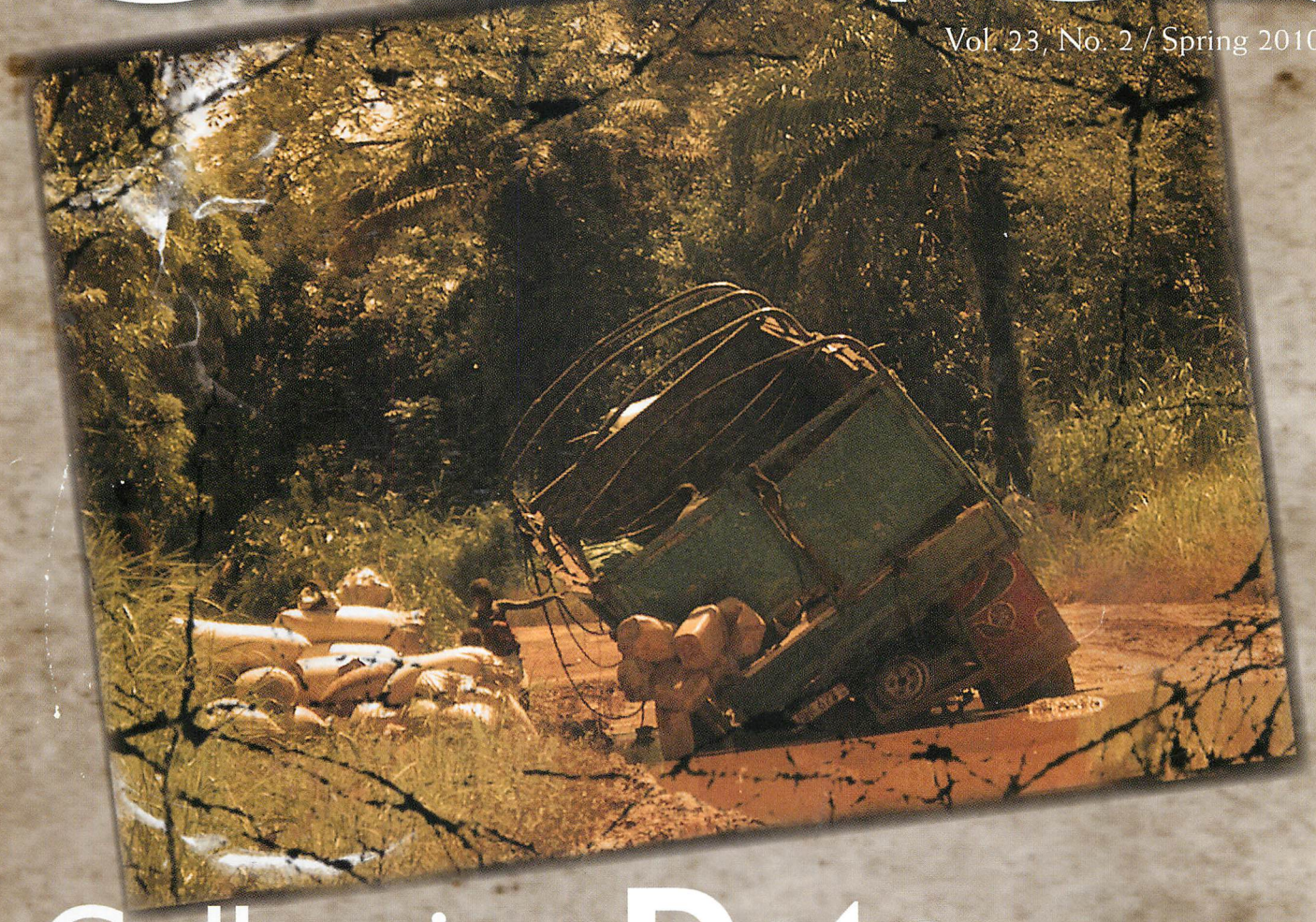
CHANCE

Vol. 23, No. 2 / Spring 2010



Collecting Data in Challenging Settings

Also...

Can People Distinguish Pâté from Dog Food?

Commentary on the Graphic Displays in the 2008 National Healthcare Quality Report and State Snapshots

02 >
0 74470 74240 1

Springer

American Statistical Association

Volume 23, Number 2, 2010

CHANCE

A Magazine of the American Statistical Association

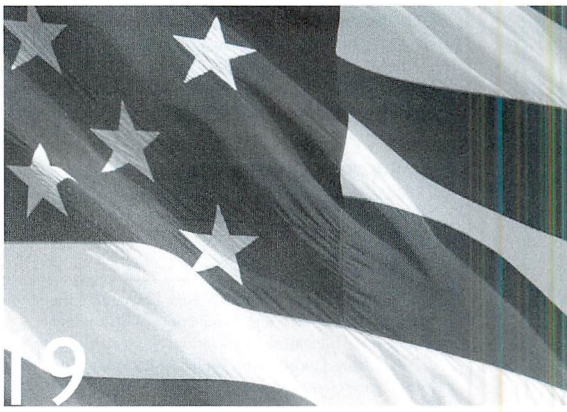
Articles

Columns

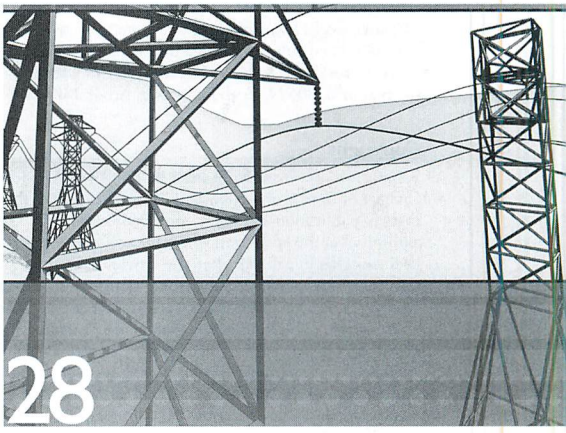
Departments

Abstracted/indexed in *Academic OneFile*, *Academic Search*, *ASFA*, *CSA/Proquest*, *Current Abstracts*, Current Index to Statistics, *Gale*, *Google Scholar*, *MathEDUC*, *Mathematical Reviews*, *OCLC*, *Summon by Serial Solutions*, *TOC Premier*, *Zentralblatt Math*

Cover image: **Photo Courtesy of Jana Asher**



Holmes Finch compares and contrasts item response models using formulas, pictures, numerical examples, and data on gender differences in opinions on women in politics.

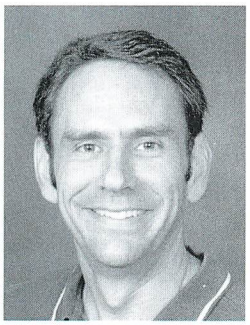


Bernard Dillard uses a discrete wavelet transformation to analyze electricity consumption data for multiscale statistical process control with the aim of avoiding energy interruptions.

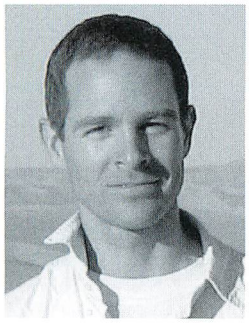
John Bohannon, Robin Goldstein, and Alexis Herschkowitsch report their scientific study comparing human-grade food items.

About the Authors

Jana Asher, an ASA Fellow, is executive director of StatAid, a nonprofit statistical research and consultation organization, and an internationally recognized expert on the collecting and analyzing human rights violations data.

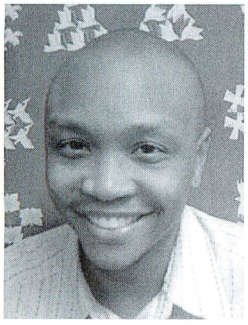


Holmes Finch teaches courses in statistical and research methodology as well as psychometrics and educational measurement. His research interests are in the area of latent variable modeling and involve issues in psychometrics.



John Bohannon is a visiting scholar at the Harvard University Program in Ethics and Health. After completing a PhD in molecular biology at the University of Oxford in 2002, he focused on bioethics as a Fulbright fellow in Berlin.

Robin Goldstein is the founder and editor-in-chief of the *Fearless Critic* book series and co-author of *The Wine Trials 2010*. He earned his AB in philosophy and neuroscience from Harvard University and a JD from Yale Law School.



Bernard Dillard is assistant professor of mathematics at Fashion Institute of Technology in New York City. He teaches statistical analysis, data analysis for business applications, and quantitative methods.

Alexis Herschkowitsch is the managing editor of the *Fearless Critic* book series and co-author of *The Wine Trials 2010*. She earned her BA from the University of Texas at Austin.

Editor's Letter

Mike Larsen,
Executive Editor

Dear Readers,

This issue of *CHANCE* begins with an article by Jana Asher on collecting data in challenging settings. In particular, Asher describes her experiences conducting in-person survey interviews in East Timor. She gives us personal anecdotes, practical statistical advice, and an interesting story.

Qi Zheng explains the origins of the Luria-Delbrück distribution and its role in studying evolutionary change in *E. coli*. The statistical reasoning underlying the phenomenon has a connection to the distribution of slot machine returns.

Holmes Finch's article, "Using Item Response Theory to Understand Gender Differences in Opinions on Women in Politics," compares and contrasts item response models and how they describe a data set. The models are explained using formulas, pictures, and examples.

In Volume 22, Number 4, Jürgen Symanzik proposed a puzzle based on 10 data points and a set of seven instructions. Contest winner Stephanie Kovalchik, a graduate student at UCLA, provided a solution in the form of an amusing letter and an illustrative graphic. The 10 data values were flight times in seconds recorded on the log 10 scale of the Space Shuttle Challenger. Brad Thiessen earned honorable mention for his graph that included temperature and historical facts.

Bernard Dillard asks, "Who turned out the lights?" We are all concerned with energy demand and production. Bernard uses a discrete wavelet transformation to analyze electricity consumption data measured on a frequent time scale. The fit of the model is used in multiscale statistical process control. The ultimate goal is to be able accurately predict points of extreme energy demand and respond appropriately.

Students in virtually all statistics courses learn something of least squares estimation when studying prediction of an outcome from an explanatory variable. Ivo Petras and Igor Podlubny ask whether there is a reasonable alternative to the default criterion. "Least circles" is presented for your consideration.

To introduce students to concepts of design of experiments, instructors sometimes have students conduct taste tests of various food items, such as gummy bears (see Vol. 23, No. 1). John Bohannon, Robin Goldstein, and Alexis Herschkowitsch compared dog food and pâté. Really, they did. Read about their design and the results in this issue.

Ronald Smeltzer shows us an early time-line bar graph by Philippe Buache depicting the water level of the Seine River in Paris from 1760 to 1766. The picture creatively and effectively depicts data in print before the advent of the modern printing techniques that we enjoy today.

Howard Wainer, in his *Visual Revelations* column, writes about the graphics in the 2008 National Healthcare Quality Report and State Snapshots. Usefully and accurately displaying information graphically is important and challenging. Wainer makes suggestions for improving some of the displays.

Continuing a series of articles on postage stamps, Peter Loly and George P. H. Styan discuss stamps issued in sheets with 5x5 Latin square designs. Color versions of the stamps, as well as previous articles on stamps, are available online at *www.amstat.org/publications/chance*.

Jonathan Berkowitz's puzzle celebrates the 2010 Winter Olympics, which was held in his home city of Vancouver, British Columbia. The puzzle, titled "Employs Magic," is actually five smaller puzzles, each a cryptic five-square of 10 words.

Mark Glickman's *Here's to Your Health* column will appear in the next issue.

In other news, the Executive Committee of the ASA met recently and made decisions that impact *CHANCE*. First, the committee voted to continue *CHANCE* for another three years in both print and online versions. The next executive editor will serve 2011–2013. I'll enjoy reading *CHANCE* in the years to come. Second, the Executive Committee voted to make the online version of *CHANCE* free to the ASA's certified student members. This is a great development, because students are potential long-term subscribers and future authors. They also can be inspired by the significant role that probability and statistics can play in major studies and activities. I hope that other professionals will be motivated to submit articles to *CHANCE* to entertain and influence this group.

I look forward to your suggestions and submissions.

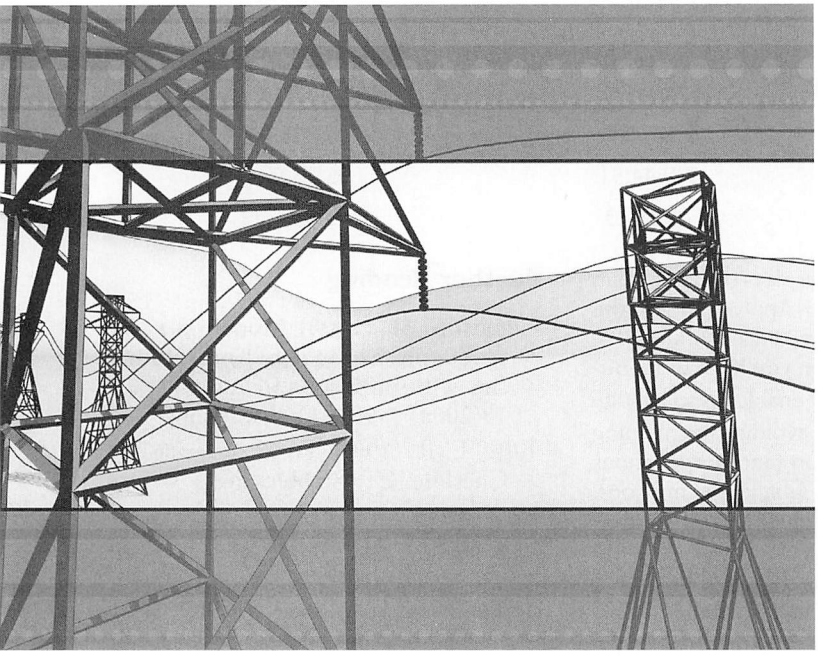
Enjoy the issue!
Mike Larsen

Through your email you can get a table of contents notification for *CHANCE*. Go to *www.springer.com/mathematics/probability/journal/144* and add your email address in the box that says "Alerts For This Journal". The web site also has a place you can recommend *CHANCE* to your library.

Hey, Who Turned Off the Lights?

A look at electricity consumption

Bernard Dillard



Scene One: August 14, 2003. New York City (NYC). A hot, humid dog day. You've traveled to the city of dreams to see a matinee of "The Lion King" on Broadway. Horns blowing. Skyline breathtaking. People scrambling. In the theater, you settle into your orchestra seats and great singing and dancing commence. In a magical moment, Rafiki prepares to lift Simba and Nala's newborn cub as next in line to rule the Pride Lands when suddenly the power goes out. "Please exit the theater."

Scene Two: April 12, 2004. Los Angeles International Airport (LAX). You've finally booked that trip to Hawaii. Sitting in the coach section in the back of the plane by the restroom, you calm yourself: The flight will not be that long, and soon you'll be walking on the black sand at Big Island. You settle into your seat and close your eyes. "Attention passengers, there will be a (long) delay due to a power outage in one of the control towers. Feel free to get up and use the restroom while we wait."

Scenes like those described above were two of the many that could have occurred during actual power outages in the two largest U.S. cities. Theories abound as to why those blackouts happened during those times. One theory concerning the 2004 blackout includes a bird sitting on a power line. But that theory could not account for simultaneous outages at the Bellagio Hotel in Las Vegas. Many scholars maintain these outages had to do more with an overloading effect of electricity consumption on a power grid— much like overuse of power in a home causes a fuse to blow.

Many posit that, if appropriate measures had been in place to detect signs of excessive electricity use, these power failures could have been avoided through the use of cutting-edge statistical monitoring techniques. Hence, in a time where an understanding of electricity consumption data and its relation to other factors are pivotal to various aspects of American life, including national security, proper analysis of *historical* consumption and related data becomes a key element for successfully detecting future abnormalities.

The natural way for a statistician to treat electricity consumption data is as a time series. Two factors make the analysis and monitoring tasks nonstandard. First, like many modern time series, the time scale on which the data are collected is frequent. Clearly, the level of data aggregation depends on the objective of the application or analysis. In our case, we are interested in rapid detection of abnormal behavior in electric consumption or related variables (such as temperature) that could indicate a blackout is imminent. According to S. Basu and A. Mukherjee's 1999 *INFOCOM* article, "Time Series Models for Internet Traffic," traditional time-series models, such as autoregressive integrated moving average (ARIMA) models, are not useful for data measured on such a frequent scale.

The second complicating factor is we typically monitor not only the electricity consumption series but also a set of several related time series, out of a belief the other series might carry information about consumption. For example, we can monitor the weather and hypothesize that an extreme wave of cold or hot weather would lead to increased consumption. Thus, we need a method that can simultaneously monitor multiple time series and take into account the interrelations between the series.

The discrete wavelet transformation (DWT) can be used to analyze the features and structure of electricity consumption and consumption-related data measured on a frequent time scale. Multiscale statistical process control (MSSPC), which combines DWT and control chart methodology, can be used to assess abnormalities in individual time series. These techniques are illustrated in this article using data on hourly electricity consumption and temperatures in New Hampshire from August 29 to September 1, 1997. What can we learn about the time series using these methods? What do we anticipate will be possible with improved statistical methods?

Electric Consumption and Temperature Data

It is well known that fluctuation in electricity consumption depends heavily on many factors, the most important source being meteorology, and particularly temperature, as stated in R. Cottet and M. Smith's 2003 article "Bayesian Modeling and Forecasting of Intraday Electricity Load," that appeared in the *Journal of the American Statistical Association*. In most locations, although the meteorological variables that affect load can differ according to region, temperature appears to be by far the most important meteorological factor in most locations. Consequently, we study the consumption behavior using not only electricity load data but also relevant temperature data.

The consumption data in this analysis is provided by the New Hampshire Electric Co. (*www.seattlecentral.org/qelp/sets/042/042.html#About*). It records the electric consumption over the course of four days from one delivery point in New Hampshire at the end of August 1997. The electricity consumption load was measured in kilowatts per hour (kwh) and was recorded over the period of 96 hours: from 12:52 a.m. on August 29, 1997, to 11:52 p.m. on September 1, 1997. Figure 1 describes the consumption series graphically by using a time plot.

The time plot reveals several notable observations. The most noticeable pattern is a cyclical fluctuation, which repeats daily. Typically, early morning hours are characterized by low energy consumption, whereas during the late morning and evening hours, consumption reaches the highest values of the day. This reflects the levels of activities of most people during late and evening hours, which require more electricity usage. This pattern of usage results in a daily toothlike structure, which has been observed in other geographical areas and at other periods of time, such as in Harvey and Koopman's "Forecasting Hourly Electricity Demand Using Time Varying Splines" in the *Journal of the American Statistical Association*.

Also, on September 1, the toothlike structure differs from those of the previous days. The relative maximum on this day exceeds that of the previous three days by far. On this day, the maximum point occurs at 8:52 p.m. The electricity consumption load at this time is 2148.12 kwh, which is the highest load of the entire four days. In addition, the second highest consumption on that day (taking place at 10:52 a.m.) is higher than the highest consumption levels on the previous three days.

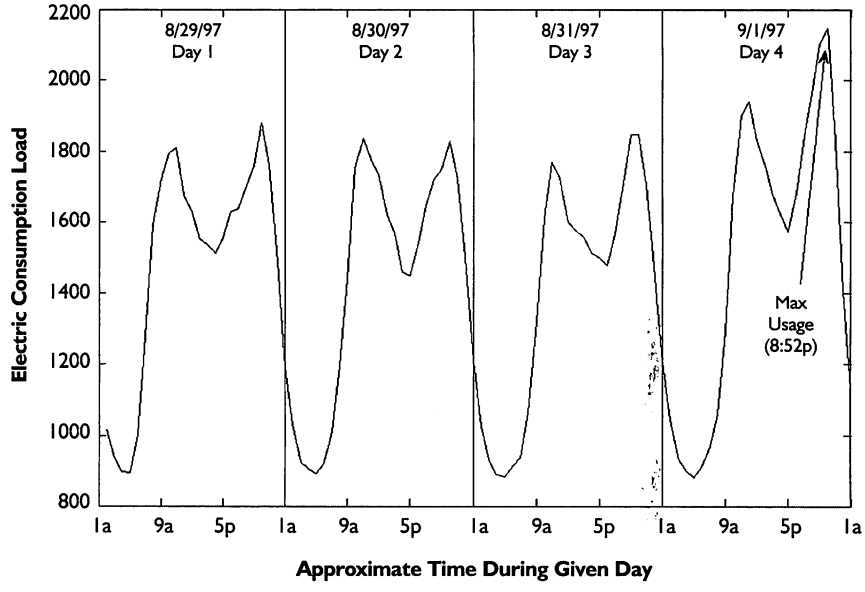


Figure 1. Graph of hourly electric consumption load over time for four days in 1997 in New Hampshire

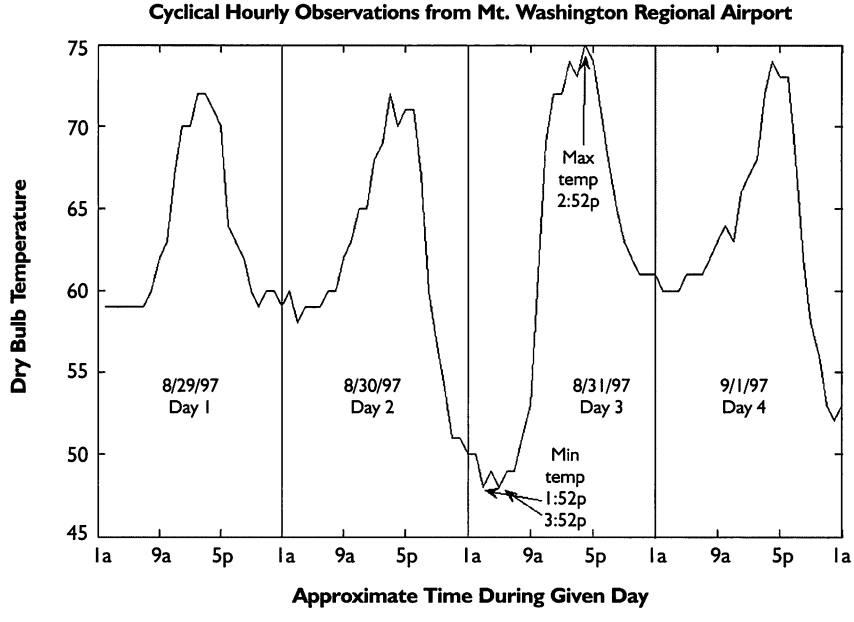


Figure 2. Graph of hourly temperatures over time for four days in 1997 in New Hampshire

Along with the consumption data, we consider temperature data for the same area and during the same time. We use hourly dry-bulb temperatures from an hourly observation table from Mount Washington Regional Airport (HIE) in Whitefield, New Hampshire, from August 29 to September 1, 1997. These data are from the National Climatic Data Center in Asheville, North Carolina (*http://lwf.ncdc.noaa.gov/oa/ncdc.html*). Temperatures are recorded hourly from 12:52 a.m. on August 29 until 11:52 p.m. on September 1. Figure 2 provides a time plot of the temperature data.

The World of the Wavelet

Meaning "small wave," the term "wavelet" refers to mathematical functions that break data down into different frequency components and that then analyze each of the frequency components with a scale-matched resolution. In their article "Wavelets fo Computer Graphics," that appeared in *IEE Computer Graphics and Applications*, E. J. Stollnitz, T. D. Rose, and D. H. Salesin describe wavelets as mathematical tools for hierarchically decomposing functions. They essentially allow functions to be described in terms of a coarse overall shape, along with details that range from broad to narrow.

Historically, wavelets have been touted as the quintessential mathematical tool for image compression. In computer science circles, they have been lauded for their ability to flexibly adapt to shapes and patterns of the original image and reconstruct them using minimal space. Through a tag-team effort of using high- and low-pass filters, wavelets produce snapshots of images while minimizing pixilated space. Because wavelets possess such a great ability to stretch and shrink, they are able to confront the task of duplicating complex pictures.

Over the last decade or so, the utility of wavelets has widened from this well-known idea of image compression to the relatively new area of anomaly detection. Even though several scholars have suggested that these mathematical tools may provide promising results for such detection, hardly any literature exists in which these methods are examined alongside age-old, more traditional approaches for detecting out-of-control processes. Simply said, since wavelets possess this uncanny ability to adapt and flex, they become ideal "spies" on the hunt for unknown aberrant behavior in a time series.

Of course, classical approaches to modeling techniques for detecting anomalies center on autoregressive moving-average (ARMA) models and Fourier analysis. Wavelet analysis, however, becomes more suitable than these traditional methods for several reasons:

1. Wavelet analysis allows us to analyze a series while simultaneously preserving temporal and spatial information. Other key methods either preserve temporal or spatial information, not both.
2. Wavelet analysis is more flexible in its monitoring of frequent data (data that is daily, hourly, etc.).
3. Wavelet analysis requires the least tweaking from the nonstatistician user; current software makes for a user-friendly environment to aid in the use of wavelet analysis.

The most prominent pattern in the temperature data is, like the consumption data, a daily cyclical pattern with highs at late afternoon and lows in the night. The highest temperature of 75°F occurs on the third day (August 31) at 2:52 p.m. (All temperatures are Fahrenheit.) Interestingly, the lowest temperature of 48° is also on the third day. These lows occur at 1:52 a.m. and 3:52 a.m. Temperatures during this four-day period are not unusual for New Hampshire during this time of the year.

The One-Dimensional Wavelet Transform

The goal of using wavelets is to turn the information of a signal into coefficients, which can be manipulated, stored, transmitted, analyzed, or used to reconstruct the original signal. From a methodological point of view, wavelet techniques offer an analysis of the series as a sum of orthogonal signals corresponding to different time scales.

From a more practical viewpoint, wavelets are used to extract information from different types of data like audio signals, images, and, more recently, over-the-counter sales and electricity consumption. The choice of wavelet used in the analysis is based on the interplay between a specific analysis goal (e.g., signal processing or monitoring) and the properties needed in a wavelet filter to achieve that goal. More on wavelets is explained in the two sidebars, "The World of the Wavelet" and "The Haar Wavelet."

By definition, the discrete wavelet transform (DWT) is described by the mathematical representation:

$$W_x(a,b) = \frac{1}{\sqrt{a}}\sum_{t=-\infty}^{\infty} x(t)\psi\left(\frac{t-b}{a}\right),$$

where $W_x(a,b)$ is the set of transformed wavelet coefficients at the appropriate approximation and detail levels, while $x(t)$ represents the data points of the original series at time t. We express $\psi(t)$ as the basic wavelet function onto which the signal is analyzed. When stretched or shrunk, $\psi(t)$ becomes, $\psi(\frac{t-b}{a})$ where "a" is the dilation parameter and "b" is the translation parameter. In the example in this paper, the basic wavelet is stretched or shrunk to the point at which it has become the Haar wavelet. "The Haar Wavelet" sidebar presents more details of the Haar wavelet function and describes a simple numerical example.

Essentially, the DWT is a decomposition of the original series into a linear combination of detail coefficients and the finest approximation coefficients. The DWT is very useful in analyzing data that has been parsed into varying (or multiscale) levels. Hence, wavelets become central to the analysis of our consumption and temperature data.

DWT of the Consumption Data

For the electricity consumption data, the signal is analyzed using the Haar, which is the most basic wavelet. Each wavelet, of course, has its own unique characteristic. The choice of wavelet depends on a few things: insight into the data (what each level captures), the goals of the monitoring experience, and ease of interpretation and generalization. Decomposing our consumption time series with the Haar becomes appropriate because the goal is to detect any sudden shifts occurring from back-to-back data points. Since the Haar's basic makeup is to average consecutive data points, its strength is its ability to pinpoint any huge jumps or dips in the data stream.

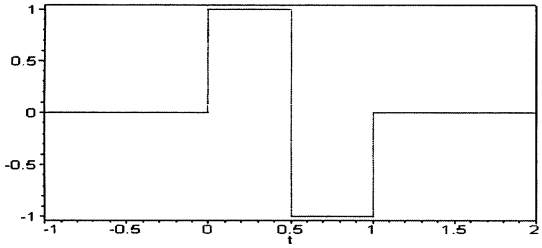
Figure 3, which can be viewed on page 30, describes the wavelet decomposition of the electricity consumption data using the Haar wavelet and five levels of decomposition. The original signal of the consumption data is given as the top graphs on either side, while the five graphs beneath represent the Haar wavelet decomposition at each of the five

The Haar Wavelet

The simplest wavelet we can use to decompose a series is the Haar wavelet. In the case of the Haar, it is useful to note its mathematical representation. Let ψ be defined by:

$$\psi(t) = \begin{cases} 1, & t \in [0, 1/2) \\ -1, & t \in [1/2, 1) \\ 0, & t \notin [0,1) \end{cases}.$$

When applied to data this wavelet performs a moving average for pairs of consecutive data points. An example is given below.



The Haar mother wavelet, $\psi(t)$

A Simple Discrete Wavelet Transformation (DWT) Example Using the Haar Wavelet

Consider the simple time series of eight observations: [9 27 30 14 20 32 50 26]. A process of averaging and differencing is employed to produce key values used in decomposing the series. Here, we apply the Haar-based DWT for only two levels to convey the basic process.

The first set of values that arise from applying the Haar to the series comes about through averaging data pairs. For example, the average of 9 and 27 gives 18. The average of 30 and 14 gives 22. Summarizing in this vein, we have 18, 22, 26, and 38. These are referred to as the approximation coefficients of the DWT. This is the first level of averaging. In signal processing, especially, we would only concern ourselves with these four approximations, which result in the notion of "downsampling," or reducing sample size. For the purposes of monitoring our data, however, it becomes appropriate to ignore downsampling and maintain all sample data points. Our goal is not to compress an image but to monitor a series. Consequently, at level A1, we preserve all data information from the original series by averaging and duplicating these newfound approximations, namely, [18 18 22 22 26 26 38 38].

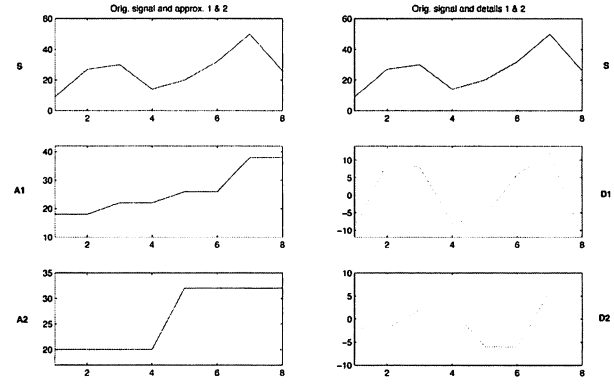


Figure 1. Original signal and DWT using the Haar wavelet

We then measure deviations by subtracting the average of the data pairs from the first data point of the appropriate pair. For example, the difference between 9 (first point of first data pair) and 18 (average of first data pair) is –9. The difference between 30 and 22 (average of second data pair) is 8. Continuing in this fashion, these deviations are –9, 8, –6, and 12 and are referred to as "detail coefficients." At level D1, these detail values are presented such that, when added back to their corresponding approximation counterparts, they give the values for the original series. Hence, the detail values are [–9 9 8 –8 –6 6 12 –12]. Essentially, we have the approximation values added to the detail values to yield the values in the original series. Or, [18 18 22 22 26 26 38 38] + [–9 9 8 –8 –6 6 12 –12] = [9 27 30 14 20 32 50 26]. Figure 1 shows a plot of the original series along with plots of the approximation (level A1) and detail (level D1) coefficients for this first-level DWT.

Now, level A1 becomes our "new" signal. We get new approximations by averaging distinct values in A1: (18+22)/2 = 20 and (26+38)/2 = 32. We report approximations as [20 20 20 20 32 32 32 32]. Instead of reporting each set of approximation values twice as in level A1, we repeat them four times in level A2 to maintain the original number of data points (no downsampling effect). We get detail values at level D2 as we did in level D1. We subtract "new" approximation values (mimicking the role of the original signal) from "old" approximation values. Or, [18 18 22 22 26 26 38 38] – [20 20 20 20 32 32 32 32] = [–2 –2 2 2 –6 –6 6 6]. Figure 1 shows these two sets of new approximation (A2) and detail (D2) coefficients.

Note that A2+D2 = A1 and A1+D1=Signal. One could report A1 or A2 to approximate the series and reduce the amount of information that is required. Or one can examine the details (D1 and D2) to look for abnormal behavior.

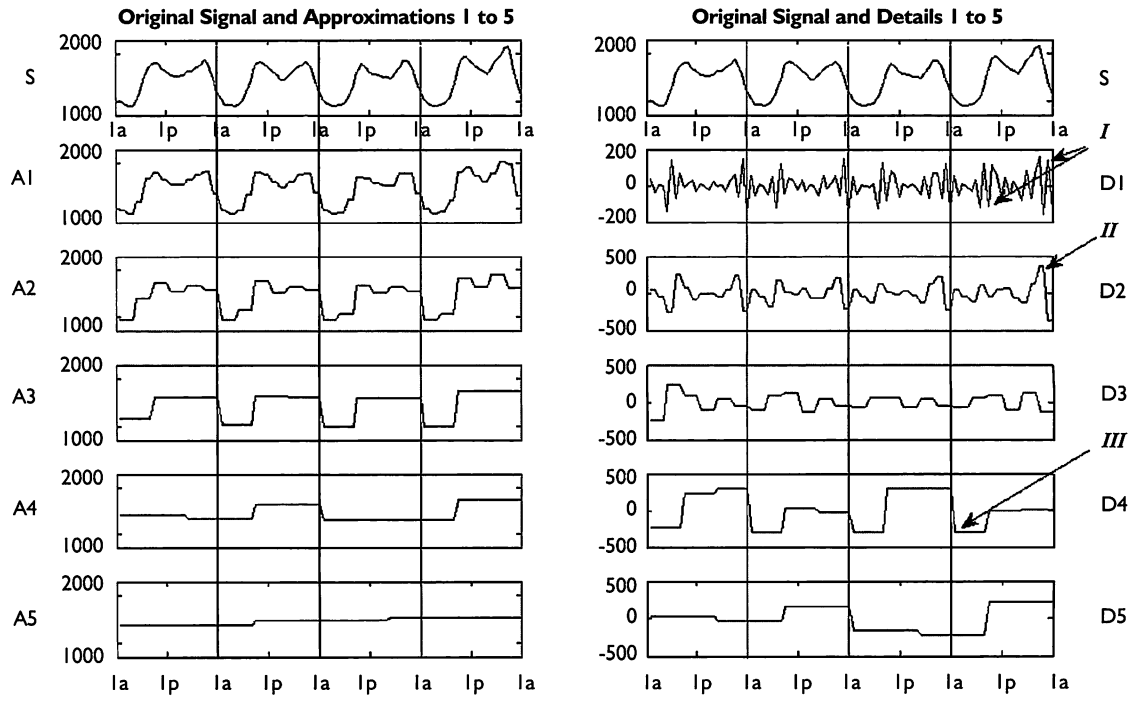


Figure 3. Discrete wavelet transformation (DWT) of electricity consumption data using the Haar wavelet

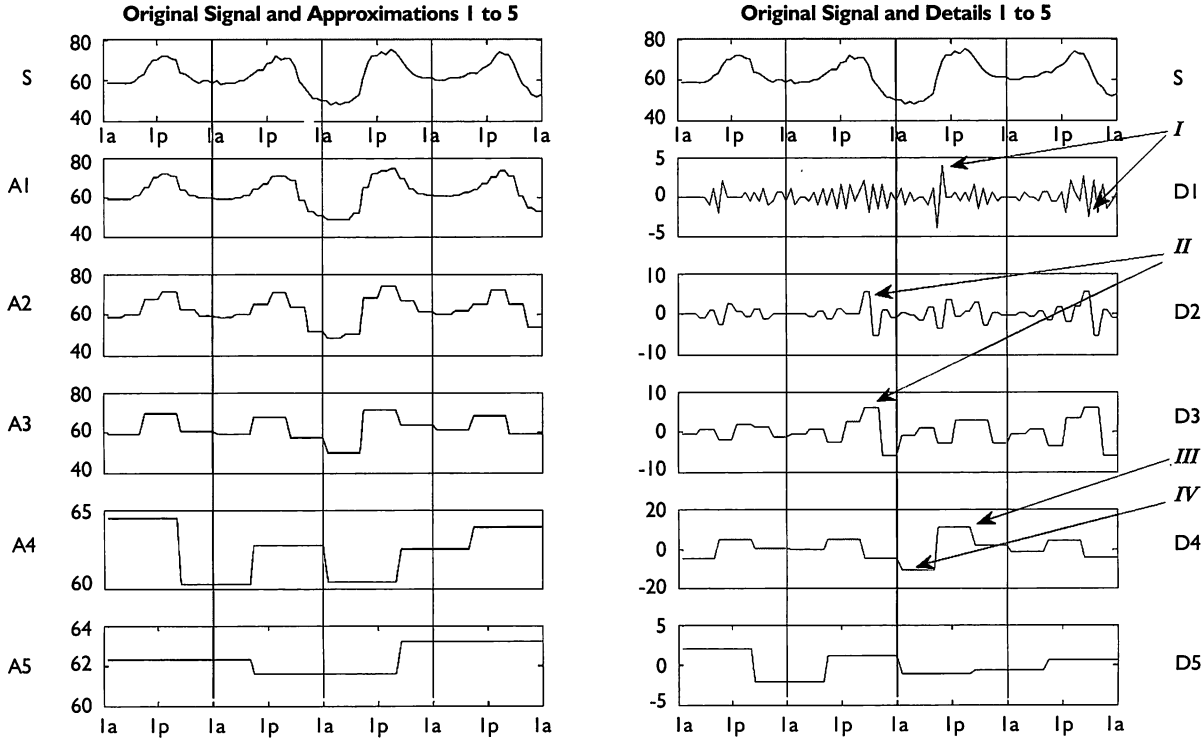


Figure 4. Discrete wavelet transformation (DWT) of temperature data using the Haar wavelet

levels. The graphs on the left correspond to the approximations, and those on the right correspond to the details. For ease of viewing, we insert vertical lines, which show approximate beginning and ending of days. We reiterate that we are using a DWT here: Our concern here is to analyze the consumption data independent of the temperature data.

We see that different features of the data are captured at different levels of the decomposition: While D4 captures the daily cycle, D3 captures the four similar toothlike structures that are apparent in the original series. At level D1, the points marked by *I* capture the most abrupt increases and decreases in consumption (day 4). At level D2, point *II* captures the absolute maximum value of the original series of 2148.12 kwh at 9:52 p.m. on day 4. At level D4, point *III* captures the absolute minimum consumption value of 882.36 kwh at 3:52 a.m. on day 4.

Regarding the choice of the number of decomposition levels, it appears that five levels capture most of the relevant information in the data, whereas further levels of decomposition contain only irrelevant noise. Technically, the fifth level of decomposition requires 2^5 data points (which we have, in fact). Decomposing into too many levels imposes more "holes" in the analysis and would sacrifice accuracy.

Of course, if there were a different consumption series, the location of *I*, *II*, and *III* would be different, depending on the traits of the different series. When the time series change, the results change. Further, if the series were too massively long, the DWT would still highlight these points, since the Haar focuses on what occurs between two consecutive points and not what happens among a massive group of points. One should remember that, at this stage, no anomaly detection has taken place. The DWT is simply a method that summarizes and breaks down the original series into details and approximation coefficients. It is true that the sum of D1+D2+D3+D4+D5+A5 yields the original value of the electricity consumption series at that particular time point for all points.

DWT of the Temperature Data

Following the same logic as stated with the consumption data, we choose the Haar wavelet to decompose the temperature series. Again, the task using this DWT is to recognize key details in the series independent of the electricity consumption series. Figure 4 describes the Haar wavelet decomposition of the temperature series. The original series is given as the top graph, and the five graphs below it represent the five-level decomposition. The graphs on the left correspond to the approximations, and those on the right correspond to the details. Daily times are inserted, which approximately correspond to 1 a.m. and 1 p.m. The decomposition of the temperature series reveals several details.

Level D1 captures the sharpest changes in temperatures (two points marked by *I*): the sharp increase on day 3 (from 61° to 69°) and the sharp fall in temperature on day 4 (from 68° to 62°). It is also interesting to see what happens at the two points marked by *II*, which occur on the second day around 6 p.m. This peak represents the difference between the approximations (on the left side) at levels 1 and 2 and at levels 2 and 3. At this point, the decomposition highlights the major and sudden decrease in temperature during this time. The approximation differences at these times become noteworthy. At level D4, the Haar captures the daily cycle, highlighting each of the four "hills" in the original series. Point *III* captures the highest peak occurring on the third day, while point *IV* captures the lowest point of the time series.

Monitoring Time Series

Traditionally, the established statistical method of control charts has been used for monitoring a process over time for the purpose of detecting abnormalities. This technique has been heavily used in many applications. It provides systematic guidelines for showing if a process is "in control" or "out of control," where an in-control process is defined as "a process that is operating with only chance causes of variation present" or one in which the chance causes of variation are an inherent part of the process, as D. C. Montgomery states in *Introduction to Statistical Quality Control*. For example, sources of variability, which cause the process to be out of control, arise from operator errors, maladjusted machines, or other defective bases.

To create a control chart for monitoring the mean of a process, independent and identically distributed (i.i.d.) samples are taken every time point from the process, and the following are computed: the process (or sample) mean, the sample size, the standard deviation of the sample average, and a constant value for the Z-statistic. We use these values to compute the upper control limit (UCL) and lower control limit (LCL), which are the statistical cutoffs for assessing whether the process is in or out of control.

Once a new sample arrives, if its mean exceeds the control limits, an out-of-control alarm is raised. Such control charts that abide according to the above principles are referred to as "Shewhart control charts," developed by Walter S. Shewhart. Conventionally, we use the constant value of "3" for our Z-statistic, which further classifies our limits in the chart as "3-sigma" control limits.

The problem with simple Shewhart charts is that they assume i.i.d. samples. In a time series, we typically have samples of size 1 (one series of measurements), and the points are autocorrelated (correlated with one another over time). Furthermore, in many cases we cannot assume, as Shewhart charts do, that the distribution of the observations is normal. On the other hand, we show in the previous section how wavelets can be used to analyze autocorrelated observations without making distributional assumptions.

A method called "multiscale statistical process control" (MSSPC), combining DWT and Shewhart control charts, was introduced by B. R. Bakshi and is primarily used in chemical engineering circles. The idea of this methodology is to decompose the signal using DWT and then to monitor the coefficients at each scale separately using a Shewhart chart. If there is an alarm at one or more scales, the original series is reconstructed from all the coefficients that exceed the thresholds and another Shewhart chart is used to monitor this reconstructed series. On the following page, Figure 5 illustrates this MSSPC process.

MSSPC sounds an alarm if a new point in the reconstructed series exceeds the control limits. Next, we illustrate this method for our two time series.

MSSPC for the Consumption Data

Figure 3 shows a DWT of the electrical data using the Haar. The UCL and LCL are computed at each decomposition level by using the standard deviation and mean of the coefficients in that level. For the approximation level, the mean is exactly the same as for the original series.

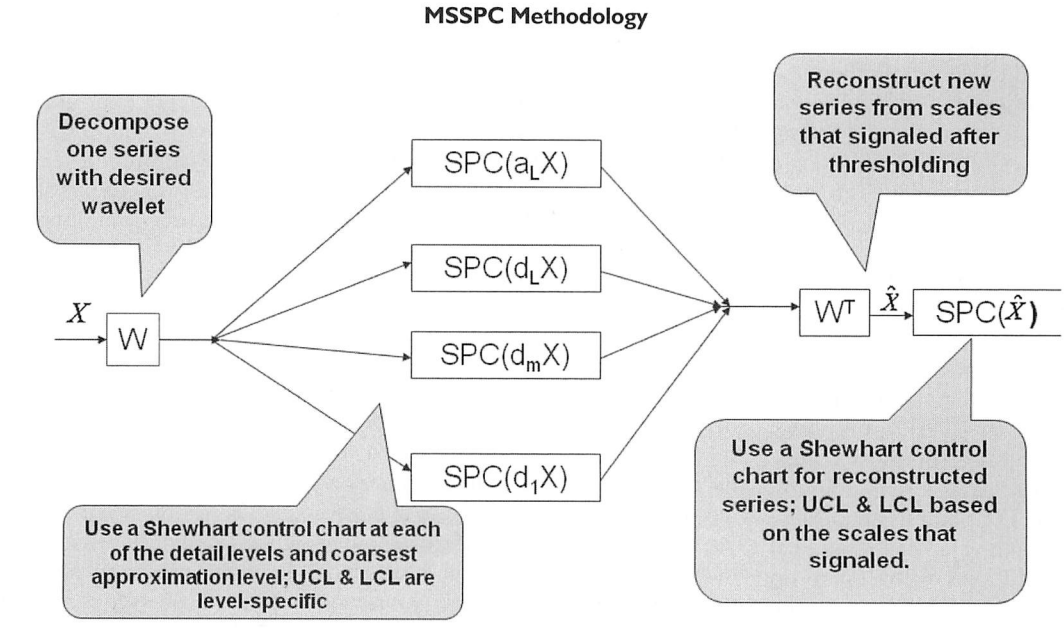


Figure 5. Illustration of MSSPC algorithm. X represents the initial series. W represents the wavelet decomposition of X. The quantity $a_L X$ represents the finest approximation level, while $d_1 X$-$d_L X$ represent the range of the detail levels. The m^{th} detail level is denoted $d_m X$. SPC denotes the implementation of a Shewhart control chart at that level. w^T represents the wavelet reconstruction onto X, which we call $\hat{X}$. Finally, we construct a control chart for the reconstructed signal, which we denote by SPC($\hat{X}$).

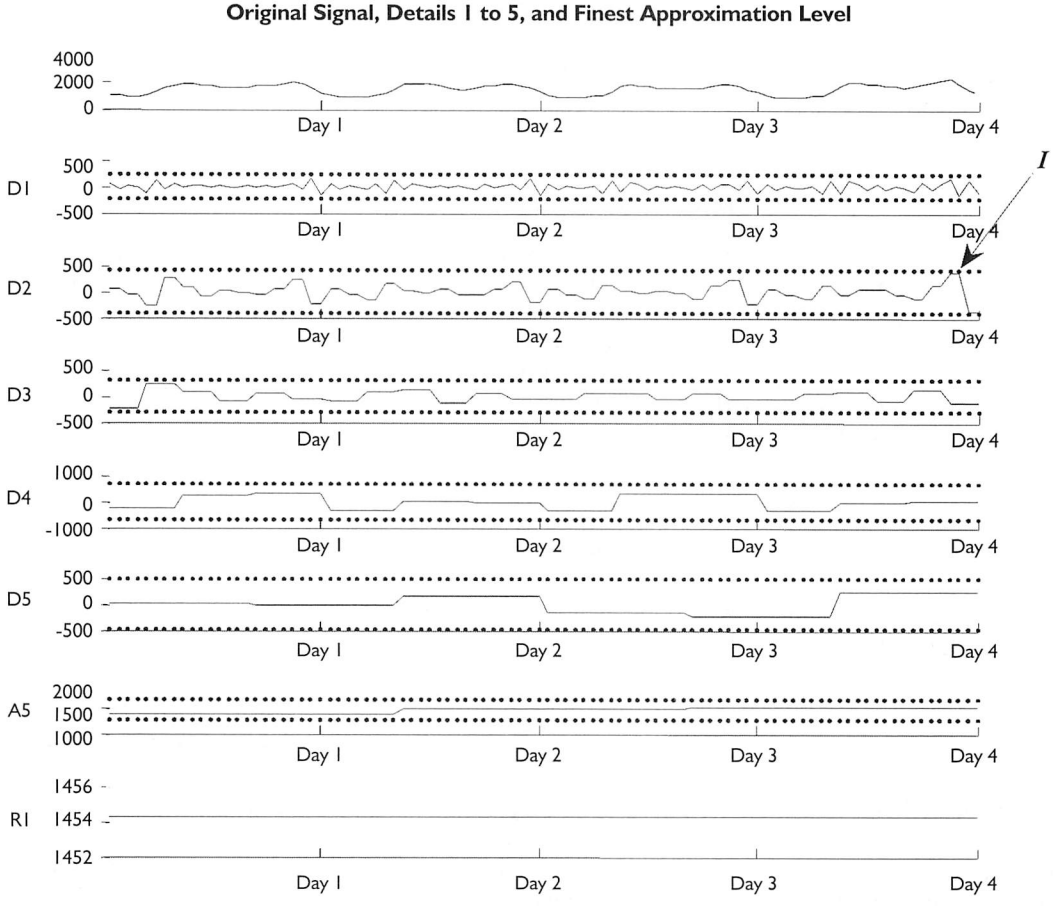


Figure 6. Application of multiscale statistical process control (MSSPC) to electricity consumption data. For each pair of lines, the top dotted line represents the upper control limit (UCL) and the bottom dotted line represents the lower control limit (LCL).

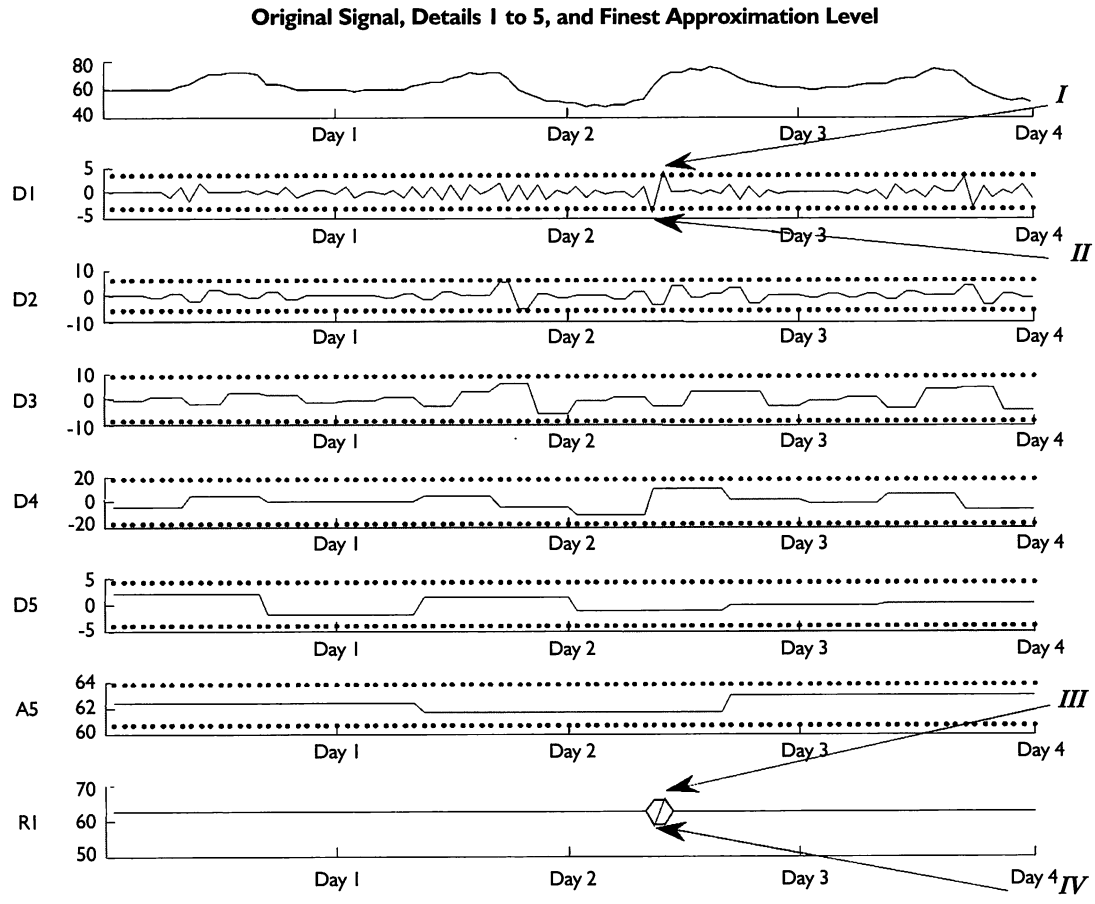


Figure 7. Application of multiscale statistical process control (MSSPC) to temperature data. For each pair of lines, the top dotted line at each level represents the upper control limit (UCL) and the bottom dotted line represents the lower control limit (LCL).

For the detail coefficients, the mean at each decomposition level is zero. The standard deviation varies per level and is estimated from the coefficients in that level.

Figure 6 shows the MSSPC method applied to the electricity data. Each of the four days is labeled as a marker in the monitoring process. We use observations only from the first three days (72 hours) to determine the UCL and LCL. Doing this allows us to create tighter UCLs and LCLs in our attempt to identify outlying coefficients. The top dotted line in each of the D1–D5 and A5 levels represents the UCL, which is given by the mean plus three times its standard deviation. The dotted line below it represents the LCL, which is given by the mean minus three times its standard deviation. We observe that in each level, there is no point exceeding the UCL or LCL. Based on our baseline stability period, the closest point that approaches either of the limit lines is point *I*, which is the greatest consumption value.

According to the control chart, however, since it still lies inside the control limits, it is not classified as a point that is out of control. Since none of the points at either of the detail levels or the finest approximation level falls outside of the control limits, then none of them falls outside the limits in the reconstruction phase. For this series, MSSPC advises us to zero-out all points at this phase and conclude that there is no anomalous behavior attached to this particular process of electricity consumption. Level R1 (the reconstructed series) illustrates this truth, as all points have been zeroed out. There are no UCL and LCL at this level because there is no standard deviation to take into account, since all of the points at each of the decomposition levels remain inside the detection limits. Hence, all sources of variation in this case are considered to be an inherent part of the process.

MSSPC for the Temperature Data

MSSPC is now applied to our temperature data to see if any points are detected that contribute to an out-of-control process. Figure 7 illustrates this process applied to our familiar temperature data points.

Again, only observations from the first three days (72 hours) are considered for determining process stability. The upper and lower dotted lines are identified as the UCL and LCL, respectively, as described before. The difference in this data set is that there are a couple of points that fall outside of the UCL and LCL. Points I and II exceed the value of the UCL and LCL for the prescribed algorithm for this decomposition level. For this detail, the algorithm gives a UCL and an LCL of 3.45 units. Points I and II are both located at +4 units and -4 units, respectively. This means that in the reconstruction of the series (R1), these

coefficients would be retained and added to any other coefficients at this same time point that lie outside of the UCL and LCL of its respective decomposition detail level. These points are denoted by III and IV.

Only the finest scale coefficient is used to compute the reconstructed signal and detection limit for the reconstructed signal. So we use information from the standard deviation in D1 to calculate the detection limit in R1. The reconstructed series then provides us with statistically based reasoning for concluding that some of the variation in the temperature series is attributable to factors other than mere chance.

The Baseline Stability Period

Although wavelets have the ability to monitor hourly data, as seen in this article, real applications typically would have several days or weeks of data available. In our case, there is no guarantee that the temperature points that were labeled as out of control would continue to be so if there had been a much longer length of stability on which the UCLs and LCLs were based. This, in fact, becomes the beauty of MSSPC. Although the algorithm is nonnegotiable with respect to its methodology, it is flexible in that its results are strictly data-driven, data-dependent, and data-sensitive. UCLs and LCLs change slightly or drastically for a different baseline period of stability.

A great deal of responsibility lies with the statistician whose task it is to analyze the history of the data and to report the baseline stability period upon which anomalous behavior will be based. As one might imagine, it becomes advantageous to revisit the data and update this stability period every so often to take into account reasonable data changes (such as expansion of customer base).

Any nonsignificant points near but not beyond the control limits will generally remain nonsignificant even if there had been a longer series, as long as the stability period remains unchanged. Hence, a change in significance or nonsignificance of out-of-control points becomes a function of the change of the baseline stability period, as opposed to a function of how long the data series itself is. Irrespective of the baseline period, we still rely heavily on the strength of the Haar wavelet to zero in on sudden and quick disparities in consecutive time points and signal an appropriate alarm. As mentioned, the Haar has a short memory, which actually serves as the strength in this particular application. This wavelet does not depend necessarily on how long a process has been "normal" or how long the baseline stability period is.

There is no definitive formula for calibrating this stability threshold. Determining the length of stability, then, should be based on a logical sense of data history within the business (such as the power company). Care should be taken in establishing this baseline period so as to minimize the number of false positives (alerts that are not really alerts) and the number of positive failures (real alerts that are not reported). Practically, businesses would probably need to hire statistical consultants to parse through data and determine an appropriate length of process stability based on each particular data set.

Practical Decisions Based on Results

What do you do with out-of-control points in the reconstruction phase? After all, what good is all the statistical talk if there is no suggested course of action? Although there are no hard-and-fast rules concerning what should be done in cases where statistically based methods are employed, we can at least make decisions informed by something other than gut feeling.

In this analysis, our concern primarily lies with values exceeding the UCL, since we were monitoring for high electricity consumption or high temperature. But we could very well have been monitoring both variables during the winter months, in which temperature values falling below the LCL for a certain set of data could signal an alarm by translating into increased consumption and possible grid system failure. Hence, using both the UCL and LCL becomes a key strategy to pinpointing such out-of-control data points.

Greater still, situations exist whereby both limits may not necessarily be needful. Only a one-sided limit may be needed, thereby requiring a less extreme high value for significance. For example, one application of this method is to the area of biosurveillance as it relates to rapid detection of a large-scale bioterrorist (such as anthrax) attack. In the last several years, scholars such as Goldenberg et al. have led efforts to monitor nontraditional data sources, such as sales of various over-the-counter (OTC) medications. In their study, they explore situations in which they monitor sales of certain grocery items that people may purchase to treat flulike symptoms.

Significant spikes in daily purchase levels above a prescribed UCL only (and not below a certain LCL level) might suggest cause for alarm and may allow for timely intervention before life-threatening spores would cause damage to a person's respiratory system. But the idea is that through real time (via UPC bar-code scanning), this data mining and monitoring technique could be applied almost immediately to serve the public's safety interest. In this instance, emphasis would be placed only on analyzing what transpires above the UCL, thus lowering our limit cutoff point and allowing for a less extreme high value for significance. Essentially, the cutoff value applied with MSSPC would depend on the data and the nature of the process to be monitored.

The results here show the implementation of MSSPC with the UCL and LCL overlaying all the data points. This simply shows the overall verdict on which points would have signaled an alarm. However, in real life, the points would be identified shortly after having been observed or recorded. Once the point was identified as being out of control, it would immediately be reported to an automated system set up to monitor quality control, which would be established by the statistician and the power grid company. In this new world of technology, little effort would have to be exerted to devise an automated monitoring system to do this and forward information to powers-that-be in control towers to have them make decisions concerning what the system has found to be out-of-control points.

Since outlying aberrant data points are strictly based on the historical baseline stability period, we would not need knowledge of the entire series to zero in on out-of-control points. The algorithm identifies and catches the point quickly. Hence, if the data were being monitored hourly or daily, the automated system could report that aberrant behavior shortly after that hourly or daily observation occurs, respectively.

In this article, we use a series having only 96 data points. In real life, we could conceivably monitor a series with indefinite length. Significance is not based on how many points are in the series: As soon as the point falls outside of the control limits, it is reported as aberrant. Being out of control is simply based on "looking back" while "rolling forward." The monitoring process would remain in effect until it needed to be updated or tweaked. The only tweaking to the algorithm would be a new change in the baseline stability period (which would be data entered by the statistician) and whether one- or two-sided limits would be used in MSSPC.

The basic idea, then, becomes to develop an automated indication system that provides an alert or flag to process operators that an observed value has exceeded prescribed thresholds and should be addressed quickly. In our scenario, an alert would be given at the time point in the temperature series that lies outside the control limits. What this would mean is that consumption should arguably be decreased within a reasonable time frame of the alert so as to avoid any possible overloading effect or blackout.

Of course, one does not make decisions solely on this algorithm. These statistical methods simply provide an objective monitoring tool, which could be used in tandem with other methods that electric companies already use to monitor their power grids. Essentially, attacking these kinds of issues using wavelet-based methods equates to dealing with such problems proactively rather than reactively.

Future Directions

This article examines wavelet-based methods for analyzing and monitoring very frequent time series. Although these methods require less parameters and assumptions than other traditional statistical monitoring methods, there are a few parameters that must be specified by the modeler. The first is the choice of wavelet. Again, the Haar is useful because of its ability for detecting sudden, abrupt changes in the series.

Since the Haar is a wavelet whose basic function is to average and subtract consecutive points in the series, it detects quickly any two consecutive points that have a large range. For gradual, slower changes in the series, this wavelet is not as useful because of its "short memory." Even if the overall series is slowly, monotonically increasing or decreasing, the Haar essentially averages out this effect. By using more complicated wavelets, we are able to capture trends and patterns of the series that are more subtle, like the situation described above.

The focus, as seen in the two previous examples, has been on analyzing each data set independent of other factors. However, this traditional DWT approach is limited in that the analysis does not capture any interaction between the series when it was clear that electricity consumption depended, among other things, on the temperature. The primary task, then, becomes to investigate both series together and to use wavelet decomposition to analyze both signals simultaneously. Ideally, we want a method that could be generalized to more than two series and that could capture not only relationships between the series at the same time points (e.g., at hour 3) but also relationships of a lagged nature. This multivariate method, called "multiscale principal components analysis" (MSPCA), seeks to capture the interrelations between the different series and any abnormalities that might occur in this relationship. This type of monitoring becomes crucial in detecting anomalies based on the interplay between series.

What do you think the results would be if both series could be monitored together in a manner that takes account of their associations with one another over time? We anticipate that we will be better able to identify out-of-control time points when monitoring both series together. A simultaneous monitoring system that would take into account changes in relations between different data sources could provide a great improvement in detecting real outbreaks and in eliminating false alarms. Such theory almost begs for great minds in the statistical field to develop the MSPCA algorithm and investigate results, comparing them to monitoring with univariate monitoring with MSSPC.

These monitoring techniques can prove invaluable for the statistician. At the end of the day, however, such wavelet-based monitoring techniques described herein can be helpful in guarding against situations that threaten normal electricity consumption load within a city. Proper use of these methods can contribute to the early detection of any ensuing electricity consumption overload and the natural response of decreasing consumption on the necessary power grid. C

Further Reading

Bakshi, B. R. 1998. Multiscale PCA with application to multivariate statistical process monitoring. *American Institute of Chemical Engineering Journal* 44:1596–1610.

Basu, S., and Mukherjee, A. 1999. Time Series Models for Internet Traffic, *INFOCOM* vol. 2, 1996: 611–620.

Cottet, R., and Smith, M. 2003. Bayesian modeling and forecasting of intraday electricity load. *Journal of the American Statistical Association* 98:839–849.

Goldenberg, A., Shmueli, G., Caruana, R. A., and Fienberg, E. S. 2002. Early statistical detection of anthrax outbreaks by tracking over-the-counter medication sales. *Proceedings of the National Academy of Sciences* 99:5237–5240.

Graps, A. L. 1995. Introduction to wavelets. *IEEE Computational Sciences and Engineering* 2(2):50–61.

Harvey, A., and Koopman, S. 1993. Forecasting hourly electricity demand using time-varying splines. *Journal of the American Statistical Association* 88:1228–1253.

Hubbard, B. B. 1998. *The world according to wavelets.* Wellesley, MA: A. K. Peters, Ltd.

Montgomery, D. C. 2004. *Introduction to statistical quality control, 5th ed.* Hoboken, NJ: Wiley & Sons.

Percival, D. B., and Walden, A. T. 2000. *Wavelet methods for time series analysis.* Cambridge, UK: Cambridge University Press.

Stollnitz, E. J., DeRose, T. D., and Salesin, D. H. 1995. Wavelets for computer graphics: A primer, Part 1, *IEEE Computer Graphics and Applications* 15(3):76–84.

You, C and Chandra, K 1999. Time Series Models for Internet Data Traffic *IEEE Computer Society*:164–171