREPARED THIS REPORT:

**Hydro-Québec**

**This is an EPRI Technical Update report. A Technical Update report is intended as an informal report of continuing research, a meeting, or a topical study. It is not a final EPRI technical report.**

## NOTE

For further information about EPRI, call the EPRI Customer Assistance Center at 800.313.3774 or e-mail askepri@epri.com.

# ACKNOWLEDGMENTS

The following organization, under contract to the Electric Power Research Institute (EPRI), prepared this report:

Hydro-Québec
75, Boulevard René-Lévesque Ouest
Montréal (Québec) H2Z 1A4
Canada

Principal Investigators
A. Bouffard, Hydro-Québec
D. Dorr, EPRI

This report describes research sponsored by EPRI.

This publication is a corporate document that should be cited in the literature in the following manner:

*Advanced Analytics for Smart Meter Data.* EPRI, Palo Alto, CA: 2018. 3002011236.

# ABSTRACT

Electric service providers across the world are uncovering new ways to leverage smart grid investments and big data technologies for improved management and visibility of the power system. Consequently, it makes sense to coordinate across and with the many solution providers offering tools and applications that promise to yield new and unique insights from utility data sets. A repository for distribution data sources has been designed to facilitate data mining, discovery, and innovation by way of a three-way collaboration between EPRI, its member utilities, and the analytics (solutions and research) community.

As an EPRI Data Mining Initiative Research Partner, the Institut de recherche d'Hydro-Québec (IREQ) has endeavored to solve two of the most important use cases associated with smart meter analytics while providing a unique perspective on the data resolution paradigm. The subject use cases include energy theft analytics and meter connectivity validation. This report covers use case selection, key findings, and takeaways.

Overall, the goals of the Data Mining Initiative are to learn what can be accomplished with existing data, identify insights from the data, and become more versatile with data mining strategies, approaches, and the basic science associated with big data. As the research arm of Hydro-Québec, IREQ has been learning how to attain more value from their smart meter deployment. The work described in this document builds upon their efforts to create a data mining and discovery repository that leverages traditionally disparate data layers in conjunction with a common information semantic layer. The repository is designed to validate solutions to key data challenges faced by electric service providers. The data sets in the repository represent multiple years of data from the advanced metering infrastructure (AMI), distribution supervisory control and data acquisition (D-SCADA) systems, other utility assets, power system models, and a diverse customer base.

**Keywords**
Smart meters
Advanced metering infrastructure (AMI)
Data analytics
Energy theft
Meter connectivity
Phase validation

**Deliverable Number: 3002011236**

**Product Type: Technical Update**

## Product Title: Advanced Analytics for Smart Meter Data

**PRIMARY AUDIENCE:** Department managers, data analysts, distribution operation staff

**SECONDARY AUDIENCE:** Asset managers, geographic information systems staff

### KEY RESEARCH QUESTION

With the advent of smart meter data and the planned use for all types of future analytics use cases, are there minimal data resolution considerations that will impact how well certain algorithms perform? If so, it is important that electric service providers understand the minimum resolution requirements prior to defining their smart meter data sampling and collection strategies.

### RESEARCH OVERVIEW

Working with EPRI, the Institut de recherche d'Hydro-Québec (IREQ) data science team evaluated two of the more important distribution analytics use cases, namely, energy theft (revenue protection) and meter-to-phase connectivity. The use cases were evaluated using real system data collected and stored in a repository acting as a test bed for data mining and algorithm development. To understand the effectiveness of the algorithms applied, the Hydro-Québec data were then compared to actual field-verified data. This report covers IREQ use case selection, key findings, and takeaways.

### KEY FINDINGS

- For correcting circuit topology records regarding meter-to-phase connectivity, the algorithms performed best when the smart meter data was provided at the native measurement resolution (no loss due to rounding or truncation) and was sampled at 15-minute intervals.
- For detecting cases of electrical noncompliance, the optimal data scenario was at the native measurement resolution (no loss due to rounding or truncation) and sampled at 30-minute intervals.
- For the two cases considered, results will certainly be dependent upon the data available and the analytics methods employed.

### WHY THIS MATTERS

The chief value from this work and the data repository that enabled this research is increased understanding of what can be accomplished with existing data. The benefits to the analytics community are the prioritization and documentation of the most important use cases for distribution power system situational awareness, which subsequently enables EPRI research partners to focus directly on areas with known value.

### HOW TO APPLY RESULTS

It is not enough to declare that smart meter data will be used to solve situational awareness use cases. Rather, there is a need to understand what data are required to enable optimal output from a given analytic approach. The better the understanding of resolution considerations for the top use cases, the more proactive a service provider can become at ensuring that future smart meter deployments or replacements are specified with the correct data resolution criteria and that data are sampled at appropriate intervals. Subsequently, knowing the use case data requirements and resolution constraints will yield superior insights from the data and lead to more actionable and trustworthy outcomes.

## LEARNING AND ENGAGEMENT OPPORTUNITIES

- This work is a derivative product of the EPRI Distribution Modernization Demonstration project and the specific activity referred to as the Data Mining Initiative. Additional information on the initiative can be obtained at http://smartgrid.epri.com/DMD-DMI.aspx.

**EPRI CONTACT:** Doug Dorr, Program Manager, ddorr@epri.com

**PROGRAM:** Distribution Program 180 (Supplemental)



**Electric Power Research Institute**
3420 Hillview Avenue, Palo Alto, California 94304-1338 • PO Box 10412, Palo Alto, California 94303-0813 USA
800.313.3774 • 650.855.2121 • askepri@epri.com • www.epri.com

# ACRONYMS

| | |
|---|---|
| AMI | Advanced Metering Infrastructure |
| ARIMA | Autoregressive Integrated Moving Average |
| ARMA | Autoregressive Moving Average |
| D-SCADA | Distribution Supervisory Control and Data Acquisition |
| DER | Distributed Energy Resources |
| DMD | Distribution Modernization Demonstration (EPRI project) |
| ESD | Extreme Studentized Deviate |
| GIS | Geographic Information System |
| IREQ | Institut de recherche d'Hydro-Québec |
| kWh | kilowatt-hour(s) |
| LOF | Local Outlier Factor |
| PQ | Power Quality |
| SAIDI | System Average Interruption Duration Index |
| SAIFI | System Average Interruption Frequency Index |
| SCADA | Supervisory Control and Data Acquisition |
| STL | Seasonal-Trend Decomposition using Loess |
| V | volt(s) |

# CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# 1
# OVERVIEW

## Solutions to Advanced Metering Infrastructure (AMI) Use Cases from the EPRI Distribution Modernization Demonstration Data Mining Initiative

As an EPRI Data Mining Initiative Research Partner, the Institut de recherche d'Hydro-Québec (IREQ) has endeavored to solve two of the most important use cases associated with smart meter analytics and to provide a unique perspective on the data resolution paradigm. The subject use cases include energy theft analytics and meter connectivity validation. As the research arm of Hydro-Québec, IREQ has been learning how to attain more value from their smart meter deployment. The work described in this document builds upon their efforts to create a data mining and discovery repository that leverages traditionally disparate data layers in conjunction with a common information semantic layer.

The use cases evaluated by IREQ discussed in this document are just two of the many AMI analytics cases that electric service providers around the world are deploying. The main consideration is not the results, but rather the idea that analytics for power system sensor data is multifaceted. In general, the industry is in the early stages of understanding which parameters and nuances associated with the data are most impactful to the outcomes.

## More on the Data Mining Initiative

Electric service providers across the world are uncovering new ways to leverage smart grid investments and "big data" technologies for improved management and visibility of the power system. It therefore makes sense to coordinate both across the industry (with peers) and with the many solution providers presently offering tools and applications that promise to yield new and unique insights from utility data sets. This repository for distribution data sources is designed to facilitate data mining, discovery, and innovation by way of a three-way collaboration among EPRI, its member utilities, and the analytics (solutions and research) community.

Overall, as shown in Figure 1-1, the goals of the Data Mining Initiative are to 1) learn what can be accomplished with existing data; 2) identify insights from the data; and 3) become more versatile with data mining strategies, approaches, and the basic science associated with big data. The repository is designed to validate solutions to key data challenges faced by electric service providers. The data sets in the repository represent multiple years of data from the AMI, distribution supervisory control and data acquisition (D-SCADA) systems, other utility assets, power system models, and a diverse customer base.

**Data Mining Initiative Core Objectives:**

- Develop and maintain a data repository where the analysts can get data sets of interest.
- Provide members and research partners with the details on each data-driven value case (or use case) where an innovative solution would provide insight and/or benefits.
- Define the internal and external data sets necessary to populate the data repository and support each use case.
- Determine suitable data size ranges to adequately evaluate emerging big data technologies. Document to the extent practicable any data ingestion, semantic, or other challenges associated with the data sets used in the initiative.
- Develop a consolidation of the most valuable use cases per utility business unit, and describe the implementation requirements to accomplish the use cases.
- Estimate the application value of attaining the various use case insights.

**Figure 1-1**
**Restatement of Data Mining Initiative core objectives**

As of the fall quarter 2017, there are more than 40 member-prioritized AMI use cases available in the repository (see examples in Table 1-1), and over 30 research partners have joined the initiative. These research partners (universities and solution providers) each have a custom statement of work and are sharing their findings with EPRI Distribution Modernization Demonstration (DMD) project members during recorded webcasts.

Overall, the benefit of this collaborative approach is threefold:

- The benefit to *EPRI members* is the ability to better understand the existing data and associated insights.

- The benefit to the *analytics community* is the prioritization and documentation of the most important use cases for distribution power system situational awareness. This enables research partners to focus directly on areas with known value.
- Toward benefit to the *public*, the enhanced partnerships established under the Data Mining Initiative will foster a better understanding of industry needs, capture leading practices in data analytics, transfer knowledge from industry experts, and accelerate dissemination of ideas and solutions to the market. For more information on the initiative visit: http://smartgrid.epri.com/DMD-DMI.aspx

**Table 1-1**
**Sample Use Cases from the Data Mining Initiative**

| **Asset Awareness** |
|---|
| Predictive Health Index for Distribution Service Transformers |
| Predictive Health Index for Non-Communicating Recloser |
| Predictive Health Index for Distribution Line Regulators |
| Distribution Capacitor Bank Problem Detection |
| **Load and Distributed Energy Resources (DER) Awareness** |
| Customer-Owned Photovoltaic Forecasting |
| Detection of Electric Vehicle Load Signatures |
| Identifying Load Abnormalities |
| Load and DER Signature Recognition |
| **Outage Awareness** |
| Sequence of Outage Events Replay |
| Leveraging AMI Meter Flags to Analyze Momentary Outages and Voltage Sags |
| Dynamic Momentary Outage Detection and Calculator |
| **System Awareness and Grid Optimization** |
| Optimal Placement of Automated Distribution Switches |
| Optimal Sizing and Placement of Distribution Capacitor Banks in Conjunction with Controllable Smart Inverters |
| Load Balancing Using SCADA and Smart Meter Data |
| Selection of Bellwether Meters for Grid Optimization Programs |
| Virtual Monitoring of Distribution Lines |
| Development of Electrical Load Model Utilizing SCADA and AMI Data |
| Near-Real-Time Measurement and Verification for Grid Optimization Programs |
| Visualization of Distribution Network Voltage Excursions |

## More on the IREQ AMI Use Case Selection

IREQ evaluated two use cases—energy theft detection and meter connectivity validation—with the specific goal of understanding the optimal data resolution and sampling rate to achieve a suitable result. The remaining sections of this document discuss that work and the results.

Note that energy theft and meter connectivity are just two of the compelling use cases associated with smart meter analytics. According to surveys and interviews conducted for the EPRI DMD project, more than 30 of the 40 member-prioritized AMI analytics use cases contribute a positive (cost-to-value) proposition for distribution service providers. A sampling of the most popular use cases can be found in Table 1-2. The listing does not include customer analytics but focuses more on power flows and power-related analytics. Most importantly, the table provides a snapshot of the cases that distribution utilities across North America are either deploying or considering for their operations, planning, and asset management areas.

**Table 1-2**
**Example AMI Use Cases of High Value to Distribution Service Providers**

| Use Case | Primary Data | Analytic Category | Summary Description |
|---|---|---|---|
| Energy theft detection | AMI/load awareness | AMI and others | Interval metering data should be analyzed for patterns indicative of energy diversion (theft). |
| Dynamic reliability metrics | Outage awareness | AMI | Using AMI last gasp and power restored data, the System Average Interruption Duration Index (SAIDI) and System Average Interruption Frequency Index (SAIFI) can be updated as often as every 15 min. Outage start/end times can be automated. |
| Remote outage and power quality (PQ) pinging | System awareness | AMI | Remote meter pinging allows call center representatives or field crews to have instant customer awareness regarding under or overvoltage or breaker trips vs. power outage. |
| Phase and load balancing | System awareness | AMI | In most three-phase systems, voltage is monitored on one phase allowing imbalances and energy losses to be hidden. The AMI enables improved resolution to voltage imbalance. |
| Feeder level load forecasts | System awareness | AMI | Given the ability to draw 15-minute or hourly load data from the AMI, it is possible to inform and to improve load shape forecasting at the feeder level. |
| Capacitor bank health | System and asset awareness | AMI | Using 10–20 AMI meters per phase and the correct type of algorithms can help auto identify capacitor bank fuse failures and other health issues. |
| Voltage reduction validation | System and asset awareness | Distribution + AMI | The AMI + SCADA can accurately evaluate the reduction in power usage for a given feeder or substation and can confirm energy savings verification results. |
| Phase validation | System awareness | Distribution + AMI | Validating meter phase connectivity and improving the accuracy of the distribution models and geographic information system (GIS) is important for reliable use of distribution management systems. |
| Call center unloading | Practices and new technology | AMI | With the AMI, the system auto determines and reports where the outages are located and provides customers with outage and estimated time to repair notifications. |

**Table 1-2 (continued)**
**Example AMI Use Cases of High Value to Distribution Service Providers**

| Use Case | Primary Data | Analytic Category | Summary Description |
|---|---|---|---|
| Transformer health and loading | System awareness | AMI | With AMI data, distribution transformers can be ranked and assessed to understand conditions such as overload time, pre-failure signatures, voltage mismatch, and more. |
| Voltage regulator health | System and asset awareness | AMI | Using AMI and PQ instrumentation, it is possible to identify voltage regulator operational issues before failures occur. |
| Energized line segments | System awareness | Distribution + AMI | The AMI supplies additional visibility on downed live conductors and live line segments that are supposedly deenergized to support public and crew safety analytics. |

# 2
# IREQ USE CASE SELECTION

The distribution system AMI forms the basis of the transformations experienced with the advent of smart power systems. Meters now provide electricity distributors with a more accurate view of system behavior. In addition, there are many operational efficiency benefits that analytics can make in relation to meter data.

The deployment of an AMI requires a major investment along with recurring costs. The technology infrastructure required to handle the massive volume of data being generated is relatively new to the field. Based on the solution chosen, compromises may have to be made regarding the sampling rate and resolution of stored data, which could have a negative impact on the results of analytics applied to the data. This is an important consideration as many of the utilities planning smart meter deployments are considering all of the peripheral analytics benefits without a clear understanding of the minimum resolution paradigm.

This report assesses the impact of two specific meter data criteria (sampling rate and resolution) on the quality of analytics results. The aim is thus to provide a comparative base for the industry to guide or justify their investments.

## Methodology

The approach involved running two algorithms developed at IREQ and applying them to data at different frequencies and resolutions—first using the native measurement resolution at 15-minute intervals from the AMI and again using artificially reduced data sampling rates and resolution. In line with the goals of the EPRI Data Mining Initiative—which is about results and insights—the subject algorithms are not discussed in this report; however, more on each algorithm is available through IREQ.

### *Description of Source Data*

The source data originated from readings at 15-minute intervals of Class 0.2 (ANSI C12.20) smart meters as well as topological data on the power system. Table 2-1 describes the data and variables that were used. It should be noted that the meter's voltage resolution is specified at 0.01 V. This figure is used throughout the report to represent the maximum or native resolution of the data. However, when considering the accuracy and reproducibility specifications of the meter, the measurement tolerance is approximately $\pm$0.48 V. Other factors may influence this tolerance as well. Whether the subject meters were producing measurements within the specified tolerance is unverified. The desired outcome from this study is to simply compare results obtained when using the native resolution from the meter to the results obtained when using a reduced resolution.

**Table 2-1**
**Description of source data**

| Data | Variables |
|---|---|
| Smart meter | • ID of anonymous customer<br>• Electricity consumption every 15 minutes (kWh)<br>– Resolution: 0.01 kWh<br>– Accuracy: 0.2%<br>– Reproducibility: 0.2%<br>• Mean voltage observed during 15 minutes (V)<br>– Resolution: 0.01 V<br>– Accuracy: 0.2%<br>– Reproducibility: 0.2%<br>• Time stamp (YYYMMDD: HH: MM: SS) |
| Topological data | • Line number<br>• Transformer number<br>• Link between service address and transformer |

### *Description of Data Generated*

To test the performance of the algorithms using various data sampling rates and resolutions, additional data sets were generated in order to simulate readings every 30 minutes and 60 minutes, respectively, at maximum resolution or resolution rounded off to the nearest unit (kWh or V). Based on the source data at 15-minute intervals, electricity consumption was added to the mean voltage data observed over 30- and 60-minute periods.

## Choice of Analytics Cases

Two analytics cases were selected that meet a generalized need within the industry. The first one allows the link between the service address (customer) and transformer (MV-LV) to be corrected. The second algorithm is used to detect cases of noncompliance (for example, electricity theft), which has considerable business value for electricity distributors.

### *Correction of Topology*

To use smart meter data with a view toward enhanced operational efficiency, the data must be cross-referenced with power system topology. Unfortunately, the databases that supply this information are often incomplete or have a non-negligible percentage of errors. One key piece of information is the link between the service address and transformer. Research showed that about 20% of transformers are incorrectly linked to at least one service address. Fortunately, thanks to analytics, the data science team was able to correct about 97% of the linking errors using the source data previously presented. The results show the impact of reducing the sampling rate and resolution of the source data.

### *Detection of Electrical Noncompliance*

Electrical noncompliances represent a significant loss for electricity distributors. It is not surprising to note that multiple solution providers are working on developing analytics to detect such noncompliances. At IREQ, data scientists developed several methods which, when combined, enable detection of anomalies that are often associated with electricity theft. Naturally, the quality of the data being used will have an impact on the benefits obtained from the analytics.

The data set used in the present analysis includes only four transformers for a total of 22 customers. The data set is limited due to the available and field-validated information. For instance, of these 22 customers, four were involved in energy theft; however, it is just as important to know that the 18 others were compliant. Even though this amount of data seems small in comparison to the total number of smart meters, researchers in other fields have successfully proven that small amounts of data will yield nearly identical results as compared to processing a much greater sample size, in this case meters. The beauty of this finding is that big data algorithms can be vetted on small data sets. For the subject analysis, using a ratio of four positives to 18 normal incidences allows the data science team to assess the performance of the algorithm in relation to false positives.

## Results

### *Correction of Topology*

To correct the link between the transformer and service address, a circuit with 116 transformers and 1,118 service addresses was used. A complete field validation of the circuit confirmed the presence of 24 linking errors. Use of the source data in its original form (maximum resolution and sampling rate) identified 23 of the 24 known linking errors (representing one false negative, or 4%) and two incorrectly identified link changes (8% false positives). In EPRI's opinion, the most important criterion is the number of false negatives. In fact, the false positives can be easily corrected through field validation of the analytics results. However, the false negatives are missed and can only be detected through a complete validation of the line. In the present case, 99.91% precision can be attained in the link between the transformer and service address with only 25 field validations (one error remaining out of 1,118 service addresses).

Table 2-2 shows the degradation in performance of the algorithm by varying the resolution and sampling interval. Numbers in green represent acceptable outcomes, while numbers in red represent unacceptable outcomes.

**Table 2-2**
**Results of corrected topology at different data resolutions and sampling intervals**

| Sampling Interval Resolution | 15 min 0.01V/kWh | 30 min 0.01V/kWh | 60 min 0.01V/kWh | 15 min 1 V/kWh | 30 min 1 V/kWh | 60 min 1 V/kWh |
|---|---|---|---|---|---|---|
| False positive | 2 (8%) | 2 (8%) | 3 (13%) | 32 (68%) | 94 (87%) | 211 (98%) |
| False negative | 1 (4%) | 3 (13%) | 4 (17%) | 9 (38%) | 10 (42%) | 19 (79%) |

It can be noted that resolution is the determining factor in the usability of the algorithm. In fact, even with a 15-minute sampling interval, as soon as the resolution is rounded to the nearest volt, the corrections that are identified no longer provide a quality topology. A breakdown of the results shows nine error cases out of 24 that were not resolved (38%) and 32 cases that were incorrectly identified.

A maximum resolution is preferable for the correction of the link between the transformer and service address. With respect to the sampling interval, everything depends on the final precision sought. Even at a sampling interval of 60 minutes, 20 cases out of 24 are detected with a minimum number of false positives.

### *Key Takeaway for Correction of Topology*

The algorithm produced the most useful results when the data resolution was maximized and data were acquired at a 15-minute sample interval. As a given utility begins to attain and validate results for its specific system, machine-learning tools are expected to improve these more generic results allowing investigators to use less resolute data to attain acceptable results.

### *Detection of Energy Theft*

To detect electrical noncompliance, four transformers associated with 22 service addresses were used. Of these 22 customers, four cases of theft were confirmed. The results are shown in Table 2-3. Numbers in green represent acceptable outcomes, while numbers in red represent unacceptable outcomes.

**Table 2-3**
**Results of the detection of electrical noncompliances at different sampling intervals and resolutions**

| Sampling Interval Resolution | 15 min 0.01V/kWh | 30 min 0.01V/kWh | 60 min 0.01V/kWh | 15 min 1 V/kWh | 30 min 1 V/kWh | 60 min 1 V/kWh |
|---|---|---|---|---|---|---|
| True positive | 4 | 4 | 4 | 3 | 3 | 3 |
| False positive | 0 | 0 | 3 | 6 | 8 | 9 |
| True negative | 18 | 18 | 15 | 12 | 10 | 9 |
| False negative | 0 | 0 | 0 | 1 | 1 | 1 |

At the maximum resolution and highest sampling rate, all of the electrical noncompliance cases were found and there were no false positives. A change in the sampling interval to 30 minutes did not appear to have any effect. However, at a sampling interval of 60 minutes, false positives tended to increase. This result is significant because each false positive requires a manual follow-up inspection. Scaling the results as if to validate the entire line (1,118 customers), three false positives translate to 152 manual inspections, resulting in needless incurred costs. Repeating the analysis using data at the nearest volt resolution, the number of false positives increased significantly and further increased as the sampling interval increased.

Noteworthy for this analysis is that the present cases involved significant obvious theft, which is generally assumed to be easier to detect as compared sophisticated theft. EPRI believes that the quality of the results would decrease even further for less obvious theft cases.

### *Key Takeaway for Detection of Energy Theft*

To detect electrical noncompliance, the optimal scenario appears to be at maximum data resolution with a 30-minute sampling interval. Of course, this scenario depends on the detection method used.

# 3
# CONCLUSIONS

To better understand why the data resolution and sampling rate impact the analytics results, an analogy can be drawn with photography. In fact, the sampling rate can be compared to the exposure time and data resolution to the intelligibility of a photograph (see Figure 3-1).

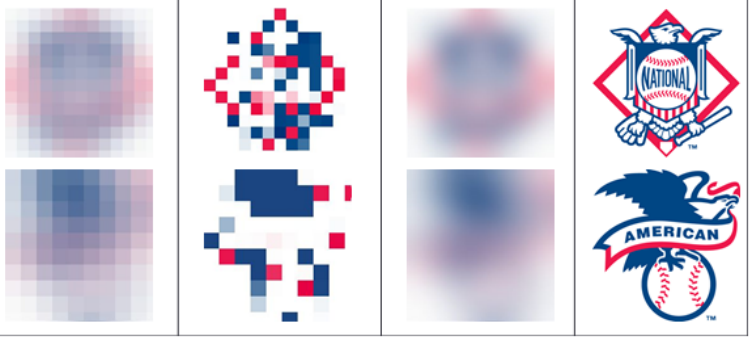
| SMART METER READINGS | PHOTOGRAPHY | SCENARIOS | | | |
|---|---|---|---|---|---|
| Sampling rate | Shutter speed (blur) | Low | High | Low | High |
| Resolution | Resolution | Low | Low | High | High |
| | |  | | | |

**Figure 3-1**
**Analogy between data and a photograph**

Certainly, the more defined the image, the greater the impact of the resolution and sampling rate degradation. The same principle applies to analytics data. Storing data generated by the AMI may entail major storage infrastructure and database costs. In order to limit expenses, an assessment could be performed to determine whether there is any benefit in reducing either the data resolution or the sampling rate. In EPRI's opinion, with today's technologies, reducing data resolution will not necessarily result in savings. In fact, it appears that data resolution is the criterion with the most influence on analytics. With respect to the sampling rate, the difference in the results between the data obtained at 15-minute intervals and the data obtained at 30-minute intervals seems negligible. Such a shift in sampling interval could result in substantial savings with respect to the information technology infrastructure, presenting an interesting course of action to consider. Naturally, the results observed are based on two specific algorithms; other data use cases would be needed for more generalizable results.

## Key Findings

Given the cost of transmitting, storing, and processing large datasets, it is not enough for a service provider to declare that it is going to use smart meter data to solve situational awareness use cases. Rather, there is a need to understand what data will be required to enable optimal output from a given analytic approach. The better the understanding of resolution considerations for the top use cases, the more proactive a service provider can become at ensuring that their future smart meter deployments or replacements are specified with the appropriate data

resolution criteria. Subsequently, knowing the use case data requirements and resolution constraints will yield much better insights from the data and lead to more actionable and trustworthy outcomes.

The primary value from this work and the data repository that enabled this research is the ability to better understand what can be achieved using existing data and what new insights the data offer. The benefit to the analytics community is the prioritization and documentation of the most important use cases for distribution power system situational awareness. This subsequently enables EPRI research partners to focus directly on areas with known value.

## Takeaways

For improved meter-to-phase connectivity accuracies, data resolution needs to be maximized and data should be acquired at a 15-minute sampling interval. In terms of the algorithms used in this analysis to detect electrical noncompliance, the optimal scenario was at highest data resolution with a 30-minute sampling interval. For the two cases considered, results will certainly depend upon the data available and the analytics methods employed.

# A
# SUPPLEMENTAL DISCUSSION ON ANOMALY DETECTION ALGORITHMS

One of the in-progress final outcomes from EPRI's DMD project is a repository of algorithms that members may find to be useful for their individual data mining activities and initiatives. To this end, EPRI is developing an algorithm repository and a theory section that should support members as they begin their analytics endeavors.

The following section suggests some of the repository content in terms of analytics theory that supports the two use cases described in this report, namely, anomaly detection as it relates to either normal energy use or to a normal meter connection. The contention here is that if data are available to define normal boundaries, anomaly detection algorithms can be used to identify those incidences that require a more thorough review.

The word "anomaly" is defined as "something that deviates from what is standard, normal, or expected." Anomaly detection refers to the practice of identifying unexpected patterns—an extremely powerful tool in terms of decision-making, data applications, and achievement of greater accuracy in forecasting future patterns or occurrences. Scientific research and development itself can be abstractly thought of as being based on these two simple steps of identifying anomalies and learning from them. When scientists understand the anomalies and the information they provide, the anomalies cease to be unexpected. The discovery of the Higgs boson subatomic particle and its integration into the standard model of particle physics is a recent example of this in practice.

In the context of algorithms and data mining, there are two main subcategories into which nearly all anomaly detection algorithms fall. The first category of algorithms can be thought of as distance or metric-based algorithms. For these algorithms to function, they must rely on certain assumptions regarding the likelihood or distribution of the observations in question. These assumptions—which are usually justified relying on the tools of probability and statistics—can be used to categorize the degree to which a given observation is expected or not expected.

1. A simple example of an anomaly detection algorithm of this type could be a density-based algorithm that makes use of the local outlier factor (LOF). The LOF can intuitively be thought of as a quantitative measure of the relative density around a given observation. The steps to compute the LOF are relatively straightforward:
2. Iterate over every observation.
3. For an observation $A$, find the $k^{th}$ nearest neighbor to the observation.
4. Compute the distance of the observation to the $k^{th}$ nearest neighbor. Define that distance to be k-distance$(A)$.
5. Define $N_k(A)$ as the set of k-nearest neighbors to $A$.
6. Define (for some distance metric $d$ and some other observation $B$,
$$\text{reachability-distance}_k(A,B) = \max\{\text{k-distance}(B), d(A,B)\}$$
7. Define the local reachability density of an observation $A$ as the following:

$$lrd(A) = \left( \frac{\sum_{B \in N_k(A)} \text{reachability-distance}_k(A, B)}{|N_k(A)|} \right)^{-1}$$

8. The local reachability densities can be compared with those of their neighbors using the following formulation to obtain the LOF:

$$LOF_k(A) = \frac{\sum_{B \in N_k(A)} \frac{lrd(B)}{lrd(A)}}{|N_k(A)|} = \frac{\sum_{B \in N_k(A)} lrd(B)}{|N_k(A)| \times lrd(A)}$$

The LOF is thus equivalent to the average local reachability density of the neighbors divided by the observation's own local reachability density.

An intuitive way to interpret the resulting scores generated with the LOF is with the following three cases:

1. $LOF_k = 1$
   - Indicates that the observation is comparable to its neighbors in relative density
2. $LOF_k < 1$
   - Indicates a much denser region (an inlier)
3. $LOF_k > 1$
   - Indicates a far less dense region (an outlier)

Now, after choosing k and computing the $LOF_k$ score for all observations, the resulting scores may look something like those in Figure A-1.

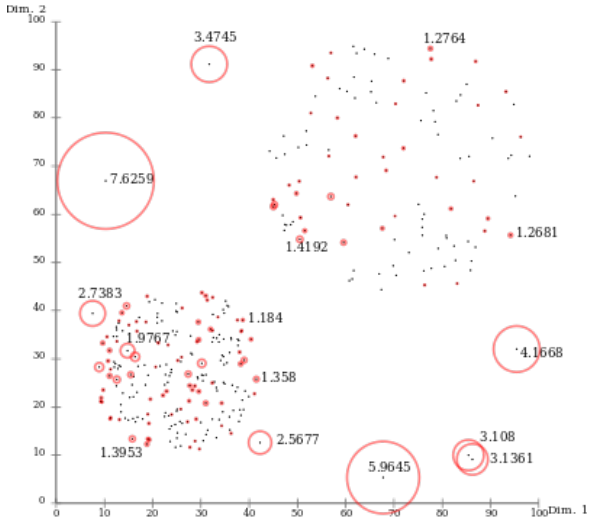



**Figure A-1**
**Computed LOF scores**

If the expectation regarding the distribution of points that must be provided to the algorithm is that these points should be clumped in a variety of different clusters corresponding to different events, then the points with the largest $LOF_k$ value may be considered as anomalies to be flagged. Such points might indicate a device malfunction or a significant low-probability event such as a total solar eclipse. On the contrary, if the expectation regarding the distribution of

points was that the points be randomly dispersed in the parameter space, then points with very small $LOF_k$ as anomalies may need to be considered. Such densely packed points might be representative of a hidden pattern or interrelatedness within the physical process that generated the data, which could prove useful if taken advantage of. An example of this might be the occurrence of foul play in a game of chance.

A similar quantitative measure of density may be found using a variety of other approaches as well. Due to the ease with which it can be run in parallel, the k-means clustering algorithm is a popular choice to characterize the relative densities for observations in large data sets.

In addition, other types of distance metrics such as Minkowsky distance, Pearson distance, or Chebychev distance may also be used in a natural manner. Depending on the assumptions of the algorithm, such alternative metrics may be more natural or provide better results as compared to those using Euclidean distance.

The second category of anomaly detection algorithms is usually applied in the context of time-series data and follows a very simple three-step procedure:

1. Forecast a signal for some point using previous data and choose the degree of confidence desired should a point appear to be anomalous.
2. Construct the confidence interval for the desired degree of confidence.
3. Check to see if the actual point value varies from the expectation (the confidence interval) enough to deem it an anomaly.

Algorithms of this type thus rely on a forecast and regression-based approach to determine whether or not a given observation is expected. The confidence interval allows for the occurrence of natural fluctuations in the signal that are associated with the reality of imperfect data measurements or are caused by chance. Even after accounting for these factors, the residuals (the differences between the actual and the predicted point values) should follow, or are statistically expected to follow, a very particular distribution. Arguably one of the most common (and powerful) implementations makes use of the median absolution deviation. Using the median absolution deviation ensures that the calculations that test the "expectedness" of observations are robust (meaning that a few very extreme points do not significantly alter the forecast or the "expectedness" of the observations). This follows from the fact that the median value of a collection of numbers is much more robust than the average value.

$$x = \{5, 5, 4, 6, 5, 4, 4, 4, 5, 7, 6\}$$

$$median(x) = 5;\ mean(x) = 5$$

Now, suppose there is some extreme observation that is observed, for example, 29.

$$x' = \{5, 5, 4, 6, 5, 4, 4, 4, 5, 7, 6, 29\}$$

$$median(x') = 5;\ mean(x') = 7$$

If the assumption regarding the observations of x over time were that each new observation be independent and tightly clustered around a single value (like a Gaussian distribution around its mean μ), then it is natural to try to estimate the value that each of the observations should be clustered around (this is what is expected). If the median is used to estimate the value of μ, then

the observations with the largest residuals are 7 and 29. (The tools of statistics can then be used to measure how unexpected these observations actually are.) On the other hand, if the mean is used to estimate the value of μ, then the observations with the largest residuals are 4, 5, and 29. Notice that the observation of the new value of 29 caused many older values that were previously expected to instead become unexpected anomalies. This is a very simple example of how a single observation can drastically affect the calculation of the expected values of observations when non-robust approaches are used.

One of the leading implementations that utilize median absolute deviation is the Anomaly Detection Library developed by Twitter (https://github.com/twitter/AnomalyDetection). The underlying algorithm is known as the Seasonal Hybrid Extreme Studentized Deviate (ESD). This algorithm actually builds off of Twitter's original Breakout Detection algorithm (https://github.com/twitter/BreakoutDetection). Without going into too much detail, the test that Twitter's algorithm utilizes to determine whether a point is an outlier is known as the Generalized ESD test. As with any approach, there are cases in which such an approach works well and other cases in which it does not work well. Depending on the problem at hand, robustness to sudden major outliers may not be what is desired.

One very common approach to forecasting time-series data is known as seasonal-trend decomposition using loess (STL). The STL method works by decomposing a time series into seasonal and trend components that are added together to obtain the resulting signal prediction. STL is very similar to the classical decomposition of a time series into seasonal and trend components; however, the key difference is the allowance for the seasonal component to change over time when STL is used. Additionally, unlike many of the other decomposition methods, STL can accommodate any type of seasonality, whether monthly, yearly, quarterly, or weekly.

Another very simple (and yet powerful) method, the autoregressive integrated moving average (ARIMA), is often used to forecast signals from time-series data. ARIMA models are actually a generalization of the autoregressive moving average (ARMA) model. ARMA models are unusual in that they require the time-series data to initially be stationary in order to function properly. As a definition, the properties of stationary time-series data necessarily do not depend on the time the series is observed. Thus, if a time series is stationary, it does not contain global trends or seasonality. ARMA methods are so called due to the fact that they regress future points with previous observations as well as previous residuals. In this way, future predictions are always based on previous predictions and the previous predictions' residuals. This is particularly desirable when rapid adaptation to temporal events is needed. At the same time, this also means that the causal events preceding a spike or drop are not always obvious. ARIMA models are basically ARMA models with one extra step—the differencing step. Using this one extra step, ARIMA models modify a given time series by taking the difference of adjacent observations until the series is stationary. After this step, an ARMA model can be fit, and it is a simple matter to work backwards to obtain the predicted signal.

The most difficult part in the application of an ARIMA model is the fact that the number of differences to perform, the number of autoregressions to regress, and the number of error coefficients to calculate and regress must be provided to the model. While there are standard methods to determine the optimal parameters for a given time series, there is also a very useful function in the R forecast library (the auto.arima function) that runs several tests to automatically determine a very good choice of parameters to use to define the ARIMA model. The ARIMA

model then provides the expectation and the associated confidence interval as the result. A common approach using ARIMA models requires the following steps

1. Construct an original model of the signal and an adjusted model of the signal using outlier points.
2. Use t-statistics to check if the adjusted model is a better fit over the original model.

Despite the differences in approaches and implementations, all anomaly detection algorithms generally follow the same structure: First, determine the expectation of an observation using prior information, and then determine whether the actual value is reasonable given this additional information as well as all prior assumptions of the problem.

**Export Control Restrictions**

Access to and use of EPRI Intellectual Property is granted with the specific understanding and requirement that responsibility for ensuring full compliance with all applicable U.S. and foreign export laws and regulations is being undertaken by you and your company. This includes an obligation to ensure that any individual receiving access hereunder who is not a U.S. citizen or permanent U.S. resident is permitted access under applicable U.S. and foreign export laws and regulations. In the event you are uncertain whether you or your company may lawfully obtain access to this EPRI Intellectual Property, you acknowledge that it is your obligation to consult with your company's legal counsel to determine whether this access is lawful. Although EPRI may make available on a case-by-case basis an informal assessment of the applicable U.S. export classification for specific EPRI Intellectual Property, you and your company acknowledge that this assessment is solely for informational purposes and not for reliance purposes. You and your company acknowledge that it is still the obligation of you and your company to make your own assessment of the applicable U.S. export classification and ensure compliance accordingly. You and your company understand and acknowledge your obligations to make a prompt report to EPRI and the appropriate authorities regarding any access to or use of EPRI Intellectual Property hereunder that may be in violation of applicable U.S. or foreign export laws or regulations.

**The Electric Power Research Institute, Inc.** (EPRI, www.epri.com) conducts research and development relating to the generation, delivery and use of electricity for the benefit of the public. An independent, nonprofit organization, EPRI brings together its scientists and engineers as well as experts from academia and industry to help address challenges in electricity, including reliability, efficiency, affordability, health, safety and the environment. EPRI members represent 90% of the electric utility revenue in the United States with international participation in 35 countries. EPRI's principal offices and laboratories are located in Palo Alto, Calif.; Charlotte, N.C.; Knoxville, Tenn.; and Lenox, Mass.

Together…Shaping the Future of Electricity



3002011236

**Electric Power Research Institute**
3420 Hillview Avenue, Palo Alto, California 94304-1338 • PO Box 10412, Palo Alto, California 94303-0813 • USA
800.313.3774 • 650.855.2121 • askepri@epri.com • www.epri.com