that the date is a Sunday. It is also possible to add local terms, such as the mean count over the previous seven days as an additional feature to allow the method to adapt better to recent local events. These additions can be very helpful, and as we shall see at the end of the chapter, regression methods that include these extra terms perform better than moving average (which had been our favorite method) on our data.

10. SICKNESS AVAILABILITY

Another way to deal with day-of-week variations is to use the sickness availability method to smooth the time series by removing noise due to the day-of-week effect. This algorithm transforms the daily counts in the time series into a daily sickness value, which is defined as the number of people getting sick every day irrespective of whether they seek health care or not. The term availability refers to the probability that a patient will seek health care during a specific day of the week; hence, there are a total of seven values of availability, one for each day of the week. The availability of a day can be thought of as the fraction of a weeks-worth of visits that get assigned to the given day. The sickness availability method is based on the intuitive assumptions that the expected count is the product of the true amount of sickness and the current day's availability.

We can estimate the expected availability for a specific day of week (dow) using the average of the availabilities on that day for the past m weeks. We can calculate the expected availability A_{dow} as:

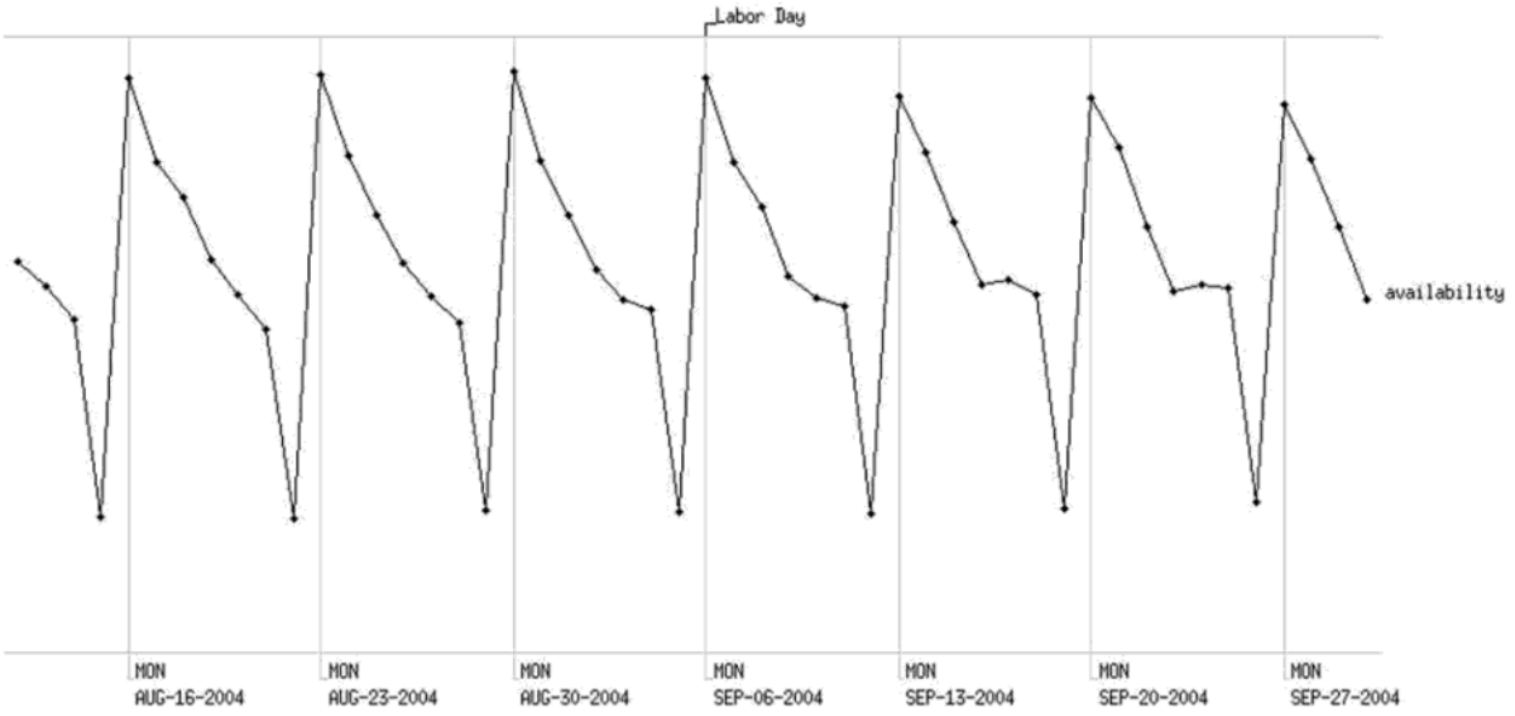
$$A_{dow} = \frac{\sum_{i=1}^{m} (C_{i(dow)} / \sum_{dow=0}^{6} C_{i(dow)})}{m} \tag{9}$$

In Eq. 9, A_{dow} refers to the expected availability for the day of week specified by the variable *dow*, which takes on seven different values ranging from 0 to 6. The term $C_{i(\text{dow})}$ is the actual number of patients that visited the ED on the particular day of week *dow* during the *i*th week in the past. Since national holidays affect the number of patients visiting the ED, weeks containing holidays are ignored completely in the availability calculations. Finally, the parameter *m* controls the smoothness of the sickness curve.

Since sickness is defined as the total number of people in the city getting sick on a particular day, we can calculate it as follows:

$$S_{today} = \frac{C_{today}}{A_{today}} \tag{10}$$

The term S_{today} in Eq. 10 refers to the number of people who are sick on the current day while the term C_{today} refers to the number of people who visited the ED on the current day. A_{today} is the probability that a patient will visit the ED for the day of week specified by *today*. Figure 14.18 shows the availability estimated in the period leading up to the synthetic ramp outbreak: it shows a consistent picture that we usually see far more patients on Mondays than Sundays, and then the visits taper off during the rest of the week. Figure 14.19 shows both the original count and the estimated sickness (count/availability). Sickness is much more stable than the original count because day of week effects have been greatly reduced. It is now possible to run a time series algorithm on the sickness values. In this case, we chose to use moving average with a window of seven days. The resulting alarms show something that was not achieved in any



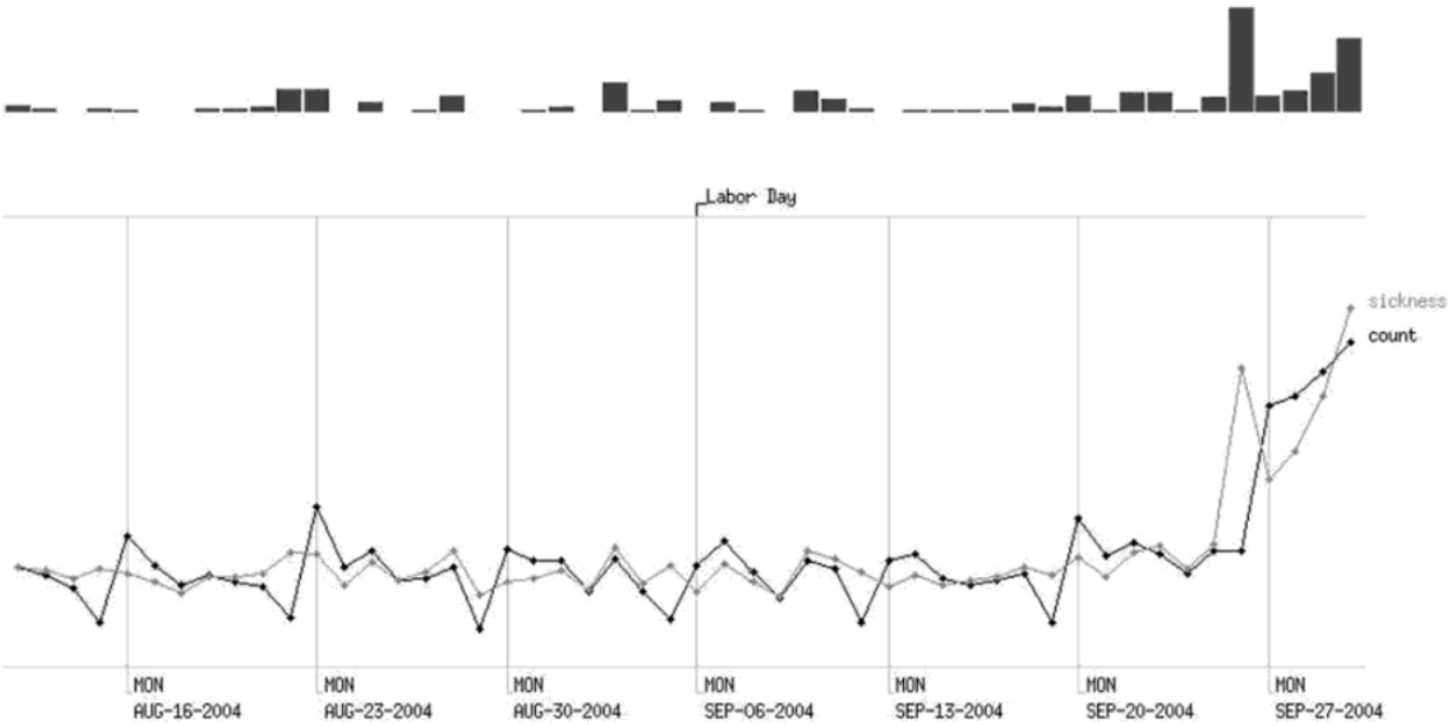


FIGURE 14.19 The black "count" line shows the raw data. The gray "sickness" line shows the sickness after the count has been divided by the corresponding availability value from Figure 14.18. Note how the sickness time series is now smoother than the original data and decorrelated with day of week. The alarm levels are derived from moving average applied to the sickness time series.

of the previous illustrations: a strong alarm resulting from the higher-than-expected counts for the Sunday.

We would like to emphasize that the sickness availability method only smoothes the time series. It is not a detection algorithm by itself. Instead, a detection algorithm, such as the control chart, moving average, or CUSUM algorithms should be used on the smoothed data.

11. FURTHER COMPARISON OF THE UNIVARIATE ALGORITHMS

We insert another copy of Table 14.1, with some of the above methods added for comparison. The newly evaluated algorithms follow:

- Regression using two features: the mean count over the past week and hours of daylight. This allows the algorithm to account for seasonal variation by putting a negative coefficient in front of hours of daylight.
- Regression using the additional feature *is_Monday*, which is set to 1 if today is Monday and 0 otherwise. This allows the algorithm to anticipate the Monday bump in physician visits and so be less prone to false positives.
- Regression using indicator variables for all days of the week except for Sunday (which would be redundant), and additionally, *hours_of_daylight* and *mean_count_over_previous_seven_days*.
- Using sickness/availability to compensate day-of-week effects, and then using the approach of comparing against yesterday. This method thus looks for jumps in the day-of-week-adjusted counts.

For this data set, with its seasonal components and day-of-week components, we see that sickness availability (to cope with day-of-week effects) combined with moving average (to cope with seasonal trends) performs well. Some of the regression methods perform almost equally well. We should note that this does not mean these methods are best in general: individual properties of individual data sets mean different approaches can be stronger for different data sets. Our only general advice is that in our experience relatively simple methods usually work at least as well as complex approaches. A second important note is that the numbers in Table 14.2 cannot be used as an estimate of how quickly real outbreaks are expected to be detected: the numbers are a function of many things, including the simulated magnitude of the outbreaks and the simulated noise levels.

12. ADDITIONAL METHODS

A complete set of approaches would require a very thick book, such as Hamilton (1994). The purpose of this chapter has been a tour of a sufficient variety of approaches to introduce the reader to some of the issues faced when choosing or implementing a time-series–based method, without going into the