October 2013

ROC-Based Model Estimation for Forecasting Large Changes in Demand

Matthew J. Schneider
Cornell University, mjs533@cornell.edu

Wilpen L. Gorr
Carnegie Mellon University

ROC-Based Model Estimation for Forecasting Large Changes in Demand

Abstract

Forecasting for large changes in demand should benefit from different estimation than that used for estimating mean behavior. We develop a multivariate forecasting model designed for detecting the largest changes across many time series. The model is fit based upon a penalty function that maximizes true positive rates along a relevant false positive rate range and can be used by managers wishing to take action on a small percentage of products likely to change the most in the next time period. We apply the model to a crime dataset and compare results to OLS as the basis for comparisons as well as models that are promising for exceptional demand forecasting such as quantile regression, synthetic data from a Bayesian model, and a power loss model. Using the Partial Area Under the Curve (PAUC) metric, our results show statistical significance, a 35 percent improvement over OLS, and at least a 20 percent improvement over competing methods. We suggest management with an increasing number of products to use our method for forecasting large changes in conjunction with typical magnitude-based methods for forecasting expected demand.

Keywords

model, estimation, demand, ROC

ROC-Based Model Estimation for Forecasting Large Changes in Demand

Matthew J. Schneider[1] Wilpen L. Gorr[2]

Abstract

Forecasting for large changes in demand should benefit from different estimation than that used for estimating mean behavior. We develop a multivariate forecasting model designed for forecasting the largest changes across many time series. The model is fit based upon a penalty function that maximizes true positive rates along a relevant false positive rate range and can be used by managers wishing to take action on a small percentage of products likely to change the most in the next time period. We apply the model to a crime dataset and compare results to OLS as the basis for comparisons as well as models that are promising for large-change demand forecasting such as quantile regression, synthetic data from a Bayesian model, and a power loss model. Using the partial area under the curve (PAUC) metric, our results show statistical significance, a 35 percent improvement over OLS, and at least a 20 percent improvement over competing methods. We suggest managers with large numbers of time series (*e.g.*, for product demand) to use our method for forecasting large changes in conjunction with typical magnitude-based methods for forecasting expected demand.

Index Terms

Management By Exception, PAUC Maximization, Large Changes, Forecasting Exceptions, ROC Curves.

I. Introduction

Demand forecasting generally is done with extrapolative time series methods, such as exponential smoothing with level, trend, and seasonal components. Time periods during which the underlying univariate model is stable and forecast accuracy is acceptable are called "business as usual" (BAU) in this paper. Highly disaggregated time series, such as for product or service demand, however are notorious for having large changes—outliers, step jumps, turning points, etc.— that cannot be forecasted using simple

[1]: Department of Statistical Science and Labor Dynamics Institute, Cornell University, Ithaca, USA: mjs533@cornell.edu

[2]: H. John Heinz III College, Carnegie Mellon University: gorr@cmu.edu

extrapolative forecast models. Thus the time series forecasting field has long recognized the importance of handling exceptions to BAU; in particular, by developing time series monitoring methods for early detection of large changes (*e.g.*, Brown, 1959; Trigg 1964).

Time series monitoring supports reactive decision making, after large changes have occurred. Better, if forecast models are accurate enough, is to forecast large changes in demand to allow proactive decision making, with a chance of preventing losses or taking advantages of potential gains. This paper proposes a new estimation method for forecast models aimed at improving forecast accuracy for large changes in demand. The new estimation method minimizes a loss function based on a receiver operating characteristic (ROC) measure, partial area under the ROC curve (PAUC). Section 2 reviews the underlying decision framework—management by exception (MBE)—and ROC methods used to implement MBE for large-change demand forecasting, including PAUC. Essentially, ROC assesses predictions for binary classification, in the case of this paper that a future period will or will not have a large increase or decrease in demand.

While central-tendency forecast error measures, such as MAD, MAPE, and MSE, are best for evaluating forecast accuracy under BAU, recent research shows that a different kind of forecast error measure is needed for evaluation under large-change conditions for demand. Gorr (2009) compared univariate versus multivariate forecast models using the same data and found that forecast performance assessed using MAPE strongly favored simple univariate methods, whereas ROC assessment strongly favored the multivariate model. The multivariate model had leading indicator independent variables capable of forecasting large changes when there were large changes in lagged leading indicators. Gorr & Schneider (2013) compared simple versus complex univariate forecast models for large changes using monthly data from the M3 competition and found that PAUC assessments for large-change forecasts showed that the complex univariate models are significantly more accurate than simple univariate models. Apparently, complex univariate models have enough additional parameters in functional forms to make them sensitive to subtle indications of rapidly changing trends.

Parker (2011) goes a step further and shows that classification performance over seven measures of classification (which included AUC but not PAUC) is best by picking the right performance measures as a loss function for estimation. In line with this result, this paper provides evidence that parameter estimates for a multivariate forecast model made using the PAUC, ROC-based loss function are much more accurate for large-change demand forecasts than those from a central-tendency-based loss function (MSE). To our

knowledge, there is no previous empirical research in time series forecasting using a ROC-based loss function for model estimation.

The general point of the emerging literature on large-change demand forecasting, including this paper, is that an organization can continue to use whatever extrapolative forecast models it prefers for BAU, but it needs a second, preemptive forecast model that takes over for large demand changes. Needed are two kinds of forecast accuracy measures (central tendency for BAU magnitudes and PAUC for classification), two kinds of forecast models (simple extrapolative for BAU and complex univariate or multivariate for MBE), and two kinds of loss functions for forecast model estimation (central tendency for BAU magnitudes and PAUC for MBE classification).

Section 2 provides motivation and background for the paper's new estimation method with an overview of MBE and ROC applied to large-change forecasting. Section 3 develops the ROC-based method for parameter estimation for a multivariate, leading indicator forecast model and develops comparison models. Section 4 describes the time series data used to calibrate the model and rolling horizon forecast experiment. Section 5 presents results comparing alternate forecast models with significance testing. Finally, section 6 concludes the paper with a summary and suggestions for future work.

II. Management by Exception for Demand Forecasting

This section provides an overview of MBE implemented with ROC analysis as applied demand forecasting. MBE and ROC provide a decision-making framework, model, and methods for managing exceptional, large-change demand conditions. The large literature on optimal inventory control models (*e.g.*, Brown, 1959) provides corresponding methods for BAU conditions, but models and methods for large-change conditions are fairly new. The key to success in this area is getting accurate forecasts, tuned for the tails of demand distributions, and that is the purpose of this paper, to the provide new estimation methods for large-change demand conditions.

MBE depends on the decision of whether or not to flag a forecast as being large-change, implemented via decision rules analogous to hypothesis testing, except that decision-rule thresholds cannot use the traditional Type I (false positive) error rates of empirical research (1 or 5 percent). Instead a false positive rate and corresponding decision-rule threshold must be determined based on cost/benefit considerations. ROC provides the decision model and methods for determining optimal false positive error rates, and corresponding decision rule thresholds.

MBE, one of the oldest forms of a management control system (Taylor, 1911; Ricketts & Nelson, 1987), provides the principle that only variances (exceptions) from usual conditions should be brought to managers' attention. All else should be handled by operational staff using standard procedures. Then managers' limited time can be devoted to decisions requiring their expertise and power on emerging problems or opportunities. One type of variance is a large change in demand for products or services (West et al., 1985; Gorr, 2009; Gorr & Schneider 2013). Demand is only partially affected by an organization's efforts, given competition in the market place, limits of marketing programs, and changing consumer tastes. Hence, large changes in demand are an important source of variances for triggering MBE reports to production and marketing managers. MBE can be reactive, based on detection with time series monitoring methods, or proactive, based on forecasting. First we discuss detection and then move on to forecasting.

Time series monitoring methods, such as the Trigg (1964) and Brown methods (1959), compute a test statistic used in a decision rule analogous to hypothesis testing, for making the binary decision. The decision rule uses a threshold value, if exceeded by the test statistic, is a signal trip or "yes," otherwise the decision is "no" there is no large change. If "yes" then the time series undergoes diagnosis, using additional data and expertise, to determine if any intervention is needed into BAU practices (*e.g.*, a new marketing plan, price decrease, product improvements, or decreased production level). Next we discuss the mechanics of implementing such decision rules using ROC.

There needs to be an external determination as to when a time series has data points considered in fact to be large changes. Such a data point is called a "positive" and is determined by a "gold standard." All other time periods are "negatives." In public health, where ROC is used extensively (*e.g.*, Pepe, 2004), the analogous problem to MBE for demand forecasting is population screening for sick individuals. To be economically feasible, screening must use inexpensive and therefore imperfect tests, but then for individuals flagged as possibly sick, one needs a "gold standard" test for determining whether individuals are really sick or not, one that generally is expensive and more invasive. For example, for prostate and breast cancer screening, the gold standard is biopsy, examining sample tissue under a microscope. Biopsy is not infallible, but is much more accurate than screening tests such as PSA level in blood samples for screening prostate cancer.

Of course, the demand forecasting problem does not have gold standard tests, such as biopsy, for large demand changes. Instead, managers must use judgment to determine changes large enough to be worth the cost of diagnosis and possible action. Gorr (2009) used a gold standard policy, that the top small

percentage of large changes in standardized time series data be considered positives, and reasoned that police officials have means to make such judgments (*e.g.*, police would like to prevent the large changes that are reported in the news media). This gold standard is applied to out-of-sample forecasts during the evaluation stage, when actual values are available. A gold standard policy avoids the alternative of applying expertise and judgment to all time series points individually to determine positives in the evaluation phase of forecasting. Cohen *et al.* (2009) took this alternative, and while effective, was very costly.

There are four outcomes for a binary decision: true positive (the signal trips and the time period is a positive), false positive (the signal trips but the time period is a negative), false negative (the signal does not trip but the time period is a positive), and true negative (the signal does not trip and the time period is a negative). Application of a decision rule with a given threshold in repeated trials over time and across time series is summarized using a contingency (or confusion table) with frequency counts of all four possible outcomes. Common statistics from this table are the true positive rate, TPR = number of true positives/number of positives, and false positive rate, FPR = number of false positives/number of negatives. The complements of these statistics are the false negative rate and true negative rate.

It is a fact that increasing the true positive rate necessarily increases the false positive rate, so that there is a trade-off to be made in determining an optimal, corresponding decision-rule threshold. This is seen in the shape of the ROC curve, which plots true positive rate versus false positive rate for all possible decision rule thresholds and is an increasing function with decreasing slope between (0,0) and (1,1). The higher the ROC curve for a model, the more accurate the binary decision model. An overall measure of the performance of a monitoring or forecast model thus is area under the ROC curve (AUC) which ranges between 0 and 1. Better in practice is partial area under the ROC curve (PAUC). This is the area for a restricted range of false positive rates, often from 0 up to 10 or 20 percent because the cost of processing signal trips for false positive rates exceeding those rates generally is excessive and/or beyond available resources. See Figure 1 in section 5 for example ROC curves and to get a sense of the kind of time series data being forecasted in this paper.

Empirical research uses traditional values, such as 1 or 5 percent, for false positive rates (Type I errors) that determine decision rule thresholds from normal or t-distribution tables. This practice implements a conservative view on accepting evidence of new theories. Business, however, needs to determine thresholds to obtain the optimal trade-off of true versus false positive rates. It is straightforward to write a utility model for the binary decision problem and to derive optimality conditions (*e.g.*, see Metz, 1978; Cohen

et al., 2009). The optimal false positive rate is determined by finding the point at which a derived straight line is tangent to the ROC curve. The slope of that line depends on the prevalence of positives and the ratio of the utility of avoiding a false negative versus the utility of avoiding a false positive. For example, Pittsburgh police officials estimated that it is 10 times more important to avoid a false negative than a false positive when monitoring serious violent crimes for large increases, and this led to a 15 percent false positive rate as optimal for time series monitoring (Cohen et al., 2009). Likewise, population screening for prostate and breast cancers have false positive rates roughly in the range of 10 to 15 percent for most parts of the world (*e.g.*, Banez et al., 2003; Elmore et al. 2002). In both crime and public health cases, the severe consequences of false negatives (not intervening when there is a large increase in serious violent crime or not catching cancer in early stages) outweighs the costs of processing false positives. So-called "A" items from ABC inventory analysis (*e.g.*, Ramanathan, 2006) are likely similar in terms of importance or consequence.

All of the framework and methods discussed in this section depend on being able to forecast large changes accurately enough. Thus the next section of this paper develops a model estimated using a loss function based on PAUC to best tune model parameters for MBE.

III. Multivariate Leading Indicator Modeling

Multivariate leading indicator models that restrict forecasting to linear predictors of the form

$$\hat{y} = X\hat{\beta} \tag{1}$$

are compared. y is the dependent vector with observations y_i for $i = 1, ..., n$ and X is the matrix of leading indicators (with time lagged values) with rows x_i. All models estimate $\hat{\beta}$ in-sample on different loss functions L which are functions of the data (y and X). Proposed models are well suited for large-change forecasting and central tendency models (ordinary least squares) provide a benchmark. First, the PAUC loss function is formally developed, then the proposed PAUC Maximization Forecast (PMF) model is developed, followed by comparison models.

For all modeling, we define the initialization set as the set of data which is used to estimate $\hat{\beta}$ and not used in forecasting. The training set is is the set of data used for model selection based on pairwise comparison of out-of-sample results in the training set only. The test set is the set of data used to evaluate all models in this paper and report the results. This paper uses rolling horizon forecasts and iteratively

conditions on all data up to time t to forecast data in time $t+1$. We define in-sample data as data up to time period t that is used to forecast out-of-sample data in time period $t+1$. Depending on the time period, in-sample data can exist in both the training and test sets, however pairwise comparisons for metaparameter selection are only performed in the training set (using the PAUC loss function as the comparison) and results are only reported for the test set. Both of these are done using out-of-sample forecasts only. See Table II.

A. PAUC Loss Function

This section develops the functional form of the 1-PAUC loss function used for estimation. A manager states the gold standard policy that transforms the decision variable, $\mathbf{y}$, into a binary gold-standard vector, $\mathbf{g}$, where a 1 indicates a positive and a 0 indicates a negative,

$$\mathbf{y} \in \mathbb{R}^n \longrightarrow \mathbf{g} \in \{0,1\}^n \tag{2}$$

. A positive is an observation, worthy of investigation and possible intervention, that we want to have flagged by a forecast. The policy is implemented using a threshold, not to be confused with decision-rule thresholds discussed below, for standardized values of the dependent variable, y^*, in our empirical application. Standardization matches the police criterion of equity for allocating resources to different regions of a city. If raw crime counts were used, all extra police resources would be allocated to the highest crime areas; whereas, with standardized crime counts any area, regardless of crime scale, with a relatively large increase in crime can get extra police resources. Section IV has details on the gold standard used in this paper.

ROC curves plot TPR versus FPR for all possible decision-rule thresholds of a given set of forecasts. ROC curves are constructed by comparing the rank of all forecasts to the gold standard vector. For forecast values $\hat{y}_i = X_i\hat{\beta}$, define the j^{th} decision rule threshold, $j = 1, 2, ...,$(1+ number of unique $\hat{y}_i$'s) corresponding to selected constants c_j's which divide the ranked $\hat{y}_i$'s. Then, a decision rule is defined under the j^{th} threshold and i^{th} observation where $\mathbf{1}_{\hat{y}_i > c_j}$ outputs a 1 if $\hat{y}_i > c_j$ or 0 otherwise where

$$DR_{i,j} = \mathbf{1}_{\hat{y}_i > c_j} \tag{3}$$

.

The resulting collection of TPRs and FPRs for all thresholds are

$$TPR_j(\hat{\beta}, X, \mathbf{g}) = (\sum_{i=1}^{n} \mathbf{1}_{DR_{i,j}=g_i=1})/(\sum_{i=1}^{n} g_i) \tag{4}$$

$$FPR_j(\hat{\beta}, X, \mathbf{g}) = (\sum_{i=1}^{n} \mathbf{1}_{DR_{i,j}-g_i=1})/(n - \sum_{i=1}^{n} g_i) \tag{5}$$

AUC is calculated as the sum of trapezoidal areas and PAUC is limited to a maximal FPR (*e.g.*, 20%) in practice.

$$AUC(\hat{\beta}, X, \mathbf{g}) = \frac{1}{2} \sum_{j=2}^{U} (FPR_j - FPR_{j-1})(TPR_j + TPR_{j-1}) \tag{6}$$

$$PAUC(\hat{\beta}, X, \mathbf{g}) = \frac{1}{2} \sum_{j=2}^{\{j:FPR_j \leq 0.20\}} (FPR_j - FPR_{j-1})(TPR_j + TPR_{j-1}) \tag{7}$$

1-PAUC is the loss function proposed in this paper for estimating forecast models used to implement MBE. Equivalent, of course, is PAUC maximization.

Explicit solutions for maximizing AUC exist under the assumption of normality (Su & Liu, 1993), but more recent research found that models which are tuned for AUC do not perform well for PAUC (Pepe et al., 2006; Ricamoto & Tortorella, 2010). PAUC maximization was recently studied in biostatistics for classifying patients as diseased or non-diseased using approximations to the PAUC function with wrapper algorithms (Wang & Chang, 2011) or boosting (Komori & Eguchi, 2010). Other biometric papers propose new PAUC maximization algorithms by using a weighted cost function with AUC and a normality assumption (Hseu & Hsueh, 2012). As such, the PAUC maximization papers concentrated on identifying the distributional differences between diseased and non-diseased populations, whereas, we use multiple time series as the dependent variable which presents challenges to boosting algorithms and sample size issues to PAUC approximations. Time series are treated as elements (versus individuals as elements) and large changes within time series are positives (versus diseased individuals as positives). Our application differs structurally since large changes can occur in any time period with any time series.

B. PAUC Maximization Forecast Model

In this section, we detail the estimation procedure used to generate forecasts for our proposed PMF model. In overview, first, in each time period t, the proposed model chooses optimal coefficients β_t^* of the leading indicators X_t which have the best PAUC for the gold standard $\mathbf{g}_t$. Next, the estimation procedure combines current and past values of the optimal coefficients iteratively using an exponential smoothing procedure, which gives less weight to older estimates. This extra step provides consistency in parameter estimates from period to period. Finally, the proposed model forecasts large changes in time period $t+1$ and as time moves forward, the model is re-estimated for each successive set of forecasts.

For the current time period t, we define the cross-sectional loss as

$$L_t = 1 - PAUC_t(\mathbf{g}_t, X_t, \beta_t) \tag{8}$$

and select

$$\beta_t^* = \arg\min_{\beta_t} L_t \tag{9}$$

which minimizes L_t or equivalently maximizes $PAUC_t$ according to an optimization procedure described below. $PAUC_t^*$ is calculated by using only functions of the in-sample vector $X_t\beta_t^*$. All unique cutoff values $c_1, c_2, ..., c_{(1+\text{number of unique values of } X_t\beta_t^*)}$ are chosen by first sorting across values within $X_t\beta_t^*$ and then averaging consecutive values which are not identical. These cutoff values represent various managerial decisions j of predicting large changes, $DR_{t,j} = \mathbf{1}_{X_t\beta_t^* > c_j}$. Then, $PAUC_t^*$ is estimated by inputting the vectors $DR_{t,1}, DR_{t,2}, ...DR_{t,(1+\text{number of unique values of } X_t\beta_t^*)}$ into the equations in the previous section.

To find optimal values of β_t^*, we employ the optim function in R, using the Nelder-Mead simplex method (which while is relatively slow, is known to be robust) for minimizing L_t (R Development Core Team, 2012). We set starting values equal to the OLS estimates of β_t and the run the optimization for a maximum of 500 iterations or until convergence. After L_t converges to a minimum, the current values of β_t are labeled β_t^* and the in-sample prediction vector $X_t\beta_t^*$ is determined to maximize $PAUC_t$.

Our early research using optim resulted in inconsistent parameter estimates from month to month. Thus, instead of using β_t^* for forecasting $t+1$, we train the forecasts over a rolling horizon of forecasts (*e.g.*, every month over several years). We incorporate a learning rate, λ, for the forecasting coefficients $\hat{\beta}_{t+1}$

which are a weighted combination of the current optimized values β_t^* and the past forecasting coefficients $\hat{\beta}_t$. Otherwise, our empirical results indicate that no past memory (*i.e.*, using only β_t^*, when $\lambda = 1$) results in and poor out-of-sample forecasts. We perform a grid search on the training set to determine the optimal $\lambda \in [0, 1]$ which represents the weighting of the optimization procedure in time period t.

$$\hat{\beta}_{t+1} = \lambda\beta_t^* + (1 - \lambda)\hat{\beta}_t$$

The resulting forecast for time period $t + 1$ and time series i with leading indicators $X_{i,t+1}$ uses only data from time period t or before and forecasts an index for a large-change:

$$\hat{g}_{i,t+1} = X_{i,t+1}\hat{\beta}_{t+1}$$

.

C. Comparison Models

The proposed PMF model is compared to several other models which incorporate the same multivariate leading indicators. The benchmark comparison model is Ordinary Least Squares (OLS) which we consider least suited to forecasting large changes. Other models appealing for large-change forecasting are also implemented. Power Loss models differ from squared error (*i.e.*, OLS) by varying the exponent of fit errors to give greater or less weight to extreme observations. Quantile regression fits the conditional quantiles (*e.g.*, median is 50%) of a given decision variable. Finally, a Bayesian technique is implemented using Markov Chain Monte Carlo (MCMC) techniques with the posterior predictive distribution (PPD) of the dependent variable. For notational simplicity, we drop the subscript t in this section. Although all model coefficients $\hat{\beta}$ are estimated on in-sample data, we choose the model metaparameters (p, τ, and quantile of the Bayesian regression) based on the PAUC loss function using out-of-sample data in the training set only. Further detail is given in the empirical application.

1) Power Loss: To estimate $\hat{\beta}$, we consider in-sample loss functions of the type

$$L = \sum_{i=1}^{n} |y_i - X_i\beta|^p$$

where $p \in [0, \infty]$. When $p = 2$, the solution solves the least squares problem, $\hat{\beta} = (X'X)^{-1}X'Y$, and the forecast $\hat{Y}$ is equal to the conditional mean, however, that interpretation is sacrificed here. Theoretically,

as $p \to 0$, the loss is 0 when all $y_i = X_i\hat{\beta}$ (*i.e.,* perfect classification) and as $p \to \infty$, the loss is equal to the maximal observational loss over i. We select

$$\hat{\beta} = \arg\min_{\beta} L$$

for each p and use the results of a grid search on training data to determine the best p for out-of-sample forecasting. We expect that large values of p should perform well in-sample if there was only one large change since the prediction will minimize the maximal distance between y_i and $X_i\hat{\beta}$ over all i. Lower values of p give increasingly less weight to the maximal observational loss (*e.g.,* $p = 0.5$ penalizes each forecast by the square root of its distance to y_i).

Although we consider many power loss models for each p, we select the best power loss model with p^* according to the PAUC loss function on the training set. The resulting model with p^* is then evaluated on the test set.

2) Quantile Regression: Quantile regression estimates $\hat{\beta}$ by minimizing

$$L = (\tau - 1) \sum_{\{I: y_i < X_i\beta\}} (X_i\beta - y_i) + (\tau) \sum_{\{I: y_i \geq X_i\beta\}} (y_i - X_i\beta)$$

where $\tau \in [0, 1]$ and represents the τ^{th} quantile. We select

$$\hat{\beta} = \arg\min_{\beta} L$$

for each τ over an equally spaced grid of 101 values. When $\tau = 0.5$, the forecast $\hat{\mathbf{y}}$ is equal to the conditional median and powers loss when $p = 1$. Low and high values of τ represent extreme quantiles of conditional distribution of $\mathbf{y}$. Although there are a variety of quantile regression models for each τ, we select the best quantile regression model with τ^* according to the PAUC loss function on the training set. The resulting model with τ^* is then evaluated on the test set.

Quantile regression can also be interpreted as varying the ratio of costs of over-forecasting and under-forecasting. Quantile regression implicitly penalizes the costs of over-forecasting (when $y_i < X_i\hat{\beta}$) and under-forecasting (when $y_i \geq X_i\hat{\beta}$) by different ratios. This can be seen by setting $\tau = \frac{c_u}{c_u + c_o}$ where c_u is the cost of under-forecasting and c_o is the cost of over-forecasting. When c_u is small compared to c_o, τ represents a low quantile and $X_i\hat{\beta}$ will be small because an over-forecast is greatly penalized. In the

case of forecasting large changes, it is not clear whether the cost of over-forecasting or under-forecasting should be of more importance since the performance of PAUC depends on the magnitude and relative rank of the forecasts. In our empirical study, we seek to determine whether quantiles aligned with higher costs of over-forecasting perform better for PAUC because incorrect over-forecasts increase the false positive rate and therefore, decrease PAUC.

3) Bayesian Regression: One advantage of Bayesian estimation is that we can generate thousands of different forecasts for a single observation y_i and subsequently, analyze the distribution of these generated forecasts (synthetic data). From this distribution, we can select a quantile of the generated forecasts to forecast a large change. In the results section, we investigate whether forecasts based on quantiles perform better for MBE. Synthetic data models capture the underlying fit of the data and allow us to generate replicates of "fake data" using MCMC samples of the regression coefficients and error variance. So, it is possible to create thousands of artificial values for each y_i. The resulting synthetic data mimics the same distribution as y_i (*i.e.,* to include variation) because the data is generated conditional on $y_i|X_i, \beta, \sigma^2$ in the Bayesian model.

For the Bayesian regression, we use the same regression equation as OLS, $y_i = X_i\beta + \epsilon_i$, but place a diffuse but proper multivariate normal prior on β with mean zero and a block diagonal covariance matrix. We assume ϵ_i is independent and identically distributed for each observation and drawn from a normal distribution with mean zero and constant variance σ^2. For the prior of σ^2, we assume an Inverse-Wishart prior with an mean of zero and a degree of belief parameter of 1. We use MCMC techniques to sample draws of β and σ^2 from their resulting posterior distributions.

To generate the synthetic data, we use 1,000 posterior samples for each parameter (β, σ^2) after a burn-in of 1,000 samples. Since β is a k-dimensional vector, 1,000 samples are generated for each component which generates a k by 1,000 matrix. Values of y_i are generated 1,000 times for each i using the available samples. Then, those values are rank ordered and the appropriate quantiles are selected. The result is that the forecasted quantiles are taken on synthetic data generated from the conditional distribution of y_i (given X_i and parameter samples). Finally, we use a grid search of the empirical quantiles to select the optimal quantile for out-of-sample forecasting. The intuition is that forecasted quantiles other than the posterior mean or median may perform better for forecasting exceptional behavior.

Although there are a variety of Bayesian regression models for each quantile, we select the best quantile according to the PAUC loss function on the training set. The resulting model is then evaluated on the test

set.

IV. Empirical Application

A. Data Source

The data used in this paper are monthly crime counts by census tract from Pittsburgh, Pennsylvania. The dependent variable is the count of serious violent crimes (homicide, rape, robbery, and aggravated assault) while the 12 leading indicators are one, two, three, and four month time lags of illicit drug 911 calls for service, shots-fired 911 calls for service, and offense reports of simple assaults (Cohen et al., 2007; Gorr, 2009). The data span January 1990 through December 2001 across 175 census tracts with 24,500 observations available out of 25,200 after dropping the beginning four month's observations used for time-lagged variables. For notation, we define $\mathbf{y}$ as the vector of violent crimes and X as the 12 column matrix of leading indicators. Table I shows the summary statistics for our data. Tract 404 represents a randomly selected low-crime area while tract 1115 is a random high-crime area. Besides overall performance of the computational experiment, we report forecast performance and gold-standard points for these two arbitrarily-chosen areas in the results section. Note that all crime counts are relatively low for monthly crime time series by census tract in Pittsburgh, making it challenging to obtain high forecast accuracy of any kind.

TABLE I
Summary Statistics of Crime Data

	Violent Crimes	Drugs	Shots	Assaults	Tract 404	Tract 1115
Min	0	0	0	0	0	2
Median	0	0	1	3	0	7
Mean	1.2	1.9	1.6	3.8	0.6	7.7
Std. Dev.	1.9	4.3	3.3	4.2	0.9	3.5
Max	29	71	50	42	4	20

B. Gold Standard Policy

We employed a standardization procedure to define gold standard large changes in violent crimes (chosen to be about three percent of all census tracts) in accordance with Gorr (2009). In each census tract, the number of violent crimes were standardized according to their past smoothed mean and variance to account for the sizable time trends in the multiple-year data. The top five standardized values across all census tracts were labeled large changes each month as a gold standard policy defining positives.

In more detail, we perform a standardization procedure on each time series (*i.e.*, census tract) which shifts and rescales the current actual value in time t, y_t by its smoothed mean m_t and variance v_t, respectively. A low smoothing constant was used to allow the estimated mean to drift with the time series, but not to change appreciably from month to month. Smoothed means tend to yield data not over dispersed so that the Poisson assumption is valid. Thus we initialize values and assume $m_t = v_t$ from a Poisson distribution assumption since the number of violent crimes follows a count distribution. For each time period, we set our standardized value $y_t^* = \frac{y_t - m_t}{\sqrt{v_t}}$ and updated the estimates of the smoothed mean and variance by the current actual value. Once all values in each time series are standardized, we select the five largest values (three percent) for each month's cross-section of census tracts to define large increases in crime.

C. Rolling Horizons

Crime forecasting for deployment of police resources needs only one-step-ahead forecasts (one-month- ahead in this case). Urban police resources are highly mobile and easily and commonly reassigned or targeted. Also, most modern urban police departments have monthly review and planning meetings by sub- region (zone or precinct) so that one-month-ahead forecasts are needed. While forecasting large decreases in crime is perhaps useful for pulling police resources away from areas, the primary interest is crime prevention and forecasting large increases. A separate study, of the same magnitude and effort for large decreases, would be necessary for large decreases but is not conducted in this paper. A growing empirical literature shows that crime prevention in this setting has at least moderate success (*e.g.*, Braga et al. 2012). We reestimate our models every month after forecasts are produced. All data up to time period t is used to forecast time period $t + 1$. Table II describes the conceptual setup.

TABLE II
PARAMETERS OPTIMIZED ACROSS DATA SETS

Data Set	Initialization Set	Training Set		Test Set	
Data Used	In-Sample	In-Sample	Out-of-Sample	In-Sample	Out-of-Sample
Type	Parameter	Parameter	Metaparameter	Parameter	Not Applicable
PMF Model	β	β	λ	β	
Power Loss	β	β	p	β	
Quantile Regression	β	β	τ	β	
Bayesian Regression	β	β	Quantile	β	
OLS	β	β	Not Applicable	β	

D. Forecast Evaluation

The PMF Model forecasts an ordinal index, $\hat{\mathrm{g}}$, where larger index values indicate that census tracts that are more likely to have large changes next month. Such a scale-invariant index is sufficient for use in decision rules for signalling large-change forecasts. On the other hand, all competing methods' have magnitude forecasts estimating demand and therefore forecasts need to be standardized according to their past mean and variance. Since standardization is not scale-invariant, this transformation changes each method's PAUC. If no standardization of magnitude-based methods were performed, competing methods would always forecast large changes for the most violent census tracts because their magnitudes are higher. Empirically, standardization improved PAUC performance for magnitude-based methods.

Our data consisted of 175 time series with 136 months each. We used 44 months of data to initialize each method in the initialization set, the next 24 months for the training set, the next 12 months to burn in the exponential smoothing procedure for the gold standard policy, and the final 60 months for evaluation in the test set. All model parameters were chosen via grid search in the 24 months of the training set. Grid searches were performed over 101 values of the learning rate λ for the PAUC Maximization Forecast model, the power p for the power loss model, the quantile τ for quantile regression, and the quantile of the PPD for the Bayesian model. Rolling horizon forecasting on out-of-sample data was performed in the training set to select the best values of these metaparameters. The last 60 months of data in the test set was used for evaluation of out-of-sample forecasting and the results are presented in the next section. Forecasts were made by each model with a rolling horizon of one-month, consistent with decision making in crime forecasting. All models were re-estimated at every forecast origin, however, only forecasts of the magnitude-based methods were standardized and their coefficients were not adjusted.

V. Results

We summarize all 60 months of out-of-sample forecasts for each model with a single ROC curve. Each ROC curve represents 10,500 forecasts (60 months times 175 census tracts) to forecast 300 large changes. The 300 large changes consisted of the top 5 large changes for each month. Smoothed ROC curves up to a false positive rate of 20% (PAUC's relevant range) are shown in Figure 1 where the PMF model is seen to strongly dominates competing methods. This means that the proposed model forecasts the most gold standard large changes at any given false positive rate. For example, this figure shows that the proposed model has double the number of true positives at a false positive rate of 10%.

Fig. 1. Smoothed ROC Curves Over Five Years in Test Set

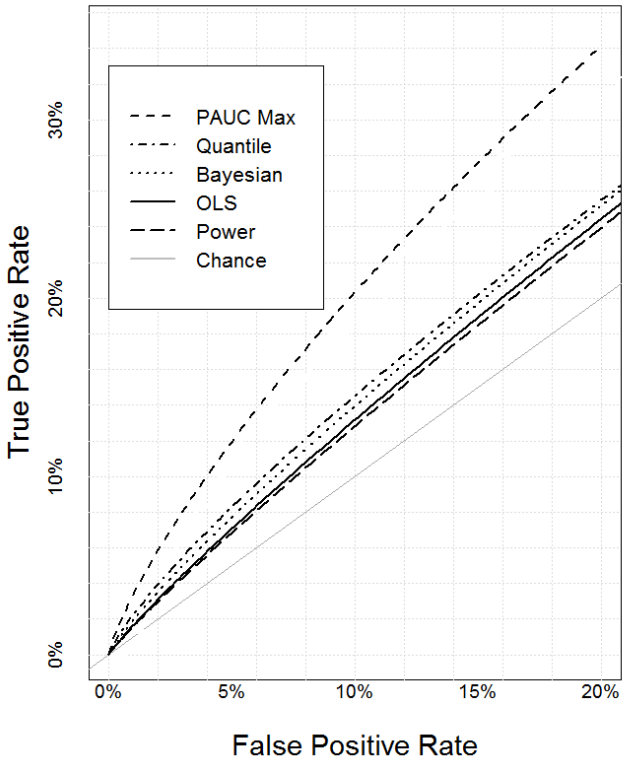


Table III presents corresponding results for each model. In order to test for statistical significance between two correlated ROC Curves, we used 1,000 bootstrap samples with the R package pROC (Robin et al., 2012). We compared the PAUC of the proposed model to other models. Each bootstrap sample randomly selects the same number of large changes and observations as the original data, and 1,000 PAUCs are generated. Differences are calculated and their standard deviation divides the PAUC difference of two original ROC Curves to generate the test statistic found in the table. The PAUC Maximization Forecast model had a statistically greater PAUC than all other models at the 5% alpha level in the test set.

TABLE III
FIVE-YEAR ROC CURVE FOR THE TEST SET

Model	PAUC	Test Stat.	One Sided p-value	Percent Improvement
PAUC Maximization	.0359	-	-	
Quantile Regression	.0297	1.74	.041	20.5%
Power Loss	.0285	2.05	.020	25.7%
Bayesian Method	.0283	2.04	.021	26.8%
OLS with Leading Indicators	.0266	2.51	.006	35.0%

Using PAUC methodology, quantile regression models with low quantiles forecasted large changes more

accurately than models with high quantiles in the training set. Specifically, we found that for forecasting the top three percent of large changes, quantile regression did the best in the training set when it implicitly assigned a cost of over-forecasting 12 times greater than a cost of under-forecasting. This represented the $\tau^* = .08$. Additionally, a Bayesian method based on empirical quantiles of sampled synthetic data and the power loss method had a higher PAUC than OLS, but not statistically so. Power Loss had poorer performance than OLS at higher powers in the training set. $p^* = .68$ was selected for use in the test set since it had the maximum PAUC in the training set.

We also evaluated forecasts with the continuous measure of correlation to the gold standard vector (*i.e.*, 1 if large change, 0 otherwise) in the test set. Table IV shows that all correlations were significantly different than zero at the 5% alpha level and the PMF Model had the highest correlation. Correlations are all positive indicating that higher forecasts are more closely associated with large changes. However, correlations are very low because 97% of the gold standard vector was zero (*i.e.*, not a large change) as defined in our gold standard policy.

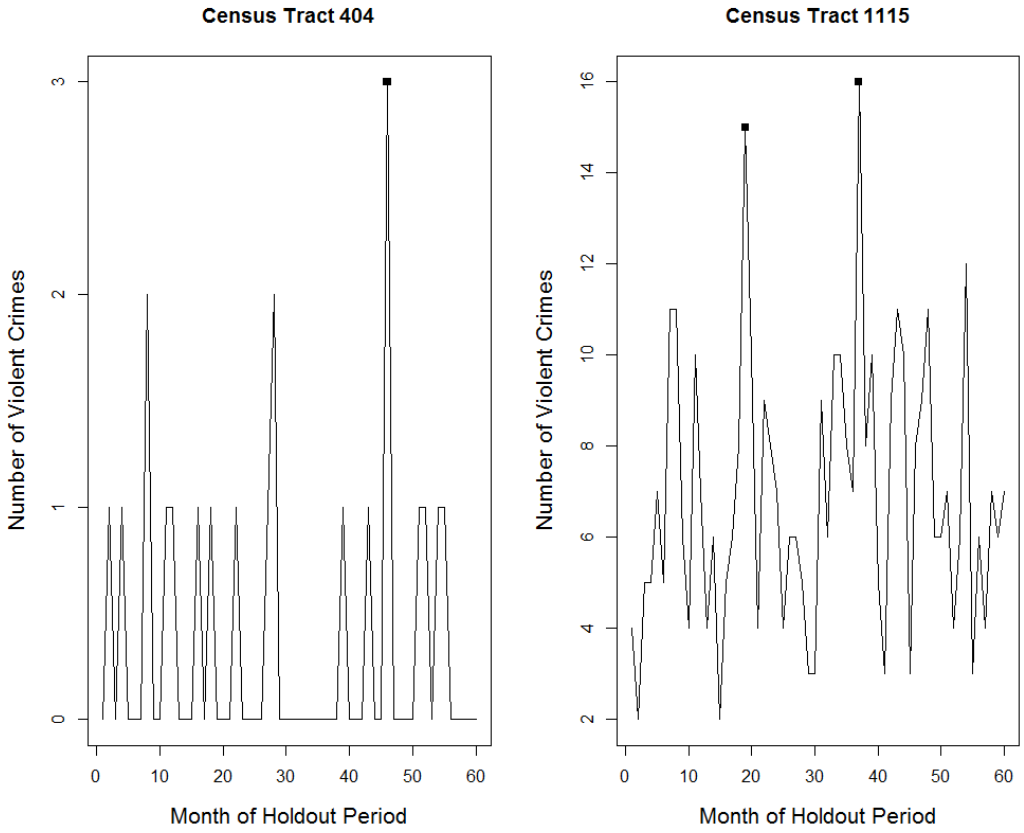
TABLE IV
FIVE-YEAR CORRELATION OF GOLD STANDARD TO FORECASTS, TEST SET

Model	Correlation	p-value
PAUC Maximization	5.51%	0.011
Quantile Regression	2.41%	0.009
Power Loss	2.50%	0.014
Bayesian Method	2.56%	0.011
OLS with Leading Indicators	2.48%	0.000

Over the five-year test set, we show the actual number of violent crimes for the two sample census tracts previously summarized in Table I, in Figure 2. The black point markers in Figure 2 are large-increases positives from our gold standard policy. For Census tract 404, the PMF Model's highest index forecast (0.102) occurred in month 46 at the only large increase and therefore, our proposed method results in zero false positives for a decision rule using a cutoff of 0.102. However, OLS's standardized forecast during month 46 (0.113) is the 11th highest forecast among the five years. Therefore, if a manager used a cutoff of .113 for OLS, it would result in one true positive and ten false positives. In sum, for the Census tract 404, the PMF model outperforms OLS in terms of forecasting the large increase. On the other hand, Census tract 1115 is a high-crime area and the two large changes identified occurred in months 19 and 37, and had 15 and 16 violent crimes, respectively. The PMF Model forecasted the two large increases as the 29th and 30th highest forecasts in Census tract 1115, respectively. OLS's standardized forecasts

placed these two large increases as the 21st and 40th highest forecasts for Census Tract 1115. Therefore, although OLS had less false positives for forecasting the first large increase, the PMF Model had less false positives (28) for forecasting both large changes compared to OLS (38).

Fig. 2. Actual Violent Crimes for Two Census Tracts in the Test Set



VI. Conclusion

This is the first paper to use partial area under the curve (PAUC) from receiver operating characteristics (ROC) analysis as the basis for a loss function to estimate forecast model parameters (1 – PAUC is the loss function used). PAUC tunes forecast models to the tails of product or service demand distributions, thereby substantially increasing large-change forecast accuracy (with statistical significance) over models estimated using MSE or other central tendency measures as the loss function. The PAUC-based model is also statistically and substantially more accurate than other comparison models that can be tuned or used for large-change forecasting. The forecast model of this paper is multivariate with leading indicators able to forecast large changes when there are large changes in the lagged indicators. The same loss function and optimization methods can be applied to any forecast model, including the complex univariate models

found superior to simple univariate models for large change forecasting by Gorr & Schneider (2013). Our findings confirm previous research which says that models which include loss functions for the accuracy error metric desired perform the best (Parker, 2011).

Accurate large-change forecasting is the key to proactive management-by-exception (MBE) for managing product inventories and marketing programs. The MBE principle states that only variances (exceptions) should be brought to the attention of managers, with business-as-usual decisions handled by staff using standard procedures. Thus the decision to be made for MBE is binary, whether or not a forecast is large enough to bring to management for further analysis and possible interventions. Such decisions are made using decision rules analogous to hypothesis testing, however, it is necessary to select decision rule thresholds for MBE and ROC enables one to estimate and compute an optimal decision rule (and corresponding false positive rate).

The implications of MBE and demand forecasting is that firms should continue to use their current extrapolative forecast models for business-as-usual conditions, but also implement a large-change forecast model, such as developed in this paper. When a large-change decision rule fires for a demand time series, the business-as-usual forecast is preempted with management review of the product.