Forecast accuracy measures for exception reporting using receiver operating characteristic curves

Wilpen L. Gorr

School of Public Policy and Management, H. John Heinz III College, Carnegie Mellon University, United States

Abstract

The exception principle of management reporting suggests that, under ordinary conditions, operational staff persons make decisions, but that the same staff refer decisions to upper-level managers under exceptional conditions. Forecasts of large changes or extreme values in product or service demand are potential triggers for such reporting. Seasonality estimates in univariate forecast models and leading independent variables in multivariate forecast models are among the approaches to forecasting exceptional demand, a forecast activity that this paper identifies as requiring new accuracy measures based on the tails of sampled forecast error distributions, rather than conventional measures which use the central tendency. For this purpose, the paper introduces the application of the receiver operating characteristic (ROC) framework, which has been used for the assessment of exceptional behavior in many fields. In a case study on serious violent crime in Pittsburgh, Pennsylvania, the simplest, non-naïve univariate forecast method is best for forecasting ordinary conditions 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, proposed by the father of scientific management, 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 managers of routine work and makes the best use of their limited time for addressing 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 correspond to relatively large changes or extreme demand. So, for example, ordinary conditions might correspond to foreseeable demand, such as that extrapolated using a time trend and...
seasonality. However, a product that has a 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 the detection of exceptions that have already occurred, whereas proactive MBE anticipates exceptions with forecasts. While forecast errors are necessarily much larger than detection errors, proactive payoffs are also much higher than reactive payoffs. Managers have a chance to prevent large losses or take advantage of major opportunities.

The inability to extrapolate exceptional conditions accurately using demand time series data is one trigger for reactive MBE. Time series monitoring methods that use this approach (e.g., Brown, 1959, 1963; Trigg, 1964) have the objective of identifying exceptional conditions as soon 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, not the corresponding actual demand values and forecast errors. In this case, the manager must specify decision rules with control limits on forecasted demand, as in the case study of this paper.

The question naturally arises as to how to best evaluate forecasts for exceptional versus ordinary conditions. I claim that the central tendency of functions of sampled forecast errors (e.g., means or medians of 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 the 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, in press) provides similar measures for reactive MBE. Note that, in the literature, forecast accuracy concerns measures of how close forecasted values are to the actual values. Here I use “forecast accuracy” to mean the degree to which forecasts are correct, whether in terms of being close to actual values, within a specified range of values such as exceptionally high, or in the correct qualitative category.

Section 2 reviews the literature 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; the forecast models and methods; and the experimental design. Section 5 presents the results of the case study. Finally, Section 6 provides a summary and discussion.

2. Literature

This section opens with a brief review of conventional forecast accuracy 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, whereas new forecast accuracy measures are needed for exceptional conditions. The review then moves on to MBE and the role of population screening in triggering upper-management action. Finally, the section concludes with a review of the ROC framework extracted from Cohen et al. (in press), for the convenience of the reader, and extended for the purpose of this paper.

2.1. Central tendency forecast error measures

The most 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 (MdSAPE). These measures, and several additional variations, have various properties and limitations (e.g., Armstrong & Collopy, 1992; Fildes, 1992).

More recently, researchers have proposed additional scaled measures that facilitate cross-sectional comparisons. Hyndman and Koehler (2006) proposed the Mean Absolute Scaled Error (MASE), which divides MAD by the mean of in-sample first differences of the time series. Valentin (2007) applied this measure in evaluating forecasting performance in a business context. Kolassa and Schütz (2007) divide MAD by the historic mean of the time series, leading to a measure comparable to the MAPE (replacing the actual value in the denominator with its estimated mean). This measure, denoted here as MADS, is appropriate for cross-sections of micro-scale time series with many zero-value (intermittent) observations. I use MADS in this paper’s case study for the comparison with ROC-based measures.

2.2. The case for new forecast accuracy measures for exceptional conditions

Under ordinary conditions, simple patterns are discernable, such as time trends and seasonality. Forecast models incorporate these patterns in an attempt to estimate actual values for use in short-term planning for production and distribution. Hence, the forecast error (actual minus forecasted value) is the most appropriate underlying measure for ordinary conditions. In addition, the central tendency is the single best measure for comparing forecast errors from alternative ordinary-conditions models. Therefore, it seems safe to argue that central tendency forecast accuracy measures are best for ordinary demand conditions.

While forecast errors for exceptions are reflected in central tendency measures, those values are likely to be overwhelmed by the relatively large number of ordinary values (“stew made with one elephant and one rabbit tastes like elephant”). The case study in this paper demonstrates this point, where central tendency and exceptional forecast accuracy measures individually determine opposite models as best, in some sense. Hence, conventional measures do not distinguish well between models intended for forecasting exceptional conditions. Furthermore, the underlying decision regarding exceptional product demand values is binary: either there is an indication that future demand will be exceptional or not. This is beneficial because it is easier to attain an acceptable accuracy for binary outcomes than for continuous error measures of the same variables. Therefore, it seems clear that we need new measures for exceptions forecasting, ones that focus on the tails of forecast error distributions.

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 being a common component of management information systems (e.g., Ackoff, 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, Aronson, Liang, & McCarthy, 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 (e.g., Banez et al., 2003; Elmore et al., 2002). ROC analysis, to be discussed next, is widely used in medical screening and many other fields (e.g., Swets, Dawes & Monahan, 2000).
Table 1
Contingency table and measures.

<table>
<thead>
<tr>
<th>Test Positive</th>
<th>Test Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP</td>
<td>FN</td>
</tr>
<tr>
<td>FP</td>
<td>TN</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>True Positives</th>
<th>False Positives</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP</td>
<td>FP</td>
</tr>
<tr>
<td>FN</td>
<td>TN</td>
</tr>
</tbody>
</table>

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

2.4. ROC framework

Peterson, Birdsall and Fox (1954) conceived the theory of detectability as the task of discriminating between “signal plus noise” and “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 the 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 (Swets, 1988). See this table for notation and further definitions.

Three statistics from the columns panel of Table 1 provide the information traditionally used in the literature by researchers to compare the accuracy of alternative tests: the 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 the 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 after the actual demands that were forecasted are experienced, as is 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 & Thompson, 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 those in Fig. 1 (as opposed to fitted curves) have many small 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. 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, thus making the test better than chance. A ROC curve, if better than chance, has a positive but decreasing slope as the false positive rate, FPR, increases. The closer the curve is to the upper left corner of the ROC chart (or the greater the area under the ROC curve), the more...
accurate the classifier. Any point below and to the right of another point is dominated by the former. The ROC curves in Fig. 1, with calibration of single smoothing, are better than chance, but not highly accurate for this challenging application.

The ROC curve makes it clear that there is a tradeoff in the selection of a point from the curve for the implementation of a test: the higher the true positive rate or TPR (desirable outcome), the higher the false positive rate or 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 may be desired if the exceptional condition is important enough and resources are available to handle the volume of exceptional cases for follow-up and 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., 2003; Elmore et al., 2002).

A simple assessment of benefits enables the analyst to find the optimum tradeoff using a ROC curve (Cohen et al., in press; DeNeef & Kent, 1993; Metz, 1978). Domain experts merely have to estimate the ratio of how many more times it is 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 error-prone. Small changes in the underlying assessments can lead to large ranges of FPR values possibly designated as optimal. Instead, for such a case, I suggest that managers match control limits to the resources available for MBE, based on plots of the effort rate measure of Table 1. See the example plots in Section 5.

3. Decision rules for the forecasting of exceptional demand conditions

Suppose that an organization forecasts time series, $y_{it}$, for $i = 1, \ldots, I$ products or services. Historical time series data, including possibly independent-variable time series for multivariate models, for $t = 1, \ldots, T$, are used to make forecasts $F_{iT+m}$, where $m = 1, \ldots, M$ 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 decision rules only for exceptionally high values, the side of primary interest to police, but rules for exceptionally low values follow 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 allow the simplicity of setting a single gold-standard cutoff and control limit across time series of varying scales (e.g., for all neighborhoods in a police jurisdiction). This has merit for the crime case study of Section 4, in that it treats both low and high crime areas the same. For example, if crime count data and forecasts were used with a constant gold cutoff and control limit, then mostly high crime neighborhoods, or simply large neighborhoods, would have the most exceptions, and would thereby draw management attention mostly to those areas at the expense of areas with lower-volume time series.
Without standardization, the decision maker would have to set gold-standard cutoffs and control limits for each cross-sectional time series, which seems impractical for many settings.

The decision maker specifies the gold standard cutoff as a Z-value, $Z_u$, so that the fraction $\alpha$ of actual demand values in the corresponding upper tail of each $y_i$, distribution defines the positives for ROC analysis and the exceptional conditions for upper-level management decision making. For example, for police I claim that this cutoff will be quite high, corresponding to crime flare-ups that are the subject of local news coverage. In any event, this choice will be made using informed judgment.

Suppose that rolling forecasts, $F_{it}$, are available for $t = t_1, \ldots, T$, where $t = 1, \ldots, t_1 - 1$ served as the estimation data for the earliest forecasts for time series $i = 1, \ldots, I$. Finally, suppose that $F_{it}'$ are standardized values for $t$ in $t_1, \ldots, T$. The decision maker must specify a control limit, $L$ (ultimately determined from ROC analysis), then the decision rule for triggering exception reports is:

$$
\text{If } F_{iT+1}' \geq L, \\
\text{issue an exception report for product } i = 1, \ldots, I.
$$

While the standardized forecasts in rule (1) use data for $t_1, \ldots, T$ for standardization, standardized demand $y_{iT+1}'$ for accuracy assessment, after the fact, uses data from $t_1, \ldots, T + 1$.

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 out of the four positives.

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 at different times, and are shown with the larger point markers). The data 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 they are 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, thereby possibly preventing the corresponding 5 out of 7 part 1 violent crimes in December 1995 that are within the designated 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 Comstat 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 reviews 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
Fig. 2. Sample time series for part 1 violent crimes: census tract 1114 in Pittsburgh with the gold standard cutoff (average of 6 positives per month), a stochastic decision variable from the robust improper linear model, and the control limit (0.2 false positive rate) displayed. Notes: Gray areas indicate months under exception reports. There are four positives above the gold standard cutoff line and gray halos on three of those points are true positives.

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.
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 et al., in press), but given adequate forecasts would also be candidates for using proactive MBE for the prevention of crime flare ups.

Monthly crime counts by geographic area are among the Compstat performance measures. Most important are the 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 population approximately 4000 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 the forecast requirements of 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 the top management can allocate resources for prevention.

4.2. Univariate forecast methods

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 the time trend, however, are too small to produce exceptionally large forecasts.

Gorr et al. (2003), using the MAPE forecast accuracy measure, found 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., Bunn & Vassilopoulos, 1993; Withycombe, 1989), that group estimates of seasonality applied to individual products are more accurate for forecasting than the individual product-level seasonality. Perhaps, though, the situation is reversed for exceptions forecasting, because district-level seasonality has more variation than group seasonality. Hence, I retest group versus individual district seasonality in this paper. Tuning the smoothing methods by varying smoothing constants has little effect on the forecast performance for exceptional conditions (as was confirmed by experimentation).

This paper thus includes the single (smoothed level) and Holt (smoothed level and time trend) exponential smoothing methods, with smoothing factors optimized for each forecast origin over the in-sample data, minimizing the MSE for one-step-ahead forecasts. The 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 a time trend.

4.3. Distributed lag model

Perhaps more valuable for management than large seasonal-based changes 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 winter holiday season). In contrast, exceptional crime increases due to gang rivalries, illicit drug dealing increases, influxes of illicit handgun suppliers, etc. are much more difficult to detect, and 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. The distributed lag model thus 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 et al., 2007). These variables enter the model for estimating and forecasting 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., districts that touch an observation district at either points or lines are considered contiguous) that includes barriers, such as 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 whether 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.

The standardization for each variable used time series smoothed with single smoothing and a low smoothing constant (0.05) — thereby allowing smoothed means to drift with changing time series, but to maintain high seasonal variations as exceptional. I found the Poisson assumption to work well with smoothed means, and the series were not over-dispersed. Hence, I standardized each data point of a variable by subtracting its corresponding smoothed mean and dividing by the square root of the same smoothed mean.

To include time lagged information, I also smoothed the resultant leading indicator index with single exponential smoothing, but with a large smoothing factor (0.50), to leave approximately four lagged months to comprise most of the index, and declining weights (0.5, 0.25, 0.125, and 0.0625 respectively). This provided time for increased lesser crimes to “harden” into increased serious violent crimes.

Note that the resulting robust model 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 forecasting by the selection of a proper control limit via the ROC methodology, it cannot be used to make conventional forecasts.

4.5. Experimental design

I used a rolling-horizon experimental design based on a five-year moving window of historical data for the 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 2520 patrol district observations and 10,320 census tract observations.

There were 84 months forecasted, from January 1995 through December 2001, across 42 patrol districts and 172 census tracts, giving a total of 3528 patrol district forecasts and 14,448 tract forecasts per forecast method.
5. Results

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

5.1. Central tendency forecast accuracy

Table 2 reports the 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, which uses the most recent data point as the forecast, and the LAG 12 method, which uses the data point that is 12 months before the forecast point as the forecast. The latter is an attempt to account for seasonality, and is commonly used by the police in Compstat reporting.

The cells of Table 2 contain the MADS of 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 amount a forecast method is worse 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 78.4 for census tracts. I believe that the accuracy at the tract level is too poor for use by police, and the patrol district accuracy is also 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. Exceptions forecast accuracy

This section provides ROC results for part 1 violent crimes in Pittsburgh’s census tracts. The 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 in this paper to conserve space. I have chosen restrictive gold standard cutoffs for each forecast method – yielding an average of six positives per month (prevalence = 0.0343) – because the 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 an 0.05 smoothing constant, and with variance estimates equal to the 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 exception reports were triggered by either the lagged, smoothed leading-indicator index or a second lag of the same quantity that exceeded control limits. Similarly, decision rules for the causal model and two univariate forecasts used “or” conditions with the forecast, and the first and second lag of forecast to lengthen the period for the “hardening” of violent crimes. While this practice increases false positives, it also increases true positives, resulting in 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 range of relevant false positive rate (FPR) values (0–0.25); for example, subjective assessments by crime analysts in Pittsburgh led to the selection of 0.16 FPR for reactive MBE (Cohen et al., in press). Neither the random walk nor the LAG12 naïve methods are included in Fig. 4 because they are both very poor performers for exceptional conditions forecasting.

In this FPR range, all four forecast methods are better than chance (yielding higher true positive or 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
Table 2
Scaled MAD forecast accuracy: Pittsburgh, part 1 violent crime.

<table>
<thead>
<tr>
<th></th>
<th>Precincts</th>
<th>Patrol districts</th>
<th>Census tracts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Holt smoothing/District seasonality</td>
<td>1.15</td>
<td>1.20</td>
<td>1.35</td>
</tr>
<tr>
<td>LAG12</td>
<td>1.49</td>
<td>1.37</td>
<td>1.25</td>
</tr>
<tr>
<td>Random Walk</td>
<td>1.22</td>
<td>1.29</td>
<td>1.19</td>
</tr>
<tr>
<td>Causal Model</td>
<td>1.54</td>
<td>1.19</td>
<td>1.11</td>
</tr>
<tr>
<td>Single smoothing/City seasonality</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Minimum MADS</td>
<td>18.9</td>
<td>44.0</td>
<td>78.4</td>
</tr>
<tr>
<td>N</td>
<td>504</td>
<td>3528</td>
<td>14,448</td>
</tr>
<tr>
<td>Districts</td>
<td>6</td>
<td>42</td>
<td>172</td>
</tr>
</tbody>
</table>

Fig. 4. ROC curves for the relevant FPR range: one-month-ahead forecasts for part 1 violent crimes in Pittsburgh census tracts with a prevalence of 6 positives on average per month ($P = 0.0343$).

Fig. 5 and 6 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 from 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.

Fig. 6 displays better-than-chance ratios for forecasting months with high values for the leading indicator index using the decision rule for part 1 violent crimes. The leading indicator crimes are also highly desirable targets for police prevention and suppression, so the good performance seen in Fig. 5 for this case is a welcome side benefit. Not surprisingly, here the robust method dominates because of its construction, with the causal model next, followed by single smoothing, and last (as always), comes Holt with district seasonality. A similar figure to Fig. 6 (not included here) shows that the decision rule for part 1 violent crime is also mildly helpful at forecasting exceptional values of part 1 property crime (burglaries, larcenies, and motor vehicle theft), another desirable target for police prevention and a side benefit. Again, the robust model dominates.
In summary, the ROC measures clearly portray the benefits of forecasting exceptional crime conditions over ranges of interest to managers. Considering the importance of preventing large increases of part 1 violent crimes, I believe that an early warning system based on forecasts of these crimes will be desirable for municipal police departments. It would also provide the side benefit of identifying large increases in leading indicator and part 1 property crimes. There is a significant cost in processing false positives, but most likely it is well worth it. Note that, because there tend to be series of exception reports in a neighborhood, 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 for “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 a lower payoff.

6. 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 routine 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 have 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 previously 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 the ROC analysis on exceptional crime forecasts based on seasonality and 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.

As a final note, 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 Nos. 98-IJ-CX-K005 and 2001-CX-0018 awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. The points of view in this document are mine and do not necessarily represent those of 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 the data used in this research. I wish to thank Professors Jacqueline Cohen, Robyn Dawes, Michael DeKay, and Daniel Neill, 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


**Wilen L. Gorr** is Professor of Public Policy and Management Information Systems at the H. John Heinz III College, School of Public Policy and Management of Carnegie Mellon University. His current research is on forecasting crime for the deployment of police and evaluating interventions aimed at reducing or preventing crime. He has published in the *International Journal of Forecasting*, *Management Science*, *Geographical Analysis*, *Criminology*, and other leading journals.