LLNL-TR-514611

# Understanding Diurnal Patterns in Wind Power Generation Data

M. Ndoye, C. Kamath

November 17, 2011

## Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

# Understanding Diurnal Patterns in Wind Power Generation Data

Mandoye Ndoye and Chandrika Kamath
Lawrence Livermore National Laboratory
Livermore, CA 94551

***Abstract*—Integrating wind energy on the power grid is a challenging task given the intermittent nature of these resources. When the percentage of wind energy was small, control room operators did not have a major problem in scheduling wind resources. However, as this percentage has increased, it is clear that the operators will need more accurate forecasts, as well as any additional information they can exploit, to make better-informed scheduling decisions. In this paper, we investigate diurnal patterns observed in wind-power time series data. Using actual wind generation data from two sites, we try to identify these patterns, understand them better, and determine if it is possible to use weather conditions in the vicinity of the wind farms to predict the pattern for a day. Such analysis could provide insights useful in scheduling wind resources on days with inaccurate forecasts.**

***Index Terms*—Wind power generation, diurnal patterns, prediction**

## I. Introduction

RENEWABLE resources, such as wind, are providing an increasing percentage of our energy requirements. However, integrating wind energy on the power grid is challenging for several reasons. Control room operators find it difficult to schedule wind power as it is an intermittent resource. They typically use 0-6 hour ahead forecasts, along with the actual generation in the previous hours, to determine the amount of energy to schedule for the hours ahead. These forecasts are obtained from numerical weather prediction simulations or based on estimates of wind speed in the region of the wind farms. However, the forecasts can be inaccurate, especially for ramp events, where the generation suddenly increases or decreases by a large amount in a short time.

In our previous work [1], [2], we analyzed ramp events and identified important weather conditions associated with them. The control room operators could then monitor these variables to determine if a day was likely to have ramp events. In the current work, we are interested in the situation where the energy forecasts are inaccurate. In such cases, the control room operators consider the energy generation for the previous few days and hours, and based on their experience and expertise, estimate the energy they should schedule for the upcoming hour. However, as this approach can be somewhat ad-hoc, we want to determine if there are ways to improve it.

In discussions on scheduling wind resources with operators at Southern California Edison, we had observed that there appeared to be a diurnal pattern in the generation for the previous days. A closer examination of historical data confirmed the presence of these patterns. The generation may be low and flat on days with little wind, or it may be high and flat on days when the wind speed is at a sustained high level for most of the hours in the day. Or, the generation may be high in the early hours, drop down to near zero by noon, and rise again in the late evening. It is obvious to ask if there is a limited number of these patterns for the wind generation at a site? If so, can we use the expected weather conditions to predict what kind of pattern is likely for the day?

In this paper, we analyze historical data to understand these diurnal patterns better and to determine if we can use weather conditions to provide the control room operators additional information they can exploit to schedule wind energy on the grid. We start in Section II by describing the wind and weather data used in our work. Next, in Section III, we outline our approach to identifying the daily patterns. Section IV describes the results using our test-bed data sets and we conclude in Section V with a summary and ideas for future work.

## II. Data description

We conduct our analysis using data from two regions - the Tehachapi Pass in Southern California and the Columbia Basin region on the Oregon-Washington border. The wind generation data are available at 15 minute intervals for the Tehachapi Pass and at 5 minute intervals for the Columbia Basin region. While the weather data are available at different temporal resolutions from several meteorological towers in the two regions, we focus on daily averages to remain consistent with the daily patterns in the generation data.

### *A. Wind generation data*

In our study, we use data for the years 2007-2008 from wind farms in Tehachapi Pass that feed into the grid through Southern California Edison (SCE) and the 2007-2009 data from farms in the Columbia Basin region which are part of the Bonneville Power Administration (BPA) balancing area [1], [3]. We chose data from the recent past as any analysis of these data is likely to be more relevant. Also, the last few years have seen a large increase in installed wind power, which makes this analysis timely. For example, in the BPA balancing area, the installed wind capacity has increased from 700 MW in 2006-2007 to over 1300 MW in 2008 and more than 2600 MW in 2009 [4], [5].

The Columbia Basin data available for the period 2007-2009 are the total generation from all the wind farms in the

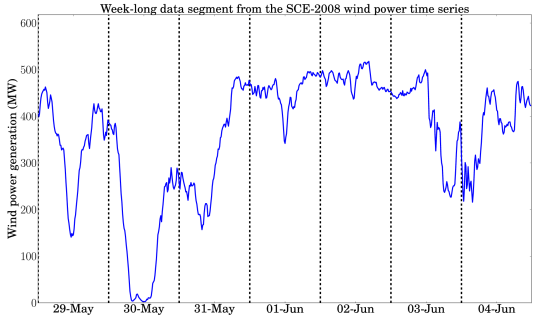


Fig. 1. A week-long segment for wind generation: SCE, May-June 2008.

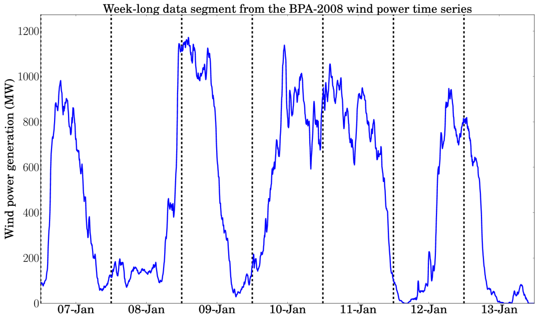


Fig. 2. A week-long segment for wind generation: BPA, January 2008.

BPA balancing area [6], sampled at 5 minute intervals. There are missing values in the data - if values were missing for one or two consecutive intervals, they were filled-in using interpolation, while longer periods were replaced by "-9999" to indicate such values for future processing.

The Tehachapi Pass wind generation data are sampled more coarsely than the Columbia Basin data. These data are available at 15 minute intervals for the Vincent and Antelope regions. As these regions are close by, their wind generation is very similar, and we consider the sum of the generation in our analysis. Also, the generation from the Antelope region occasionally had small negative values which were replaced by zero before being added to the data from the corresponding interval from the Vincent region.

Figure 1 shows the wind power generation for SCE for a week in May-June, 2008. In this short segment of the data, there are two discernible patterns. The generation on 29 May and 30 May starts high at midnight, drops by the middle of the day and then rises again in the afternoon. A similar pattern, though less pronounced, is also seen in the generation for 31 May. The generation on 1 June and 2 June are both somewhat flat, though the 1 June data also has a drop around mid-day. It is harder to categorize the pattern for 3 and 4 June, 2008, at least visually by looking at the original data.

Figure 2 shows the wind power generation for BPA for a week in January 2008. In this segment, the diurnal patterns are initially less obvious. However, a closer look indicates that several days have a flat pattern for most of the day, with a rise or fall at the start or end of the day. For example, 8 January is an initial low flat, followed by a rise late in the day. 11 January is an initial high flat (with some large dips), followed by a fall late in the day. 13 January is a fall by mid-morning, followed by a low flat for the rest of the day, while 12 January is a low flat until early afternoon, when the generation rises. The generation for 7 January and 9 January are also very similar (though one is a slightly shifted version of the other). 9 January could also be considered as a high flat, which drops by mid-day.

### B. Weather data

For the weather data, we used the Remote Automated Weather Station (RAWS) data for Oregon and Southern California available from the Western Regional Climate Center (http://wrcc.dri.edu). For each region, we started by considering weather stations near the area of the wind farms and selected those which had the fewest missing values. For the Columbia Basin region, four sites (Locks, Patjens, Umatilla, and Wasco) had no missing values and were considered in our analysis (see Figure 3). For the Tehachapi Pass region, three sites (Bearvalley, Jawbone, and Piutes) met our criterion of no missing values and were therefore used in the analysis (see Figure 4).

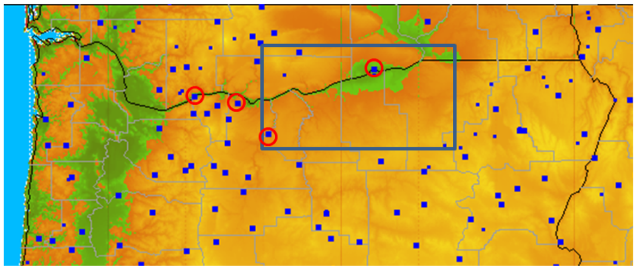


Fig. 3. The Oregon-Washington border region, where the square box indicates the region of the wind farms in the BPA BA. The small squares indicate the meteorological tower locations from WRCC. The four circles indicate the specific sites chosen in our analysis, which are at the following latitude/longitude: Locks (45.669444, 121.881667); Patjens (45.322222, 120.925); Umatilla NWR (45.916667 119.566667); Wasco (45.61,121.33).

As discussed in our earlier work [2], the data from each weather station comprises of 28 variables. Some are irrelevant to the analysis, such as the day of the year, while others are either missing, such as the barometric pressure, or are correlated to other variables. For example, the fuel temperature and the soil temperature are correlated to the air temperature. When all such variables are removed, we are left with the following seven variables:

```
1   Solar Rad. total kW-hr/m2
2   Speed average m/s
3   Wind dir vector deg
4   Speed Gust m/s
5   Air temp Average deg C
6   Relative humidity Average percent
7   Precipitation Total mm
```

For each region, the variables from the selected weather stations were appended to form one long vector which represented the values of the weather conditions in the region for

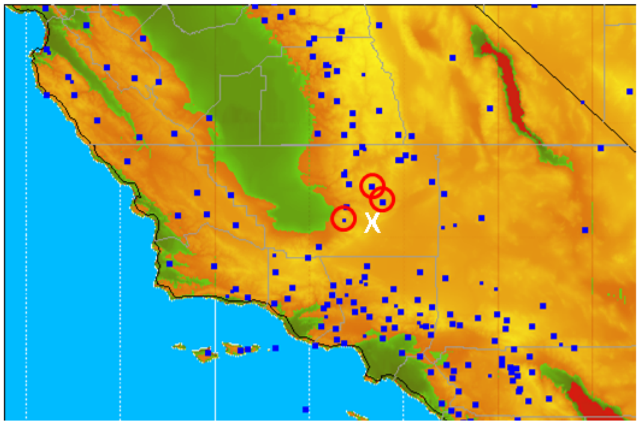


Fig. 4. The Southern California region, where the white cross indicates the Tehachapi Pass area. The small squares indicate the meteorological tower locations from WRCC. The three circles indicate the specific sites chosen in our analysis, which are at the following latitude/longitude: Jawbone (35.294722,-118.226389); Bearvalley (35.139722, -118.625); and Piutes (35.431667, -118.329722), with Tehachapi Pass located at (35.102222, -118.282778).

that day.

## III. Analysis of the Data

Our first task in the analysis was to see if we could identify the patterns in the wind generation data. An examination of the data indicated that we could consider the signal as being composed of several components. For example, the segments in Figures 1 and 2 indicate that there is i) a high frequency noise component; ii) short term variations that last from several minutes to a couple of hours; and iii) a longer-term trend signal that represents the diurnal pattern. Thus, to identify the diurnal patterns in the data, we need to pre-process the data to remove the high-frequency noise and the short-term variation. In the case of BPA data, we also need to address the issue of increasing installed capacity during the years of analysis.

### *A. Accounting for increasing installed capacity at BPA*

In our analysis of BPA data, we observed that the maximum wind generation during a reasonably long period was linearly correlated to the installed capacity during that period. Thus, to account for the increasing installed capacity from 2007 to 2009, we normalized all measurement points to a common nominal capacity $C_0$ by scaling each point by $C_0/C$, where $C$ is the installed capacity for the data point and $C_0$ = 2617.0 MW was chosen as the largest installed capacity during the years of analysis.

### *B. Removing the noise*

Next, we remove the high frequency component, which represents the measurement noise in the time series data, by using a Fourier/frequency decomposition. Representing the data as $\{y(n) : n = 0, 1, \ldots, N-1\}$, we obtain the corresponding discrete Fourier transform coefficients as

$$Y_k = \sum_{n=0}^{N-1} y(n) \cdot e^{-i\frac{2\pi kn}{N}}, \; k = 0, 1, \ldots, N-1. \qquad (1)$$

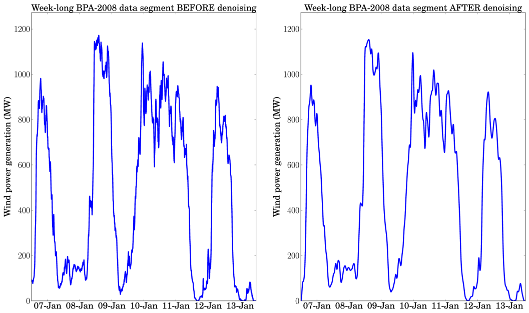


Fig. 5. Reducing the noise in the time series: (left) original signal from BPA, January 2008; (right) denoised version.

To reduce the noise, we reconstruct the signal using the frequency components associated with the $K$ largest coefficients:

$$\tilde{y}(n) = \frac{1}{N} \sum_{k=0}^{K-1} Y_k \cdot e^{i\frac{2\pi kn}{N}}, \; n = 0, 1, \ldots, N-1. \qquad (2)$$

We choose $K$ to preserve a prescribed percentage $\theta$ of the original time series energy:

$$\|\tilde{y}(n)\|^2 = \theta\% \cdot \|y(n)\|^2 \qquad (3)$$

In our experience, a value of $\theta$ in the range between 95 to 99 worked well for our data. Figure 5 shows the results from denoising a data segment using the above procedure, where $\theta$ was chosen equal to 99. We used $\theta = 98$ and $\theta = 99$ to denoise the year-long observations for SCE and BPA, respectively.

We note that our denoising approach is akin to the use of principal component analysis to remove noise.

### *C. Removing short-term variation*

Once we have reduced the noise, we observe that there are short-term variations in the data, which last from several minutes to an hour or two. For example, in Figure 5, the denoised signal can be considered as composed of four peaks and four valleys, starting with the peak on 7 January and ending with the valley on 13 January. However, there is variation in the generation in the time interval around these peaks and valleys. Some of the variation is relatively small in magnitude, for example, in the first valley in 8 January, while in other cases, the variation can be large, as in the peak that spans 10-11 January. These short-term variations can hinder the identification of diurnal patterns, and they must be removed. There are several ways in which this can be done.

A simple approach is to start by visually inspecting the data and identifying the time range for these variations. Then, by smoothing the signal with a Gaussian filter, whose standard deviation $\sigma$ is appropriately selected within the above range, we can obtain a signal that makes it easier to identify the diurnal patterns. Figure 6 illustrates this approach using a week-long data segment. The data are smoothed using five Gaussian filters whose $\sigma$-values range from one to five hours in one-hour increments. When $\sigma$ is equal to four or five hours, we observe that the smoothed signal is devoid of short-term

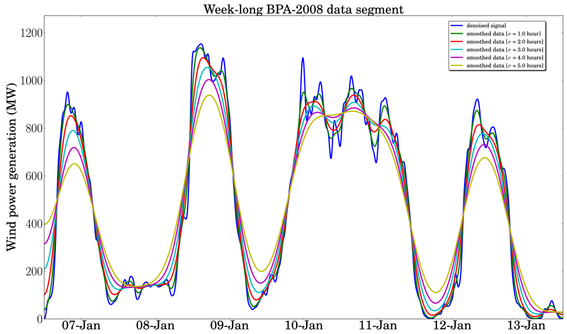


Fig. 6. Smoothing of a week-long data segment from BPA, January 2008, using five Gaussian filters with $\sigma$-values ranging from one to five hours in one-hour increments.

variations and captures the trend in the data necessary to subsequently identify the diurnal patterns. Based on this visual analysis, we smoothed the entire SCE and BPA data using a $\sigma$ of 2.5 hours and 4.5 hours, respectively.

We also considered other approaches that remove the short term variation by directly determining the intrinsic scale of the data. The application of a scale-space approach [7], [8] indicated that to obtain the intrinsic scale, we needed to use a $\sigma$ of 2.25 hours for SCE and 4 hours for BPA. We also considered a multi-scale analysis of the data using the undecimated wavelet transform [9], [10] with the quadratic spline wavelet. Our analysis indicated the intrinsic scales to be at $k$ = 4 and $k$ = 6, corresponding to structures in the time scale ranges between 1.5 to 3 hours for SCE, and between 3 and 6 hours for BPA, respectively. We observe that the three methods all provide similar measures for the intrinsic scale of the data. Also, our approach to removal of the short-term variation includes the reduction of the measurement noise; treating the denoising step separately allows us to identify appropriate follow-on steps for analysis.

### D. Identifying diurnal patterns

Having reduced the noise and removed the short-term variation in the data, the next step is to identify the diurnal patterns. We observed that a day might typically be composed of time periods with increasing generation, or decreasing generation, or flat generation, where the power generation remains roughly constant. A diurnal pattern could then be composed of these "sub-patterns" suitably concatenated. Our task thus reduces to identifying periods with flat, up, or down generation, and determining in what combination they occur during a day.

An obvious solution approach is to identify peaks and valleys in the data and use them to identify first the "sub-patterns" and then the patterns for each day, as follows:

**Finding peaks and valleys:** Let the smoothed wind power data be denoted by $\mathbf{s} = \{s(n) : n = 0, 1, \ldots, N-1\}$, the current data point by $s(n)$, the candidate peak by $s^{(p)}(m)$, and the candidate valley by $s^{(v)}(m)$. We start by selecting a threshold $T$ which represents the minimum amount of change necessary to define a point as a genuine peak or valley. Algorithm 1 describes how we find the peaks and valleys by starting in a `seek-a-peak` state, tagging the first data point as a candidate peak, and traversing the sequence $\mathbf{s}$ while alternating between the `seek-a-peak` and `seek-a-valley` states.

**Algorithm 1** *Peak-valley finding algorithm*

```
if in seek-a-peak state then
    if $s(n) > s^{(p)}(m)$ then
        $s^{(p)}(m) \leftarrow s(n)$
        $n \leftarrow n + 1$
    end if
    if $s(n) + T < s^{(p)}(m)$ then
        $s^{(p)}(m)$ is declared a peak
        $s^{(v)}(m) \leftarrow s(n)$
        $n \leftarrow n + 1$
        switch to seek-a-valley state
    end if
end if
if in seek-a-valley state then
    if $s(n) < s^{(v)}(m)$ then
        $s^{(v)}(m) \leftarrow s(n)$
        $n \leftarrow n + 1$
    end if
    if $s(n) - T > s^{(v)}(m)$ then
        $s^{(v)}(m)$ is declared a valley
        $s^{(p)}(m) \leftarrow s(n)$
        $n \leftarrow n + 1$
        switch to the seek-a-peak state
    end if
end if
```

**Identifying patterns using peaks and valleys:** Having identified the peaks and valleys in the entire time series, we now focus on the segment for each day. To determine the "sub-patterns" that comprise the generation for the day, we need to identify if the first or last data-point is a peak or a valley within the context of a day-segment. We accomplish this with Algorithm 2 where the first and last data-points for the day are denoted by $\texttt{edge}_1$ and $\texttt{edge}_2$, respectively, and the first and last extrema-points inside a day-segment by $\texttt{extrema}_1$ and $\texttt{extrema}_2$, respectively.

Following the assignment of the end-points of the day-segment, we can now tag the sub-intervals in a day with an up ('U') or down ('D") sub-pattern, based on whether the sub-interval started with a valley (peak) and ended in a peak (valley). A day-segment will have either no tags or a concatenation of 'U' and 'D' tags. Given a maximum number of tags, $M$ (2 in our case), we assign a pattern to each day as follows:

- **no tags** - A day with no tags is assigned the `flat` pattern: The range of values in such a segment is less than the threshold $T$.
- **one to $M$ tags** - A direct correspondence exists between the sequence of tags and the assigned pattern: A day-segment with tag-sequence 'U' is assigned to the `up` pattern, a day-segment with tag-sequence 'D' is assigned to the `down` pattern, a day-segment with the tag-sequence 'D'-'U' is assigned to the `down-up` pattern, and so on.

**Algorithm 2** *End-point assignment algorithm*

```
if no peak/valley inside day-segment then
    if $\texttt{edge}_1 > \texttt{edge}_2 + T$ then
        $\texttt{edge}_1$ is a peak, $\texttt{edge}_2$ is a valley
    else if $\texttt{edge}_2 > \texttt{edge}_1 + T$ then
        $\texttt{edge}_2$ is a peak, $\texttt{edge}_1$ is a valley
    else
        neither $\texttt{edge}_1$ nor $\texttt{edge}_2$ is a peak/valley
    end if
else
    if $\texttt{extrema}_1$ is a peak, $\texttt{extrema}_1 > \texttt{edge}_1 + T$ then
        $\texttt{edge}_1$ is a valley
    else if $\texttt{extrema}_1$ is a valley, $\texttt{extrema}_1 < \texttt{edge}_1 - T$ then
        $\texttt{edge}_1$ is a peak
    else if $\texttt{extrema}_2$ is a peak, $\texttt{extrema}_2 > \texttt{edge}_2 + T$ then
        $\texttt{edge}_2$ is a valley
    else if $\texttt{extrema}_2$ is a valley, $\texttt{extrema}_2 < \texttt{edge}_2 - T$ then
        $\texttt{edge}_2$ is a peak
    else
        neither $\texttt{edge}_1$ nor $\texttt{edge}_2$ is a peak/valley
    end if
end if
```

- **more than $M$ tags** - A day segment with more than $M$ tags is assigned to the pseudo-pattern `others`.

**Choosing the value of the threshold, $T$:** The assignment of patterns to days depends on the choice of the threshold. Varying the threshold changes the number of detected peaks/valleys, which, in turn, modifies the pattern assignments for certain days. To mitigate this, we choose a threshold that achieves maximal stability; that is, small deviations from the threshold value lead to minimal changes in the number of peaks/valleys detected. Figure 7 shows the number of peaks/valleys detected as the threshold is varied from a small value to half the maximum value of the time series. The function has a profile with two regimes, decreasing exponentially at first and then linearly. We choose as threshold the value of $T$ that lies between the two regimes as this choice maximizes stability while minimizing the number of relevant peaks/valleys that could be missed. This threshold is 100 for the data in Figure 7. Threshold values of 100 and 300 were used for the SCE and BPA data, respectively.

**Enhancing the pattern assignment:** Once the pattern for each day in the 2007-2008 data for SCE and the 2007-2009 data for BPA has been assigned, we verify the assignments by selecting days at random and visually inspecting the original generation for the day to determine if the assignment is correct. As is often the case in these situations, while the pattern for some days is very clear, it can be ambiguous for other days. We saw an example of this earlier in Figure 2, where the diurnal pattern for 7 January is 'up-down', but one could question if the pattern for 9 January should be 'down' (as it starts from a high and falls to near zero by early afternoon), or 'flat' (as it is high and flat for several hours), or 'up-down' (as it is slightly shifted version of the pattern on 7 January 2008).

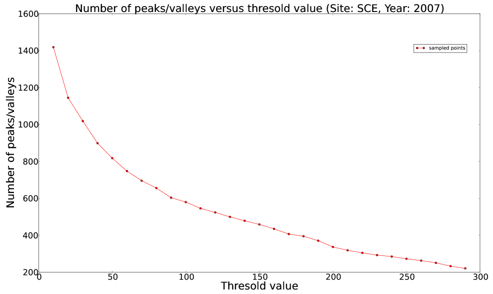


Fig. 7. Number of peaks/valleys as function of threshold value $T$: SCE 2007 dataset.

In such cases, we have several options we could pursue. One would be to create additional patterns, such as 'flat-up', 'down-flat', 'flat-down', 'up-flat', or even 'flat-down-up-flat' (for the pattern seen for 1 June 2008 in Figure 1). We could also have split the flat pattern into a “high-flat: and a “low-flat” to account for the magnitude of the generation. However, having a multitude of diurnal patterns would mean fewer examples of each pattern, making it difficult to train a classifier to predict the pattern. So, we focused on six patterns in our study: 'flat', 'up', 'down', 'up-down'. 'down-up', and 'others', where the last category was composed of patterns that did not belong to the first five categories. Since we found several days where the pattern was flat for a large percentage of the day, with a sharp rise or fall at the start or end, we chose to label these patterns as 'flat', instead of “down” or “up”. The rationale was that weather conditions, being daily averages, would be predictive of the pattern that was prevalent for a majority of the day.

We also observed that sometimes, a pattern for a “day” might be best assigned by including a small shift in the time series. For example,the pattern for 9 January 2008 in Figure 2 could be considered an “up-down” pattern if we took into account the rise in generation late in the day of 8 January. However, in our assignment of patterns to days, we chose to ignore these effects.

### E. *Predicting the patterns*

Once we have the pattern associated with each day, we can combine it with the weather data described in Section II-B to create the data set for prediction. The SCE data set consists of 731 days, each represented by 21 weather conditions (i.e., the features) and the pattern associated with the day. The BPA data set has 1036 days described using 28 features and the associated pattern. Our task now is to determine if we can build a predictive model to assign a pattern to a day given the weather conditions for the day. We note that since we are working with historical data, we are using the actual weather conditions for each day. In practice, if we are successful in building a model to predict the diurnal pattern, we would be using the forecast weather conditions for the day.

In our work, we use an ensemble of decision trees to create the predictive model. This ensemble is generated by introducing randomization at each node of the tree in two ways [11]. We first randomly sample the examples at a node and select a fraction (we use 0.7) for further consideration. Then, for each feature, instead of sorting these examples based on the values of the feature as would be done at a node of a tree [12], we create a histogram, evaluate the splitting criterion at the mid-point of each bin of the histogram, identify the best bin, and then select the split point randomly in this bin. We use the Gini splitting criterion described in [12]. The randomization is introduced both in the sampling and in the choice of the split point and the use of the histograms speeds up the creation of each tree in the ensemble.

## IV. Experimental Results

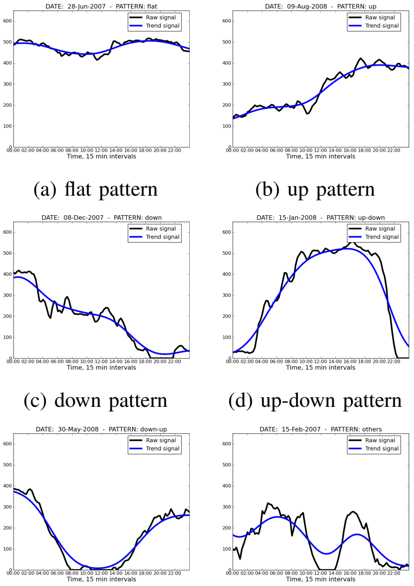
We next describe the result of our analysis for the SCE and BPA data. First, in Figures 8 and 9, we show examples of the different types of patterns found in the wind power generation data in these two regions. The curves in black are the original generation, while the blue curve is the trend curve after removal of the short-term variation. We selected patterns that illustrate our labeling procedure as well as the challenges. For example, while many of the flat patterns are near zero for most of the day, others have some variation, as shown in Figure 8(a), or are flat for most, but not all, of the day, as shown in Figure 9(a). These latter examples of the 'flat' pattern vary from the 'up' or 'down' patterns as they have the flat part present for a larger percentage of the day. Needless to say, there is some subjectivity in the labeling of these patterns, though care was taken to be consistent in the labeling. The 'others' pattern also has some interesting behavior. Most of these are composed of three or more 'sub-patterns'. In particular, we found that the SCE region had distinct patterns of the form 'down-up-down-up' or 'up-down-up-down', as shown in Figure 8(f), which were not seen in the BPA data. While interesting, there were too few occurrences of these patterns to assign them to a class of their own.

We also obtained some interesting insights during our use of the Fourier transform in denoising. Figure 10 displays the well-defined components of the frequency domain representation of the 2007 data from SCE and BPA (the results for the other years are similar). In the case of SCE, there are three peaks (indicated by arrows in the figure), at frequencies 0 Hz, 0.00001158 Hz, and 0.00002316 Hz. The latter two frequencies correspond to intervals of 23.99 hours and 11.99 hours, respectively. In contrast, BPA has only two peaks at 0 Hz and 0.00001158 Hz. The third peak in the SCE data is reflective of the patterns composed of concatenations of four sub-patterns.

Next, in Tables I and II, we list the percentages of the different patterns for the two sites for the years of the analysis. We find that the 'flat' and the 'down-up' patterns occur more frequently in the SCE region accounting for nearly 75% of the days, while the 'flat' pattern dominates in the BPA region, accounting for an average of 57% of the days. We also observe that in BPA, none of the other patterns dominate, though 'up-down' and 'others' tend to occur less frequently.

We also investigated the monthly distribution of the different patterns to see if there is a seasonal variation as shown in Figures 11 and 12. The SCE data showed a reduction in the occurrences of the 'flat' pattern during the summer months and a corresponding increase in the 'down-up' pattern. A similar, though less pronounced, behavior is seen in the BPA region. We also observe that the years 2007 and 2009 are similar for BPA, while 2008 is a bit different.

Finally, we investigated the possibility of predicting the diurnal pattern based on the weather conditions. The clear monthly variation that we see in the SCE data indicates that such a prediction might be possible, though it is unclear if this will hold for the BPA site as well. Table III indicates the percentage error rate in prediction obtained using 5 runs of five-fold cross validation with an ensemble of 10 trees. For

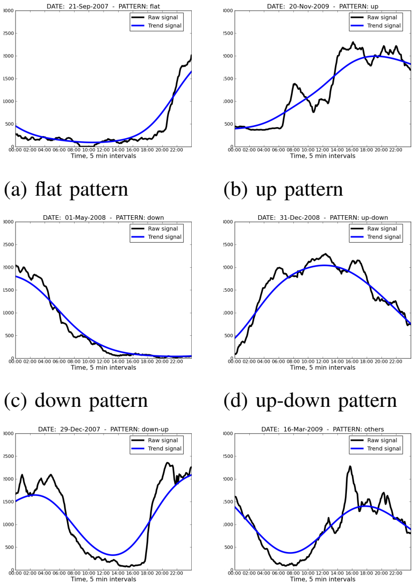


(a) flat pattern (b) up pattern

(c) down pattern (d) up-down pattern

(e) down-up pattern (f) others pattern

Fig. 8. Sample diurnal patterns from SCE data.



(a) flat pattern (b) up pattern

(c) down pattern (d) up-down pattern

(e) down-up pattern (f) others pattern

Fig. 9. Sample diurnal patterns from BPA data.

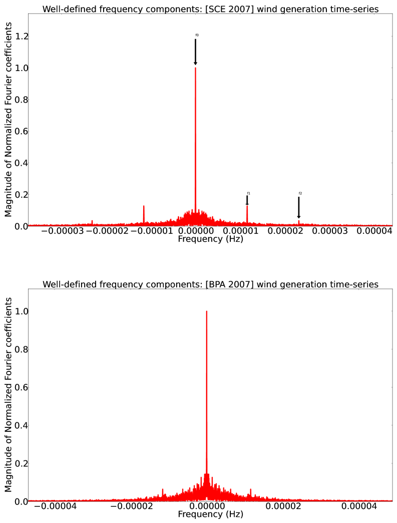


Fig. 10. Fourier domain representation of the 2007 SCE (top) and BPA (bottom) data.

TABLE I
PERCENTAGE OF PATTERNS IN THE SCE 2007-2008 DATASETS.

| Pattern | SCE-2007 | SCE-2008 |
|---|---|---|
| *flat* | 46.02 % (168 days) | 49.45 % (181 days) |
| *up* | 8.76 % (32 days) | 8.46 % (31 days) |
| *down* | 3.28 % (12 days) | 2.45 % (9 days) |
| *up-down* | 3.83 % (14 days) | 4.09 % (15 days) |
| *down-up* | 28.21 % (103 days) | 26.50 % (97 days) |
| *others* | 9.86 % (36 days) | 9.01 % (33 days) |

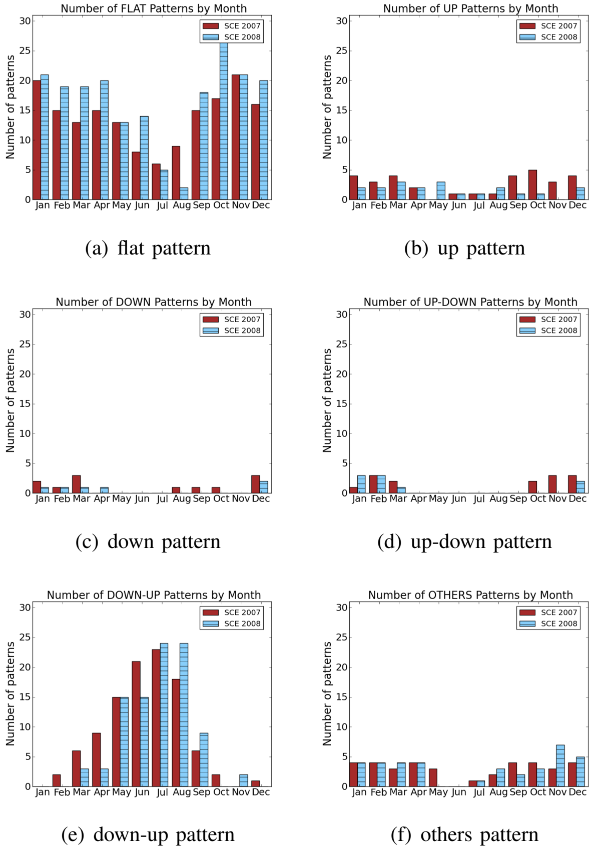
each site, we consider two cases - one with all the patterns and the other with just the majority pattern(s) and the remaining patterns all labeled as others. So, for SCE, we have 'flat', 'down-up', and 'others', while for BPA we have 'flat' and 'others'. Since there are relatively small numbers of the non-majority patterns at the two sites, it is not clear if these numbers are sufficient for the decision tree ensemble to learn the patterns. The second experiment allows us to determine if we can at least predict the majority classes with a low error rate.

The results in Table III indicate that we can reduce the error rate for both SCE and BPA data by combining the non-majority classes with the 'others' pattern. As mentioned earlier, we suspect that this is due to an insufficient number of examples of the non-majority patterns. When we consider all patterns, the accuracy for the SCE site is better than for the BPA site most likely because the two majority patterns in SCE constitute a larger percentage (75%) of the total sample than the one majority pattern in BPA (57%). We also observe that the accuracy for the two sites is comparable when the non-majority patterns are merged with the 'others' pattern.

While these prediction results are encouraging, we believe that there are several ways in which this study can be improved. First, we need to consider the data for additional years so we can have a reasonable number of examples of the minority classes. With a larger data set and more examples of each class, we may be able to define each pattern better, leading to fewer mis-labeled days. Second, we need to revisit the patterns identified to see if there are some patterns that occur frequently enough that they should be considered as a separate pattern. For example, the pattern in Figure 9(a), though labeled as a 'flat' as the generation is flat most of the day, could form a category of its own if, with a larger data set, we found that there were enough examples of such patterns. We may also find that the few distinct patterns of the form 'down-up-down-up' or 'up-down-up-down' found in SCE should be considered a separate class, while the 'up-down' pattern occurs rarely enough, even in a large data set, to be considered in the 'others' category. And finally, the use of high quality weather data from additional sites might also improve the accuracy of prediction.

TABLE II
PERCENTAGE OF PATTERNS IN THE BPA 2007-2009 DATASETS WITH THE CORRESPONDING NUMBER OF DAYS IN PARENTHESIS.

| Pattern | BPA-2007 | BPA-2008 | BPA-2009 |
|---|---|---|---|
| *flat* | 60.27 % (220) | 53.69 % (196) | 57.80 % (211) |
| *up* | 8.49 % (31) | 15.34 % (56) | 9.58 % (35) |
| *down* | 10.41 % (38) | 12.05 % (44) | 8.21 % (30) |
| *up-down* | 4.38 % (16) | 5.47 % (20) | 6.84 % (25) |
| *down-up* | 12.05 % (44) | 9.31 % (34) | 11.23 % (41) |
| *others* | 4.38 % (16) | 4.38 % (16) | 6.30 % (23) |



(a) flat pattern

(b) up pattern

(c) down pattern

(d) up-down pattern

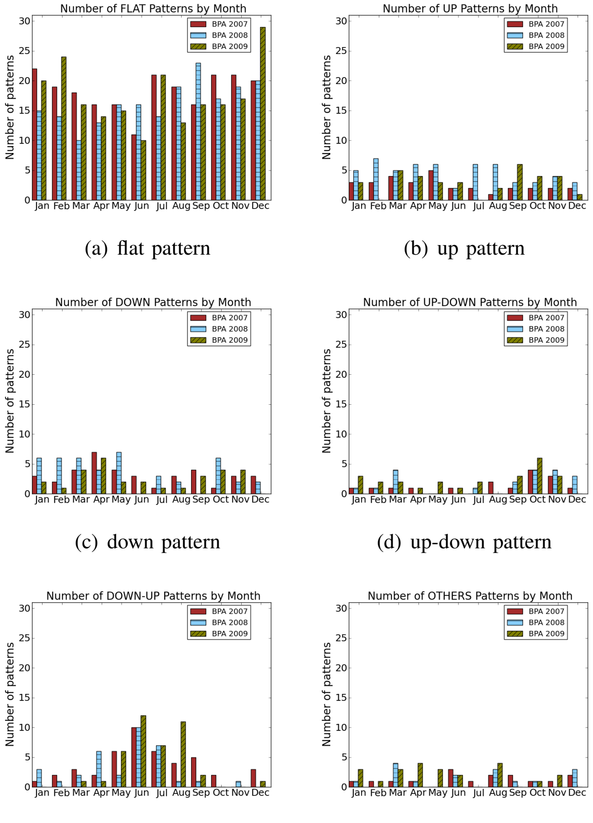
(e) down-up pattern

(f) others pattern

Fig. 11. Monthly distributions of patterns for the SCE datasets.

## V. CONCLUSIONS

In this paper, we considered the problem of identifying diurnal patterns in wind power generation data from two sites - Tehachapi Pass in Southern California and the Columbia Basin



(a) flat pattern

(b) up pattern

(c) down pattern

(d) up-down pattern

(e) down-up pattern

(f) others pattern

Fig. 12. Monthly distributions of patterns for the BPA datasets.

TABLE III
PERCENTAGE ERROR RATE IN PREDICTING THE DIURNAL PATTERNS USING THE WEATHER CONDITIONS

| Site and patterns | Percentage error rate |
|---|---|
| SCE, all patterns | 31.09 % |
| SCE, ’flat’, ’down-up’, and rest merged into ’others’ | 24.77% |
| BPA, all patterns | 37.88% |
| BPA, ’flat’ and rest merged into ’others’ | 22.21% |

region at the Washington-Oregon border. We found that it was indeed possible to identify patterns in the data, though, given the size of our data sets, some patterns occurred infrequently. We also found that for SCE, there was a seasonal dependence of the two majority patterns. Though our data sets were small, we also obtained encouraging results in predicting the patterns using weather conditions in the region around the wind farms. The next step would be to repeat the analysis with data from additional years; this would not only provide the benefits of a larger data set, but also mitigate any effects of yearly changes in the weather.

## ACKNOWLEDGMENT

This work was funded through the Renewable System Interconnect Program in the Office of Energy Efficiency and Renewable Energy at the US Department of Energy. We thank Min-Lin Cheng (SCE), Robert Farber (SCE), John Pease (BPA), and the Western Regional Climate Center for access to the data used in this study.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

## REFERENCES

[1] C. Kamath, “Understanding wind ramp events through analysis of historical data,” in *Proceedings, IEEE PES Transmission and Distribution Conference*, 2010, available at http://ckamath.org/publications_by_project/windsense.

[2] ——, “Associating weather conditions with ramp events in wind power generation,” in *Proceedings, IEEE PES Power Systems Conference and Exposition*, March 2011, available at http://ckamath.org/publications_by_project/windsense.

[3] ——, “Using simple statistical analysis of historical data to understand wind ramp events,” Lawrence Livermore National Laboratory, Tech. Rep., February 2010, available at http://ckamath.org/publications_by_project.

[4] “Balancing act: BPA grid responds to huge influx of wind power,” Bonneville Power Administration Fact Sheet. [Online]. Available: http://www.bpa.gov/corporate/pubs/fact_sheets/08fs/Wind-Balancing-act-Nov2008.pdf

[5] “How BPA supports wind power in the pacific northwest,” Bonneville Power Administration Fact Sheet. [Online]. Available: http://www.bpa.gov/corporate/pubs/fact_sheets/09fs/BPA_supports_wind_power_for_the_Pacific_Northwest_-_Mar_2009.pdf

[6] “BPA wind projects map: Current and proposed wind project interconnections to BPA transmission facilities.” [Online]. Available: http://www.transmission.bpa.gov/PlanProj/Wind/documents/map-BPA_wind_interconnections.pdf

[7] A. P. Witkin, “Scale-space filtering,” in *8th International Joint Conference on Artificial Intelligence*, Karlruhe, Germany, 1983, pp. 1019–1022.

[8] T. Lindeberg, *Scale-space Theory in Computer Vision*. Kluwer Academic Publishers, 1994.

[9] C. S. Burrus, R. Gopinath, and H. Guo, *Introduction to Wavelets and Wavelets Transforms: A Primer*. Prentice Hall, 1997.

[10] S. Mallat, *A Wavelet Tour of Signal Processing*. Academic Press, 1999.

[11] C. Kamath, E. Cantú-Paz, and D. Littau, “Approximate splitting for ensembles of trees using histograms,” in *Proceedings, Second SIAM International Conference on Data Mining*, 2002, pp. 370–383.

[12] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Boca Raton, FL: CRC Press, 1984.