International Journal of Forecasting 29 (2013) 274–281

Contents lists available at SciVerse ScienceDirect

International Journal of Forecasting

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

Large-change forecast accuracy: Reanalysis of M3-Competition data using receiver operating characteristic analysis

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

[a] School of Public Policy and Management, H. John Heinz III College, Carnegie Mellon University, Pittsburgh, PA 15213, United States
[b] Department of Statistical Science, Cornell University, Ithaca, NY 14850, United States

ARTICLE INFO

Keywords:
Forecasting
ROC
M3-Competition
Exceptions reporting
Large-change forecast accuracy

ABSTRACT

This paper applies receiver operating characteristic (ROC) analysis to micro-level, monthly time series from the M3-Competition. Forecasts from competing methods were used in binary decision rules to forecast exceptionally large declines in demand. Using the partial area under the ROC curve (PAUC) criterion as a forecast accuracy measure and paired-comparison testing via bootstrapping, we find that complex univariate methods (including Flores-Pearce 2, ForecastPRO, Automat ANN, Theta, and SmartFCS) perform best for this purpose. The Kendall tau test of dependency for PAUC and a judgmental index of forecast method complexity provide further confirming evidence. We also found that decision-rule combination forecasts using three top methods generally perform better than the component methods, although not statistically so. The top methods for forecasting large declines match the top methods for conventional forecast accuracy in the M3-Competition's micro monthly time series, and therefore, evidence from the M3-Competition suggests that practitioners should use complex univariate forecast methods for operations-level forecasting, for both ordinary and large-change forecasts.

1. Introduction

According to the management by exception (MBE) principle (Taylor, 1911), operations-level staff should make resource-allocation decisions for the production of goods or services under ordinary conditions; however, under exceptional conditions, staff should defer to higher-level management. This approach makes the best use of top managers' limited time, allowing them to deal with the difficult cases and the broader lines of strategies and policy making. In the case of product or service demand forecasting, one type of exception is a large forecasted change from current demand. If a forecasted change exceeds a predetermined threshold level, then the demand forecasting system should issue an exception report, calling for diagnosis by staff members and possible action by upper management.

Gorr (2009) introduced receiver operating characteristic (ROC) curves as an accuracy framework for time series forecasting in support of MBE. The ROC framework analyzes the tails of forecast error distributions for exceptional demand conditions; on the other hand, traditional forecast error measures (such as the MAPE and MSE) place the most weight on the centers of forecast error distributions, and therefore are best suited for ordinary demand conditions.

The "gold standard" for the assessment of a forecast method in this paper is the actual change in demand, available ex post. For example, as a policy, managers may wish to review the top few percent of actual decreases (or increases), as defined by a cutoff quantile point of the gold standard distribution. If a decision rule's threshold is crossed (i.e., the rule "fires") and it identifies an actual large change, the result is a "true positive", otherwise it is a

* Corresponding author.
E-mail address: gorr@cmu.edu (W.L. Gorr).

0169-2070/$ – see front matter
doi:10.1016/j.ijforecast.2012.12.001

"false positive". Other outcomes are "true negative", where both the forecasted and actual changes are ordinary, and "false negative", where the actual change was large but the forecasted change was ordinary.

Gorr (2009) defined gold standard values as those values which are *extreme* in regard to the standardized time series of data; for example, the top five percent of standardized time series values. We can refer to this definition as "absolute", because it references the entire time series, whereas the *current* paper's definition is "relative", because it references only the last historical data point of a time series. The absolute definition is preferable when there are large costs involved in adjusting from a baseline or average level of production. One such example is the neighborhood crime level, where a flare-up above the baseline crime pattern comes to the attention of news reporters. The increased fear and lost confidence in police by the public are large societal costs, in addition to losses by crime victims.

The relative definition for time series gold standards, introduced in this paper, is preferable in circumstances where there are high costs in changing the production technologies from current levels, coupled with a potential for avoiding future costs from holding excess inventory or not meeting customer demands. Examples are when additional machines need to be set up for increased demand or employees must be laid off for decreased demand. Because the focus is on large changes from current production levels, the decision horizon must be in the very short term of one or two steps ahead, and the first step ahead is the more important. For example, managers might expect large changes in six months or a year and be able to track and adjust to such changes incrementally, but a large change in the next time period requires swift and substantial changes in their plans.

This paper applies ROC analysis to the M3-Competition data and its univariate forecast methods. A key question is whether complex univariate forecast methods perform better than simple ones under ROC measures, similar to the case of Gorr (2009), who compared complex multivariate models to simple univariate methods for short-term forecasting and found complex methods to be more accurate. Most of the literature over the past 30 years has supported the use of simple univariate methods for ordinary conditions (e.g., the *M*-competitions). This paper provides additional evidence that complex forecast methods are significantly more accurate than simple methods for exceptions forecasting, and specifically for univariate methods.

We also investigate whether a combination forecast leads to an increased accuracy for exceptions forecasting. For forecasting ordinary conditions, combinations are averages or weighted averages of individual forecasts. In contrast, combination forecasts for exceptions use "or" or "and" logical connectors for individual-forecast-method decision rules. For example, the best-performing combination forecast method in this paper requires that the combination decision rule fire if the decision rule of any of the component forecast methods fires (with "or" connectors).

Also new in this paper is the application of the partial area under the ROC curve (PAUC) as the forecast accuracy measure for exceptions forecasting. A statistical test for differences in PAUC using paired comparisons and accounting for correlated data is also included. The total area under the ROC curve has the interpretation of being the probability that a decision rule will signal a randomly-chosen positive instance as being higher than a randomly-chosen negative instance (Fawcett, 2006).

Section 2 provides a brief literature review of forecast error measures and competitions. Section 3 covers the experimental design for our reanalysis of M3 data, and Section 4 provides results. Finally, Section 5 concludes the paper.

2. Literature review

In this section we review the M3-Competition and its analysis of forecast accuracies, especially in regard to micro monthly time series. We also review the literature on the statistical tests available for comparing ROC curves.

2.1. M3-Competition data and forecast methods

Operations and marketing managers forecast individual products or product families in an attempt to meet demand. Hence, we limit this study to the micro monthly time series of the M3-Competition, which best match this decision setting. While both the M1- and M3-Competitions included micro time series, we only use M3 data in this paper. The M3-Competition has a wider range of univariate methods, and includes more complex ones than M1. Furthermore, Koning, Franses, Hibon, and Stekler (2005) provide a judgmentally-derived complexity ranking for the M3 forecast methods, made by three forecasting experts, which we relate to forecast accuracy. We averaged the complexity ranks across the experts and rescaled ties, to yield the average ranks in Table 1. To learn more about the forecast methods in Table 1, see Table 2 of Makridakis and Hibon (2000, p. 456).

2.2. M3-Competition results

A major conclusion of the M3-Competition is that "Statistically sophisticated or complex methods do not necessarily produce more accurate forecasts than simpler ones". Micro monthly data, however, are one case in which complex forecast methods are more accurate than simple ones. For example, Table 2 summarizes the best and worst performing methods for micro monthly data according to four forecast error measures used in the M3-Competition (Koning et al., 2005). All of the best methods are complex, except for Theta, which has a mid-range complexity. All of the worst methods are simple, except for the Box–Jenkins methods, which also have mid-range complexity. Apparently micro monthly time series data have patterns that complex methods are able to estimate and make good use of under ordinary conditions. The question is whether complex models are also better for exceptional conditions. For example, when change is in progress, neural networks have pattern recognizers that can turn on model components selectively in order to capture and extrapolate the change, and expert systems can switch to more reactive models.

Table 1
Average rank of experts' judgmental assessment of forecast method complexity.

Forecast method	Expert 1	Expert 2	Expert 3	Average rank
Naïve2	1	1	1	1.0
Single	2	2	2	2.0
Holt	3	3	3	3.0
Robust-Trend	4	4	5	4.3
Winter	6	5	5	5.3
Dampen	5	6.5	7	6.2
PP-autocast	8	6.5	8	7.5
Theta sm	7	8	9	8.0
Comb SHD	10	9	5	8.0
Theta	9	10	14	11.0
BJ automatic	11.5	11	11.5	11.3
Autobox1	11.5	13	11.5	12.0
Autobox3	18.5	13	11.5	14.3
Autobox2	18.5	13	11.5	14.3
ARARMA	13	15	17.5	15.2
SmartFCS	15	18	17.5	16.8
Flores-Pearce 2	15	18	17.5	16.8
Flores-Pearce 1	15	18	17.5	16.8
ForecastPRO	17	18	17.5	17.5
Forecast X	21	18	17.5	18.8
RBF	20	22	21.5	21.2
AutomatANN	22	21	21.5	21.5

Table 2
Best and worst forecast methods for M3 micro monthly time series data.
Source: Taken from Koning et al. (2005).

Error measure	Best four forecast methods (in order)
sMAPE	SmartFCS, Theta, AutomatANN, ForecastPRO
Median sAPE	SmartFCS, Theta, AutomatANN, ForecastX
RMSE	Theta, SmartFCS, ForecastX, ForecastPRO
Error measure	**Worst four forecast methods (in order)**
sMAPE	Robust-Trend, Naive2, Single, ARARMA
Median sAPE	Robust-Trend, Naive2, ARARMA, Single
RMSE	Robust-Trend, Naive2, RBF, Autobox

2.3. ROC statistical tests

Cohen, Garman, and Gorr (2009) and Gorr (2009) provide reviews of ROC curves and analysis applied to time series data monitoring and forecasting respectively. Hence, this section only summarizes the ROC literature in regard to additional material on statistical tests which are introduced in this paper for time series testing.

Area under curve (AUC) is the total area under an ROC curve over the entire false positive rate (FPR) range of 0–1. The higher the AUC, the better the forecasting method (or other test mechanism). AUC can be computed using the trapezoidal rule, given a comprehensive set of FPR, TPR pairs, or by computing the nonparametric Wilcoxon statistic, as shown by Hanley and McNeil (1982). The Wilcoxon statistic can be used to calculate the standard error of the AUC for statistical tests (Hanley & McNeil, 1982). Alternatively, the standard error and AUC can be determined using the DeLong, DeLong, and Clarke-Pearson (1988) method.

Partial area under curve (PAUC) is the area under an ROC curve for a specified FPR range, generally starting at zero. In many situations, a decision maker has a maximum FPR threshold which he or she is not willing to exceed, and PAUC represents this case. PAUC can be computed using the trapezoidal rule, and bootstrapping can be used to compute its standard error.

Parametric and nonparametric statistical tests for comparing the AUCs of two ROC curves with correlated data are described by Hanley and McNeil (1983). Forecasting competitions generally include correlated data, because alternative forecast methods are applied to the same cross-section of time series. The available tests require standard errors calculated by either the Dorfman and Alf (1969) maximum likelihood program or the Wilcoxon statistic, and a correlation coefficient calculated by the Pearson product-moment correlation method or the Kendall tau rank correlation coefficient. Bootstrapping removes the need for a covariance estimate, and accounts for correlated data (Janes, Longton, & Pepe, 2009).

3. Experimental design

We used ROC analysis to study the large-change forecast accuracy for 1- and 2-month-ahead forecasts. This section describes how we processed the 474 time series of the M3 micro monthly data to create empirical ROC curves. The time series tend to be declining at the forecast origin, so we focused on exceptional declines—certainly a major concern for managers in firms selling products or services.

This section discusses how we standardized our data to facilitate the cross-sectional specification of decision rule limits, as well as how we tabulated the results to produce ROC curves.

3.1. Standardizing forecasted change and its gold standard

The notation for the time series, forecasts, time series changes, and forecasted change is as follows.

Cross-section of actual time series:

$Y_{it}\,(i = 1, \ldots, I; t = 1, \ldots, T + m)$, where i is a time series; t is time; T is the single, fixed forecast origin of the M3-Competition; and m is the forecast horizon (here we use $m = 1$ and 2 only).

Set of alternative forecast methods $j = 1, \ldots, J$ and forecasts:

$$F_{ijt}\,(i = 1, \ldots, I; j = 1, \ldots, J; t = T + m).$$

Forecasted change:

$$\text{ForecastDelta}_{ijT+m} = F_{ijT+m} - Y_{iT}$$
$$(i = 1, \ldots, I; j = 1, \ldots, J; m = 1 \text{ or } 2).$$

The gold standard for comparison with forecasted changes is the true, ex post value for a 1-month-ahead or 2-month-ahead forecast, minus the last realization in the estimation data set:

$$\text{Delta}_{iT+m} = Y_{iT+m} - Y_{iT} \quad (i = 1, \ldots, I; m = 1 \text{ or } 2).$$

We need to standardize each ForecastDelta and Delta in order to remove scale and control variation, in a way analogous to the computation of z-scores. Then we can use the same standardized threshold values of decision rules for each time series (as is done with t-statistics or normal tables). While we can estimate the sample mean and standard deviation for Deltas, there is a limitation in standardizing ForecastDelta. The M3-Competition had a

single forecast origin for each time series and a single set of corresponding forecasts for $m = 1, \ldots, 18$, so there is only a sample of size one for each F_{ijT+m}. If the competition had used a rolling or expanding horizon design with many forecast origins, we could estimate the mean and standard deviation of ForecastDelta for each series and m. However, for the M3-Competition we must use an approximation, which is facilitated by the way ROC curves are constructed. We need only assume that ForecastDeltas are proportional to Deltas by the forecast method in the sample of time series. Then we can normalize ForecastDeltas by the mean and standard deviation of the deltas (see the Appendix).

3.2. Gold standard cutoff

Deltas that were a specified number of standard deviations below the mean were considered true large change values ("positives" in regard to ROC). We specified three cutoffs of −1.28, −1.65, and −2.33 standard deviations below the mean, corresponding to 10%, 5%, and 1% quantile points of the delta distribution if it were normally distributed. For first differences (Delta_{iT+1}) of the 474 time series, there are 110, 74, and 24 positives for the 10%, 5%, and 1% cutoffs respectively. For second differences (Delta_{iT+2}), the corresponding numbers of positives are 89, 57, and 27. Even after removing spurious, regression-to-the mean cases in Section 3.3, the numbers of positives are higher than for a normal distribution because of the "fat" lower tail (and thin upper tail) of the distribution, and also because our standardization is approximate. Regardless, it is important to analyze more than one gold-standard cutoff in order to examine how the forecast performance varies with the definition of positives. For example, Gorr (2009) found the ROC performance to improve with more extreme definitions, probably because the most extreme cases are the easiest to distinguish from the rest of the distribution.

3.3. Regression to the mean

It is necessary to control for regression-to-the-mean behavior in cases when the exceptional values do not persist (i.e., they are outliers) and the time series patterns return to the mean of the series. For large declines, the problem occurs when the time series has a high outlier that returns to the mean. Take the case of 1-step-ahead forecasts. Any non-responsive forecast method or model has a good performance for the data point following the outlier, spuriously inflating AUC or PAUC measures. The actual data point returns to the mean, while the unresponsive forecast method never left the mean. So $\text{ForecastDelta}_{ijT+1}$ fires a decision rule, testing positive, and Delta_{iT+1} is a positive, yielding a spurious true positive. The same is true for m-step-ahead forecasts, m steps after an increasing outlier. The 1-step-ahead forecasts for the three thresholds of −1.28, −1.65, and −2.33 have 47 out of 110, 17 out of 74, and 3 out of 24 regression cases, respectively. The 2-step-ahead forecasts have 30 out of 89, 15 out of 57, and 3 out of 23 regression cases. Any forecasts affected by regression to the mean were therefore removed from analysis by reclassifying the corresponding time periods as ordinary instead of exceptional.

3.4. Forecast performance

For every threshold, standardized ForecastDeltas less than or equal to the z-value threshold were considered to signal a large decrease (test positive). Forecasts methods with test positives in a series that had an actual positive are true positives. Otherwise, the forecast method provided a false positive. This process was repeated for a maximum of 475 z-value thresholds, occurring at the boundaries of the 474 ranked normalized ForecastDeltas, thus spanning all possibilities for the construction of ROC curves.

True Positive Rates (TPRs, number of true positives divided by number of positives) and False Positive Rates (FPRs, number of false positives divided by number of negatives) were computed to obtain increasing two-dimensional points (FPR, TPR) for each method. The connection of these points created each method's empirical ROC curve, and statistical tests from Section 2 were applied.

4. Results

We decided to limit our analysis to the PAUC measure for false positive rates between 0.0 and 0.2, believing that this would include the range in which most managers would be comfortable operating. Note that it is common to use larger false positive rates (which are the same as type I error rates) in practice than those used in theory testing (e.g., see Cohen et al., 2009), depending on the cost of false negatives, the prevalence of positives, and the resources available for diagnosis and follow-up to test positives.

4.1. Partial area under curve

We compare the large-change forecast performances of forecast methods using a non-parametric bootstrap approach for paired comparisons between PAUCs. One-sided p-values were computed for each PAUC threshold's top performing method to see whether it was statistically better than the other methods. We use 1000 bootstrap samples for each pair of methods.

See Table 3 for results at the 0.05 significance level for 1-step-ahead forecasts over the FPR range of 0.0–0.2. We only included methods in the comparison that have complexity scores in Table 1 (dropping AAM1 and AAM2), and, in addition, we dropped Rule-Based Forecasting because it was designed for annual data and we are analyzing monthly data.

In general, complex methods performed significantly better than simple methods. Automat ANN, Flores-Pearce 2, ForecastPRO, SmartFCS, and Theta were in the set of methods which were either best or not significantly different from the best for all three cutoff points used for gold standards. All but Theta are complex methods, with subjective scores from Table 1 of 16.8 or higher. Theta has mid-range complexity, with a score of 11.0. SmartFCS, with a complexity score of 16.8, was in the significantly better methods for the 95% and 90% gold standard cutoffs, and BJ automatic (mid-range complexity score of 11.3) joins the significantly better set for the 90% gold standard cutoff. Note in Table 3 that the more extreme the positive cases are

Table 3
Paired comparisons with the top forecasting method of PAUC for FPR range 0.0–0.2 using bootstrapping: 1-step-ahead forecasts.

99% gold standard	PAUC	*p*-value	95% gold standard	PAUC	*p*-value	90% gold standard	PAUC	*p*-value
Flores/Pearce 2	0.145		Theta	0.107		Automat ANN	0.071	
Automat ANN	0.123	0.112	Automat ANN	0.104	0.396	SmartFCS	0.069	0.431
Theta	0.122	0.086	ForecastPRO	0.104	0.374	Theta	0.067	0.275
ForecastPRO	0.125	0.082	SmartFCS	0.102	0.315	ForecastPRO	0.065	0.225
Dampen	0.112	0.045	Flores/Pearce 2	0.100	0.210	Flores/Pearce 2	0.065	0.203
Single	0.104	0.024	Theta sm	0.088	0.017	BJ automatic	0.059	0.080
SmartFCS	0.119	0.021	BJ automatic	0.089	0.015	Forecast *X*	0.056	0.044
Autobox2	0.099	0.016	ARARMA	0.081	0.008	Naïve 2	0.048	0.028
Holt	0.101	0.015	PP-autocast	0.089	0.005	ARARMA	0.053	0.026
PP-autocast	0.105	0.010	Forecast X	0.084	0.005	Dampen	0.054	0.021
Winter	0.099	0.010	Flores/Pearce 1	0.084	0.004	Flores/Pearce 1	0.054	0.018
ForecastX	0.100	0.008	Autobox2	0.078	0.002	PP-autocast	0.053	0.014
BJ automatic	0.098	0.003	Dampen	0.083	0.002	Autobox3	0.049	0.014
Flores/Pearce 1	0.099	0.003	Autobox3	0.075	0.001	Robust-Trend	0.043	0.012
ARARMA	0.085	0.003	Holt	0.072	0.000	Theta sm	0.052	0.012
Naïve 2	0.076	0.001	Winter	0.070	0.000	Autobox2	0.051	0.008
Theta-sm	0.089	0.000	Single	0.070	0.000	Holt	0.044	0.002
Autobox3	0.074	0.000	Autobox1	0.062	0.000	Winter	0.042	0.001
Autobox1	0.059	0.000	Naïve 2	0.053	0.000	Single	0.038	0.000
Robust-Trend	0.025	0.000	Robust-Trend	0.040	0.000	Autobox1	0.035	0.000

(i.e., the more stringent the gold standard cutoff), the better the PAUC performance, which is similar to the findings of Gorr (2009).

For the 2-step-ahead forecasts, the significantly better forecasting methods (in order of decreasing PAUC values) are as follows:

99% gold standard: Automat ANN, Theta sm, Flores/Pearce 2, ForecastPRO, BJ automatic, Theta, SmartFCS, and Autobox2
95% gold standard: ForecastPRO, Flores/Pearce 2, Theta sm, Theta, SmartFCS
90% gold standard: Theta sm, Flores/Pearce 2, ForecastPRO, SmartFCS.

Most of these methods were in the significantly better sets for 1-step-ahead forecasts, but Theta sm and Autobox 2 show up as new for 2-step-ahead forecasts. Flores/Pearce 2 and ForecastPRO are in every significantly better set, while SmartFCS is close behind in all but one of those sets.

A ROC curve is a plot of the true positive rate (TPR) versus the false positive rate (FPR), obtained by varying the threshold level of an exceptions decision rule. Fig. 1 displays a selection of ROC curves for 1-step-ahead forecasts and the 95% gold standard case from Table 3. The ROC curves for other cases are similar qualitatively. Shown are three top-performing methods (all complex) and three simple smoothing methods. Also shown is the line representing a chance decision mechanism.

For a given FPR, the method with the highest ROC curve is best, having the highest TPR. At 0.01 FPR there is no difference in performance, but by 0.05 FPR, the complex methods have a TPR range of approximately 0.32–0.42 while the simple methods have a range of 0.15–0.28. At 0.10 FPR, the complex methods have a TPR range of 0.61–0.63, while the simple methods only have a range of 0.40–0.49. Thus, the complex methods have much better performances than the simple methods. At the maximum FPR rate in Fig. 1, the best method finds just over 80% of the positive cases (gold standard large decreases). Note that Dampen has a better performance than Single or Holt because, as was shown by Snyder and Koehler (2008), it "··· possesses a special capacity to adapt to structural change without direct intervention".

Table 4
Kendall tau test.

	99% gold standard	95% gold standard	90% gold standard
One step ahead	tau = 0.197 *p*-value = 0.241	tau = 0.464 *p*-value = 0.005	tau = 0.535 *p*-value = 0.001
Two steps ahead	tau = 0.432 *p*-value = 0.009	tau = 0.379 *p*-value = 0.023	tau = 0.411 *p*-value = 0.013

4.2. Complexity

This section investigates the effect of forecast method complexity on ROC performance, measured by PAUC, over the M3 micro monthly time series. We eliminated Rule-Based Forecasting from the analysis because it is an annual time series method, whereas the micro-level data analyzed in this paper are monthly. We also dropped the Naïve method because it yields 0 change when comparing the forecasts to the last historical value, as well as the AAM1/AAM2 methods, which were not ranked by the experts for complexity in Table 1. We expected the relationship between complexity and PAUC to be positive.

Table 4 contains the results of applying Kendall's tau with a two-sided test and a 0.05 significance level in regard to the dependence of PAUC for the FPR range of 0–0.20 on average rank for complexity in Table 1. The cases included are the three gold-standard cutoffs for defining positives and 1- and 2-month-ahead forecasts. Five out of six cases test significant at the 0.05 level or better, thus providing further evidence that the complex forecast methods are best for the large-change forecast accuracy for the M3 micro monthly time series.

4.3. Decision-rule combination forecasts

It is well-known that a simple average combination of methods' forecasts often forecasts more accurately than the component methods (e.g., Clemen, 1989). We

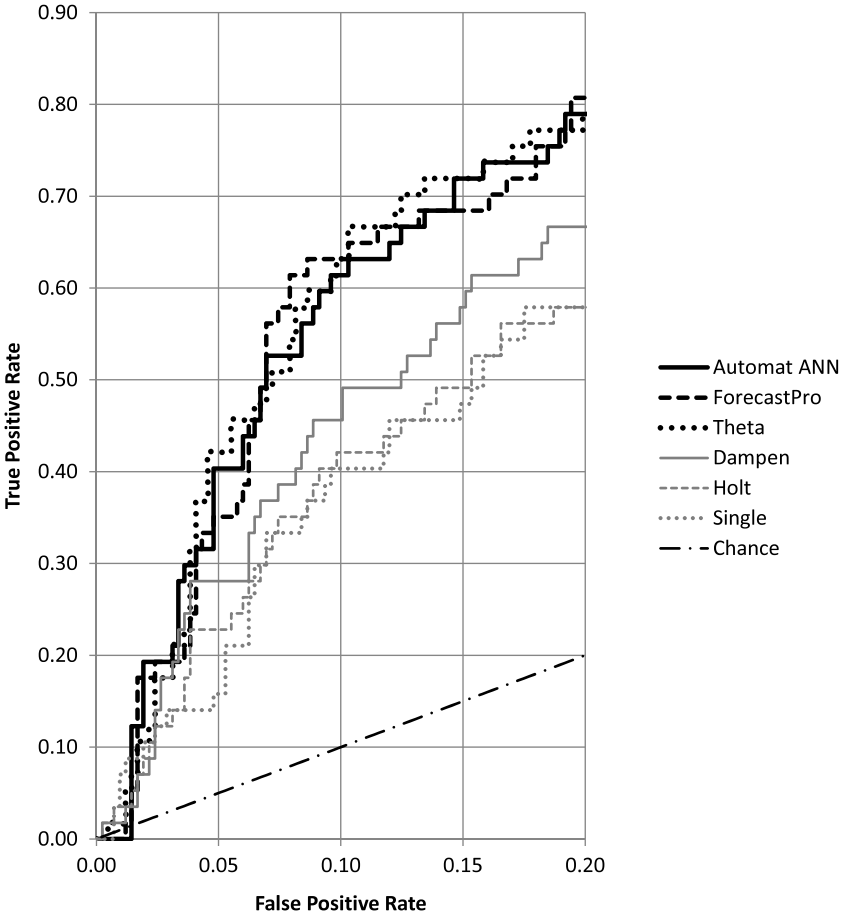

Fig. 1. ROC curves for a selection of M3-Competition methods, micro monthly time series, 1-month ahead forecasts, 95% gold standard cutoff.

propose combination forecasts for exceptions forecasting that combine decision rules instead of forecasts. For a decision-rule combination forecast with a fixed number of component forecast methods, if a prescribed number of component methods' rules fire (test positive), then the composite decision rule fires. The benefit of such a rule could be either to make more conservative decisions, reducing false positives, or to be more inclusive and liberal, adding more test positives—depending on whether "and" or "or" logical connectors are used for the component rules.

We created three combination forecasts, each with the same three top-performing, complex forecast methods: ForecastPRO (expert system), Automat ANN (neural network), and Theta (decomposition method). Because each of the component methods has a different modeling approach, this combination promises to maximize the information available for forecasting exceptions. The first combination rule (Min) tests positive whenever any one of the three component methods tests positive. The second (Median) is a median combination forecast that tests positive whenever two of the three component methods test positive. Finally, the third (Max) tests positive when all of the top three methods test positive.

Table 5 gives the results, using the same paired comparison test as in Table 3. Here, we limit the comparisons to the three combinations and their three component forecast methods, to see whether combinations can improve the forecast accuracy over their components. The liberal Min combination is best for all three 1-month-ahead cases and one of the three 2-month cases, but not significantly so at the 0.05 significance level. ForecastPRO is best in the other two 2-month-ahead cases. Thus, rule-combination forecasts are promising: for forecasting large changes for important micro monthly time series, we recommend the Min rule-combination forecast.

5. Conclusion

This paper has applied ROC analysis to the M3-Competition's micro monthly time series for 1- and 2-month-ahead forecasts. Using the partial-area-under-the-curve (PAUC) criterion, paired comparison testing via bootstrapping, and the Kendall tau we found that complex methods perform best for forecasting large declines in these time series, which tended to decline as a group over time. The classification of top methods matches that obtained using conventional forecast accuracy methods in the M3-Competition: complex methods forecast these series better than simple ones.

We also found that a rule-combination forecast, testing positive whenever any one of the three decision rules in the combination method fired, performed better than the component methods, but not statistically significantly so.

Table 5
Paired comparisons with the top forecasting method of PAUC for FPR range 0.0–0.2 using bootstrapping: 1- and 2-step-ahead forecasts for three rule-combination forecasts and their component forecast methods.

99% gold standard	PAUC	*p*-value	95% gold standard	PAUC	*p*-value	90% gold standard	PAUC	*p*-value
One month ahead								
Min	0.134		Min	0.111		Min	0.075	
Median	0.128	0.302	Max	0.108	0.343	Max	0.070	0.188
ForecastPRO	0.125	0.261	Median	0.108	0.328	Automat ANN	0.071	0.171
Theta	0.122	0.179	Theta	0.107	0.247	Theta	0.067	0.075
Max	0.120	0.135	ForecastPRO	0.104	0.169	ForecastPRO	0.065	0.040
Automat ANN	0.123	0.057	Automat ANN	0.104	0.079	Median	0.066	0.033
Two months ahead								
Min	0.141		ForecastPRO	0.123		ForecastPRO	0.092	
Median	0.136	0.343	Median	0.120	0.325	Min	0.090	0.389
ForecastPRO	0.127	0.183	Min	0.119	0.347	Median	0.087	0.188
Theta	0.121	0.097	Theta	0.117	0.260	Theta	0.083	0.109
Automat ANN	0.129	0.078	Max	0.105	0.043	Automat ANN	0.077	0.076
Max	0.116	0.076	Automat ANN	0.094	0.013	Max	0.079	0.048

Thus, the evidence from the M3-Competition suggests that operations managers should use complex methods such as Theta, a neural network, ForecastPRO, or SmartFCS for forecasting both ordinary and large-change demand data.

Acknowledgments

We wish to thank Michele Hibon for providing M3 data and insights about those data; Robin et al. (2011) for his R package, pROC, and guidance in its methodology for calculating PAUCs and their standard errors; Hyndman (2011) for his R package, Mcomp, which allowed us to implement M3 time series data easily in R; and the anonymous referees for insightful comments, careful editing, and suggestions, including the suggestion to include the Kendall tau test.

Appendix. Standardizing the forecasted change of M3-Competition time series for ROC analysis

A limitation of the M3-Competition data is that we have only one 1-step-ahead forecast from each method for each series considered, and therefore cannot estimate the mean and variance of forecasts by method and series directly. Assume, however, that a method's forecasts have the same mean as a series (they are unbiased forecasts), but have a variance proportional to the data, dependent on the forecast method, but with a proportion which is fixed across series for a given forecast method.

Consider the following notation:

Cross-section of actual time series: $Y_{it}(i = 1, \ldots, I; t = 1, \ldots, T + m)$, where T = single, fixed forecast origin of M-Competition and m = forecast horizon (here we use $m = 1$ only).

Set of alternative method $j = 1, \ldots, J$ *forecasts*: $F_{ijt}(i = 1, \ldots, I; j = 1, \ldots, J; t = T + m)$.

Sample statistics: Mean M_{it} and Variance S_{it}.

Standardized actuals: $Y'_{it} = (Y_{it} - M_{it})/\text{sqrt}(S_{it})$.

Decision thresholds U and L (fixed across time series):

If $Y'_{it} \geq U$ then high test positive (signaling a large increase);

if $Y'_{it} \leq L$ then low test positive (signaling a large decrease);

if $L < Y'_{it} < U$ then test negative.

Assume that the time series are stationary and that the mean of forecast method j for series i is M_{it} (independent of the forecast method) but that the standard deviation is $k_j\text{sqrt}(S_{it})$. Thus, we have:

Standardized forecasts: $F'_{ijt} = (F_{ijt} - M_{it})/k_j\text{sqrt}(S_{it})$.

Decision rules with control limits U and L (same across series and forecast methods because of standardization):

If $F'_{ijT+1} = (F_{ijT+1} - M_{it})/k_j\text{sqrt}(S_{it}) \geq U$ then high test positive or

if $F'_{ijT+1} = (F_{ijT+1} - M_{it})/\text{sqrt}(S_{it}) \geq k_jU$ then high test positive (use this rule for all series).

If $F'_{ijT+1} = (F_{ijT+1} - M_{it})/k_j\text{sqrt}(S_{it}) \leq L$ then low test positive or

if $F'_{ijT+1} = (F_{ijT+1} - M_{it})/\text{sqrt}(S_{it}) \leq k_jL$ then low test positive (use this rule for all series).

Take the case of forecast method j and low test positives. This is the key point: the right-hand *value*, k_jL, is of no concern because we process over values $\min[(F_{ijT+1} - M_{it})/\text{sqrt}(S_{it})]$ to $\max[(F_{ijT+1} - M_{it})/\text{sqrt}(S_{it})]$ with a grid to create the ROC curve for forecast method j. Regardless of the value of k_j, the ROC methodology produces a valid ROC curve and threshold values for the false positive rate. In other words, we do not depend on specific values for the right-hand side, but instead enumerate a grid of possible values for it, depending on the sample of left-hand-side values, from the smallest to the largest possible threshold values that do not result in constant decisions (all test positive or all test negative).

References

Clemen, R. T. (1989). Combining forecasts: a review and an annotated bibliography. *International Journal of Forecasting*, *5*(4), 559–583.

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.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating curves: a nonparametric approach. *Biometrics*, *44*, 837–845.

Dorfman, D. D., & Alf, E., Jr. (1969). Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals—rating method data. *Journal of Mathematical Psychology*, *6*, 487–496.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*, 861–874.
Gorr, W. L. (2009). Forecast accuracy measures for exception reporting using receiver operating characteristics curves. *International Journal of Forecasting*, *25*, 48–61.
Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*, 29–36.
Hanley, J. A., & McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, *148*, 839–843.
Hyndman, R. J. (2011). Data from the *M*-competitions. Comprehensive R Archive Network. http://cran.r-project.org/web/packages/Mcomp/.
Janes, H., Longton, G. M., & Pepe, M. S. (2009). Accommodating covariates in ROC analysis. *Stata Journal*, *9*(1), 17–39.
Koning, A. J., Franses, P. K., Hibon, M., & Stekler, H. O. (2005). The M3 competition: statistical tests of the results. *International Journal of Forecasting*, *21*, 397–409.
Makridakis, S., & Hibon, M. (2000). The M3 competition: results, conclusions and implications. *International Journal of Forecasting*, *16*, 451–476.
Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., & Sanchez, J.-C. et al. (2011). Display and analyze ROC curves. Comprehensive R archive network. http://cran.r-project.org/web/packages/pROC/pROC.pdf.
Snyder, R. D., & Koehler, A. B. (2008). Incorporating a tracking signal into a state space model. *International Journal of Forecasting*, *25*, 526–530.
Taylor, F.W. (1911). *Shop management*. Project Gutenberg eBook. Available from: http://www.gutenberg.org/dirs/etext04/shpmg10.txt.

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 Ph.D. candidate in the Department of Statistical Science at Cornell University. His research interests include statistical database privacy, quantitative marketing, and decision-based estimation.