Chapter 23:
Time Series Forecasting

Wilpen L. Gorr

H. John Heinz III College
Carnegie Mellon University
Pittsburgh, Pennsylvania

Andreas M. Olligschlaeger

TruNorth Data Systems
Baden, Pennsylvania

Table of Contents

Chapter 23:
Time Series Forecasting

Introduction

This chapter presents time series methods useful for tactical deployment of police resources. The methods answer the following questions:

- What crime levels are expected in police zones, patrol districts, census tracts, or other areas of a jurisdiction given past time trends?

- Are there any new crime patterns, large increases or decreases, starting up in the jurisdiction?

The first question is answered using extrapolative time series models to forecast expected crime levels by geographic area. The second uses observed departures from the expected crime levels as the basis for detecting new crime patterns on an early warning basis. Automation is important for such work. For example, if a police organization has 100 census tracts and 10 crime types that it wishes to track, then it would have 1,000 time series to analyze—too much work for manual visual inspection. This chapter presents standard models and methods long-used in industry for time series forecasting and detection, making them available in highly optimized and automated computer code tailored for crime analysis.

Time series forecasting can be complex and require sophisticated software and highly-trained analysts. The good news here, however, is that the forecasting literature on operations management applications has shown that simple methods are as accurate as or more so than complex methods. Most influential have been the so called "M-Competitions" in which forecasters forecasted time series data without seeing the future data and independent judges analyzed forecast accuracy with the future data (see Makridakis et al.,1982 and Hibon & Makridakis, 2000).

This chapter presents simple methods that are easy to understand and use. These methods are self-adaptive to changing conditions, taking care of themselves in the dynamic setting of cities and actions of criminals and police. The implementation in *CrimeStat* optimizes forecast model parameters in extensive but fast algorithms, thereby making the module easy to use. The detection component has parameters that cannot be readily optimized, but the chapter provides default values from a research study on crime detection (Cohen, et al. 2009). Finally, all areas of interest such as all patrol districts in a jurisdiction are processed in a single run, again making it efficient for analysts. Outputs are easily displayed in GIS as choropleth maps. On these maps it is

desirable for analysts to also display individual crime points when zooming into areas of interest for detailed diagnosis (see Gorr and Kurland, Chapter 2, 2012). Out of the entire jurisdiction, the automated detection methods bring to attention the areas needing further analysis. Then the crime analyst zooms into those areas, one-by-one, and studies the detailed crime patterns.

The chapter starts with overviews of time series data and extrapolative forecast methods. It then presents details on exponential smoothing models, which are among the simplest but most accurate forecasting methods, along with classical decomposition for estimation of seasonal adjustments. Early detection of time series pattern changes is the final topic, covered first at the conceptual level and then as implemented in *CrimeStat*. Exponential smoothing forecasts are the basis for detection, so that all the methods in this chapter work together.

Time Series Data

This chapter uses univariate time series data, meaning that for a given observation unit, there is only a single variable:

$$y_{it} \; i = \; 1, \dots, m \text{ and } t \; = \; 1, \dots, T \tag{23.1},$$

where
   i = area identifier (e.g., patrol district, census tract)
   m = total number of areas
   t = time period such as month serial number
   T = most recent time period, called the "forecast origin."

Service Demand

For example, in the private sector, y_{it} might be defined as the product demand by sales territory represented by the number of units of product sold per time period. For police the corresponding variable could be the number of crime incidents of a particular type per time period, usually counts of offense report incidents or computer aided dispatch (CAD) calls for service. Arrest and other incident-related data are less useful for tactical deployment of police resources. Offense reports are official records of crimes having been committed and can be either Uniform Crime Reporting (UCR) hierarchy-based in the U.S., with only the most serious crimes included, or incident-based with individual records for each crime type committed in an incident being included. Which one to use depends on the particular need required. For example, for dispatching purposes UCR data may be sufficient to determine priority of a call for service.

The question arises, should one use raw crime counts per area and time interval or crime rates which are crime counts per time interval and divided by the population at risk? The answer

for tactical deployment of police is to use raw crime counts which directly determine the size of effort needed for police deployment. Crime rate is somewhat academic in the sense that it tells something about behavior of criminals and the effects of congestion, valuable for insight and understanding, but not the needed measure for resource allocation.

Fixed time and observation units

The time interval of observations in the forecasting area can be any unit of time from hours to days, weeks, months, quarters, or years. For example, electric load forecasting needs hourly forecasts. Time series analysis (but not forecasting) of crime patterns can benefit from estimating hourly seasonality factors of crime for week days versus weekend days. While average hourly variations in crime over the 24 hours of a day can be informative, there is not a large enough sample size in crime data by zone or patrol district to yield accurate hourly forecasts.

Generally the smallest time intervals possible for crime forecasting are weeks or months, but then the time series data is still very noisy. The number of days per month of course varies, and some organizations adjust monthly time series data to the average number of days per month —multiplying by (365/12)/(days in a month), or doing a similar calculation with the number of work days (trading days) per month. This is not necessary if including seasonality in a forecast model, because the seasonal factor automatically includes an appropriate adjustment. Most crime time series data is seasonal.

The data used in this chapter might better be called "space and time series data" because for crime in a police jurisdiction there are administrative and other areas of interest, each having crime time series data and needing forecasts. The administrative areas include "zones" each with a police station, commander, field officers, and so forth operating semi-autonomously. A zone is partitioned into "patrol districts" each with a single patrol unit assigned three shifts per day. Other areas of interest include census reporting areas, such as census tracts in the U.S. Census areas which generally have a target population (e.g., 4,000 for census tracts) and are neighborhoods with common socio-economic patterns. Often in the U.S. patrol districts are made up of one or a few census tracts.

Whatever areas are of interest, the needed aggregate space and time series data can easily be generated using GIS:

1. Geocode crime incidents using street addresses.

2. Spatially join the geocoded crime incidents to polygons of the area of interest.

3. Using the date of incident, create variables for year and week or month.

4. Count crime incidents by area, year, and week or month to form the crime space and time series data.

The extrapolative forecast models of this chapter make minimum use of the spatial arrangement of the areas being forecasted. Multivariate models, not covered in this chapter, can use crime counts nearby a particular area as part of the model. Here, however, the only use of data outside of a particular area being modeled is in estimation of seasonal factors. As explained below, estimating seasonality using data from the entire jurisdiction is generally better than doing so for each individual district (whether zones or patrol districts).

Limitations

A major tradeoff associated with crime space and time series data is that police need forecasts for areas as small as possible so as to target resources precisely. However, the smaller the area, the fewer the crimes and the less reliable are the model estimates and forecasts. Past work suggests that for areas as small as census tracts, it is only possible to forecast high volume crimes or crime aggregates (e.g., all serious property crimes) with useful accuracy at the monthly level (Gorr, Olligschlaeger, & Thompson, 2003). Low volume crimes, such as homicides, cannot be forecasted accurately at all, even for an entire jurisdiction.

The inability to forecast small areas accurately is not a big a limitation as it seems, because of research on crime hot spots. Typically 50% of crime occurs in only 5% or less of the area of a city, in micro-area hot spots (e.g., Weisburd & Green, 1995, Weisburd et al, 2004; see Chapters 7, 8 and 9 of the *CrimeStat* manual). So if one has an early detection of a large crime increase, it likely will be in small areas. Then patrols and other police resources can be directed to the emerging hot spot.

Extrapolative Time Series Forecasting

Industry very likely uses more computer machine cycles for extrapolative forecasting than any other statistical method. Every week and month, firms forecast demand by product (or service type) and sales territory using mostly exponential smoothing models. The crime forecasting problem is analogous, with the count of crime incidents representing service demand for police. This section starts at the beginning of time series forecasting and detection, with general terms and notation used in the area before moving on to specific models.

Terms

While there are many specialized terms for specific forecast approaches and models, the general set, however, is not that large. Below is a collection of general terms and notation. A good free reference and textbook is online (Hyndman & Athanasopoulos, 2012).

- *Univariate forecast model* is one that uses time series data for the variable of interest (dependent variable only with no independent variables other than the time index itself and seasonality; see below in this list).

- *Causal or multivariate forecast model* is one that has true independent variables in addition to the dependent variable. Often multiple regression models are used for this category. This chapter does not include any such models but see Chapters 15-22 for a discussion of regression and discrete choice modeling. Much of the forecasting needs for operations management are met with univariate forecast models. Forecasts are often needed for one week or month ahead and it is difficult to beat the accuracy of univariate forecast models in the short run.

- *Forecast origin* is *T* in notation (23.1). It is the most recent data point.

- *Steps ahead* is how many time intervals into the future corresponds to a forecast. Most tactical needs for police are met with a one-step-ahead forecast.

- *Forecast horizon* is the maximum number of steps ahead, *K*, made from a forecast origin.

- *Trace forecast* is the full set of forecasts for a particular origin for each step ahead out to the horizon. For example, if one were making a trace forecast with monthly data and a 12 month horizon, forecasts would be made for each step ahead, 1, 2, and so forth up to 12. Generally, forecast errors increase as step ahead increases.

- *Level* is the current estimated mean of a time series, denoted in this chapter by a_{it} for area i at time t. Notice that level varies with time because this chapter uses exponential smoothing methods in which model parameters, such as a_{it}, self-adapt to changing time series patterns. Simple exponential smoothing, to be discussed in depth in the next section, has only a_{it} as its parameter and thus just estimates the current mean of a time series:

$$\bar{y}_{it} = a_{it} \text{ for t } = 1, \dots, \text{T.} \tag{23.2}$$

- **_Trend_** is the estimated change per time interval in the mean of a time series, moving ahead from the level. Here trend, denoted as b_{it} for area i, is also a varying parameter. This term is added to the simple exponential smoothing model to yield the Holt exponential smoothing forecast model.

$$F_{iT+k} = a_{iT} + b_{iT}k \text{ for } k = 1, \dots, K. \tag{23.3}$$

  The fitted model at time t is still $\bar{y}_{it} = a_{it}$ because a_{it} is the current level of the time series at time t. The slope, b_{it}, only comes into play when forecasting by adding the expected change.

- **_Seasonality_** is the adjustment made for each time observation for seasonal patterns such as when, for example, crime is low in February and high in July relative to the time series trend line. For weekly data, there are 52 seasonal adjustments, S_j with j= 1, ..., 52. Likewise for monthly seasonal adjustments there are 12 seasonal adjustments. Seasonal adjustments can be additive or multiplicative. Additive seasonal adjustments are affected by the scale (volume) of data at an observation unit. So a low crime rate patrol sector would have a small seasonal adjustment for any time period but a high crime rate patrol sector would have a large adjustment. In contrast, multiplicative seasonality is unitless, having values such as 1.20 for a 20 percent increase for a summer month and 0.80 for a 20 percent decrease for a winter month relative to the time trend. For space and time series, it is desirable and necessary to use multiplicative seasonality so that seasonality estimated using all crime data of a jurisdiction can be used for any sub-area of the jurisdiction. Therefore, this chapter uses multiplicative seasonality. If time period _t_ is in season _j_, seasonality is denoted as $S_{ij(t)}$ and the Holt model and forecast with seasonal adjustments are

$$\bar{y}_{it} = S_{ij(t)} a_{iT} \text{ for } t = 1, \dots, T \tag{23.4}$$

$$F_{iT+k} = S_{ij(T+k)}(a_{iT} + b_{iT}k) \text{ for } t = 1, \dots, T \tag{23.5}$$

  The method by which this model is estimated, including seasonal factors, is covered in later parts of this section.

- **_Fit error_** is $e_{it} = y_{it} - \bar{y}_{it}$ where y_{it} is data from t = 1, ..., T. Typically parameters such as a_{it} and b_{it} are estimated by finding values that minimize the sum of squared fit errors, $\Sigma\, e_{it}^2$ over all historical data.

- ***Hold-out sample*** is data used to estimate the forecast accuracy of a forecast model. The steps are to use data from t = 1, …, T to estimate model parameters (such as a_{it} and b_{it}) and to make forecasts F_{T+k} for k = 1, …, K. The hold out sample in this case is y_{iT+k} for k = 1, …, K and it cannot be used in parameter estimation. Researchers set aside (or hold out) the end of a time series and forecast that data as if it were not available. Then with forecasts made, they compare the forecast and hold out sample data to calculate forecast errors.

- ***Forecast error*** is $e_{iT+k} = y_{iT+k} - F_{iT+k}$ where y_{iT+K} is data from a hold out sample or, for contemporaneous forecasts (made in real time), is simply actual values realized after the forecast is made. Given a sample of forecast errors, researchers create summaries using measures such as the mean squared forecast error or mean absolute forecast error.

Extrapolative methods

There are two main approaches to extrapolative models. The first, time trends as used in exponential smoothing, estimates a time trend model and seasonal adjustments. To forecast, one merely continues (extrapolates) the most recently-estimated trend line into the future and makes corresponding seasonal adjustments.

The second approach estimates correlations of the dependent variable with its past values as well as other correlations. For example, if there was a recent run of large crime counts in a time series, an autocorrelation model tends to keep the run going. The Box-Jenkins model uses this approach (Box et al., 2008). Box-Jenkins has some limitations for practice. The first is that the method tends to be complex with many parameters being estimated and with several steps to the procedure. Also, the preponderance of comparative studies in the literature, including the M-Competitions, have provided evidence that the simpler time trend models are just as accurate if not more so than Box-Jenkins models. Box-Jenkins tends to overfit noisy data (such as crime data at the patrol sector level), thereby leading to less accurate forecasts. Therefore, this chapter uses trend models.

With exponential smoothing as the approach, the question arises as to how to estimate seasonality. An extension of simple and Holt smoothing is the Holt-Winters forecast model that simultaneously estimates level, trend, and seasonal factors. An alternative, and the one taken in this chapter, is to:

A. Estimate seasonal factors from the raw time series data, as a preliminary step, using Classical Decomposition (Hyndman & Athanasopoulos, section 6-3, 2012);

B. Deseasonalize the raw data (in the case of multiplicative seasonal factors) by dividing each time series data point by its appropriate seasonal factor;

C. Estimate the simple or Holt exponential smoothing model using the deseasonalized data;

D. Make extrapolative forecasts; and

E. Reseasonalize the forecasts.

This approach provides alternatives for estimating seasonal factors. The obvious one to try is to estimate seasonal factors for each area or district. A problem with this approach, however, is that a season only occurs once a year. For example, if there are five years of weekly data and one wants to estimate the seasonal factor for week 21 (or any other week), then there are only five observations. This makes seasonality the least precisely estimated parameters in extrapolative models. Often, more accurate seasonality estimates can be obtained and therefore more accurate forecasts can be had by pooling data across areas. For example, one could estimate separate seasonal factors for residential versus commercial areas and then use the resulting estimates for each residential and commercial district's time series.

A better alternative is to estimate seasonality using all of the data of a jurisdiction. There is evidence that this is the best alternative for crime forecasting (Gorr, Olligschlaeger, & Thompson, 2003). Jurisdiction-level seasonal adjustments are smaller than those for the district-level, closer to overall mean levels, which, in turn, provides greater forecast accuracy. District-level seasonal factors are overly-influenced by individual data points of the small sample sizes of districts.

Simple Exponential Smoothing

This model estimates the time-varying mean of a univariate time series. Perhaps confusing is that there are two types of parameters in the model. First is the mean of the series, estimated by a_{it}. The second is the smoothing parameter, α, that determines how quickly or slowly the method adapts to changes in the mean of the series over time. Each historical data point in the time series is weighted in a sum to estimate a_{it}, with the weights declining exponentially with the age of the data from the forecast origin. The weights should approximately sum to 1.0. For small α, the weights drop off slowly with data age, retaining more of the history in the estimate and making the mean estimate change slowly. For larger α, the weights drop off quickly, retaining relatively little of the historical data and making the mean estimate change quickly. The closed form version of the estimated mean is as follows:

$$a_{iT} = \alpha y_{iT} + \alpha(1\text{-}\alpha)y_{iT-1} + \alpha(1-\alpha)^2 y_{iT-2} + \alpha(1-\alpha)^3 y_{iT-3} + \cdots \qquad (23.5)$$

where $0 < \alpha < 1$.

For example, for $\alpha = 0.05$, this expression is

$$a_{iT} = 0.05 y_{iT} + 0.0475 y_{iT-1} + 0.045125 y_{T-2} + 0.042869 y_{iT-3} + \cdots \qquad (23.5a)$$

but for $\alpha = 0.25$ it is

$$a_{iT} = 0.25 y_{iT} + 0.1875 y_{iT-1} + 0.140625 y_{T-2} + 0.105469 y_{iT-3} + \cdots \qquad (23.5b)$$

If T=60 (for example, five years of monthly data) the sum of weights for α=0.05 is 0.954 (and as T gets larger, the sum approaches 1) and for α=0.25, it is already 1.000. So a_{iT} is a weighted average of all the historical data.

Equation 23.5 is rarely used to calculate the current mean, a_{iT}. Instead an equivalent, recursive form is used:

$$a_{it} = \alpha y_{it} + (1-\alpha)a_{it-1} \text{ for } t = 1, \ldots, T. \qquad (23.6)$$

The only issue with this form is that it needs initial values for a_{i0}. Generally y_{i1} or the average of the first few observations is used for this purpose. After a brief "burn in" period, equation (23.6) forgets the initial value and tracks the mean of the series, so the choice of initial values is not critical. They just have to be reasonable.

Without a seasonal model component, the forecast for simple exponential smoothing is:

$$F_{iT+k} = a_{iT} \text{ for } k = 1, \ldots K. \qquad (23.7)$$

In other words, the forecast is a straight line, a constant, for any future time period.

For noisy data, such as crime counts in patrol districts, and short-term forecasts of one or a few steps ahead, it is hard to beat forecasts from (23.7). If, however, the forecasts go out to six months or a year ahead, then a time trend term can improve forecast accuracy, if there is a time trend in the data. Therefore a section below introduces the Holt exponential smoothing model which includes a time trend.

Selecting a value for the smoothing constant, α

Simple smoothing has two things going for it. First is that a_{iT} is a measure of central tendency, a mean, and it is almost always more accurate over the long-haul to forecast using means. Second is that a_{iT} is self-adaptive to changes in the mean, albeit with a time lag. The larger α is, the smaller the lag to estimating a time varying mean, but also larger is the chance that a_{iT} will overly respond to noise in the time series, depart from the underlying mean, and thus forecast poorly. So there is tradeoff, a balancing act in getting α just right.

The traditional criterion for computing parameters is to minimize the sum of squared fit errors, and that is what is done in *CrimeStat*. Unfortunately the corresponding functional form is highly nonlinear, so there is not a closed form solution or equation for computing the optimal α. Instead, it is easy enough to try a grid of α values (for example, 0.01, 0.02, …, 0.99 in *CrimeStat*), compute the sum of squared fit errors for each grid value, and choose α that has the minimum.

Figures 23.1 through 23.3 show the effects of different smoothing parameter values for estimating the time series mean from the sample data provided in *CrimeStat* for this chapter. The data are from census tract 20100, Pittsburgh, Pennsylvania (a high crime area) and are monthly time series for serious violent crimes (homicide, rape, robbery, and aggravated assault) between 1990 and 1999. Figure 23.1 shows smoothed values with near-optimal α = 0.15. The smoothed values track the mean of the series well. Jurisdiction-level seasonality, estimated from all of Pittsburgh's serious violent crimes (see the Classical Decomposition section), is included in the smoothed values, accounting for much of the month-to month variation.

Figures 23.2 and 23.3 are extreme cases for α values to make a point. Figure 23.2 has smoothing for α = 0.01, resulting in very slow adaptation and a very long memory for old data values. Values that are too low for *α* cannot keep up with the changes in this time series so that forecasts based on them would do poorly. In contrast, Figure 23.3 has smoothing for α = 0.99 so that there is practically no memory of historical values. The smoothed values are very close to being the raw data with no smoothing. These smoothed values are too noisy and would forecast poorly. So an optimal α value between these two extremes is needed for forecasting, 0.15 as shown in Figure 23.1.

Straw man forecasts

Clearly simple exponential smoothing is simple, but there are even simpler methods. However, to justify its use, simple exponential smoothing has to forecast more accurately than these simpler methods. Comparative research on forecast methods, as published in journal articles such as in the *International Journal of Forecasting* or the *Journal of Forecasting*, thus
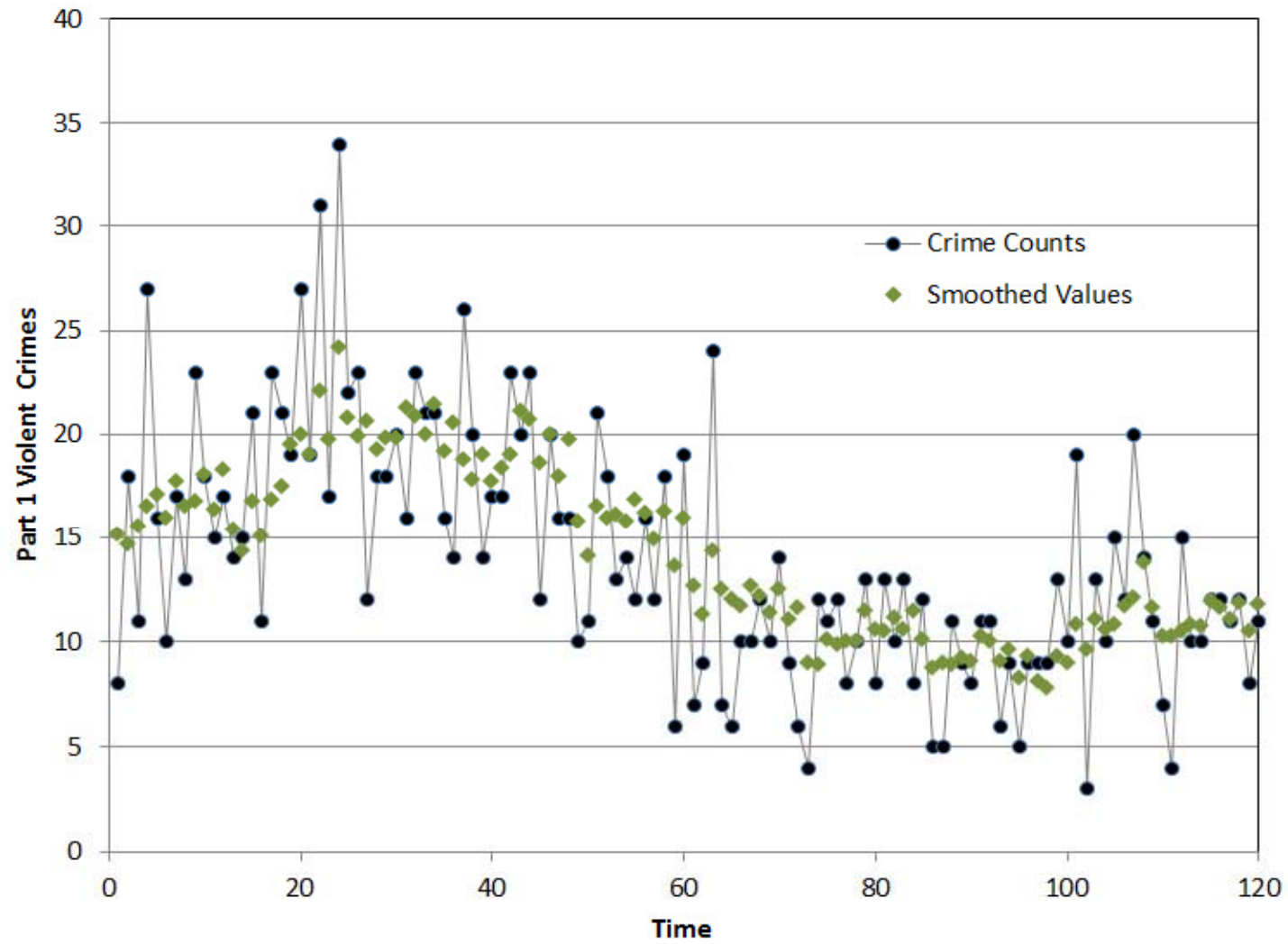
Figure 23.1:
Simple Smoothing with α=0.15
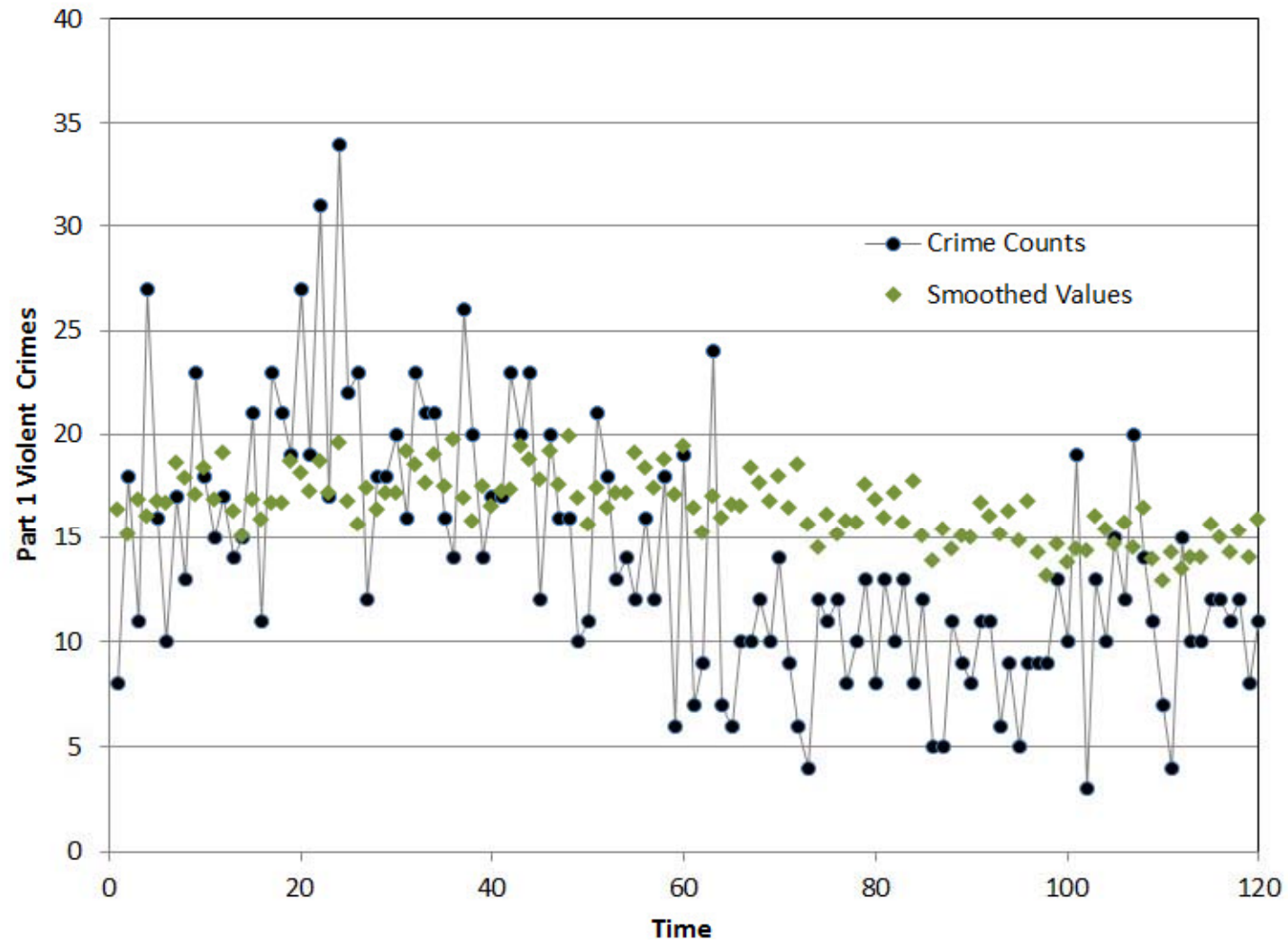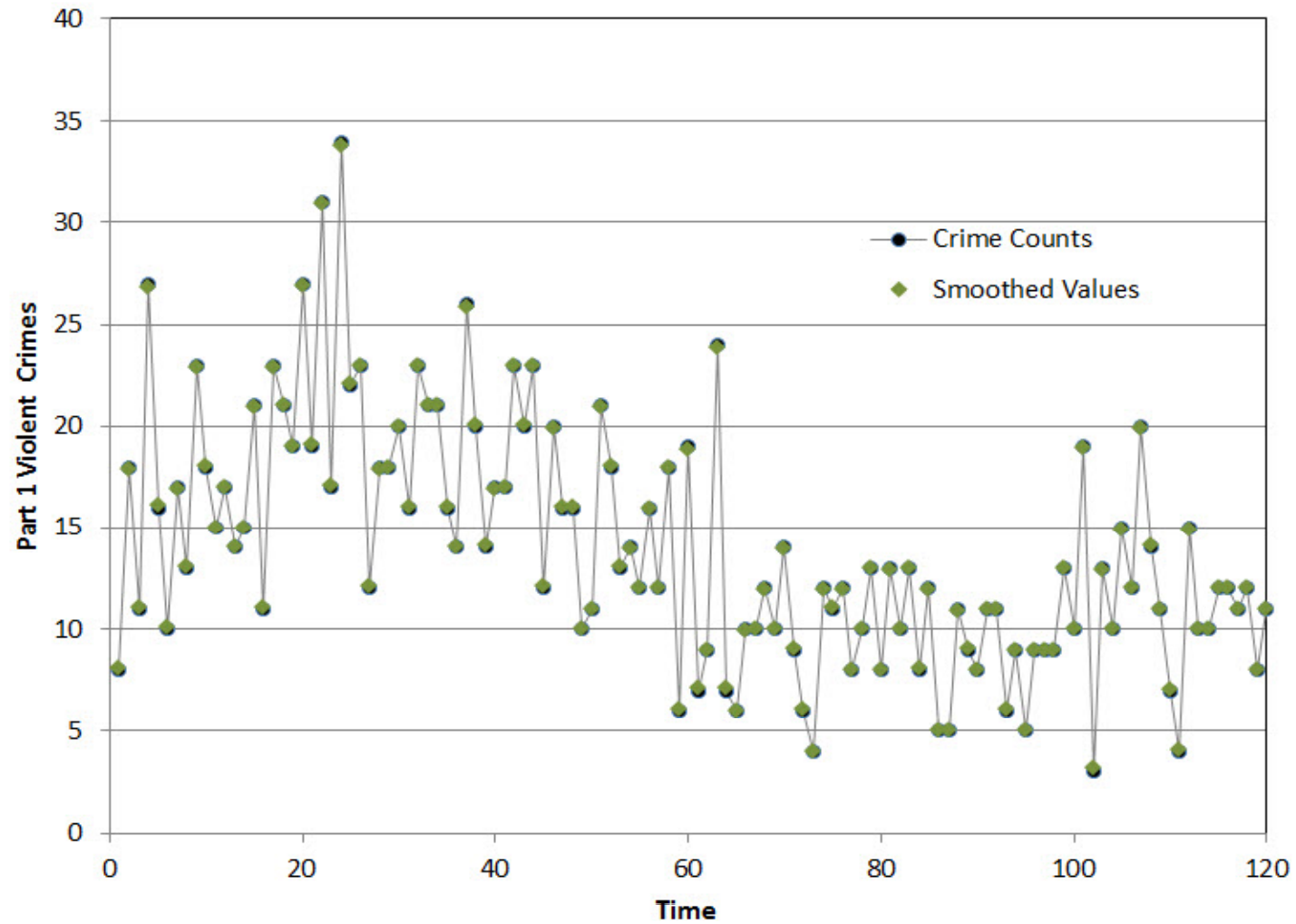
Figure 23.2:
Simple Smoothing with α=0.01

Figure 23.3:
Simple Smoothing with α=0.9

always include *straw man* methods. Hold-out sample forecast accuracy is compared on the same data between multiple forecast models/methods, including straw man methods.

The simplest method, for the case of no seasonality, is the naïve method forecast, using the last historical data point (the forecast origin) as the forecast:

$$F_{iT+k} = y_{iT} \text{ for all k.} \tag{23.8}$$

This method, also called the random walk forecast, is sometimes hard to beat, for example, in attempting to forecast stock market prices. At other times, when there is a time trend, smooth mean over time, or seasonality, the naïve method is easy to beat. Gorr, Olligschlaeger, and Thompson (2003) provided evidence that naïve forecasts and other data value-based forecasts (such as using the same month's data from the previous year) forecast very poorly for crime data.

A second straw man is the sample mean of the time series as the forecast:

$$F_{iT+k} = \frac{1}{T}\sum_{t=1}^{T} y_{it}. \tag{23.9}$$

Closest to simple exponential smoothing is the mean of a moving window of recent data, for example with the sum in (23.9) over t=T-w+1, …, T where *w* is the window length. This mean is also self-adaptive to changes in the mean of the time series, but has a larger time lag than exponential smoothing. The choice of a value for *w* is similar to that for α: too large a value makes the mean unresponsive and too small a value makes the mean unreliable.

Holt Exponential Smoothing

This model retains a_{it} and introduces a second model parameter, b_{it}, that is the coefficient of the time index of the time series as in (23.3). The Holt recursive equations for estimation are as follows:

$$a_{it} = \alpha y_{it} + (1-\alpha)(a_{it-1} + b_{it-1)}) \tag{23.10}$$

$$b_{it} = \beta(a_{it} - a_{it-1}) + (1-\beta)b_{it-1} \tag{23.11}$$

where a_{it} is the current level of the time series at time t, b_{it} is the time trend slope used in making forecasts, α is the smoothing parameter for the level with $0 < \alpha < 1$, β is the smoothing parameter for the trend with $0 < \beta < 1$, and the estimated model at the forecast origin, T, is

$$\bar{y}_{iT} = a_{iT}. \tag{23.12}$$

The forecast equation is:

$$F_{iT+k} = a_{iT} + b_{iT}k \text{ for all k.} \hspace{4cm} (23.13)$$

It is worth reiterating that equation (23.10) estimates the current mean or level of the time series while equation (23.11) estimates the trend, or change in the series for each step ahead forecast. This is in contrast to a linear regression model with the time index as the independent variable, $\bar{y}_{it} = b_0 + b_1 t$ where b_0 is the intercept term, the value of $\bar{y}_{it}$ at t is b_0. Parameter a_{it} in (28.10) is not an intercept term, but is the mean of the time series at time t.

The Holt model needs initial values, a_{i0} and b_{i0}. For the former, one can use the first observation or average of the first few observations as in simple exponential smoothing. For the latter, one can use the difference y_{i2}-y_{i1} or the average of the first few such differences. Again, as long as reasonable values are used, Holt will soon forget the initial values and be on track with the mean parameter values.

Holt parameter estimation also uses a grid search, but over the two-dimensional α and β space. For example, in *CrimeStat*, if there are 100 values to try for each smoothing parameter, then all possible pairs need to be tried with 10,000 pairs in total for the optimization. This is easily and quickly done by *CrimeStat* every time it forecasts using the Holt model.

Classical Decomposition: Seasonality

This section covers one of the earliest and most robust methods for estimating seasonality in a time series, Classical Decomposition. It is a separate procedure for estimating seasonality from raw time series data, so it can be applied just for the sake of understanding seasonal patterns that are part of a time series whether there is a time trend or not in the series. There are no smoothing or other parameters as part of the method, just straightforward calculations.

As stated earlier, CrimeStat uses the multiplicative form of seasonality, a dimensionless factor for each season that is valuable for cross-sectional data, such as in the case of crime space and time series data with its several or many zones, patrol districts, or census tracts of interest. The basic idea is to create an observation of seasonality for each data point in the series. With monthly data and the month of July for example, all of the July observations of seasonality are collected over the series and averaged to yield the July seasonality estimate.

The approach to creating a seasonality estimate for, say, July 2012 is to center an average of crimes on July that is one full year long with July in the middle. The average is an estimate of the mean of the series with seasonality removed, because the entire year is included in the average. Then the observation for multiplicative seasonality is the crime count actual value for

July 2012 divided by its centered average. The only problem with this procedure is that the number of seasons per year is usually an even number (4 for quarters, 12 for months, and 52 for weeks) so there is no natural center of a data window. Therefore, a simple adjustment is made, including an extra data point on each end of the window.

The Detection Problem

Exponential smoothing provides a relevant model and estimation method for time series that are predictable. As long as the data being smoothed do not change abruptly, exponential smoothing provides good forecasts. It is difficult to forecast abrupt changes, so often the best that can be done is to detect them as soon as possible after they have occurred or started to occur. That is the purpose of this section, to provide a method that works in partnership with exponential smoothing and extrapolative forecasts for detection of abrupt or large changes.

This section uses a world view that has two states: (a) *business-as-usual* which has time trends that can be accurately extrapolated and (b) *exceptions* which are departures from business-as-usual including outliers, step jumps, and turning points when a trend reverses direction. In crime space and time series data, a large crime increase is caused by some change in the underlying criminal element, for example, formation of a gang rivalry, sales by a new illegal gun dealer, parole of serial criminals who continue crime careers, and so forth. In some cases it can be due to the withdrawal of police resources, such as when a special enforcement program ends. Large decreases are also of interest and may be due to special police enforcement programs.

Time series detection methods merely draw attention to areas where there is evidence that an exception is occurring. It is up to the crime analyst to diagnose a detected area, to determine the nature of the problem if one is thought to exist and to recommend interventions.

Counterfactuals

To detect a change in a crime space and time series, we need a basis for comparison, that which would have happened had it been business-as-usual instead of an exception. This is called the "counterfactual forecast" and is provided by extrapolative forecasts. Suppose that one has data up to y_{iT}. Then one makes an extrapolative forecast, F_{iT} using data from $t = 1, \ldots, T-1$ and computes the counterfactual forecast error, $e_{iT} = y_{iT} - F_{iT}$. Similar to hypothesis testing, if e_{iT} (and other recent counterfactual forecast errors) is large enough, then there is evidence that there is an exception. If the change is more moderate, the tracking signal to be described next accumulates consistent counterfactual forecast errors (e.g. all positive) over several successive time periods to also provide evidence of an exception.

Tracking Signals

Detection methods calculate tracking signals, or test statistics. When the signal gets large there is evidence that there is exceptional behavior in a time series. This requires one to choose a threshold value for a "signal trip" a topic covered later in this section.

A simple tracking signal (and one of the oldest to be used) is the cumulative sum of errors:

$$CUSUM_{wiT} = \sum_{t=T-w+1}^{T} e_{it}.$$ (23.14)

where w is the window length of the summation. If there is a large error in one direction (say positive) or there is a series of medium-sized errors in one direction, then CUSUM may be large enough to signal an exception.

Standardized data is created by subtracting the mean from a sample of data and dividing by a measure of spread, such as the sample standard deviation. The advantage of working with standardized data is that if there are many samples to be examined, such as all patrol districts in a police jurisdiction, then one can use a single threshold value for a "signal trip" indicating exceptional behavior. However, with raw data for each area (or zone), one would have to choose a separate threshold for each area depending on the scale of the crime problem in each area.

The counterfactual forecast error, e_{it}, has an expected value of zero, so it already behaves like the numerator of standardized data. The k-period Brown tracking signal (Brown, 1959) divides CUSUM by an alternative measure of spread to the standard deviation, the simple smoothed mean of absolute counterfactual forecast errors (called the *mean absolute deviation*):

$$w - period\ Brown = \left| \frac{CUSUM_{wiT}}{MAD_{iT}} \right|$$ (23.15)

where

$$MAD_{iT} = \beta |e_{iT}| + (1 - \beta) MAD_{iT-1}$$ (23.16)

and 0<β<1 is a smoothing parameter

While β is a symbol used in 23.16 and also for Holt smoothing, they are two different parameters. One problem with w-period Brown is that after an exception has occurred, the signal often has to be manually reset to a low, *business-as-usual* value. Otherwise, the signal

continues to be large indicating exceptional behavior that may already have passed. Trigg (1964) thus proposed a modification to smooth the numerator as well as the denominator so that it is self-adaptive, resetting itself. Trigg used a common smoothing parameter for both the numerator and denominator while others, including McClain (1988), found evidence of better performance with separate smoothing parameters for the numerator and denominator. Now Trigg is calculated as:

$$Trigg_{iT} = \left| \frac{E_{iT}}{MAD_{iT}} \right| \tag{23.17}$$

where

$$E_{iT} = \alpha e_{iT} + (1 - \alpha)E_{iT-1} \tag{23.18}$$

and $0 < \alpha < 1$ is a smoothing parameter.

Note that while the smoothing parameter is denoted by α here, it is different than the parameters also called α for simple and Holt smoothing. While Trigg and Brown methods are similar in performance, Trigg is more convenient to use and so *CrimeStat* uses it.

Decision Rules

Tracking signals are implemented with decision rules such as the following:

If $Trigg_{iT} \geq L$ then issue a signal trip report (23.19)

Else do nothing (23.20)

where L is a threshold level to be chosen by the decision maker. While similar to decision rules used in statistics for hypothesis testing, there are important differences.

In the academic world of theory building and model testing, L is chosen to yield traditional type I error levels of 1 or 5 percent. A type I error occurs when the signal trips but in fact there is no exceptional behavior, in other words, a *false positive* occurs. These error levels are conservative so as not to claim to have evidence that a theory is true unless the evidence is strong—scientists are skeptical. A type I error rate (or *false positive rate*) of 5 percent means that 5 percent of the *negatives* (periods without exceptions) are falsely signaled as positives, which is a waste of resources if the decision maker takes action.

In management of events such as crimes, however, the false positive rate needs to be chosen to fit the situation. Larger false positive rates are often desired; for example, they are approximately in the range of 10 to 15 percent for cancer screening in the U.S. because society

values early detection of cancer (and therefore more successful treatments on the average) much more than the consequences of false positives (pain and wasted cost of biopsies that show no cancer present; see Banez et al. 2003, Elmore et al. 2002). In other words, a *false negative*, a positive that is missed by the decision rule with no signal trip, is much worse than a false positive in the cost to society. For example, with crime, an area that is experiencing a large crime increase but goes undetected is costly. It is better to accept a larger number of false positives in this case than to fail to detect an area which shows a real increase in crime (false negative). Crime analysts, when interviewed by Cohen et al. (2009), stated that in their judgment it is 10 times more important to avoid a false negative than a false positive when monitoring crime time series for exceptions.

A *true positive* is the case where there is an exception (disease or flare up in crimes) and the decision rule (23.19) signals it. The *true positive rate* is the percentage of all positives (cases where disease or crime flare up exists, for example) that are signaled by the decision rule—values in the range of 60 to 90 percent should be attainable for crime data. However, there is a trade-off. To increase the true positive rate, one must also increase the false positive rate. One increases both rates by making the threshold level, L, smaller. The optimal level, L, depends on three things. One makes L smaller if prevalence is relatively high (the fraction of all cases that are positives), the benefits of finding a positive are high, and the costs of a false positive is relatively low. Benefits and costs for police are likely similar to those of physicians screening for cancer because the benefits of preventing crimes is high and costs are incurred anyways but with efforts redirected to areas with potentially better results.

There is a formal decision framework for choosing L called *receiver operating characteristics* (ROC), a name that comes out of the signal processing field where signals are received with equipment. Cohen et al. (2009) provide an overview of this framework applied to crime data. While the concepts are good, it is impractical for police departments to carry out a formal optimization of choosing values for smoothing parameters for equations (23.16), (23.19) and L, except perhaps for the largest organizations. It is necessary for crime analysts to label points in a sample of time series that they believe to be positives (true large change points) and then independently run monitoring through the sample in simulation mode. Then all possible threshold level values and smoothing parameter values are assessed in a grid search and optimization model to provide optimal values.

Instead, this chapter recommends default values L=1.5, α=0.9, and β=0.15. We also suggest trying L=1.75 and L=2.0 for more conservative values, providing fewer signal trips. The values for α and β are optimal ones from Cohen et al. (2009) for the Pittsburgh monthly crime data, with α=0.9 being very reactive to the signal and β=0.15 more slowly updating the spread measure. Most of the value of the Trigg signal comes from keeping the measure of spread up to date so that L functions correctly (see Cohen et al., 2009).

Figures 23.4 through 23.7 illustrate Trigg time series monitoring for two Pittsburgh serious violent crime, monthly time series: for census tract 20100 as seen earlier in Figures 23.1 through 23.3, a high crime area, and for census tract 50900 a more moderate crime area. All figures use default values of α=0.9, and β=0.15 for the Trigg signal and 0.015 optimal smoothing parameter for simple smoothing with jurisdictional-level seasonality. Figure 23.4 uses the default value L=1.5 for tract 20100. One can see that Trigg with these values does a good job of signaling exceptional values, both high and low values. Figure 23.5 shows the effect of raising L to the more conservative 2.0. Only three months are signaled as exceptional. Perhaps this value of L is too conservative. Figure 23.6 uses L=1.5 for tract 50900. Again there are many signal trips, but they appear to provide good information. Finally Figure 23.7 substitutes the conservative L=2.0 which again appears too conservative in this case.

These cases suggest that once a signal is tripped police should provide extra resources to the area for three or more additional months, for example, because of the four instances in Figure 23.6 of repeated exceptions with zero to two months between exceptions.

Conclusions

This chapter has presented models and methods for time series forecasting of crime applied to all subareas of a police jurisdiction. The subareas can be as small as patrol districts or census areas such as census tracts. It is best to forecast aggregates of crime types, for example all serious violent crimes, all serious property crimes, all disorder crimes, etc. so as to increase the sample sizes for each time period and allow for reasonably accurate estimation. Time periods used in *CrimeStat* for time series include weeks and months. Day periods are too short to provide adequate sample sizes per period for accurate estimation of model parameters.

This chapter defined relevant terms, provided two univariate time series models (one for a time-varying mean and a second for a time-varying time trend that includes a slope/growth term for forecasted values), exponential smoothing methods for estimating model parameters, and extrapolative forecasts from the models. Exponential smoothing automatically learns and adjusts models for smooth changes in time trends and is perhaps the most-widely used time series method in industry for operations management.

Seasonality for week of the year (1–52) or month of the year (1–12) is estimated separately, using Classical Decomposition, a method provided by the U.S. Census Bureau. *CrimeStat* uses a multiplicative form, seasonal factors, that is valuable for cross-sectional data such sub areas of a police jurisdiction where the scale of crime varies considerably. The factors are dimensionless and represent percentage changes to time series trends due to season, such as a 25 percent increase in crime level in July.

Figure 23.4:
Trigg Signal Trips with Simple Smoothing
Jurisdiction Seasonality & L=1.5
(Pittsburgh Tract 20100)
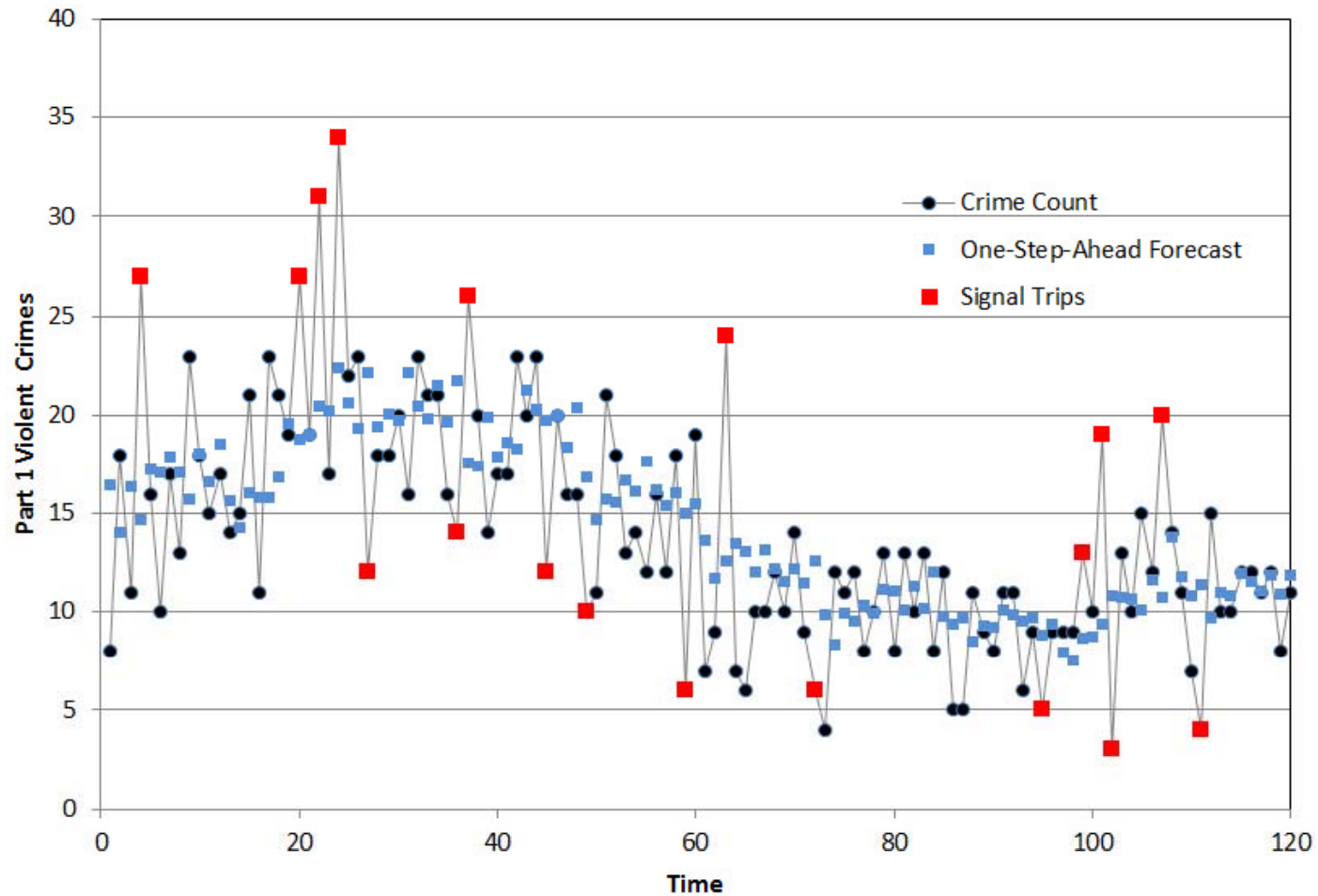
Figure 23.5:
Trigg Signal Trips with Simple Smoothing
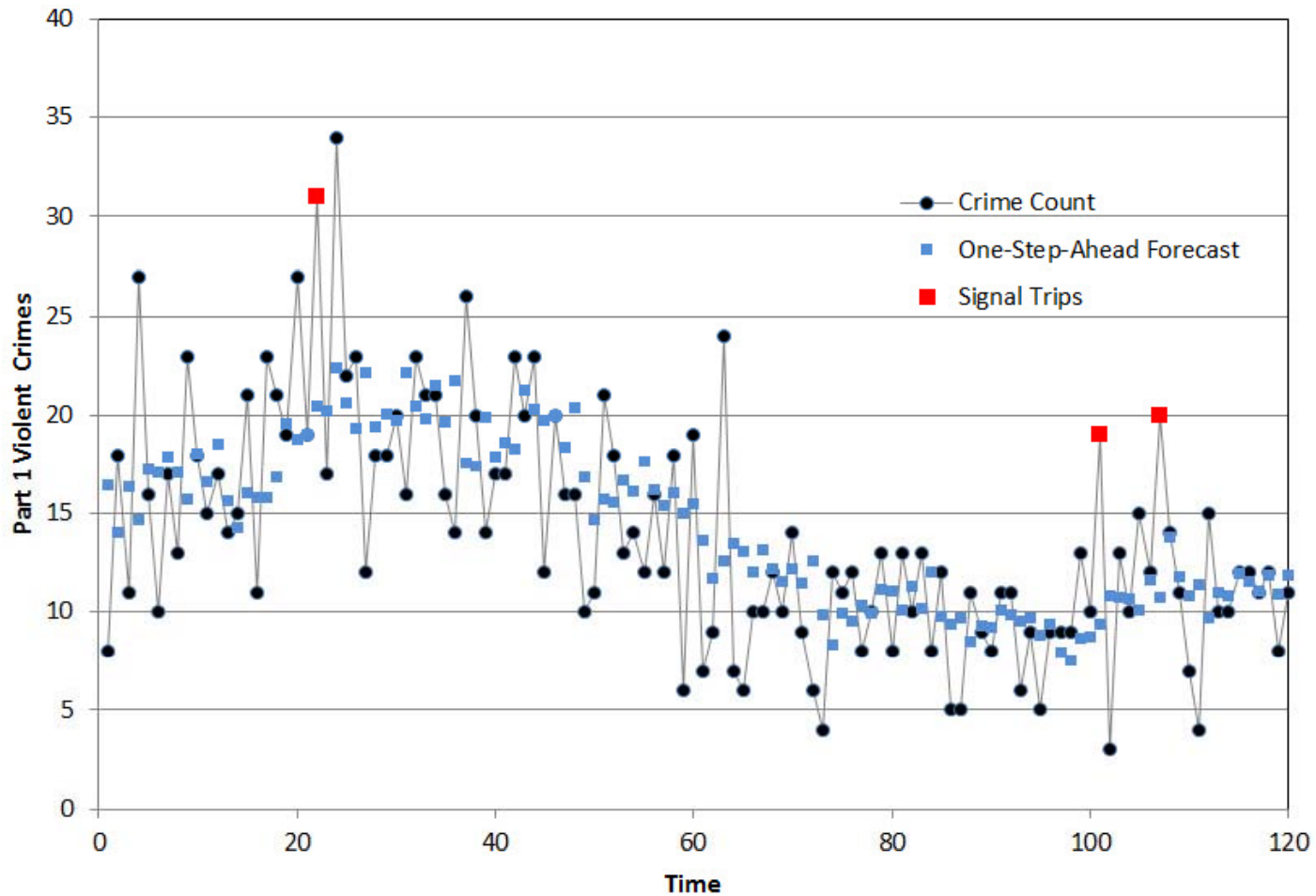## Jurisdiction Seasonality & L=2.0
(Pittsburgh Tract 20100)

Figure 23.6:
Trigg Signal Trips with Simple Smoothing
Jurisdiction Seasonality & L=1.5
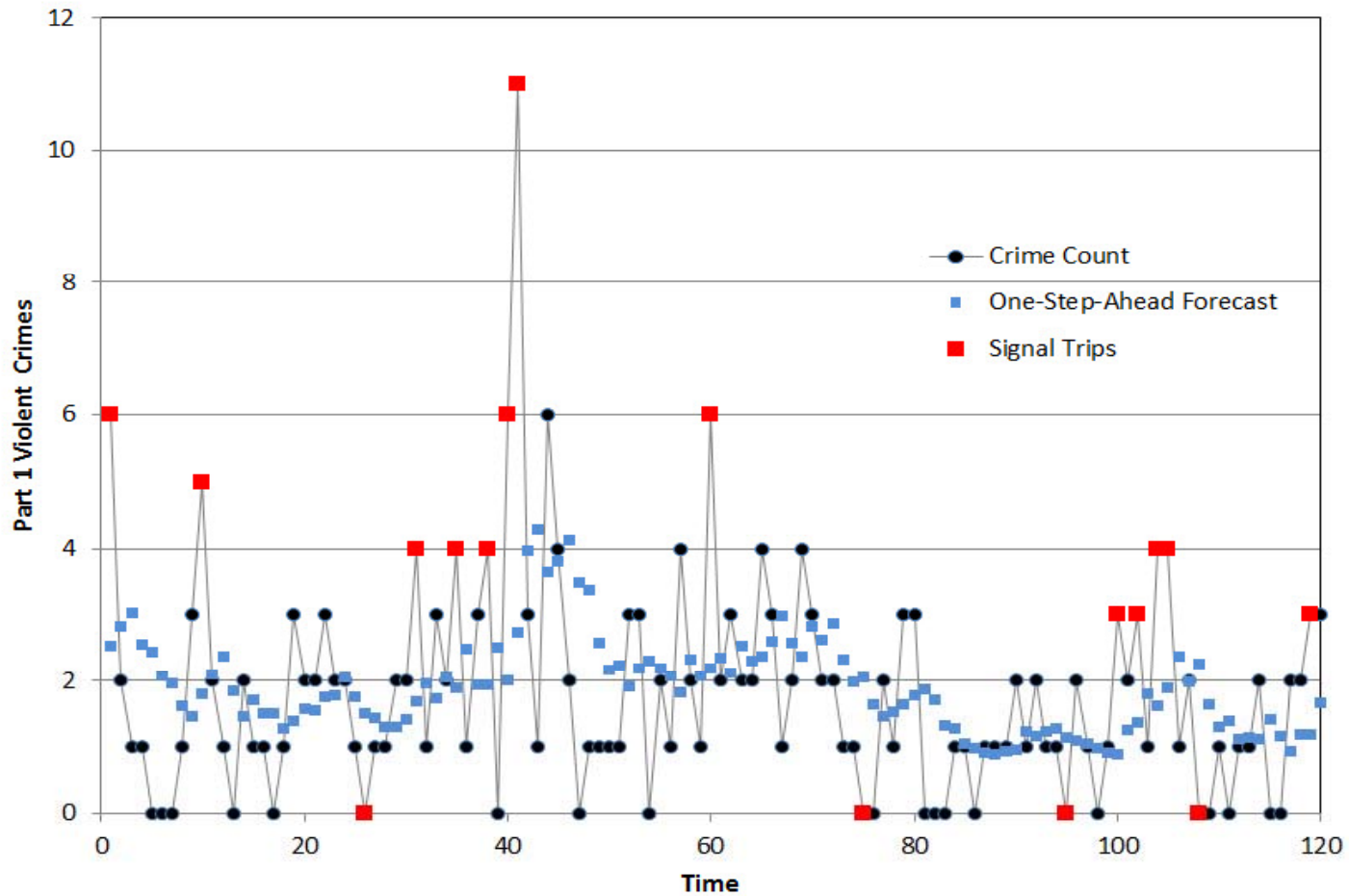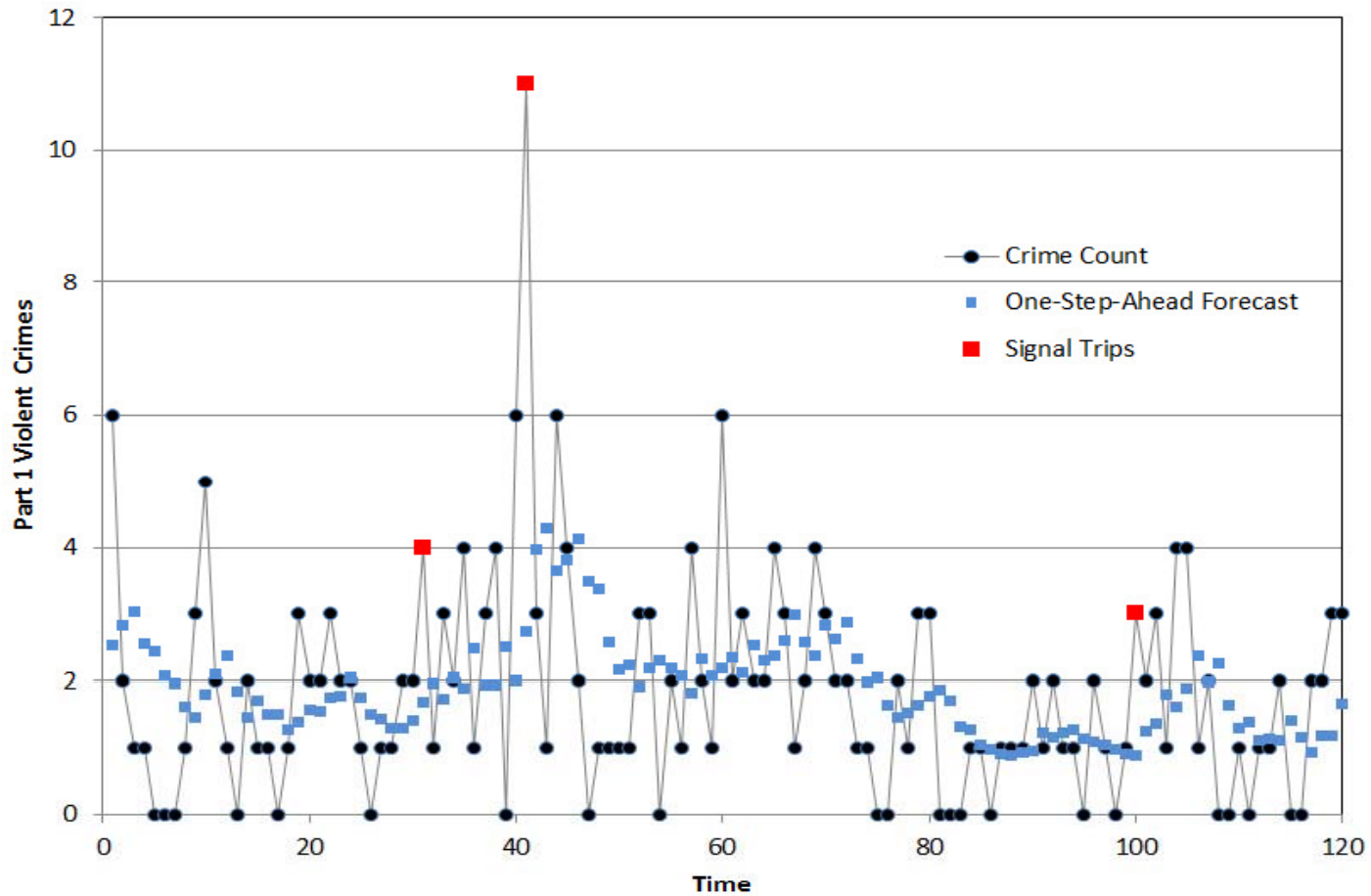### (Pittsburgh Tract 50900)

Figure 23.7:
Trigg Signal Trips with Simple Smoothing
Jurisdiction Seasonality & L=2.0
(Pittsburgh Tract 50900)

Forecasts are extrapolations of estimated time series. They are values from the trend line extended into the future with adjustments made for seasonal effects. *CrimeStat* automatically optimizes the fit of time trend parameters and batch forecasts time series for all subareas of interest, making it convenient and efficient for the crime analyst.

Perhaps more valuable than the forecasts themselves is the ability to use forecasting as a means for detecting large changes in crime time series. *CrimeStat* uses a signal tracking mechanism, the Trigg time series signal, to automatically detect large changes in crime time series such as crime flare ups. The objective is to detect large changes as soon as possible. The basis for comparison is the counterfactual forecast, an extrapolative forecast made for the last data point of a time series, which represents the crime level expected given "business-as-usual" or no change in time series pattern. In *CrimeStat*, every time series in a police jurisdiction gets an assessment for every historical data point including the very last, as to whether it appears to be ordinary or exceptional.

References

Banez, L., Prasanna, P., Sun, L., Ali, A., Zhiqiang, Z., & Adam, B. (2003). Diagnostic potential of serum proteomic patterns in prostate cancer, *The Journal of Urology*, 170, 442–446.

Box, E.P., Jenkins, G.M., & Reinsel, G.C. (2008). *Time series analysis: forecasting and control*, Wiley: Hoboken.

Brown, R. G. (1959). *Statistical forecasting for inventory control*, New York: McGraw-Hill.

Cohen, J., Garman, S. & Gorr, W. L. (2009). Empirical calibration of time series monitoring methods using receiver operating characteristic curves, *International Journal of Forecasting*, 2009, 25(3), 484–497.

Elmore, J. G., Miglioretti, D. M., Reisch, L. M., Barton, M. B., Kreuter, W., & Christiansen, C. L., (2002). Screening mammograms by community radiologists: Variability in false positive rates. Journal of the National Cancer Institute, 94, 1373–1380.

Gorr, W. L. & Kurland, K. S. (2012). *GIS Tutorial for Crime Analysis*, Esri Press, Redlands.

Gorr, W. L., Olligschlaeger, O. M. & Thompson, Y. (2003). Short-term forecasting of crime, *International Journal of Forecasting, Special Section on Crime Forecasting*, 19(4), 579–594.

Hibon M. & Makridakis S. (2000). The M3 Competition: results, conclusions and implications, *International Journal of Forecasting*, 16, 451-476.

Hyndman, R. J. & Athanasopoulos, G. (2012). *Forecasting: principles and practice: An online textbook,* http://otexts.com/fpp/.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E. & Winkler, R. (1982), The accuracy of extrapolation (time series) methods: Results of a forecasting competition, *Journal of Forecasting*, 1, 111-153.

McClain, J. O. (1988). Dominant time series monitoring methods, *International Journal of Forecasting*, 4, 563–572.

Trigg, D. W. (1964). Monitoring a forecasting system, *Operational Research Quarterly*, 15, 271–274.

References (continued)

Weisburd, D. L., Bushway, S., Lum, C., & Yang, S. (2004). Trajectories of crime at places: a longitudinal study of street segments in the city of Seattle. Criminology 42:283−321.

Weisburd, D.L. & Green, L. (1995). Policing drug hot spots: The Jersey City drug market analysis experiment, *Justice Quarterly* 12-711−736.