International Journal of Forecasting 31 (2015) 253–262

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

ROC-based model estimation for forecasting large changes in demand

Matthew J. Schneider[a,*], Wilpen L. Gorr[b]

[a] Samuel C. Johnson Graduate School of Management, Cornell University, Ithaca, USA
[b] H. John Heinz III College, Carnegie Mellon University, USA

ARTICLE INFO

Keywords:
Management by exception
PAUC maximization
Large changes
Forecasting exceptions
ROC curves

ABSTRACT

Forecasting for large changes in demand should benefit from an estimation that differs from that used for estimating mean behaviors. We develop a multivariate forecast model whose continuous forecasts are used as test statistics in decision rules to make binary (yes/no) forecasts for large changes in time series data. The model is fit based upon a penalty function that maximizes the partial area under the ROC curve (PAUC) along a relevant range of false positive rates, and can be used by managers who wish to take action on the small percentage of products whose demand is likely to change the most in the next time period. We apply the model to a crime dataset and compare the results to OLS, as a basis for comparisons, as well as to models that have shown themselves to be promising for large-change demand forecasting, including logistic regression, quantile regression, synthetic data from a Bayesian model, and a power loss model. Using the PAUC metric, our proposed forecasting model's out-of-sample performance shows statistical significance, a 35% improvement over OLS, and at least a 20% improvement over other competing methods. We suggest that managers with large numbers of time series (e.g., for product demand) should use our method to forecast large changes preemptively, in conjunction with magnitude-based methods for forecasting the expected demand.

1. Introduction

In general, demand forecasting is performed using 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 the forecast accuracy is acceptable are referred to in this paper as "business-as-usual". Highly disaggregated time series, for variables such as product or service demand, however, are notorious for having large changes – outliers, step jumps, turning points, etc. – that cannot be forecasted using simple extrapolative forecast models. For example, exponential smoothing forecast methods, as commonly used in demand forecasting, damp out unexpected large changes and for the most part ignore them. Thus, the time series forecasting field has long recognized the importance of handling exceptions to business-as-usual; and in particular by developing time series monitoring methods for the early detection of large changes (e.g. Brown, 1959; Trigg, 1964).

These methods provide test statistics for binary decision-making, yes/no a large change just occurred in a time series. A signal trip, a "yes", triggers human intervention and problem solving, drawing on additional data, qualitative information, experience, and a manager's power to intervene (e.g., for a large, unexpected demand decrease, with incentives to buy, an advertising campaign, a cut in

[*] Corresponding author.
E-mail addresses: mjs533@cornell.edu (M.J. Schneider), gorr@cmu.edu (W.L. Gorr).

http://dx.doi.org/10.1016/j.ijforecast.2014.07.002
0169-2070/© 2014 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.

production, etc.). This is an instance of "management by exception" (MBE), a management control system discussed in the second section of this paper. In contrast, we have inventory "service levels" being used to manage randomness under business-as-usual conditions. The service level is a percentile from the forecast distribution under stable conditions, and is used to hold the probability of a stock-out at an arbitrarily-chosen small value. It is part of a standard protocol implemented by computers and not requiring human intervention.

Time series monitoring supports reactive decision making, after large changes have occurred. However, if the forecasting models are accurate enough, it is better 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 provides a new estimation method for models which can forecast large demand changes. As the fit error measure, it uses the partial area under the curve (PAUC) of the receiver operating character (ROC) framework. PAUC bases estimation on behaviors in the tails of demand forecast distributions, while conventional fit measures (e.g., MSE) focus on the central tendency. Of course, a sensitivity to the tails of distributions is important for more accurate forecasts of large changes. Another point that is important and unique to the ROC framework is the fact that managers choose the optimal false positive rate for binary decision rules — possibly over 0.10 for important products — as a function of their preferences and the accuracy of the forecast models. The more important and critical a product is for the health of a firm, the higher the optimal false positive rate. (In contrast, science uses traditional, conservative false positive rates of 0.01 or 0.05 for theory testing.) However, the more accurate the large-change forecast model, the smaller the optimal false positive rate (Cohen, Garman, & Gorr, 2009).

The large-change models in this paper are multivariate and include demand leading indicators. These models have the ability to forecast large demand changes when leading indicators have large changes. Comparison models, which also seem promising for forecasting large changes (e.g., logistic regression), use the same multivariate specification. All comparison models make magnitude demand forecasts and are also used in binary decision rules to (1) trigger their use preemptively, replacing extrapolative forecasts, and (2) trigger MBE. The PAUC model, however, only produces an index that is correlated with demand, and so does not produce magnitude forecasts, but it has a binary forecast accuracy that is far superior to those of the comparison models for triggering MBE.

The large-change forecast implementation of MBE is as follows. (1) Managers decide which products to include in MBE (e.g., the "A" products of ABC inventory policy; Ramanathan, 2006). (2) All products are forecasted using the usual extrapolative methods and with corresponding service levels input into economic order quantity models to manage randomness under business-as-usual. (3) When a large-change binary decision rule trips, it triggers MBE and the use of a large-change magnitude forecast as part of problem solving. (4) After a time series again becomes stable, forecasting returns to extrapolation, but with adjustments if there is a new demand regime.

This paper is an extension of our past work on the application of the ROC framework to large-change forecasting, and therefore we begin by briefly reviewing past work. Gorr (2009) provided evidence that OLS leading-indicator forecasts for crime, similar to those used in this paper and with the same data, are less accurate than any of the exponential smoothing models tried when evaluated using central tendency forecast error measures (MAPE and MSE), but are the most accurate as an indicator of large changes using receiver operating characteristic (ROC) accuracy measures. Gorr and Schneider (2013) compared simple and complex univariate forecast models for large changes using monthly data from the M3 competition, and found that the binary, yes/no forecast accuracy of large-change forecasts of complex univariate models is significantly better than that of simple univariate models. Apparently, complex univariate models have functional forms that are sensitive to subtle indications of rapidly changing trends. The binary forecast error measure used was PAUC.

Parker (2011) went one step further and showed that the classification performance over seven measures of classification (which included AUC but not PAUC) is best when picking performance measures that are also used as loss functions for estimation. In line with this result, this paper provides evidence that the proposed multivariate forecast model estimated by minimizing 1-PAUC is much more accurate, as assessed by PAUC for large-change binary forecasts, than models estimated using a central-tendency-based loss function (MSE). To the best of our knowledge, there has been no previous empirical research in time series forecasting using a ROC-based loss function for model estimation.

Section 2 provides the motivation and background for the paper's new estimation method, with an overview of MBE, and Section 3 reviews ROC applied to large-change forecasting. Section 4 develops the ROC-based method of parameter estimation for a multivariate, leading indicator forecast model, and develops comparison models. Section 5 describes the time series data (monthly crime counts by census tract for a city) used to calibrate the model and the rolling horizon forecast experiment. Section 6 presents results comparing alternate forecasting models, with significance testing. Finally, Section 7 concludes the paper with a summary and suggestions for future work.

2. Management by exception for demand forecasting

Management by exception (MBE), one of the oldest forms of a management control system (Ricketts & Nelson, 1987; Taylor, 1911), lays down 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 analytical resources, expertise, and power for emerging problems or opportunities. One type of variance is a large change in the demand for products or services (Gorr, 2009; Gorr & Schneider, 2013; West, Harrison, & Mignon, 1985). Demand is only partially affected by an organization's efforts, given competition in the market place, the limits of marketing programs, and changing consumer tastes. Hence, large changes in demand are an important source of variance for triggering MBE reports to production and marketing managers.

This paper uses time series data on the occurrence of crime within city neighborhoods (census tracts) and the corresponding demand for police services — law enforcement and the prevention of crimes. It is common for police to use simple rules, based on crime counts, to screen the neighborhoods of a jurisdiction for large increases in crime and for possible interventions. Neighborhoods have baseline crime levels, and when there is a sizable increase in crime it becomes newsworthy, spreads through informal social networks and the news media, and greatly increases fear (e.g., Skogan, 1986; Skogan & Maxfield, 1981). Crime analysts working within police departments diagnose large crime increases by drilling down into detailed crime reports, studying patterns on crime maps, talking with field officers, using information from confidential informants, conducting surveillance, etc. They and their precinct commander craft interventions that are intended to prevent crime increases. The criminology literature has identified over 60 different steps and approaches for solving crime problems (Clarke & Eck, 2009). Alternative solutions can include increased numbers of targeted patrols in forecasted hot spot areas, foot patrols, the use of confidential informants, educating place managers, modifications to built environments, zero tolerance to all infractions of the law, building partnerships with citizens, etc.

An analogy to MBE in organizations is the public-health screening of populations for disease (e.g. Banez et al., 2003; Elmore et al., 2002; Pepe, 2004). For the sake of economic feasibility, screening uses relatively inexpensive (but therefore fallible) tests for the detection of disease in individuals. For an individual who is flagged as possibly having the disease (say, cancer), next is diagnosis: biopsy and detailed imagery to determine whether the person actually has the disease, and if so, the disease's stage. Finally, treatments (types and dosages) are designed for the stage of the disease, determined in part by the characteristics (weight, age, general health, etc.) and wishes of the individual. Finally, treatments are administered and adjusted as needed.

3. ROC analysis

There needs to be an external determination as to when a time series has data points that are considered to be large change points. Using ROC terminology, such a data point is called a "positive", and is determined by a "gold standard". All other time periods are "negatives". For example, for prostate and breast cancer screening, the gold standard is a biopsy, examining sample tissue under a microscope. Biopsies are not infallible, but are much more accurate than screening tests such as PSA levels in blood samples for detecting prostate cancer.

Of course, the demand forecasting problem does not have gold standard tests, such as biopsies, for large change points. Instead, managers must use judgment to determine which changes are large enough to be worth the cost of diagnosis, analysis, and possible action. Gorr (2009) used a gold standard policy that the top small percentage of large changes in standardized time series data were considered to be positives, and reasoned that police officials have the means of making such judgments (e.g., police want 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 it was more precise, it was judged too costly for implementation by police.

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). The 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 the false positive rate, FPR $=$ number of false positives/number of negatives. The complements of these statistics are the false negative rate and the 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 the true positive rate versus the false positive rate for all possible decision rule thresholds, and is an increasing function with a decreasing slope between $(0, 0)$ and $(1, 1)$. The higher the ROC curve for a model, the more accurate the binary decision model. Thus, an overall measure of the performance of a monitoring or forecast model is the area under the ROC curve (AUC), which ranges between 0 and 1. In practice, the partial area under the ROC curve (PAUC) is better. This is the area for a restricted range of false positive rates, often from 0 up to 10% or 20%, because the cost of processing false positives exceeding those rates is generally excessive and/or beyond available resources. See Fig. 1 in Section 6 for example ROC curves.

Empirical research uses traditional values, such as 1% or 5%, for false positive rates (Type I errors) that determine decision rule thresholds from normal or t-distribution tables. This practice implements a conservative view of the acceptance of evidence on new theories. Business, however, needs to determine thresholds that will 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 Cohen et al., 2009; Metz, 1978). The optimal false positive rate is determined by finding the point at which a derived straight line is tangential 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 to the utility of avoiding a false positive. Generally, the utility ratio is determined through expert judgment, and substitutes for objective cost/benefit assessments. For example, Pittsburgh police officials judged that it is 10 times more

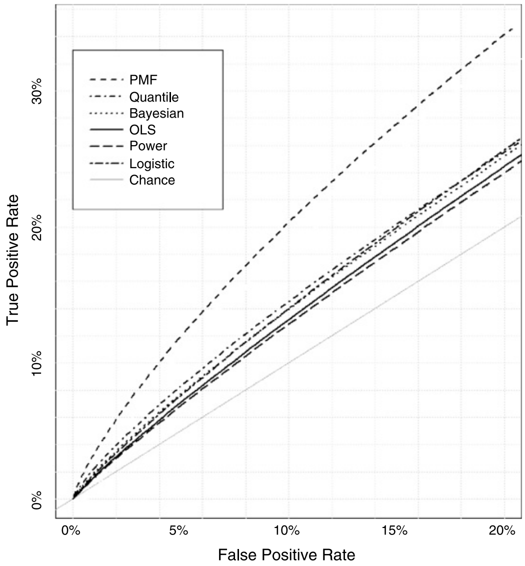

Fig. 1. Smoothed ROC curves over five years in the test set.

important to avoid a false negative than a false positive when monitoring serious violent crimes for large increases, and this led to a 15% false positive rate being considered optimal for time series monitoring (Cohen et al., 2009). Likewise, population screening for prostate and breast cancers has false positive rates in the range of roughly 10% to 15% for most parts of the world (e.g. Banez et al., 2003; Elmore et al., 2002). In both the 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) outweigh the costs of processing false positives. "A" items from ABC inventory analysis are likely to be similar in terms of importance or consequence.

4. 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 $i = 1, \ldots, n$, 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). The 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 developed formally, 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 that is used to estimate $\hat{\beta}$ and not used in forecasting. The training set is the set of data that is used for model selection based on pairwise comparisons of out-of-sample results in the training set only. The test set is the set of data that is 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 for forecasting data at time $t+1$. We define in-sample data as the data up to time period t that are 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 2.

4.1. 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, **y**, into a binary gold standard vector, **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 the decision-rule thresholds discussed below, for standardized values of the dependent variable, y^*, in our empirical application. Standardization matches the police criterion of equity in 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 standardized crime counts mean that any area with a relatively large increase in crime can get extra police resources, regardless of the crime scale. Section 5 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 jth decision rule threshold, $j = 1, 2, \ldots$ (1 plus the number of unique $\hat{y}_i$s), corresponding to selected constants c_j which divide the ranked $\hat{y}_i$s. Then, a decision rule is defined under the jth threshold and ith observation, where $\mathbf{1}_{\hat{y}_i > c_j}$ outputs a 1 if $\hat{y}_i > c_j$ and 0 otherwise, where

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

The resulting collection of TPRs and FPRs for all thresholds are

$$\mathrm{TPR}_j(\hat{\beta}, X, \mathbf{g}) = \left(\sum_{i=1}^{n} \mathbf{1}_{\mathrm{DR}_{i,j}=g_i=1}\right) \Big/ \left(\sum_{i=1}^{n} g_i\right) \tag{4}$$

$$\mathrm{FPR}_j(\hat{\beta}, X, \mathbf{g}) = \left(\sum_{i=1}^{n} \mathbf{1}_{\mathrm{DR}_{i,j}-g_i=1}\right) \Big/ \left(n - \sum_{i=1}^{n} g_i\right). \tag{5}$$

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

$$\mathrm{AUC}(\hat{\beta}, X, \mathbf{g}) = \frac{1}{2}\sum_{j=2}^{U}(\mathrm{FPR}_j - \mathrm{FPR}_{j-1})(\mathrm{TPR}_j + \mathrm{TPR}_{j-1}) \tag{6}$$

$$\mathrm{PAUC}(\hat{\beta}, X, \mathbf{g}) = \frac{1}{2} \sum_{j=2}^{\{j:\mathrm{FPR}_j \le 0.20\}} (\mathrm{FPR}_j - \mathrm{FPR}_{j-1})(\mathrm{TPR}_j + \mathrm{TPR}_{j-1}). \tag{7}$$

1-PAUC is the loss function proposed in this paper for estimating the forecasting models used to implement MBE. Of course, PAUC maximization is equivalent.

Explicit solutions for maximizing AUC exist under the assumption of normality (Su & Liu, 1993), but more recent research has found that models which are tuned for AUC do not perform well for PAUC (Pepe, Cai, & Longton, 2006; Ricamato & Tortorella, 2010) because the ROC curves for alternative methods can cross each other. PAUC maximization was studied recently in biostatistics for classifying patients as diseased or non-diseased using approximations of the PAUC function with wrapper algorithms (Wang & Chang, 2011) or boosting (Komori & Eguchi, 2010). Other biometrics papers have proposed new PAUC maximization algorithms using a weighted cost function with AUC and a normality assumption (Hsu & Hsueh, 2013). As such, the PAUC maximization papers have concentrated on identifying the distributional differences between diseased and non-diseased populations, whereas we use multiple time series as the dependent variable, which poses 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 being positives). Our application also differs structurally, since large changes can occur at any time period in any time series.

4.2. PAUC maximization forecast model

In this section, we detail the estimation procedure used to generate forecasts for our proposed PAUC Maximization Forecast (PMF) model. Note that because the PAUC-based loss function does not reference dependent variable values directly in its training, PMF only estimates an index for the dependent variable, not its expected value. As an 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 vector 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 ensures consistency in the parameter estimates from period to period. Finally, the proposed model forecasts large changes in time period $t+1$, and the model is re-estimated for each successive set of forecasts as time moves forward.

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

$$L_t = 1 - \mathrm{PAUC}_t(g_t, X_t, \beta_t) \tag{8}$$

and select

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

which minimizes L_t or, equivalently, maximizes PAUC_t, according to the optimization procedure described below. PAUC_t^* is calculated using only functions of the in-sample vector $X_t\beta_t^*$. Let n_t be the number of unique values of $X_t\beta_t^*$. Then, all unique cutoff values $c_1, c_2, \ldots, c_{(1+n_t)}$ are chosen by first sorting across values within $X_t\beta_t^*$, then averaging consecutive values which are not identical. These cutoff values represent various managerial decisions j for predicting large changes, $\mathrm{DR}_{t,j} = \mathbf{1}_{X_t\beta_t^* > c_j}$. Then, PAUC_t^* is estimated by inputting the vectors $\mathrm{DR}_{t,1}, \mathrm{DR}_{t,2}, \ldots, \mathrm{DR}_{t,(1+n_t)}$ 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 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 then run the optimization for a maximum of 500 iterations or until convergence. After L_t has converged 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 so as to maximize PAUC_t.

Our early research using optim resulted in inconsistent parameter estimates from month to month, due to the sparsity of our data. 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 poor out-of-sample forecasts. We perform a grid search on the training set to determine the optimal $\lambda \in [0, 1]$ that 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:

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

4.3. 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 suitable for forecasting large changes. Other models that are appealing for large-change forecasting are also implemented, including the logistic regression. Power loss models differ from squared error models (i.e., OLS) in varying the exponent of the fit errors to give more or less weight to extreme observations. Quantile regression fits the conditional quantiles (e.g., the median is 50%) of a given decision variable. A Bayesian technique is implemented using Markov Chain Monte Carlo (MCMC) techniques with the posterior predictive distribution (PPD) of the dependent variable. Finally, a logistic regression estimates the probabilities of a large-change forecast directly, and is used in decision rules with threshold probabilities in order to provide binary forecasts.

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 the 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.

4.3.1. Logistic regression

To estimate $\hat{\beta}$, we consider the log likelihood function (Casella & Berger, 2002),

$$L = \sum_{i=1}^{n} \left[\log(1 - F_i) + g_i \log\left(\frac{F_i}{1 - F_i}\right)\right],$$

where $F_i = F(X_i\beta) = \frac{\exp(X_i\beta)}{1+\exp(X_i\beta)} = \rho_i$ and $g_i \sim$ Bernoulli (ρ_i). The values of $\hat{\beta}$ are chosen to maximize the log likelihood. The logistic regression is a specific case of a generalized linear model which is useful for modeling binary data.

4.3.2. 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, this 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 the 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 from 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.

4.3.3. 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 τth 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 the 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 the costs of over- 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 penalized greatly. In the case of forecasting large changes, it is not clear whether the cost of over- 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, thus decreasing PAUC.

4.3.4. Bayesian regression

One advantage of Bayesian estimation is that we can generate thousands of different forecasts for a single observation y_i, then analyze the distribution of these generated forecasts (synthetic data). From this distribution, we can select a quantile of the generated forecasts to forecast large changes. 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 the error variance. Thus, it is possible to create thousands of artificial values for each y_i. The resulting synthetic dataset mimics the distribution of y_i (i.e., to include variation) because the data are generated conditional on X_i, β and σ^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 that 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 a mean of zero and a degree of belief parameter of one. We use MCMC techniques to sample draws of β and σ^2 from their resulting posterior distributions.

To generate the synthetic data, we use 1000 posterior samples for each parameter (β, σ^2) after a burn-in period of 1000 samples. Since β is a k-dimensional vector, 1000 samples are generated for each component, which generates a $k \times 1000$ matrix. Values of y_i are generated 1000 times for each i using the available samples. Then,

Table 1
Summary statistics of the crime dataset.

	Violent crimes	Drugs	Shots	Assaults	Tract 404	Tract 1115	Tract 1405
Min	0	0	0	0	0	2	0
Median	0	0	1	3	0	7	2
Mean	1.2	1.9	1.6	3.8	0.6	7.7	2.7
Std. Dev.	1.9	4.3	3.3	4.2	0.9	3.5	1.9
Max	29	71	50	42	4	20	9

Table 2
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	
Logistic regression	β	β	Not applicable	β	

those values are ordered by rankings 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 expectation is that forecasted quantiles other than the posterior mean or median may perform better for forecasting exceptional behaviors.

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.

5. Empirical application

5.1. 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, Gorr, & Olligschlaeger, 2007; Gorr, 2009). The data span the period from January 1990 to December 2001 across 175 census tracts, with 24,500 observations available out of 25,200 after dropping the first four months' 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 1 shows the summary statistics for the data. Tract 404 represents a randomly selected low-crime area, while tract 1115 is a random high-crime area. Tract 1405 is a medium-crime area that exhibits an early step jump. In addition to the overall performance of the computational experiment, we also report forecast performances and gold standard points for these three randomly-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 a high forecast accuracy.

5.2. Gold standard policy

We employed a standardization procedure to define the gold standard for large change points in serious violent crimes (chosen to be about 3% of all census tracts), in accordance with Gorr (2009), in order to provide equity in crime prevention across high and low crime-level neighborhoods. In each census tract, the number of violent crimes was standardized based on past smoothed means and variances so as to remove all scale from the time series. The top five standardized values across all census tracts were labeled large changes each month as the gold standard policy defined 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 at time t, y_t, by its smoothed mean m_t and variance v_t, respectively. A low smoothing constant was used in order 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 that are not over-dispersed, so that the Poisson assumption is valid. Thus, we initialize values and assume $m_t = v_t$ from the Poisson distribution assumption, because 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 update the estimates of the smoothed mean and variance by the current actual value. Once all of the values in each time series have been standardized, we select the five largest values (3%) for each month's cross-section of census tracts to define points of large increases in crime.

5.3. Rolling horizons

Crime forecasting for the deployment of police resources needs only one-step-ahead forecasts (one-month-ahead forecasts in this case). Urban police resources are

Table 3
Five-year ROC curve out-of-sample forecast accuracy.

Model	PAUC	Test stat.	One-sided p-value	Percent improvement
PMF	0.0359	–	–	–
Quantile regression	0.0297	1.74	0.041	20.5%
Power loss	0.0285	2.05	0.020	25.7%
Bayesian method	0.0283	2.04	0.021	26.8%
Logistic regression	0.0275	2.08	0.019	30.5%
OLS with leading indicators	0.0266	2.51	0.006	35.0%

highly mobile and are 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 for that purpose as well. While forecasting large decreases in crime is perhaps useful for removing police resources from areas, the police's primary interest is in crime prevention and the forecasting of large increases. A separate study, of the same magnitude and effort as for large increases, would be necessary for large decreases, but is not conducted in this paper. A growing body of empirical literature shows that crime prevention in this setting has at least a moderate success (e.g., Braga, Papachristos, & Hureau, 2014). We reestimate our models every month after forecasts have been produced. All data up to time period t are used to forecast time period $t + 1$. Table 2 describes the conceptual setup.

5.4. Forecast evaluation

The PMF model forecasts an ordinal index, $\hat{g}$, where large index values indicate census tracts that are more likely to experience 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 the demand, and therefore their forecasts need to be standardized. 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, thereby ignoring the equity criterion for allocating resources. Empirically, standardization improved the PAUC performance for magnitude-based methods.

Our dataset consists of 175 time series with 136 months each. We use 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 a grid search in the 24 months of the training set. Grid searches were performed over 101 values of the learning rate λ for the PMF 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 was performed on out-of-sample data in the training set to select the best values of these metaparameters. The last 60 months of data in the test set were used for the evaluation of out-of-sample forecasting, and the results are presented in the next section. Forecasts were produced 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.

6. Results

We summarize all 60 months of out-of-sample forecasts for each model using a single ROC curve. Each ROC curve represents 10,500 forecasts (60 months × 175 census tracts) for forecasting 300 large changes. The 300 large changes consist of the top five largest change points for each month. Smoothed ROC curves up to a false positive rate of 20% (PAUC's relevant range) are shown in Fig. 1, where the PMF model is seen to dominate competing methods strongly. This means that the proposed model forecasts the most gold standard large changes at any given false positive rate.

Table 3 presents corresponding results for each model. In order to test for statistical significance between two correlated ROC curves, we used 1000 bootstrap samples with the R package pROC (Robin et al., 2011). 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 1000 PAUCs are generated. Differences are calculated, and their standard deviation divides the PAUC differences of two original ROC curves to generate the test statistic found in the table. The PMF model was found to have a statistically greater PAUC than all other models at the 5% alpha level in the test set.

Using the PAUC methodology, logistic regression has a performance similar to that of OLS. The reason for the logistic regression's poor performance is probably that large changes are rare events and the logistic regression is known to perform poorly in such cases (King & Zeng, 2001). 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 3% of large changes, quantile regression did best in the training set when it implicitly assigned a cost of over-forecasting that was 12 times greater than the cost of under-forecasting. This represented $\tau^* = 0.08$. In addition, a Bayesian method based on empirical quantiles of sampled synthetic data and the power loss method had higher PAUCs than OLS, but not statistically so. Power loss had a poorer performance than OLS at higher powers in the training set. $p^* = 0.68$ was selected for use in the test set, since it had the maximum PAUC in the training set.

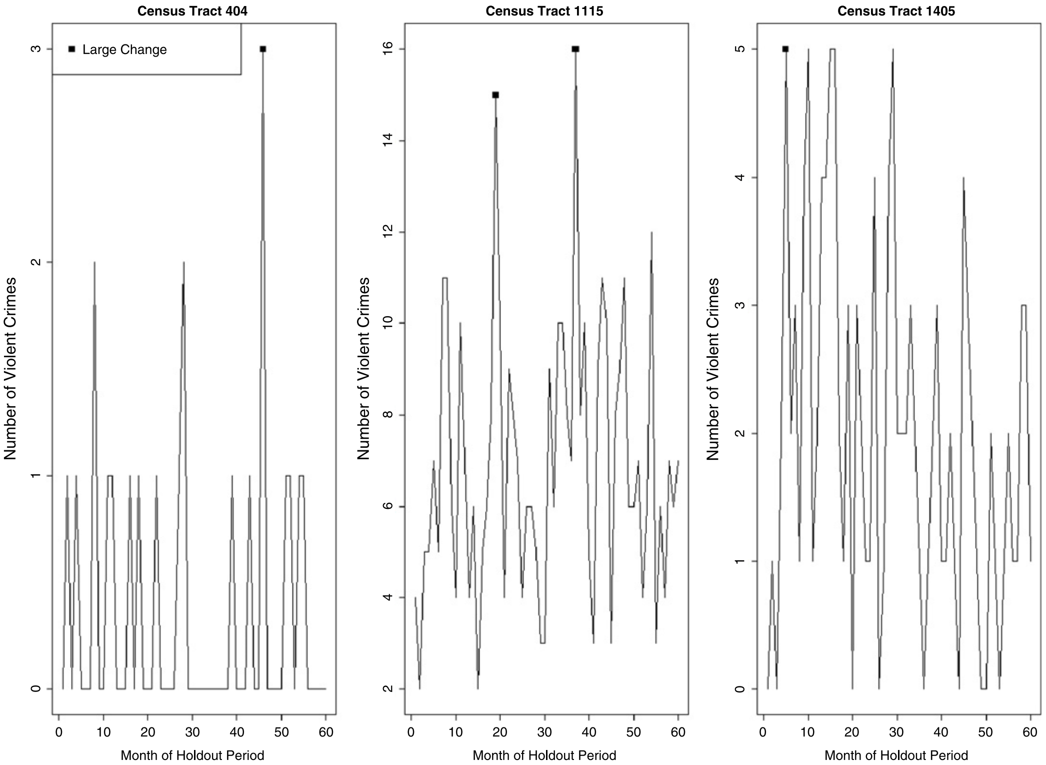

Fig. 2. Actual violent crimes for two census tracts in the test set.

Over the five-year test set, we show in Fig. 2 the actual numbers of violent crimes for the sample census tracts previously summarized in Table 1. The first two time series have outliers and the last time series has a step jump. (For example, Cohen et al., 2007, had police experts classify highly disaggregated monthly crime data, over-sampled from the 20% highest crime areas of two cities including Pittsburgh, into business-as-usual versus large changes that were worthy of analysis and possible intervention beyond normal police practices. Of the 1620 data points evaluated, 7.6% had large changes: 2.7% had step jumps and 4.9% had outliers.) The black point markers in Fig. 2 are large-increase 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 0.113 for OLS, it would result in one true positive and ten false positives. In summary, for 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 fewer false positives for forecasting the first large increase, the PMF model had fewer false positives (28) for forecasting both large changes compared to OLS (38).

7. Conclusion

This is the first paper to use the partial area under the curve (PAUC) from the receiver operating characteristics (ROC) framework as the basis of a loss function for estimating 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 the large-change forecast accuracy relative to models estimated using the 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, including logistic regression, which predicts the probability of a large change based on maximum likelihood estimation. However, unlike logistic regression, our proposed model only places weight on those large change points that are below a certain false positive rate. The forecast model in this paper is multivariate with leading indicators, and is 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 forecasting model, including the complex univariate models that were found by Gorr and Schneider (2013) to be superior to simple univariate models for large change forecasting. Our findings confirm previous research which says that models that are estimated with the same loss function as is used for the forecast accuracy assessment are the most accurate forecasters (Parker, 2011).

Accurate large-change forecasting is the key to the proactive management-by-exception (MBE) of 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 being handled by staff using standard procedures. Thus, the decision to be made for MBE is binary, namely whether or not a forecast is large enough to take to management for problem solving, including using a magnitude forecast for the large-change point, diagnosis of the causes and nature of the large-change point, and crafting a solution from alternatives. The screening of demand time series is carried out using decision rules that are 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 the corresponding optimal false positive rate).

The implication 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 that developed in this paper. When a large-change decision rule fires for a demand time series, the business-as-usual forecast is replaced with a management review of the corresponding product, including an appropriate large-change magnitude forecast.

Acknowledgments

BCS-0941226 and the Labor Dynamics Institute (Cornell University, School of Industrial and Labor Relations) were the funding source for the first author.

References

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

Braga, A. A., Papachristos, A. V., & Hureau, D. M. (2014). The effects of hot spots policing on crime: an updated systematic review and meta-analysis. *Justice Quarterly*, *31*, 633–663.

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

Casella, G., & Berger, R. L. (2002). *Statistical inference*. Cengage Learning.

Clarke, R. V., & Eck, J. E. (2009). *Crime analysis for problem solvers in 60 small steps*. Center for Problem Solving Policing. U.S. Department of Justice.

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

Cohen, J., Gorr, W. L., & Olligschlaeger, A. M. (2007). Leading indicators and spatial interactions: a crime forecasting model for proactive police deployment. *Geographical Analysis*, *39*, 105–127.

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

Gorr, W. L. (2009). Forecast accuracy measures for exception reporting using receiver operating characteristics curves. *International Journal of Forecasting*, *25*, 48–61.

Gorr, W. L., & Schneider, M. J. (2013). Large-change forecast accuracy: reanalysis of M3-Competition data using receiver operating characteristic analysis. *International Journal of Forecasting*, *29*, 274–281.

Hsu, M. J., & Hsueh, H. M. (2013). The linear combinations of biomarkers which maximize the partial area under the ROC curves. *Computational Statistics*, *28*, 1–20.

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, *9*, 137–163.

Komori, O., & Eguchi, S. (2010). A boosting method for maximizing the partial area under the ROC curve. *BMC Bioinformatics*, *11*(1), 314–330.

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, *8*, 283–298.

Parker, C. (2011). An analysis of performance measures for binary classifiers. In *2011 IEEE 11th International Conference on Data Mining (ICDM)* (pp. 517–526). IEEE.

Pepe, M. S. (2004). *The statistical evaluation of medical tests for classification and prediction*. New York: Oxford University Press.

Pepe, M. S., Cai, T., & Longton, G. (2006). Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics*, *62*(1), 221–229.

R Development Core Team (2012). *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, ISBN: 3-900051-07-0, URL http://www.R-project.org.

Ramanathan, R. (2006). ABC inventory classification with multiple criteria using weighted linear optimization. *Computers and Operations Research*, *33*(3), 695–700.

Ricamato, M. T., & Tortorella, F. (2010). Combination of dichotomizers for maximizing the partial area under the ROC curve. In *Structural, syntactic, and statistical pattern recognition* (pp. 660–669). Berlin, Heidelberg: Springer.

Ricketts, J. A., & Nelson, R. R. (1987). Management-by-exception reporting: an empirical investigation. *Information and Management*, *12*, 235–246.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*(1), 77.

Skogan, W. G. (1986). Fear of crime and neighborhood change. *Crime and Justice*, *8*, 203–229.

Skogan, W. G., & Maxfield, M. G. (1981). *Coping with crime: individual and neighborhood reactions*. Beverly Hills, CA: Sage Publications.

Su, J. Q., & Liu, J. S. (1993). Linear combinations of multiple diagnostic markers. *Journal of the American Statistical Association*, *88*(424), 1350–1355.

Taylor, F. W. (1911). *Shop management*. Project Gutenberg eBook. Available from: http://www.gutenberg.org/dirs/etext04/shpmg10.txt.

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

Wang, Z., & Chang, Y. C. I. (2011). Marker selection via maximizing the partial area under the ROC curve of linear risk scores. *Biostatistics*, *12*(2), 369–385.

West, M., Harrison, P. J., & Mignon, H. S. (1985). Dynamic generalized linear models and Bayesian forecasting. *Journal of the American Statistical Association*, *80*, 73–83.

Wilpen L. Gorr is Professor of Public Policy and Management Information Systems at the School of Public Policy and Management, H. John Heinz III College, Carnegie Mellon University, where he is also Faculty Chair of the Master of Science in Public Policy and Management degree program. Wil is a past Editor of IJF, a past member of the IIF board, and a Fellow of the IIF.

Matthew J. Schneider is a visiting scholar at the Samuel C. Johnson Graduate School of Management and recently completed his Ph.D. in the Department of Statistical Science at Cornell University. His research interests include statistical database privacy, quantitative marketing, and decision-based estimation.