Visualizing Big Energy Data

Solutions for This Crucial Component of Data Analysis

By Rob J. Hyndman, Xueqin (Amy) Liu, and Pierre Pinson

VISUALIZATION IS A CRUCIAL COMPONENT OF data analysis. It is always a good idea to plot the data before fitting models, making predictions, or drawing conclusions. As sensors of the electric grid are collecting large volumes of data from various sources, power industry professionals are facing the challenge of visualizing such data in a timely fashion. In this article, we demonstrate several data-visualization solutions for big energy data through three case studies involving smart-meter data, phasor measurement unit (PMU) data, and probabilistic forecasts, respectively.

Digital Object Identifier 10.1109/MPE.2018.2801441
Date of publication: 18 April 2018

Visualizing Smart-Meter Data

Smart grid initiatives worldwide have deployed millions of smart meters into the electric grid. A small-to-medium-sized utility company could have thousands of meters spread across its territory, recording electricity demand at hourly or sub-hourly intervals. But how should one actually plot data on thousands of smart meters, with each comprising thousands of observations over time? We cannot simply produce time plots of the demand recorded at each meter, due to the sheer volume of data involved.

One approach is to convert each long series of demand data to a single two-dimensional (2-D) point that can be plotted in a simple scatter plot. In that way, all the meters can be seen in the scatter plot; outliers can be detected, clustering

1540-7977/18©2018IEEE

can be observed, and any other interesting structure can be examined. We will present a solution to this problem by first converting the data from each smart meter into a series of probability distributions, which are then used to compute pairwise distances between load profiles. The households are embedded in 2-D space to enable simple but informative plots to be constructed.

Irish Smart-Meter Data

To illustrate, we will use data collected during a smart-metering trial conducted by the Commission for Energy Regulation (CER) in Ireland. For demonstration purposes, we will use measurements of half-hourly electricity consumption gathered from 500 residential consumers over 535 consecutive days. Every meter provides the electricity consumption between 14 July 2009 and 31 December 2010. Many days in the series have periods of missing data. The CER data set does not account for energy consumed by heating and cooling systems. Either the households use a different source of energy for heating, such as oil and gas, or a separate meter is used to measure consumption due to heating. Installed cooling systems are not reported in the study.

Data from two smart meters are shown as time-series plots in Figure 1. While it is obvious that these meters have very different demand patterns, it is not possible to say much more; the time-of-day and day-of-week patterns are hidden due to the volume of data, and even the median demand is not clear from such plots.

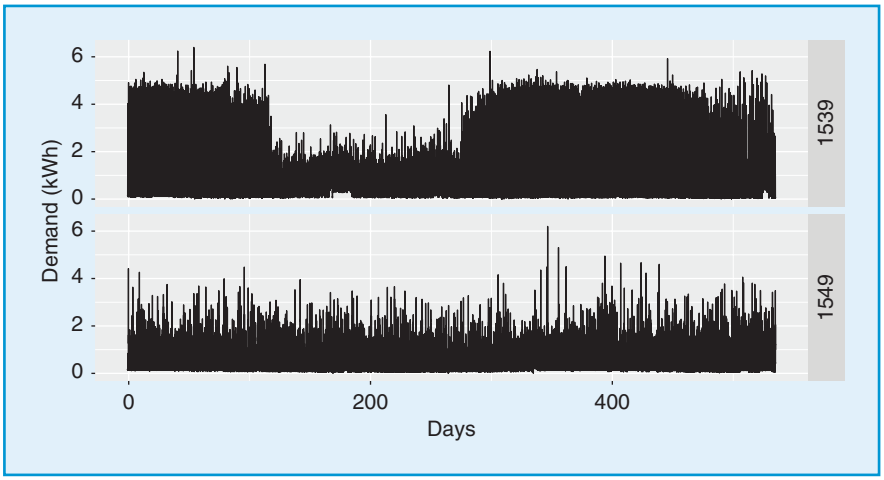

figure 1. Two examples of smart-meter demand from the CER data set.

Percentiles by Time of Week

One way to see intraday and intraweek patterns is to plot the demand against the time of the week, rather than against the time since the beginning of data collection. Figure 2 shows the same data displayed in Figure 1 but as a scatter plot against the time of the week. Now, the morning and evening peaks for meter 1539 become clear, and it is also apparent that meter 1549 has a different pattern on weekends than on weekdays.

To further examine the intraday load profiles, we can leverage the concept of percentile, which describes the distribution of the observations. The tenth percentile, for example, is the value below which 10% of the observations may be found. Widespread percentiles indicate widespread observations. On the other hand, depending upon the thickness of the percentiles on a plot, they may be overlapping each other, indicating that the observations are close to each other. In the extreme case where all observations are identical, the percentiles are identical too.

Overlaid on the individual demand data, Figure 2 also shows some percentiles of the demand distributions as they vary by half hour and day of the week, allowing us to see $48 \times 7 = 336$ probability distributions per household. For some periods, such as early morning around 4 a.m. for meter 1549, the selected percentiles are indistinguishable, indicating similar load levels because electricity consumption during sleeping hours is low and relatively certain. The evening hours (e.g., hours 18–24) are showing widespread percentiles, as the result of varying electricity consumption activities.

The percentiles are smoothed a little over time by combining neighboring half hours. For example, the percentiles for half-hour h are estimated using the data for half-hours $h-1$, h, and $h+1$. This is equivalent to a form of kernel smoothing across half hours. We use a simple estimate of each percentile to compute these curves. In this context, simple estimates are better than kernel density estimates (or some other more sophisticated estimate of the

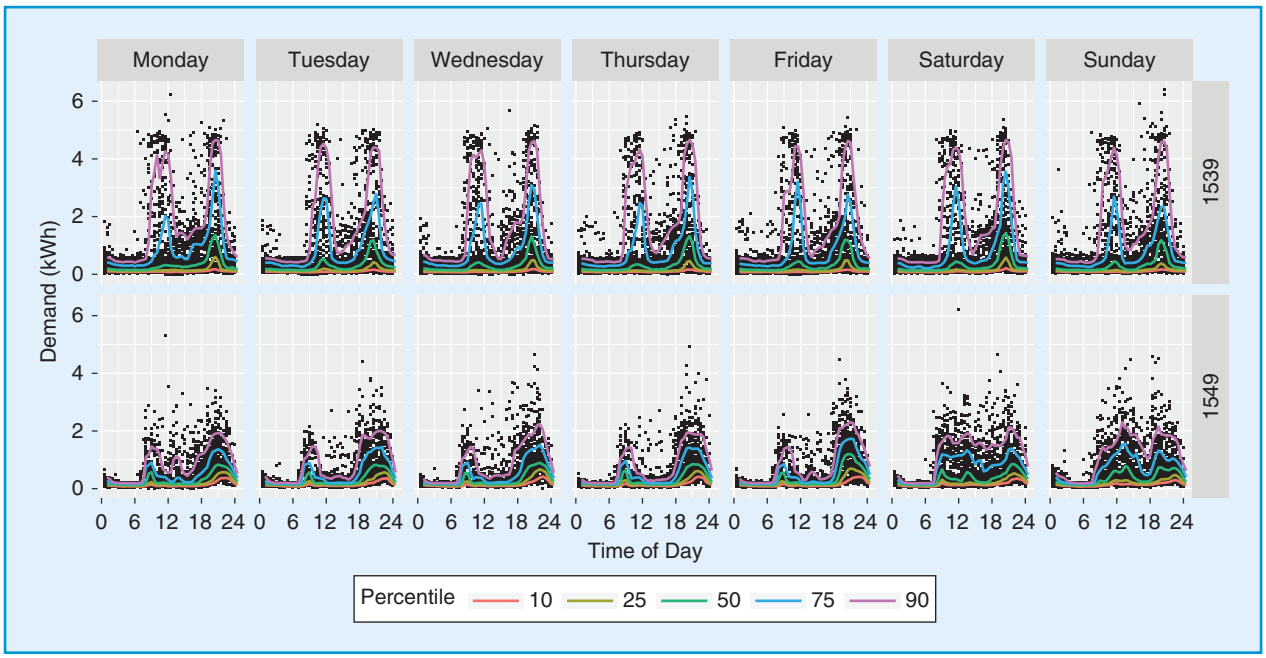

figure 2. The demand plotted against the time of the week for two smart meters from the CER data set.

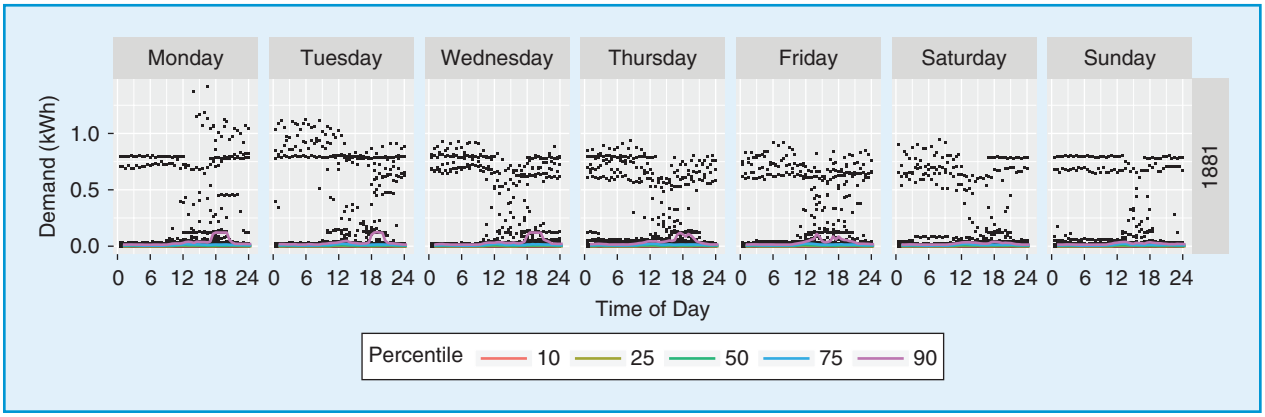

figure 3. The demand distribution of the least typical household out of the 500 smart meters included in the analysis.

distribution) because the data set contains a large number of zeros, making the distribution a mixture of discrete and continuous components. The high skewness of the data, and the nonnegative nature of demand, makes it problematic to use kernel density estimates.

There are several advantages to working with percentiles rather than the data directly. Problems with missing observations and the specific timing of household events (e.g., parties) are avoided, and attention is focused on the typical behavior of a household throughout the week. Although only five percentiles are shown in Figure 2, we actually compute percentiles for probabilities 1, 2, …, 99%.

Typical and Anomalous Households

To study the whole group of household demand distributions, we will first compute the differences in electricity consumption patterns between pairs of households. Statistically speaking, we call these differences distances. Note that the distance used here refers to the distance between two probability distributions rather than the physical distance between two houses. One way to measure the distance between two distributions is the Jensen–Shannon divergence. We have 336 probability distributions per household, one for each half-hour period of the week, so we have 336 Jensen–Shannon distance measures for each pair of households. We can measure the overall distance between the distributions from two households by summing these 336 Jensen–Shannon distance measures. In this way, we can find the distance between each pair of households in the data set.

From these pairwise distances, we can compute a measure of the typicality of a specific household by seeing how many similar houses are nearby according to the Jensen–Shannon divergence. If there are many households with similar probability distributions, the typicality measure will be high. But if there are few similar households, the typicality measure will be low. This gives us a way to find anomalies in the data set, which are the smart meters corresponding to the least typical households. The most anomalous (i.e., least typical) household is shown in Figure 3. This is clearly a very strange demand distribution, with extremely low demand almost all of the time, reflected by almost overlapping percentiles.

Visualization via Embedding

The pairwise distances between households can also be used to create a plot of all households together. If we compute 99 percentiles for 48 half hours per day and seven days a week, each of the household distributions can be thought of as a vector in K-dimensional space where $K = 99 \times 48 \times 7 = 33,264$. To easily visualize these, we need to project them onto a 2-D space. There are several ways of doing this, such as principal component analysis (PCA) and multidimensional scaling. We have used a Laplacian eigenmap method to keep the most similar points in K-dimensional space as close as possible in the 2-D space.

Figure 4 shows a 2-D embedding of the 500 households in this data set. The colors are taken from the measure of typicality, with the most typical 1% of points shown in red and the

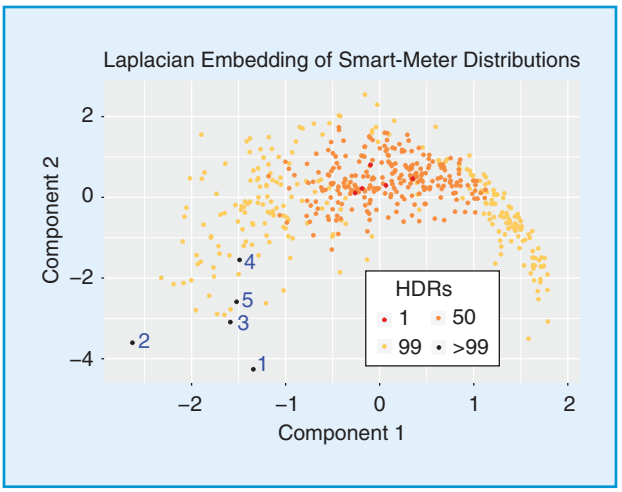

figure 4. A 2-D representation of the data from all 500 households. The most typical points are shown in red, and the most unusual are shown in black. HDR: high density region.

One approach is to convert each long series of demand data to a single two-dimensional (2-D) point that can be plotted in a simple scatter plot.

least typical 1% of points in black. The remaining points are divided into two groups with the orange points more typical than the yellow points. The blue numbers show the ranking of anomalous points. The most anomalous point (number 1) corresponds to the data shown in Figure 3.

The colors can also be interpreted as corresponding to highest density regions in the original K-dimensional space. This way of plotting the data easily allows us to see the anomalies, identify any clusters of observations in the data, and examine any other structure that might exist.

Visualizing PMU Data

Since the first prototype PMUs were developed by Virginia Tech in 1988, networked PMUs have been rapidly deployed in the last few years. As of early 2016, China and the United States have the world's largest PMU networks, with each totaling more than 2,000 PMUs in operation. Unlike the existing supervisory control and data acquisition systems that provide measurements every 2–4 s, PMUs can report data, with accurate and precise time stamps, 10–60 times per second. Consequently, we receive large volumes of high-dimensional PMU data continuously, day in and day out. Taking 30 PMUs, for example, the system operator needs to manage approximately 15 MB of data per minute, 20 GB per day, 140 GB per week or 7 TBs per year. The volume of PMU data will increase dramatically when thousands of PMUs are installed.

The problem of "too much data, too little information" must be solved, as it is becoming increasingly difficult for system operators to make use of the raw PMU data for real-time decision making. On the one hand, there is an explosion in the availability of high-rate data streams due to advances in monitoring PMU devices, leading to data overload. On the other hand, there is limited understanding on how to extract actionable information from these data-intensive monitoring devices for real-time monitoring and control purposes. Big data visual analytics offers a way forward, helping to convert these big data streams into actionable insight in real time and will aid in the development of next-generation energy management systems. In this section, we will demonstrate the most basic dimension reduction technique, PCA, as a fundamental tool for the initial steps of visualizing PMU data.

A Simple Dimension Reduction Tool—PCA

PCA, first proposed in 1901, is one of the most popular dimension-reduction techniques. Using PCA, we can remove the correlation between the variables and select only a few linearly uncorrelated variables to represent the original data. We can view the PCA as a form of orthogonal rotation, where the new axes can capture the maximum variance of the data. The orthogonal direction of the maximum variance can be identified by carrying out eigenvalue and eigenvector analysis of the covariance matrix of the sample data so that the maximum variance corresponds to the largest eigenvalues. The transformed new variables are called the principal components, while the first few principal components can explain most of the variance of the data. Thus we only require a reduced set to represent most of the information from the original data.

For event detection and diagnosis purposes, we define two statistics, T^2 and Q. T^2, constructed by the principal components, is associated with the PCA model space and represents a significant variation of the original data. Q represents the squared error of the model mismatch and the variation of the data within the residual subspace. Applying the PCA on PMU data, we can analyze many measurement sets from various locations simultaneously. We will demonstrate the elegance and the beauty of the PCA through two case studies, selected from power networks from Great Britain and Ireland.

Case 1: Visualizing Frequency Data to Distinguish Multiple Events in the Great Britain Networks

The data used here were recorded from six sites in the Great Britain networks with a 10-Hz sampling rate through the OpenPMU project, with one located in southern England, one in Manchester, and four in Orkney Islands. The well-documented event on 30 September 2012 saw a loss of load at 02:28. Later in the same day, a Great Britain–France interconnector trip event at 15:03 resulted in a Great Britain frequency drop from 49.97 to 49.60 Hz in a matter of 10 s. The initial rate of change of frequency (RoCoF) activated RoCoF-based islanding protection, erroneously disconnecting distributed generation.

We can group data from this single day into four different classes: normal data, loss of load, generation dip, and islanding event. To visualize this in Figure 5, we have plotted seven days of data randomly selected from two locations to obtain frequency coverage for normal operating conditions. It ranges from 49.8 to 50.2 Hz, represented by the black dots surrounded by the red box; this depicts the 99.9% confidence

To study the whole group of household demand distributions, we will first compute the differences in electricity consumption patterns between pairs of households.

limit. The normal data from 30 September 2012 fall in this category. In Figure 5, we have also plotted loss of load, generation dip, and islanding events from two locations. How should we interpret the patterns in this figure? Frequency is the universal parameter of the synchronous power grid, and it possesses simple and elegant characteristics. That is, the frequency data points from two locations are approximately aligned with the $y \approx x$ line. The first principal component t_1, which captures 99% of the total variance of the frequency data, is thus following this direction. In other words, we can use only one principal component to represent all frequency variables recorded across the grid.

In Figure 5, we also notice that the generation dip and loss of load events are in line with the first principal component direction, but outside the red box, with the loss of load sitting at the higher end and the generation dip sitting at the lower end. When the loss of load and generation dip events occurred in the system, the frequency variables may significantly deviate from the nominal value (50 Hz in this case) but not against each other significantly. However, for the islanding event, it is more likely that the islanded frequency deviates significantly from the rest of the system frequency and, thus, is not in line with the principal component direction. That is to say, the islanding data has its projection to the orthogonal direction to t_1 (represented by the Q axis) and is outside the red box. In comparison to the traditional time-series graph, the relative relationship of multiple events in comparison to normal operation conditions are much more straightforward, as illustrated in Figure 5.

Once an islanding event is detected in the system, the system operator will try to find out where it is located. We can accomplish this task by a simple contribution plot to visualize the contribution of individual frequency variables to the predefined PCA statistics. If the contribution of a particular frequency variable toward the Q statistic is large, an islanding site can be identified. Figure 6 illustrates variable 5 (representing PMU installed in the Orkney Island, where the islanding occurred), which dominates the contribution to Q statistic during the 9 min when it happened from 15:03:30 to 15:12:30. Both systems synchronized at 15:12:30.

Case 2: Visualizing Postdisturbance Voltage Data from Multiple Locations in the Irish Networks

We illustrate the postdisturbance voltage trajectory during an East–West Interconnector 500-MW export trip test event in the Irish network to further demonstrate PCA as a powerful dimension reduction tool for visualization.

Traditionally, the system operator will monitor the voltage traces from various locations. However, it is difficult to manage hundreds of PMUs through this traditional approach. In addition, the interaction among multiple voltage variables embedded in multiple locations is unknown. By applying PCA on the PMU data collected from 20 locations across the network, we found that three principal components are enough to monitor voltages across the entire network. The three selected principal components are capable of explaining 98% of the data variance during the

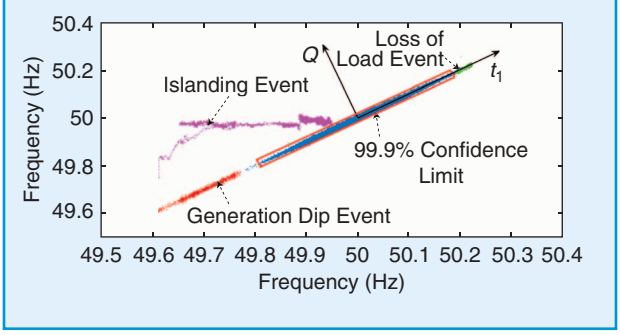

figure 5. The 2-D illustration for multiple events on 30 September 2012 recorded in the Great Britain networks. Black, blue, cyan, and purple dots represent the normal data, generation dip, loss of load, and islanding event, respectively.

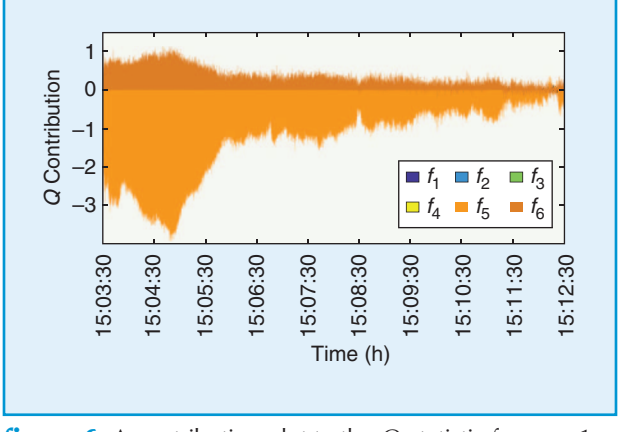

figure 6. A contribution plot to the Q statistic for case 1.

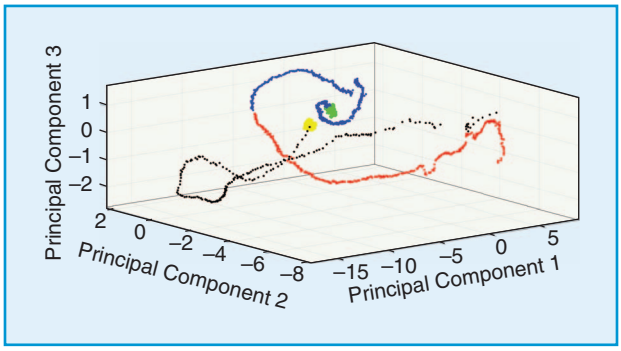

figure 7. A scatter plot of three principal components of 20 voltage variables recorded in the Irish networks for case 2.

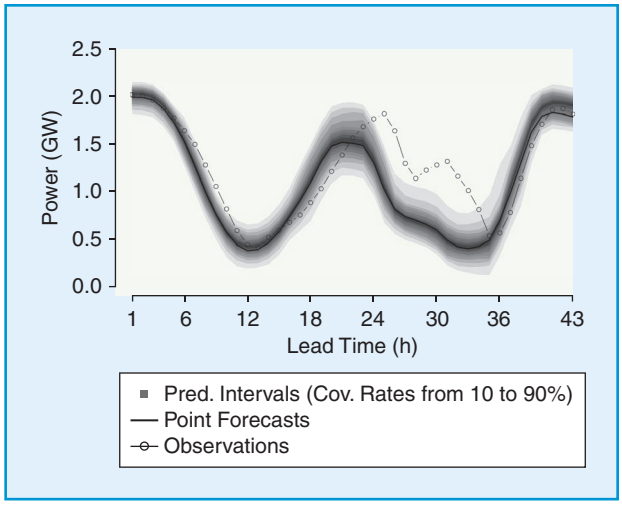

figure 8. Probabilistic forecasts represented as a river-of-blood fan chart, with a decreasing shade intensity for higher nominal coverage rate of the prediction intervals, for the whole wind power generation of western Denmark, with an hourly resolution up to nearly two days ahead.

test. As illustrated in the scatter plot of the three principal components in Figure 7, the original steady state is represented by the yellow dots; as the event progresses, it goes from the black dots to the red ones and the blue ones and finally settled to a new steady state represented by the green dots. The spiral trace indicates the oscillatory behavior during this test. The graphical visualization in Figure 7 provides a faster and easier way to interpret information, which helps reduce the decision-making time.

Visualizing Probabilistic Forecasts

While visualizing the data at the beginning of data analysis is well known to be a required step, an equally important step is visualizing the results from sophisticated models. Here we will present another case study, focusing on the visualization of forecasting results. We will use wind power forecasts as an example, although the methodology can be generally applied to other energy forecasts, such as solar power and load forecasts.

Uncertainty has always been around in power system operation and planning. For example, operational decision and control problem uncertainties originate from contingencies (generation units and lines), incomplete or erroneous overviews of the system state, and projections of future demand. Today, however, with the rapid deployment of renewable energy generation capacities throughout the world, new uncertainties are appearing that directly relate to how much power may be generated in upcoming minutes, hours, days, and beyond. Similarly, on the electricity consumption side, uncertainties are growing due to changes in consumption patterns (such as electric vehicles and more proactive consumers) but also to behind-the-meter power generation. Combined with an all-time high availability of relevant data, this has supported the increased focus on developing new approaches to analytics and forecasting for power system operations and control.

While traditional point (or single-valued) forecasts can provide the expected values for the variable of interest, probabilistic forecasts, which have been around for more than a decade, can further quantify the future uncertainties via quantiles, intervals, or probability distributions. It is challenging to visualize such uncertainties so that the probabilistic forecasts can be effectively communicated to and ultimately accepted by the business consumers of these forecasts. In this section, we will introduce and discuss alternative approaches to visualizing probabilistic wind power forecasts.

River-of-Blood Fan Chart

A prominent example of communicating probabilistic forecast information is through a "river-of-blood" fan chart, as depicted in Figure 8. An earlier version has been used, since 2005, in a significant number of technical presentations and broad-audience articles to introduce and illustrate the concept of probabilistic wind-power forecasting. This plot illustrates hourly power generation from wind power (in this case, for the whole wind power generation of western Denmark), with an hourly resolution up to nearly two days ahead. This visualization proposal is inspired by the Bank of England's probabilistic forecasts for inflation, which have published on a quarterly basis from 1996, as a pragmatic and intuitive approach to convey uncertainty information.

This so-called river-of-blood fan chart associates the traditional single-valued forecasts, relaying the mean of potential renewable power generation in the near future (formally, the conditional expectation) with a number of prediction intervals. These prediction intervals have an increasing nominal coverage rate, hence intuitively getting wider for lighter colors. For a given lead time, a prediction interval gives a range within which power generation may lie, given a certain a priori probability, i.e., its nominal coverage rate. Those prediction intervals are centered in probability on the median. The visualization appeals to both a broad audience and expert practitioners. The former may be content with a

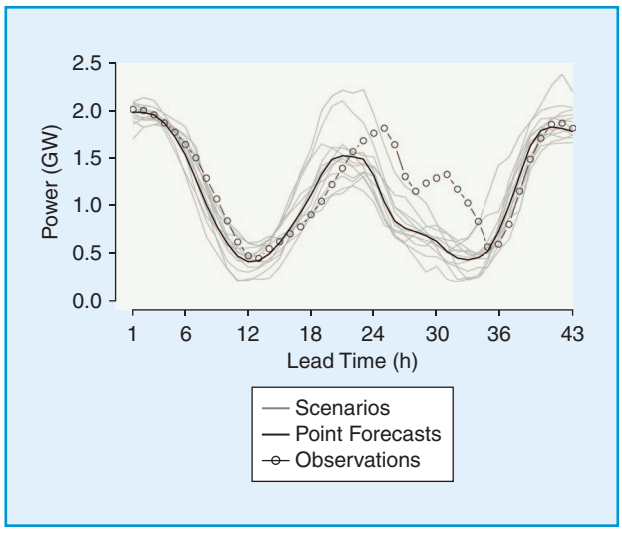

figure 9. Probabilistic forecast information conveyed by ensemble forecasts for the whole wind power generation of western Denmark.

simple and intuitive way to see how uncertain the forecasts are, while the latter is provided with enough information to reconstruct full predictive densities to be used as input to a wide range of decision and control problems in a stochastic optimization framework. Note that Figure 8 does not mean to show accurate wind power forecasts, so readers may ignore the fact that many observations are falling outside the 90% prediction interval.

Ensemble Forecasts

While the visualization in Figure 8 is appealing, it is not the only way to communicate probabilistic forecast information. Instead of the uncertainty of the future, an alternative approach provides the forecast user with a set of alternative trajectories in the future. This approach was championed by the meteorological community, which coined the term "ensemble forecast" for it. In practice, this has translated to a number of high-value applications, for instance related to trajectories of storms and cyclones and their potential impact.

For renewable energy generation, this type of representation has attracted increased interest due to the additional information it conveys, allowing the use of these alternative futures as input to existing tools for operations and control within a deterministic framework. Figure 9 depicts the ensemble forecasts that are used to convey the probabilistic forecast information for western Denmark, for a given day in the past. Since they are based on related methods, the general probabilistic information shown in Figures 8 and 9 has similarities, especially in terms of trends and uncertainty levels. However, the ensemble forecasts in Figure 9 provide additional information in terms of dependencies among lead times, which is not conveyed by the river-of-blood fan charts.

Concluding Remarks

In this article, we have offered a few examples of visualizing big-energy data. Although these examples spread across distribution (smart-meter data), transmission (PMU data), and generation (wind-power forecast data) and cover both pre- and postmodeling stages, we do not attempt to be comprehensive. There are many other insightful plots we did not present, such as maps for geospatial information (e.g., load growth and penetration of electric vehicles). Some insights are better presented dynamically via animation rather than on a static page, such as changes of load and temperature relationship over time and customer behavior changes due to the adoption of demand response programs. We hope that this article can inspire more researchers and practitioners to create effective plots from energy data.

For Further Reading

M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Comput.*, vol. 15, no. 6, pp. 1373–1396, 2003.

R. J. Hyndman and Y. Fan, "Sample quantiles in statistical packages," *Amer. Statist.*, vol. 50, no. 4, pp. 361–365, 1996.

X. Liu, D. Laverty, R. Best, K. Li, D. J. Morrow, and S. McLoone, "Principal component analysis of wide area phasor measurements for islanding detection: A geometric view," *IEEE Trans. Power Delivery*, vol. 30, no. 2, pp. 976–985, 2015.

X. Liu, J. Kennedy, D. Laverty, D. Morrow, and S. McLoone, "Wide area phase angle measurements for islanding detection: An adaptive nonlinear approach," *IEEE Trans. Power Delivery*, vol. 31, no. 4, pp. 1901–1911, 2016.

J. M. Morales, A. Conejo, H. Madsen, P. Pinson, and M. Zugno, *Integrating Renewable in Electricity Markets: Operational Problems. Series in Operational Research & Management Science*. New York: Springer Verlag, 2014.

R. Bessa, C. Möhrlen, V. Fundel, M. Siefert, J. Browell, S. H. El Gaidi, B.-M. Hodge, U. Cali, and G. Kariniotakis, "Towards improved understanding of the applicability of uncertainty forecasts in the electric power industry," *Energies*, vol. 10, no. 9, article no. 1402, 2017.

Biographies

Rob J. Hyndman is with Monash Business School, Australia.

Xueqin (Amy) Liu is with Queen's University Belfast, United Kingdom.

Pierre Pinson is with the Technical University of Denmark, Denmark.