

# **Forecast Accuracy Measures for Exception Reporting Using Receiver Operating Characteristic Curves**

By

Wilpen L. Gorr

July 20, 2008

H. John Heinz, III School of Public Policy and Management

Carnegie Mellon University

## **Abstract**

This paper identifies forecasts of exceptions in product or service demand (i.e., large changes or extreme values) as a special need in forecasting, requiring new forecast accuracy measures based on the tails of sampled forecast error distributions. For this purpose, the paper introduces application of the receiver operating characteristic (ROC) framework, which has been used to assess exceptional behavior or cases in many fields. The “exception principle” of management reporting provides the corresponding forecast requirements. Seasonality estimates in univariate forecast models and leading independent variables in multivariate forecast models are among the approaches to forecasting exceptions. In a case study on serious violent crime in Pittsburgh, Pennsylvania across small sub-areas of the city, the simplest, non-naïve univariate forecast method is best for forecasting ordinary conditions, as found in previous research using conventional forecast accuracy measures, but the most complex multivariate model is best for forecasting exceptional conditions, using ROC forecast accuracy measures.

Keywords: forecast accuracy measures, exception reporting, ROC curves, crime forecasting

## 1. Introduction

Management by exception (MBE) is based on the “exception principle” of management reporting, due to father of scientific management, Frederick W. Taylor (1911). According to MBE, organizations should be designed so that operational staff members make resource allocation decisions under ordinary or routine conditions, but refer decisions to upper-level managers under exceptional conditions. Such a design relieves higher-level managers of routine work and makes best use of their limited time for addressing the difficult cases and the broader lines of strategy or policy.

In regard to product or service demand, ordinary conditions correspond to sufficiently small changes in demand time series from period-to period; whereas, exceptional conditions are the complement with relatively-high changes or extreme demand. So, for example, ordinary conditions might correspond to foreseeable demand, such as can be extrapolated using time trend and seasonality. However, a product that has very large seasonal demand might be considered exceptional, such as a month of May peak for lawn fertilizer in the northeastern U.S. Another kind of exceptional demand (for police crime prevention) is neighborhood-level crime flare ups, the subject of this paper’s case study. A multivariate model with leading independent variables forecasts future large increases in serious violent crime when the leading variables undergo step increases.

I define *reactive* MBE as that which deals with detection of exceptions that already have occurred, while *proactive* MBE anticipates exceptions with forecasts. While forecast errors necessarily are much larger than detection errors, the payoffs from proactive management are also much higher than from reactive management. Managers have a chance to prevent large losses or take advantage of major opportunities.

The inability to extrapolate exceptional conditions accurately is one basis for triggering reactive MBE using demand time series data. Time series monitoring methods that use this approach (e.g., Brown, 1959, 1963 and Trigg, 1966) have the objective of identifying exceptional conditions as rapidly as possible after they occur. These methods use decision rules comparing a stochastic decision variable—smoothed and scaled one-step-ahead extrapolative forecast errors computed for historical data—with a pre-determined control limit. When the decision variable exceeds the control limit, an exception report is issued for analysis and possible action by upper-level management.

Decision rules for triggering proactive MBE take a somewhat different approach, because only demand forecasts are available for future time periods, and not corresponding actual demand values and forecast errors. In this case, the manager must specify decision rules with control limits for forecasted demand, as in the case study of this paper.

The question naturally arises as to how to best evaluate forecast models or methods for exceptional versus ordinary conditions. I claim that central tendency (means or medians) of functions of sampled forecast errors (e.g., squares or absolute values) is best for ordinary conditions but, in contrast, we must examine the tails of sampled forecast error distributions for exceptional conditions.

The best approach for the latter is through receiver operating characteristic (ROC) curves. The ROC framework assesses the accuracy of binary-outcome, screening systems and allows for selection of control limits for use with stochastic decision variables to balance the tradeoff between true and false positive rates inherent in the decision problem (e.g., Swets, 1988). This paper thus provides new forecast accuracy measures for proactive MBE using ROC

curves, whereas another paper (Cohen, Garman, & Gorr, 2008) provides similar measures for reactive MBE.

Section 2 reviews the literatures on forecast accuracy measures, MBE, and ROC curves. Section 3 adapts ROC curves to assessing forecast accuracy for exceptional conditions. Section 4 presents the crime case study of Pittsburgh, Pennsylvania; forecast models and methods; and experimental design. Section 5 presents the results of the case study. Finally Section 6 provides a summary and discussion.

## **2. Literature**

First in this section is a brief review of conventional forecast error measures. With some terminology and concepts thus in place, I turn briefly to make the case that conventional forecast accuracy measures are best for ordinary product demand conditions, while new forecast accuracy measures are needed for exceptional conditions. The review then moves on to MBE and the role of population screening for triggering upper-management action. Included here is the current police practice called “Compstat” that use MBE. Finally, the section concludes with a review of the ROC framework and its several components and their applications. The ROC review is extracted from Cohen, Garman, & Gorr, (2008) for the convenience of the reader and extended for the purpose of this paper.

### *2.1 Central tendency forecast error measures*

The commonly-used forecast accuracy measures are central tendency estimates of functions of forecast errors, such as the Mean Absolute Deviation (MAD), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and Median Symmetric Absolute

Percentage Error (madsAPE). These measures, and several additional variations, have various properties and limitations (Armstrong & Collopy, 1992; Fildes, 1992, and Hyndman & Koehler, 2006). For example, the RMSE has too large a variance owing to its heavy weighting of extreme forecast errors, making it unreliable. Also, it is inappropriate for comparisons of forecasts for cross-sectional or multiple time series because it has the scale of the dependent variable of each time series (Armstrong & Collopy, 1992). The MAPE, while scaleless and therefore useful for ease of interpretation with multiple time series, does not treat small errors symmetrically with large errors, is overly sensitive to very small actual values, and is undefined for the actual time series value of zero (Koehler, 2001).

More recently, researchers have proposed attractive, additional central tendency measures. Hyndman & Koehler (2006) proposed the Mean Absolute Scaled Error (MASE), which divides MAD by the mean of in-sample first differences of the time series. When this measure has values of 1 or greater for one-step-ahead forecasts, the forecast has the same or worse performance than the naive method and should not be used. Interpretation, however, for multiple-step-ahead forecasts is not as straightforward. Others scale the MAD by the historic mean of the time series leading to a measure comparable to the MAPE (substituting for the actual value in the denominator with its expectation), but that treats forecast errors symmetrically and without the problem due to low or zero actual values (Kolassa & Schutz, 2007; Valentin, 2007). This measure, denoted here as MADS, is easy to interpret, reporting the mean absolute deviation in units of time series means. It is a good option for assessing forecast accuracy in cross-sections of micro-scale time series with many zero-value (intermittent) observations, and is used in this paper.

Yet another approach to central tendency forecast accuracy is to estimate the average costs of forecast errors for inventory control and other highly-structured decision problems. Examples are Granger & Pesaran (2000) and Catt (2007). While it is difficult to estimate some of inventory costs (e.g., loss of customer good will), it may be even more difficult to estimate costs and benefits of demand forecasts for societal problems, such as law enforcement.

## *2.2 The case for new forecast accuracy measures for exceptional conditions*

Before proceeding further with the literature review, I want to make the case that new accuracy measures are needed for exceptions forecasting and that conventional forecast accuracy measures are best for ordinary conditions.

While product demand forecast errors for exceptions are reflected in central tendency measures, those values likely are overwhelmed by the relatively large number of ordinary values. Indeed, we are upset when exceptional values have too large an impact on conventional error measures, such as with the RMSE. Hence, conventional measures do not distinguish well between alternative exceptions forecast models. Furthermore, the underlying decision regarding exceptional product demand values is binary: either there is an indication that future demand will be exceptional or ordinary. Binary information is sufficient to proceed with proactive MBE, which is beneficial because it is easier to attain acceptable accuracy for binary outcomes (interval forecasts in this case) than for the continuous error measures of the same variables. So, it seems clear that we need new measures for exceptions forecasting, ones that focus on the tails of forecast error distributions.

I have already made the point that central tendency measures are dominated by ordinary values, making them relevant for this case. Under ordinary conditions, simple patterns

are discernable, such as time trend and seasonality, and corresponding forecast models incorporate these patterns in an attempt to match actual values for use in short-term planning for production and distribution. Hence forecast error is the most appropriate underlying measure for ordinary conditions. Finally, central tendency is the single best measure for comparing forecast errors from alternative ordinary-conditions models. So, it seems safe to argue that central tendency forecast accuracy measures are best for ordinary demand conditions.

### 2.3 Management by exception

Frederick W. Taylor, the father of scientific management, is credited with the "exception principle" of reporting:

... the manager should receive only condensed, summarized, and invariably comparative reports, covering, however, all of the elements entering into the management, and even these summaries should all be carefully gone over by an assistant before they reach the manager, and *have all of the exceptions to the past averages or to the standards pointed out, both the especially good and especially bad exceptions*, thus giving him in a few minutes a full view of progress which is being made, or the reverse, and leaving him free to consider the broader lines of policy.... [emphasis added] (Taylor, 1911).

MBE is pervasive today, with exception reports a common component of management information systems (e.g., Ackoff, R.L, 1967; Simons, 1991; Wetherbe, 1991).

The underlying approach of MBE is to screen populations of items using an inexpensive diagnostic test or method to identify potentially important cases for further diagnosis, analysis, and possible intervention. Screening is common today in the design of

decision support systems wherein the user first makes a wide-area scan to identify potential problems and then drills down to detailed data for diagnosis and decision making (e.g., Turban et al., 2004). The medical profession uses low-cost tests to screen populations of persons with the objective of early disease detection, thereby enabling more successful and lower cost treatments. Prostate and breast cancer screening are good examples (Banez et al. 2002, Elmore et al., 2003). ROC analysis, to be discussed next, is widely used for medical screening and many other fields (e.g., Swets et al., 2000).

#### *2.4 ROC framework*

Peterson, Birdsall, & Fox (1954) conceived the theory of detectability as the task of discriminating between “signal plus noise” from “noise alone” (Swets, 1986). Signal is any condition of an entity which we wish to detect using a screening test, classifier, or other method. Most often signal is an exceptional state, such as the presence of a disease. Any attempt to detect signal in a population may result in errors because noise can appear by chance to be signal.

“Positive” refers to actual presence of signal; whereas, “negative” is its absence. For screening, a positive test generates an exception report (signal trip) while a negative test indicates ordinary conditions. Thus there are two states of the world and two test results, leading to the common contingency table as seen in Table 1 (Swetts, 1988). See this table for notation and further definitions.

Three statistics from the columns panel of Table 1 provide information traditionally used in the literature by researchers to compare the accuracy of alternative tests: True Positive Rate,  $TPR = TP/(TP+FN)$ ; False Positive Rate (or Type I error in hypothesis testing),  $FPR = FP/(FP+TN)$ ; and Prevalence of positives,  $P = (TP+FN)/n$ , if the sampled values are randomly

drawn from the population. In addition, I include statistics in the rows panel that are relevant for managers. Of particular importance is effort rate,  $ER = (TP+FN)/n$ , which gives the fraction of all cross-sectional time series points that are under exception report status and hence relates directly to the cost of exception reporting.

To construct a ROC curve, based on the columns of Table 1, one must have a sample of cases for which the true outcomes are known. For example, a common “gold standard” or determination of positives for screening via medical imaging is analysis of tissue from biopsy or autopsy. For demand forecasting the gold standard is easy to obtain *ex post* after the actual demands that were forecasted are experienced, as explained in Section 3. ROC curves for screening tests that generate a continuous stochastic decision variable are generated by varying the control limit from the minimum through the maximum of the observed range of the decision variable. The corresponding plot of TPR versus FPR is the ROC curve. It passes through points (0,0) and (1,1).

The sample ROC curves seen in Fig. 1 are from one-month-ahead, serious-violent-crime forecasts made from a rolling horizon forecast experiment in Pittsburgh for each of the 172 census tracts comprising the city (see Sections 4 and 5 for more details). Shown are curves for two forecast methods, single exponential smoothing with deseasonalization via multiplicative factors estimated via classical decomposition using city-level time series data (Gorr, Olligschlaeger, & Thomson; 2003) and a distributed-lag causal model that uses lags of leading, lesser crimes to forecast serious violent crimes (Cohen, Gorr, & Olligschlaeger; 2007). More on these curves is below. Note that empirical ROC curves such as in Fig. 1 (as opposed to fitted curves) have step jumps because of finite sample sizes.

A line of slope 1 and no intercept in a ROC chart provides the benchmark of a chance classifier that assigns cases to the positive test state randomly. For example, suppose that one used a spinner as the screening device that has a pie-shaped area shaded, defined by a segment of its perimeter and two radii, that is 25 percent of the spinner's area. If this area represents a positive test outcome, then application of the spinner to a sample would result in 25 percent of the test positives being true positives (TP) and 25 percent of the test negatives being false positives (FP) in Table 1. The line of slope 1 is generated by varying the percentage of the spinner's shaded area between 0 and 100 percent.

If a ROC curve point lies above the benchmark line, it provides more accurate classifications than chance alone. (If a point lies below the line, such as is the case for a portion of one of the ROC curves in Fig. 1, the test can be calibrated by reversing decisions and then is better than chance.) A ROC curve, if better than chance, has positive but decreasing slope as FPR increases. The closer the curve to the northwest corner of the ROC chart (or the greater the area under the ROC curve), the more accurate the classifier. Any point to the south and east of another point is dominated by the former. The ROC curves in Fig. 1 are better than chance (including single smoothing if its curve were calibrated by reversing decisions for  $FPR \geq 0.60$ ), but not highly accurate for this challenging application. Section 5 argues nevertheless that these forecasts are useful for police management.

The ROC curve makes it clear that there is a tradeoff on selection of a point from the curve for implementation of a test: The higher TPR (desirable outcome), the higher FPR (undesirable outcome). Traditional values for FPR (Type I error) used in theory building and hypothesis testing are 0.01 and 0.05; however, for decision making, higher values can be desired if the exceptional condition is important enough and resources are available to handle the volume

of exceptional cases for follow-up and costly diagnosis. For example, in the U.S., false positive rates for breast and prostate cancer screening are as high as 0.10 to 0.15 (e.g., Banez et al. 2002, Elmore et al., 2003).

A simple assessment of benefits enables the analyst to find the optimum tradeoff using a ROC curve (Metz, 1978; DeNeef & Kent, 1993; Cohen, Garman, & Gorr, 2008). Domain experts merely have to estimate the ratio of how many more times is it valuable to avoid a false negative than a false positive and to have an unbiased assessment of positive prevalence. Then a simple graphical analysis finding the point where the corresponding optimality criterion line is tangent to the ROC curve identifies the optimum FPR and TPR pair. Finally, a reverse function determines the control limit for implementation. For the case of relatively flat ROC curves such as in Fig. 1, however, applying such an optimality criterion is overly sensitive with small changes in the underlying assessments leading to large ranges of FPR values possibly designated as optimal. Instead for such a case, I suggest that managers make judgments on the level of effort to expend on exception reporting, based on plots including the effort rate measure of Table 1. See example plots in Section 5.

An important note on product-demand forecasting and the ROC framework is the following. Conventional forecast error measures depend on the forecast accurately estimating the dependent variable value, so that forecast errors of zero are the most desirable values. This is not necessary for the ROC framework. The ROC stochastic decision variable can be in any units or scale relative to product demand and it is simply the variation in the decision variable that matters as captured through a good or optimal selection of a control limit. So, a stochastic variable that is a biased estimator of extreme product demand can nevertheless be the best for exceptional forecasts as long as it attains its extremes at the same time that product demand does.

### 3. Decision rules for forecasting of exceptional demand conditions

Suppose that an organization forecasts time series,  $y_{it}$ , for  $i = 1, \dots, I$  products or services. Historical time series data—including possibly independent-variable time series for multivariate models—for  $t=1, \dots, T$  are used to make forecasts  $F_{iT+m}$  where  $m = 1, \dots, M$  steps-ahead are forecasted. The application in Sections 4 and 5 uses only  $M=1$ , so the development below includes notation only for this case, but is easily extended by specifying an additional rule for each  $m>1$ . Furthermore, I specify a decision rule only for exceptionally high values, the side of primary interest to police, but a rule for exceptionally low values follows the same form.

While several forms of decision rules are possible, this paper uses one with standardized time series data and forecasts so as to treat all time series equitably and to facilitate setting gold-standard cutoffs and control limits across time series of varying scales (I discuss this point further below). Suppose then that  $y'_{it}$ ,  $t=1, \dots, T+1$  is the *ex post* standardized series for products  $i = 1, \dots, I$ . The decision maker specifies the *gold standard cutoff* as a Z-value,  $Z_\alpha$ , so that the fraction  $\alpha$  of actual demand values in the corresponding upper tail of each  $y_{it}$  distribution defines the positives for ROC analysis and the exceptional conditions for upper-level management decision making.

Suppose that rolling forecasts,  $F_{it}$ , are available for  $t = t_1, \dots, T$  where  $t = 1, \dots, t_1-1$  served as the estimation data for the earliest forecasts for time series  $i = 1, \dots, I$ . Finally suppose that  $F'_{it}$  are corresponding *ex ante* standardized values. The decision maker must specify a control limit,  $L$ , then the decision rule for triggering exception reports is:

$$\text{If } F'_{iT+1} \geq L \text{ then issue an exception report for product } i=1, \dots, I. \quad (3)$$

The *ex ante* standardized forecasts in rule (3) use data for  $t \leq T$  for standardization; whereas, *ex post* standardized demand  $y'_{iT+1}$  uses data through  $T+1$ .

While it is possible (and for some applications desirable) to specify different gold standard cutoffs and decision rule limits for different subsets of products, the approach above makes it easy to specify the same treatment for each product time series, regardless of scale. This has merit for the crime case study of Section 4 by treating both low and high crime areas the same. For example, if raw data and forecasts were used with a constant gold cutoff and control limit, then mostly large-volume time series would have exceptions and thereby draw management attention mostly to those areas at the expense of lower-volume time series.

Figs. 2 and 3 illustrate the application of the gold standard cutoff and decision rule of this section (see Section 5.2 for additional details). First, Fig. 2 is the monthly time series plot of standardized Part 1 Violent crimes for census tract 1114 in Pittsburgh during the period for which forecasts were made. Shown is the gold standard cutoff of 1.81 that defines an average of 6 high-value exceptions per month for the 172 census tracts. For this time series there are thus 4 positives, the points above the gold standard cutoff line. For 0.2 FPR using the robust leading indicator model of this paper (see Section 4.3), the corresponding control limit is 0.42. Any points on or above the Forecast line generated by the robust model (i.e., the stochastic decision variable) generate exception reports (shown with shading). Indeed, there tend to be exception report *episodes* of two or more consecutive months in which the leading indicators remain “stepped up”. While there are many false positives, there are also three true positives and only one false negative in this case.

Fig. 3 is a corresponding map of census tract 1114 showing streets and the locations of crimes with size-graduated point markers (some point locations had two crimes, each at different times, and are shown with the larger of two point markers). Shown are the point locations of individual Part 1 Violent crimes for December 1995—the first positive in Fig. 2. Also shown are

the leading indicator crime locations from the previous month (the most important lag), November 1995. While hypothetical, the two circles drawn on the map are nevertheless potential hot-spot areas, with spatial clusters of leading indicators, that could have been identified in November 1995 to implement preventative measures and thereby possibly preventing the corresponding 5 out of 7 Part 1 Violent crimes in December 1995 that are within the hot spots. Moreover, police might have included the area with the two additional Part 1 Violent crimes in the northeastern corner of census tract 1114 because it is on the same street as the other two hot spots, which is the main street in the neighborhood.

#### **4. Case study**

This section reviews a management approach widely used by municipal police departments and then presents the case of the Pittsburgh, Pennsylvania Bureau of Police. Included are descriptions and rationales for forecast model/method selection, including two univariate and two multivariate models, as well as other aspects of the forecast experiment design.

##### *4.1 Police Decision Making*

Major municipal police departments in the U.S. (and around the world) use the Compstat approach to decentralized management, developed by the New York City Police Department (e.g., Henry, 2003). Precinct commanders have autonomy in decision making, but are accountable to the police chief and mayor through monthly review of performance measures in open meetings held by the police chief's top-level staff. Included in the meetings is planning for the next month's deployment of resources. The process identifies problems and sets

priorities, but leaves detailed diagnosis and decision making to the precinct commanders and their staff. Currently, police use reactive MBE to respond to crime increases in small geographic areas (see Cohen, Garman, & Gorr, 2008), but given adequate forecasts would also be candidates for proactive MBE for prevention of crime flare ups.

Monthly crime counts by geographic area are among Compstat performance measures. Most important are so-called Part 1 Violent crimes (homicide, rape, aggravated assault, and robbery). Generally speaking, the smaller the geographic unit of analysis, the better for police work by more precisely directing limited police resources. The geographic areas for police work in Pittsburgh are (1) 6 precincts – large areas of a city each with a police station and commander; (2) 42 patrol districts – the territories assigned to individual patrol units within precincts; and (3) 172 census tracts – homogeneous neighborhoods of approximately 4,000 population or less that have population data collected and tabulated for the decennial census. In Pittsburgh, patrol districts are made up of one or more census tracts. The case study in this paper has forecasts for all three geographies, but only patrol districts and tracts are relevant for MBE because of the need to target police interventions to relatively small areas.

In summary, the planning and review process followed by police departments, Compstat, leads to forecast requirements: one-step-ahead forecasts of monthly counts of Part 1 Violent crimes in census tracts or patrol districts. Especially important are forecasted large increases in these time series so that top management can allocate resources for prevention.

#### *4.2 Univariate forecast methods*

The forecast models or methods used for proactive MBE must have the capacity to forecast large changes in product or service volume. For univariate, one-month-ahead crime

forecasts of serious violent crimes this capacity is obtainable from seasonal variations which can be large. One-month-ahead changes in time trend, however, are too small to produce exceptionally-large forecasts.

Gorr, Olligschlaeger, & Thomson (2003), using the conventional forecast error measure, MAPE, found that seasonality estimated from city-level time series data to yield significantly more accurate forecasts than seasonality estimated individually for each district. This is a common finding (e.g., Withycombe, 1989; Bunn, & Vassilopoulos, 1993), that group estimates of seasonality applied to individual products are more accurate for forecasting than individual product-level seasonality. Group seasonality, while less variable than product seasonality, is more reliable and in balance reliability is needed more than high variability for accurate product forecasts. Perhaps, though, the situation is reversed for exceptions forecasting. Hence I retest group versus individual seasonality in this paper. Tuning smoothing methods by varying smoothing constants has little effect on forecast performance for exceptional conditions.

This paper thus includes single (smoothed level) and Holt (smoothed level and time trend) exponential smoothing methods with smoothing factors optimized for each forecast over in-sample data, minimizing the MSE for one-step-ahead forecasts. Estimates included monthly seasonality via multiplicative classical decomposition with both city-wide group and separate factors for each district. The group seasonal estimates are used to deseasonalize time series data for estimation by single exponential smoothing and then to reseasonalize forecasts — corresponding to the most accurate forecasts from past work. The district seasonal estimates are used similarly with Holt exponential smoothing. This combination gives the maximum capacity for large-change forecasts, by including individual zone seasonality estimates and time trend.

### *4.3 Distributed lag model*

Perhaps more valuable than seasonal-based large changes for management are those obtained from leading indicators, if available. With limited expertise and relative ease it is possible to foresee large seasonal changes (e.g., for crime in Pittsburgh, there are seasonal peaks in summer and prior to the holiday season). In contrast, exceptional crime increases due to gang rivalries, illicit drug dealing increases, influx of illicit handgun suppliers, etc. are much more difficult to detect. Resulting later increases in serious violent crime can come as total surprises.

Fortunately, environmental criminologists have made significant advances in the last few decades in modeling criminal behavior with theories including broken windows, routine activities, that criminals are generalists, distance to crime, crime displacement, etc. These theories in sum suggest that certain lesser crimes should lead serious crimes in time. In addition to theoretical behavior, the leading indicator crimes are an order of magnitude more voluminous than the Part 1 Violent crimes, so by chance alone when a new criminal element enters a neighborhood, one expects to see increases in leading indicators before Part 1 violent crimes. See Cohen, Gorr, & Olligschlaeger (2007) for a review of the environmental crime literature and specification of a corresponding distributed lag model.

The distributed lag model uses 13 lesser crimes (e.g., simple assaults, criminal mischief, and disorderly conduct) and 2 citizen calls for service (illicit drug dealing and shots fired) as leading indicators (see Cohen, Gorr, & Olligschlaeger, 2007). These variables enter the model to estimate and forecast Part 1 Violent crime in two ways. First, each district has four time lags of independent variables. This structure is necessary to allow time for crimes to “harden”, once begun as an increase in leading indicators. Second, each district has four time and space lags of independent variables. These lags are sums of leading-indicator variables in districts

contiguous to an observation district, employing a queens-case contiguity matrix (i.e., either districts that touch an observation district at points or lines are considered contiguous) that includes barriers such as for rivers that prevent spatial interaction. These space and time lags (for time lags 1 through 4) account for the effects of police actions that may displace crime to the observation district, or in the opposite direction portray crime opportunities in nearby districts that may pull crime away from the observation district.

I use a linear specification with OLS estimation for the model. A Poisson or robust model estimation yields the same coefficient estimates but different standard error estimates for coefficients. Because the latter are not used in this work, OLS estimation is sufficient.

#### *4.4 Robust Improper Linear Model*

The distributed lag model has 122 independent parameters for estimation, but more than adequate sample sizes for this purpose. Nevertheless, it is interesting to see if a much simpler leading indicator model can forecast accurately. Hence, I have included a robust, improper linear model (Dawes, 1979). It includes the five strongest leading indicators—simple assaults, criminal mischief, disorderly conduct, drug calls for service and shots-fired calls for service—with time series standardized for each variable and district and then simply averaged to yield a leading indicator index. No time and space lags are used.

Standardization used time series smoothed with single smoothing and a low smoothing constant (0.05)—thereby allowing the smoothed mean to drift with changing time series but to maintain high seasonal variations as exceptional. With this dynamic, tracking mean I found that the Poisson assumption to work well and the series not over-dispersed. Hence I used the smoothed mean also as the estimate of the smoothed variance (directly smoothing squared

deviations tended to “hang up” on exceptional data points, whereas the mean estimates largely ignored exceptional values).

I also smoothed the one-month-lagged index with single exponential smoothing and a *large* smoothing factor (0.50) to leave approximately four lagged months to comprise most of the index, but with declining weights (0.5, 0.25, 0.125, and 0.0625 respectively). This provided time for lesser crimes to “harden”.

Note, as discussed in Section 2.3, that the resulting smoothed, leading-indicator index is not scaled to the level of the corresponding Part 1 Violent crime dependent variable. Hence, while it is possible to use and evaluate this index for exceptional crime level forecasting by selection of the proper control limit via standard ROC methodology, I cannot use it to make conventional forecasts nor evaluate it using conventional forecast error measures.

#### *4.5 Experimental design*

I used a rolling-horizon experimental design based on a five-year moving window of historical data for estimation of all methods and models, starting in January 1990. At each forecast origin, I re-estimated each model and method—using the corresponding five-year historical data window—and forecasted one-month-ahead, out-of-sample. For each data window, there were 2,520 patrol district observations and 10,320 census tract observations.

There were 84 months forecasted for January 1995 through December 2001, across 42 patrol districts and 172 census tracts for a total of 3,528 patrol district forecasts and 14,448 tract forecasts per forecast method.

## 5. Results

This section has results for the Pittsburgh crime case study, including conventional and ROC forecast accuracy measures. While reported results are only from a portion of all work conducted, they are robust and hold across all variations conducted.

### 5.1 Central tendency forecast accuracy

Table 2 reports central tendency forecast accuracy across all three geographies in Pittsburgh (precincts, patrol districts, and census tracts) using MADS. In addition to the forecast methods described in Sections 4.2–4.4, the table includes two naïve methods: the random walk, that uses the most recent data point as the forecast, and the LAG 12 method, that uses the data point that is 12 months older than the forecast point as the forecast. The latter is an attempt to account for seasonality and is commonly used by police in Compstat meetings.

The cells of Table 2 have the MADS for the corresponding row's forecast method divided by the best (minimum) MADS method, which is single smoothing/city seasonality for all three geographies. Thus each cell reports the times worse a forecast method is than the best method. Finally, the rows are sorted in descending order by the census tract column. Note that the causal model specification for precincts does not include any time and space lags because precincts are too large for spatial interactions to matter.

From the minimum MADS row, it is clear that accuracy decreases substantially as the geographic scale gets larger (and individual district areas get smaller), from 18.9 for precincts, to 44.0 for patrol districts, and to 78.4 for census tracts. I believe that accuracy at the tract level is too poor for use by police and also the patrol district accuracy is questionable. Census tracts

yield highly disaggregated crime time series and Part 1 Violent crimes are relatively low-volume, making this a challenging forecast problem.

### *5.2 Exceptional Forecast Accuracy*

This section provides ROC results for Part 1 Violent crimes in Pittsburgh's census tracts. Performance for patrol districts, while having differences from census tracts of potential interest to police officials, is comparable in many ways to that of census tracts, and so is not included to conserve space in this paper. I have chosen restrictive gold standard cutoffs for each forecast method—yielding an average of six positives per month (prevalence=0.0343)—because ROC performance for tracts is best at low prevalence levels and also because I believe that police will want to focus only on very large and evident changes in crime levels. Such changes represent perceivable losses in public safety at the neighborhood level, often highlighted by news reporters calling for crime prevention and police intervention.

Note that all standardized values used in decision rules were calculated from smoothed means and variances, from single smoothing and 0.05 smoothing constant and with variance estimates equal to mean values under the Poisson assumption. Also, the decision rules had additional features beyond the basic rule (3). The robust model decision rule has an “or” condition in which either the lagged, smoothed leading-indicator index or a second lag of the same quantity that exceeded control limits triggered exception reports. Similarly, decision rules for the causal model and two univariate forecasts used “or” conditions with the forecast, first lag of forecast, and second lag of forecast to lengthen the period for “hardening” of violent crimes. While this practice increases false positives it also increases true positives for a net gain.

Fig. 4 is a blow-up of the ROC curve of Fig. 1, showing all four forecast methods described in Sections 4.2–4.4 and a relevant range of FPR values ( 0 to 0.25); for example, subjective assessments by crime analysts in Pittsburgh led to the selection of 0.16 FPR for reactive MBE (Cohen, Garman, & Gorr, 2008). Neither the random walk nor the LAG12 naïve methods are included in Fig. 4 because they both are very poor performers for exceptional conditions forecasting.

In this FPR range, all four forecast methods are better than chance (yielding higher TPR values than the chance line). The causal model dominates until just under 0.25 FPR. Next are single smoothing/city seasonality and the robust model which roughly tie, followed by Holt smoothing/district seasonality which improves above 0.15 FPR. It is interesting that the causal model (the most complex model) is best for exceptional forecasting while single smoothing/city seasonality (the simplest non-naïve method) is best for the conventional error measure (MADS). Also, the additional independent variables of the causal model (an additional 10 leading indicators plus all of the time and space lags) and optimal estimation via OLS lead to an improvement over the robust model. Holt with individual district (census tract) seasonality remains the worst method for both ordinary and exceptional forecasting—apparently district seasonality is too unreliable for use in either case.

Figs. 5–7 provide additional comparisons of the four forecast methods, but from the manager’s perspective, varying the level of effort used for work under exception reports. For example, an effort rate of 0.10 corresponds to 10% (17.2) of census tracts under exception reporting on average per month. First, Fig. 5 restates the results in Fig. 4 regarding true positives for Part 1 Violent crimes in terms of the better than chance versus effort rate. The results are qualitatively the same as those in Fig. 4.

Interesting new results are in Figs. 5 and 6. First in Fig. 5 are the better-than-chance ratios for detecting months with the leading indicator index of the top five leading indicators, using the same decision rule as for Part 1 Violent crimes, and the same gold standard (average of 6 per month). The leading indicator crimes are also highly desirable targets for police prevention and suppression, so the good performance in this case is a welcome side payment. Not surprisingly, here the robust method dominates because of its construction, with the causal model next, followed by single smoothing, and last (as always) Holt with district seasonality.

Lastly, Fig. 6 checks to see if the Part 1 Violent crime decision rule identifies any Part 1 Property crimes (burglaries, larcenies, and motor vehicle theft), another desirable target for police prevention. This crime type shares some seasonal patterns and leading indicators with Part 1 violent crimes, so that there is some potential for forecasting exceptional Part 1 Property crimes. Again using the same gold standard cutoff design of an average of six exceptions per month for Part 1 Property crimes, the robust model dominates, but the performance is much less than for the other crime types considered. Notably, the causal model is only as good as chance for Part 1 Property crimes, demonstrating how OLS and additional leading indicators are able to target Part 1 Violent crimes as intended.

In summary, the ROC measures clearly portray the benefits of forecasting exceptional crime conditions over ranges of interest to managers. When considering the importance of preventing large increases of Part 1 Violent crimes, the ability to forecast these crimes better than chance, plus the side payment identifying large increases in leading indicator crimes, I believe that an early warning system based on such forecasts will be desirable for municipal police departments. There is a significant cost in processing false positives, but likely it is well worth it. Note that because there tend to be series of exception reports, that the fixed cost of setting up a

prevention treatment for a district can be distributed over several exception points leaving just the variable costs accumulating. Finally, I have done no accounting of “near misses” of Part 1 Violent crime levels near to but below the gold standard cutoff. Such cases are also worth prevention efforts, albeit with lower payoff.

## **5. Conclusion**

In this paper, I have identified ordinary and exceptional conditions as two states of the world faced by upper-level managers. In general, managers would like subordinates to handle routinized decision making under ordinary conditions, but would like to handle the exceptional conditions themselves—with demand forecasting being one determinant of the state of their world. Reactive management by exception (MBE) uses experienced large changes in demand time series data whereas proactive MBE uses forecasted large demand changes.

I claimed in this paper that the central-tendency forecast accuracy measures in the literature are best for ordinary conditions and that the forecast field has not had forecast accuracy measures for exceptional conditions. Hence, I introduced the receiver operating characteristics (ROC) framework to provide new forecast error measures based on the tails of forecast error distributions for this purpose. ROC curves portray the tradeoff that managers must make to identify exceptional conditions: To forecast more exceptional conditions successfully, one must also experience more false positives. I demonstrated ROC analysis of exceptional crime forecasts based on seasonality or leading indicators. The most complex model, a distributed-lag model specified via behavioral theories, dominates simpler univariate models for exceptional forecasts. In reverse and in contrast, the simplest, non-naïve univariate forecast method—single smoothing/city seasonality—is best for ordinary conditions.

The ROC literature is very large and well developed, offering much more for time series forecasting than presented in this initial paper. For example, a paper under development (Gorr & Schneider, 2008) applies hypothesis tests available in the STATA statistics package for partial area under the curve (PAUC), a measure that provides overall comparisons between alternative forecast models for relevant portions of ROC curves (such as in Fig. 2).

Lastly, it appears that exceptional conditions forecasting and ROC analysis provide new opportunities to build and profitably employ multivariate causal forecast models. This paper benefitted enormously from social science theories on the behavior of criminals. Applications in other areas might likewise draw on behavioral theories from marketing and other disciplines. I suggest that spatial diffusion theory may be relevant for product forecasting with leading indicators, with sales territories high in spatial hierarchies leading other territories. For example, fashion, crime trends, and many other phenomena start in major coastal cities and work their way down spatial hierarchies and inland.

## **Acknowledgements**

Funding for this research was provided by Grants No. 98-IJ-CX-K005 and 2001-CX-0018 awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. Points of view in this document are mine and do not necessarily represent the U.S. Department of Justice. Any errors in this paper are my responsibility. I am grateful to Chief Nate Harper of the Pittsburgh Bureau of Police for data used in this research. I wish to thank Professors Alfred Blumstein, Jon Caulkins, Jacqueline Cohen, Robyn Dawes, Daniel Neill Michael, and DeKay of Carnegie Mellon University; Dr. Ned Levine of Levine & Associates;

the associate editor; and three anonymous referees for their insightful comments on my research. A version of this paper was presented at the International Symposium on Forecasting in 2008.

## References

- Ackoff, R. L. (1967). Management misinformation systems. *Management Science*, 14, Application Series, B147-B156.
- Armstrong, J. S. & Collopy, F. (1992). Error measures for generalizing about forecasting methods: Empirical comparisons. *International Journal of Forecasting*, 8, 69–80.
- Banez, L., Prasanna, P., Sun, L., Ali, A., Zhiqiang, Z., Adam, B., Mcleod, D., Moul, J., & Srivastava, S. (2003). Diagnostic potential of serum proteomic patterns in prostate cancer. *The Journal of Urology*, 170, 442–446.
- Brown, R. G. (1959). *Statistical forecasting for inventory control*. New York: McGraw-Hill.
- Brown, R. G. (1963). *Smoothing, forecasting and prediction of discrete time series*. Englewood Cliffs, N.J.: Prentice-Hall.
- Bunn, D.W. & Vassilopoulos, A.I. (1993). Using group seasonal indices in multi-item short-term forecasting. *International Journal of Forecasting*, 9, 517–526.
- Catt, P.M. (2007). Assessing the cost of forecast error. *Foresight: The International Journal of Applied Forecasting*, 7, 5–10.
- Cohen, J., Garman, S. & Gorr, W. L. (2008). Empirical calibration of time series monitoring methods using receiver operating characteristic curves. *International Journal of Forecasting*, to appear.

- Cohen, J., Gorr, W.L. & Olligschlaeger, A. (2007). Leading indicators and spatial interactions: A crime forecasting model for proactive police deployment. *Geographical Analysis*, 39, 105–127.
- Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, 34, 571–582.
- DeNeef, P. & Kent, D. L. (1993). Using treatment-tradeoff preferences to select diagnostic strategies: Linking the ROC curve to threshold analysis. *Medical Decision Making*, 13, 126–132.
- Elmore, J. G, Miglioretti, D. M., Reisch, L. M., Barton, M. B., Kreuter, W., Christiansen, C. L., & Fletcher, S.W. (2002). Screening mammograms by community radiologists: variability in false-positive rates. *Journal of the National Cancer Institute*, 94, 1373–1380.
- Fildes, R. (1992). The evaluation of extrapolative forecasting methods. *International Journal of Forecasting*, 8, 81–98.
- Gorr, W. L., Olligschlaeger, A., & Thompson, Y. (2003). Short-term forecasting of crime. *International Journal of Forecasting*, 19, 579–594.
- Granger, C. W. J. & Persaran, M. H. (2000). Economic and statistical measures of forecast accuracy. *Journal of Forecasting*, 19, 537–560.
- Henry, V.E. (2003). *The COMPSTAT paradigm : Management accountability in policing, business, and the public sector*. Flushing, NY: Looseleaf Law Publications.
- Hyndman, R. J. & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22, 679–688.

- Koehler, A. B. (2001). The asymmetry of the sMAPE measure and other comments on the M3-competition. *International Journal of Forecasting*, 17, 570–574.
- Kolassa, S., & Schutz, W. (2007). Advantages of the MAD/MEAN ratio over the MAPE. *Foresight: The International Journal of Applied Forecasting*, 6, 40–43.
- Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, 8, 283–298.
- Peterson, W., Birdsall, T.G., & Fox, W.C. (1954). The theory of detectability. *Transactions of the IRE Professional Group on Information Theory*, 4, 171–212.
- Simons, R. (1991). Strategic orientation and top management attention to control systems. *Strategic Management Journal*, 12, 49–62.
- Swets, J. A. (1986). Indices of discrimination or diagnostic accuracy: Their ROCs and implied models. *Psychological Bulletin*, 99, 100–117.
- Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science* 240, 1285–1293.
- Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Better decisions through science. *Scientific American*, 283, 82–87.
- Taylor, F. W. (1911). *Shop Management*. Project Gutenberg EBook available from <http://www.gutenberg.org/dirs/etext04/shpmsg10.txt>.
- Trigg, D. W. (1964). Monitoring a forecasting system. *Operational Research Quarterly*, 15, 271–274.
- Turban, E., Aronson, J. E., Liang, T., & McCarthy, R. V. (2004). *Decision support systems and intelligent systems*. 7th Edition, Upper Saddle River: Prentice Hall.

Valentin, L. (2007). Use scaled errors instead of percentage errors in forecast evaluations.

*Foresight: The International Journal of Applied Forecasting*, 7, 17–22.

Wetherbe, J. C. (1991). Executive information requirements: Getting it right. *MIS Quarterly*, 15, 51–65.

Withycombe, R. (1989). Forecasting with combined seasonal Indices. *International Journal of Forecasting*, 5, 547–552.

Table 1  
Contingency table and measures.

	Positive	Negative	
Test Positive	TP	FP	TP+FP
Test Negative	FN	TN	FN+TN
	TP+FN	FP+TN	n

TP = True Positives  
 FP = False Positives  
 FN = False Negatives  
 TN = True Negatives  
 n = sample size

Column measures

**Prevalence:**  $P=(TP+FN)/n$   
**(Column) True Positive Rate:**  $TPR=TP/(TP+FN)$   
**False Positive Rate:**  
 $FPR=FP/(FP+TN)$

Row measures

**Effort Rate:**  $ER=(TP+FP)/n$   
**Row True Positive Rate:**  
 $RTPR=TP/(TP+FP)$   
**Chance TP:**  
 $CTP=P(TP+FP)$   
**Better-Than-Chance True Positive Ratio:**  
 $BTPR=TP/CTP=(1/P)RTPR$

Table 2

Scaled MAD Forecast Accuracy: Pittsburgh, Part 1 Violent Crime

	Precincts	Patrol Districts	Census Tracts
Holt Smoothing/District Seasonality	1.15	1.20	1.35
LAG12	1.49	1.37	1.25
Random Walk	1.22	1.29	1.19
Causal Model	1.54	1.19	1.11
Single Smoothing/City Seasonality	1.00	1.00	1.00
Minimum MADS	18.9	44.0	78.4
N	504	3,528	14,448
Districts	6	42	172

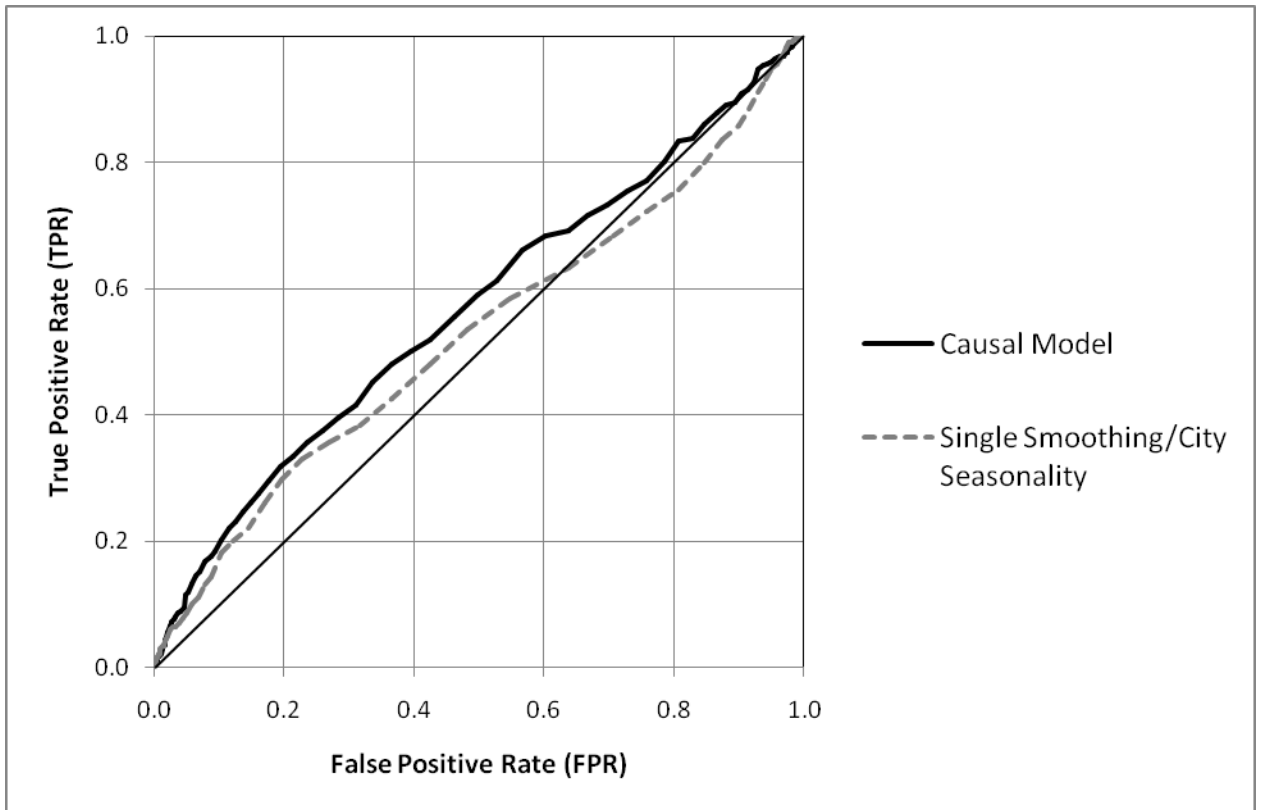
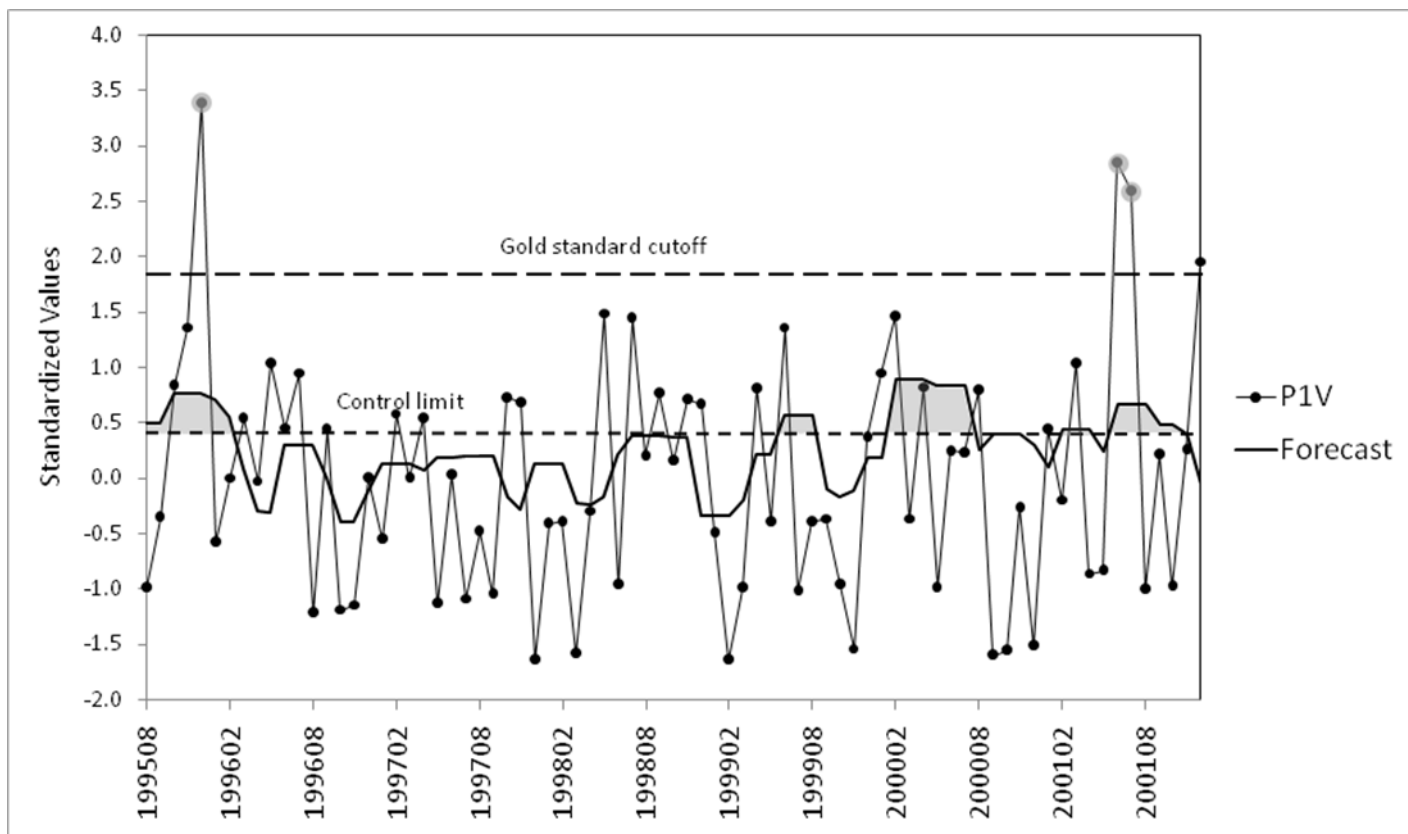


Fig. 1. Sample ROC Curves: One-month-ahead forecasts for Part 1 Violent Crimes in Pittsburgh census tracts with prevalence of 6 positives per month average ( $P=0.0343$ ).



Notes: Gray areas indicate months under exception reports. Gray halos on the three points above the gold standard cutoff are true positives.

Fig. 2. Sample time series for Part 1 Violent crimes: Census tract 1114 in Pittsburgh with gold standard cutoff (average 6 positives per month), stochastic decision variable from robust improper linear model, and control limit (0.2 false positive rate) displayed.

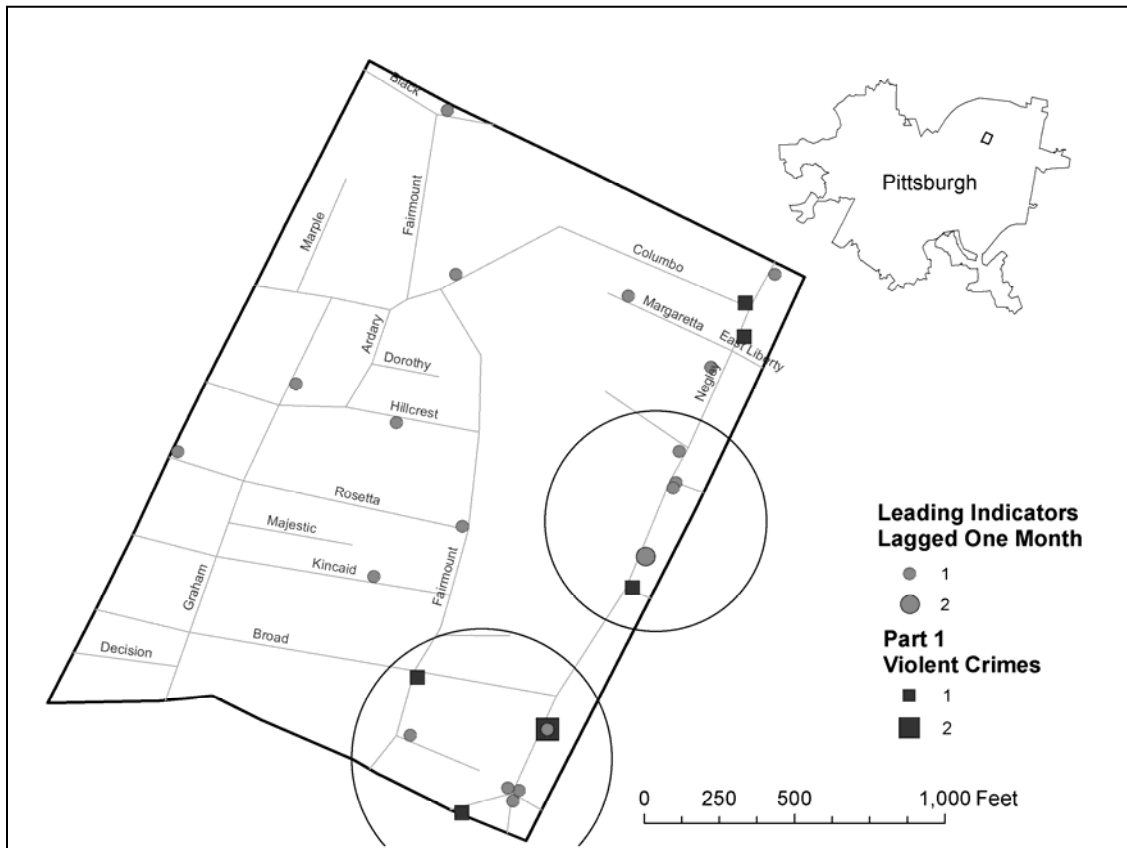


Fig. 3. Map of Pittsburgh census tract 1114 showing Part 1 Violent crimes for December 1995, five major leading indicator crimes for November 1995, and two potential leading indicator hot spots for November 1995.

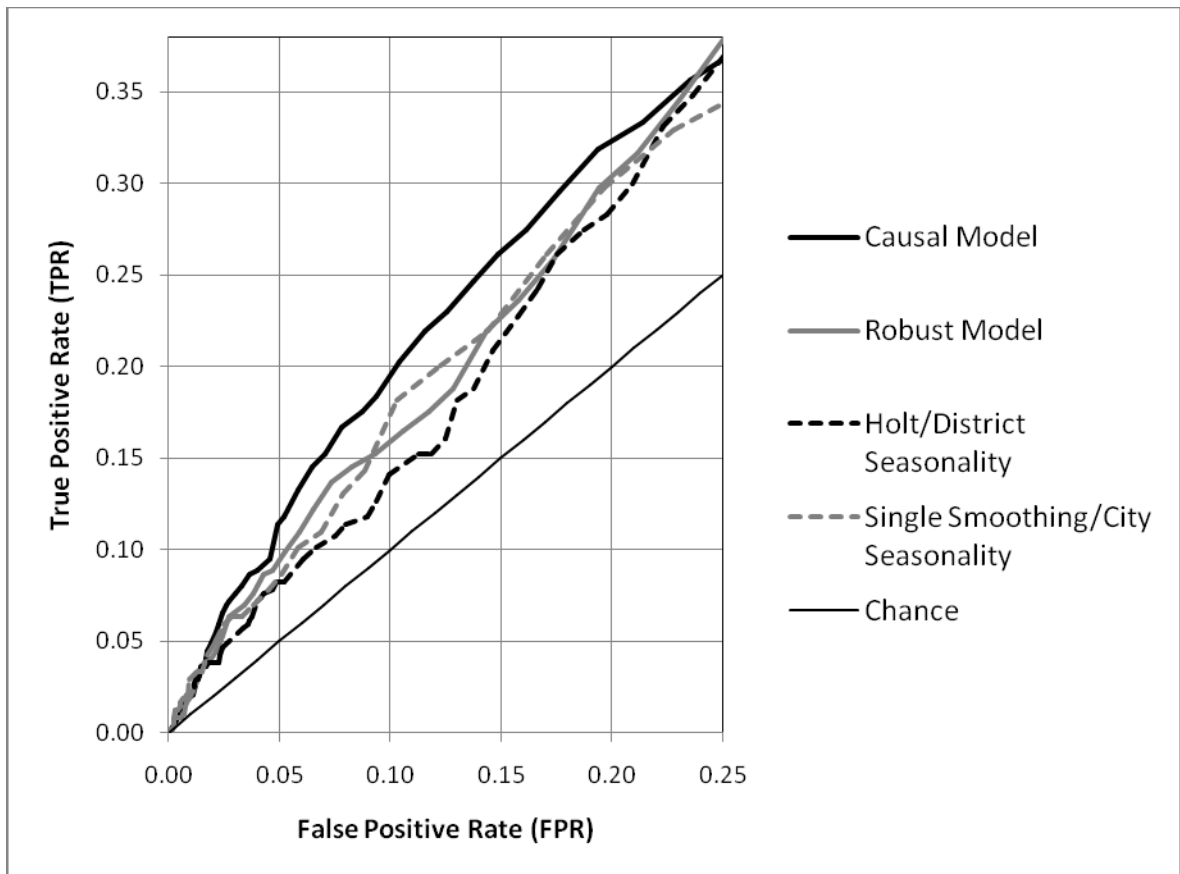


Fig. 4. ROC Curves for relevant FPR range: One-month-ahead forecasts for Part 1 Violent Crimes in Pittsburgh census tracts with prevalence of 6 positives average per month ( $P=0.0343$ ).

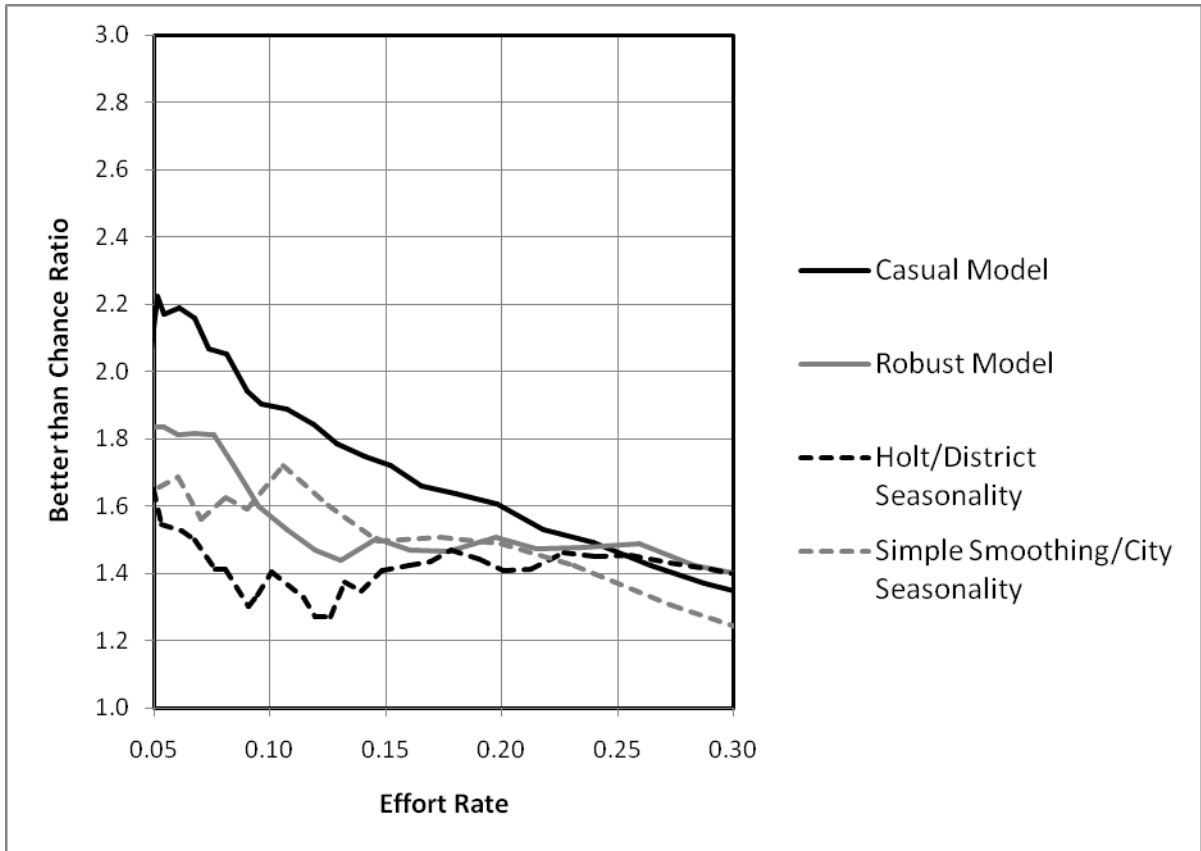


Fig. 5. Better-than-chance true positive ratio for Part 1 Violent crimes in Pittsburgh census tracts with prevalence of 6 positives average per month ( $P=0.0343$ ).

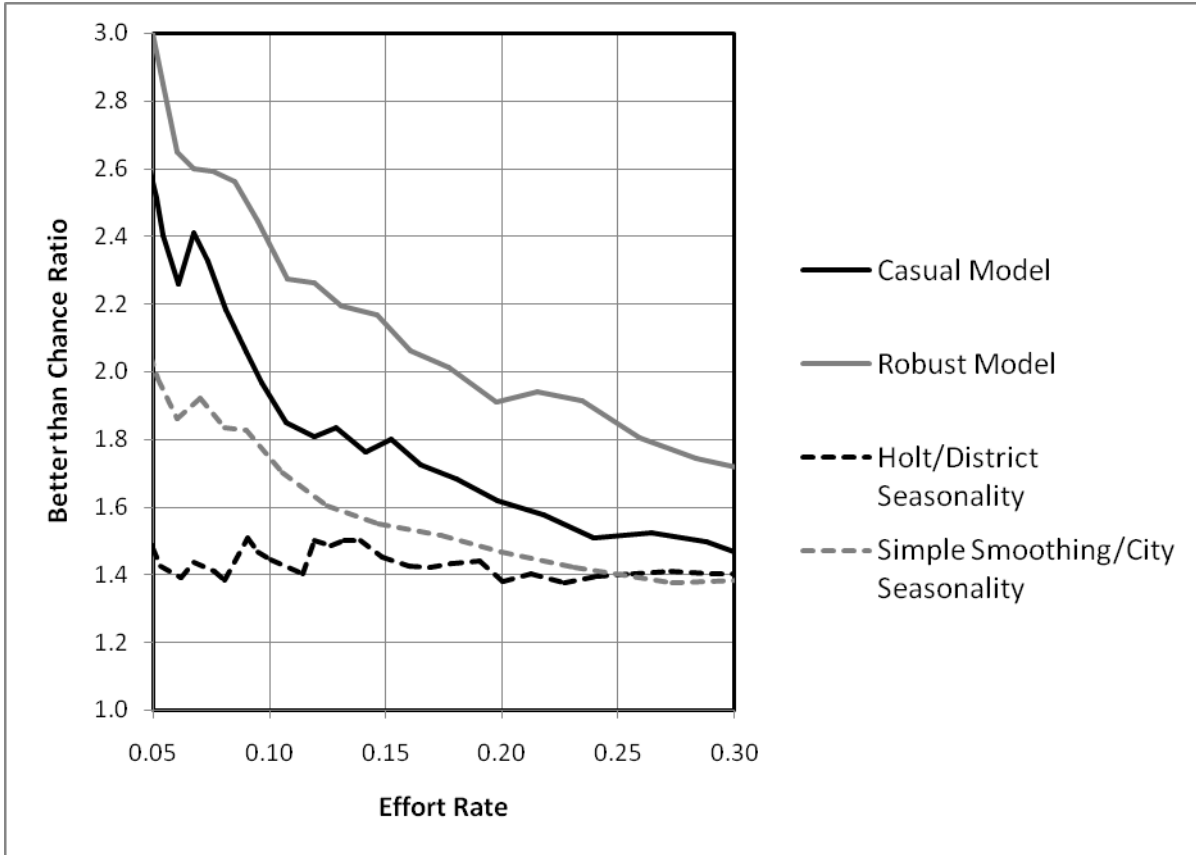


Fig. 6. Better-than-chance true positive ratio for leading indicator index in Pittsburgh census tracts with prevalence of 6 positives average per month ( $P=0.0343$ ).

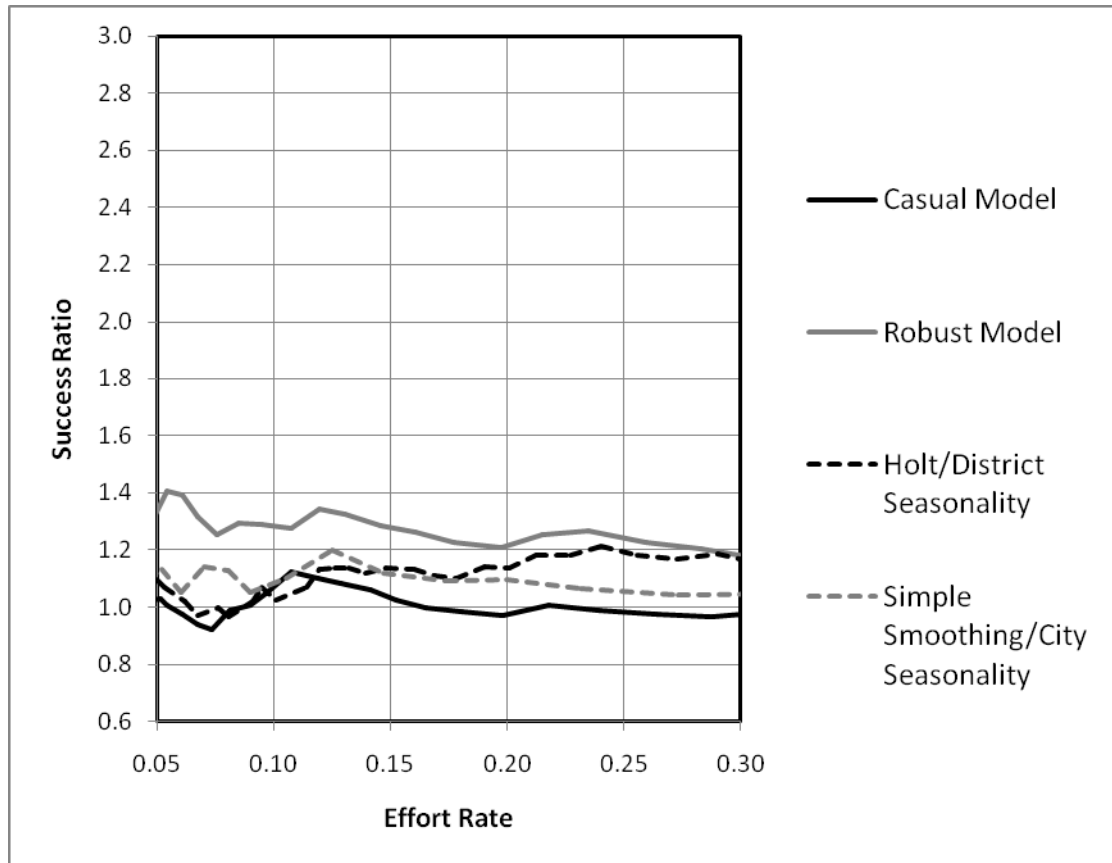


Fig. 7. Better-than-chance true positive ratio for Part 1 Property crimes in Pittsburgh census tracts with prevalence of 6 positives average per month ( $P=0.0343$ ).