

Using Receiver Operating Characteristic Analysis to Evaluate
Exceptions Forecast Accuracy: Reanalysis of M3-Competition Data on
Micro Monthly Time Series

By

Wilpen L. Gorr¹

Matthew J. Schneider²

June 18, 2008

¹H. John Heinz School of Public Policy and Management, Carnegie Mellon University,
Pittsburgh, PA 15213

²Pardee RAND Graduate School, RAND, Santa Monica, CA 90401

Abstract

This paper applies ROC analysis to M3 Competition micro monthly time series for one-month-ahead forecasts. Using the partial-area-under-the-curve (PAUC) criterion and paired comparison testing via bootstrapping, we find that complex methods perform best for forecasting large declines in these time series, which tended as a group to decline 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. A regression model of PAUC on a judgmental index for forecast method complexity and a dummy variable for Box-Jenkins methods provides further confirming evidence. We also found that a combination forecast, consisting of the simple arithmetic average of the top three methods, to perform better than the component methods as well as combinations constructed from “OR” rules with any one or pair of methods signaling exceptions.

Key Words: Forecasting, ROC, M-Competition, Exception Reporting

Introduction

Under management by exception (MBE) (Taylor, 1911), operations-level staff make resource-allocation decisions under ordinary conditions for production of goods or services. Under exceptional conditions, however, staff defer to higher-level management for decision making. In the case of product or service demand forecasting, an exception is a forecasted large change from current demand level. If a forecasted change exceeds a predetermined threshold level, then the demand forecasting system issues an exception report, calling for diagnosis and possible actions by upper management.

Recently, Gorr (2007) introduced receiver operating characteristic (ROC) curves as an accuracy measure for time series forecasting in support of MBE. An ROC curve is a plot of true positive rate (TPR) versus false positive rate (FPR) obtained by varying the threshold level of the exception decision rule. The “gold standard” for assessment of a forecast method is actual change in demand, available ex post. For example, management may wish to review top 10 percent actual increases (and the same percentage of actual decreases), corresponding to a cutoff, fractal point of the gold standard distribution. If an exception report identifies an actual large change, the result is a “true positive”, otherwise it is a “false positive”.

This paper applies ROC analysis to M-Competition data. A key question is whether complex forecast models perform better than simple models under the ROC measures, as was the case in Gorr (2008) and contrary to forecasting for ordinary conditions. With ROC measures, we also investigate whether a combination forecast leads to increased accuracy for exceptional forecasts. Finally, this paper pursues the area under the curve (AUC) and (partial area) under the curve PAUC measures for ROC curves. Several statistical tests are available for differences in AUC and PAUC and we apply them to the forecasting problem.

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

2. Literature Review

In this section we review the literature on the M-Competitions and their analysis for forecast accuracy, especially in regard to micro monthly time series. We also review the literature on statistical tests available for comparing ROC curves.

2.1 M-Competitions

M-Competitions (e.g., see Makridakis & Hibon, 2000, Hyndman & Koehler, 2006)) have been analyzed using central tendency error measures including the symmetric mean absolute percentage error (sMAPE), average ranking, median symmetric absolute percentage error (Median sAPE), root mean squared error (RMSE), Percent Better, and mean absolute scaled error (MASE). These measures provide information on accuracy over the entire distribution of

sample forecasts via central tendency measures. ROC analysis, carried out in this paper, focuses on the tails of the sample forecast distributions.

We limit attention to monthly, micro-level times series, corresponding to the setting of MBE. Clearly, operations management concerns micro-level time series and monthly data better correspond to this setting than quarterly and annual data. While both the M1 and M3 competitions have appropriate time series, and we have analyzed both, we report only on M3 data in this paper. The M3 competition has a wider range of univariate methods and especially more complex ones than M1. Furthermore, Koning et al. (2005) provides a subjective index of complexity for the M3 forecast methods, which we relate to ROC and central tendency performance. We averaged the index across three experts and rescaled tied ranks to yield the data in Exhibit 1. Methods that experts ranked as tied were given averaged values of the corresponding missing ranks.

Exhibit 1.

Average rank of experts' judgmental assessment of forecast method complexity

Forecast Method	Keith Ord	Robert Fildes	Spyros Makridakis	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
Theta	9	10	14	11.0
Comb SHD	10	9	5	11.3
BJ Automatic	11.5	11	12.5	11.7
Autobox1	11.5	13	12.5	12.3
Autobox3	18.5	13	12.5	14.7
Autobox2	18.5	13	12.5	14.7
ARARMA	13	15	17.5	15.2
Smart FCS	15	18	17.5	16.8
Flores-Pearce2	15	18	17.5	16.8
Flores-Pearce1	15	18	17.5	16.8
Forecast X	21	18	17.5	18.8
RBF	20	22	21.5	21.2
Forecast Pro	17	18	17.5	21.2
AutomatANN	22	21	21.5	21.5

One further limitation is to limit analysis in this paper to the one-month forecast horizon. It is

well known that forecast accuracy is best for this horizon plus many operations management applications depend most heavily on the first step ahead forecast.

2.2. Error Measure Used in the M3-Competition

Central tendency error measures used with the M3-Competition data have documented advantages and disadvantages. We briefly review each measure and its merits.

The sMAPE fixes the problems with the mean absolute percentage error (MAPE) and avoids large errors when the actual value approaches 0 or the actual value is greater than the forecasted value (Hibon & Makridakis, 2000). On the other hand, the sMAPE has a heavier penalty when forecasts are high compared to when forecasts are low and can be asymmetric (Koehler, 2001).

The second error measure, the Median sAPE is not influenced by extreme values and more robust than MAPE (Hibon & Makridakis, 2000).

The third error measure, RMSE, measures magnitude of error, is preferred by practitioners (Armstrong & Collopy, 1992). Alternatively, the RMSE is more sensitive to outliers (Hyndman & Koehler, 2006) and highly unreliable (Armstrong & Collopy, 1992).

The fourth error measure in the M3-Competition, percent better, is a reliable measure and immune to outliers (Armstrong & Collopy, 1992). The disadvantage is that it does not recognize the amount of improvement at all and ignores magnitude.

Hyndman & Koehler(2006) introduced an additional error measure called the mean absolute scaled error (MASE), which scales the absolute error based on the mean absolute error from a naïve or other benchmark forecast method. The MASE is easily interpretable with a value of one or greater indicating that a forecast method's errors are on average smaller than a benchmark method's errors. Also, the MASE is less sensitive to outliers and captures seasonality or trend in series (Kolassa & Schutz, 2007).

2.3 M3-Competition Results and Conclusions

A major conclusion of the M3-Competition is “Statistically sophisticated or complex methods do not necessarily produce more accurate forecasts than simpler ones” (Hibon & Makridakis, 2000). While true, a more precise statement would include monthly micro data as a case in which complexity does pay off. For example, Exhibit 2 summarizes best and worst performing methods for monthly micro data according to four central tendency error measures (Armstrong, 2008). All of the best methods are complex except for Theta, which has mid-range complexity. All of the worst methods are simple, except Box Jenkins methods which also have mid-range complexity.

Exhibit 2.**Best and worst forecast methods for M3 micro, monthly time series data**

Error Measure	Best Four Forecast Methods (in order)
sMAPE	SmartFCS, Theta, AutomatANN, ForecastPRO
Average Ranking	Theta, SmartFCS, 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
Average Ranking	Robust-Trend, Naive2, Single, ARARMA
Median sAPE	Robust-Trend, Naive2, ARARMA, Single
RMSE	Robust-Trend, Naive2, RBF, Autobox

2.4 ROC Statistical Tests

Cohen et al. (2006) and Gorr (2007) provide detailed descriptions of ROC curves and analysis applied to time series data monitoring and forecasting. Hence this section only summarizes the ROC literature in regard to additional material on statistical testing.

Area under curve (AUC) is the total area under an ROC curve over the entire false positive rate (FPR) range of 0 to 1. AUC can be computed using the trapezoidal rule with a comprehensive set of FPRs and TPRs or by computing the Wilcoxon statistic, as shown by Hanley and McNeil (1982). In addition, the nonparametric Wilcoxon statistic can be used to calculate the standard error of the AUC for statistical tests (Hanley and McNeil, 1982). Alternatively, the standard error and AUC can be determined using the DeLong, DeLong, & Clarke-Pearson (1988) method.

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

Parametric statistical tests for comparing the AUCs of two ROC curves with correlated data were described by Hanley and McNeil (1983). These tests require a covariance estimate and standard errors, calculated by the Dorfman and Alf (1969) maximum likelihood program or the Wilcoxon

statistic. Bootstrapping removes the need for a covariance estimate and accounts for correlated data (Janes, Longton, & Pepe, 2005).

For statistical comparisons between PAUCs of two partial ROC curves with correlated data, Wald test results can be used based on the bootstrapped standard errors (Janes, Longton, & Pepe 2005). The M3-Competition is a case with correlated data—with alternative methods applied to the same data—which we analyzed using the non-parametric bootstrap approach for comparison between PAUCs.

3. Experimental Design

This section describes how we processed the M3 micro monthly data to create empirical ROC curves. Included are how we standardized data to facilitate cross-sectional specification of decision rule limits as well as how we tabulated a set of contingency tables.

3.1 Diagnostic Measure

All first deltas (forecast horizon one actual value minus last actual value) were standardized using the mean and standard deviation of each series's past deltas. The time series tended to be declining at the forecast origin, so we focused on exceptional declines. First deltas that were 1.28 standard deviations below the mean of all the past deltas were considered a large change or exceptional (positives). Of the total 474 monthly, micro time series, 95 were exceptional conditions and 379 were not.

3.2 Forecast Measure

All first forecast deltas (forecast horizon one minus last actual value) for each forecast method were normalized according to the mean and standard deviation of each series' past deltas. 101 z-value thresholds were established using the inverse of the standard normal cumulative distribution from a p-value of 0 to a p-value of 1, with a difference of .01 between each.

For every threshold, normalized first forecast deltas less or equal to the z-value threshold were considered to be a signaled large change or signaled positive. Normalized first forecast deltas greater than the z-value threshold were considered to be a signaled non-large change or signaled negative. Forecasts methods with signaled 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 all 101 z-value thresholds.

True Positive Rates (TPRs, number of true positives divided by number of actual positives) and False Positive Rates (FPRs, number of false positives divided by number of actual negatives) were computed to obtain 101 monotonically increasing two-dimensional points (FPR, TPR) for each method. The connection of these points created each method's ROC curve and statistical tests were applied from Section 2.

4. Results

We decided to limit analysis to the PAUC measure for false positive rates between 0.00 and 0.20, believing that this would include the range with which most managers would be comfortable.

4.1 Partial Area Under Curve

Over the FPR range of 0.00 to 0.20, Theta was the top performer and had an area that was statistically greater than all but five methods. See Exhibit 5.

Exhibit 3

Paired comparisons of PAUC for FPR range 0 to 0.20 with the Theta forecasting method using bootstrapping.

<i>Forecasting Method</i>	<i>PAUC</i>	<i>t-stat</i>	<i>p-value</i>
Theta	0.128		
AutomattANN	0.126	-0.2	0.82
Flores Pearce 2	0.124	-0.6	0.54
Forecast Pro	0.123	-0.8	0.43
Smart FCS	0.121	-1.0	0.30
PP Autocast	0.118	-1.8	0.076
Dampen	0.114	-2.2	0.025
Box-Jenkins	0.114	-1.9	0.063
Flores Pearce 1	0.113	-2.2	0.028
Forecast X	0.110	-2.6	0.0085
Comb SHD	0.108	-3.1	0.0017
Autobox3	0.104	-2.6	0.0082
Holt	0.102	-3.8	0.00013
Theta SM	0.102	-3.5	0.00041
Autobox2	0.101	-3.3	0.00089
Single	0.100	-3.7	0.0002
Winter	0.100	-3.9	0.000094
AAM2	0.093	-3.7	0.0002
AAM1	0.093	-3.7	0.0002
Autobox1	0.086	-4.9	1.20E-06
Ararma	0.085	-4.7	2.80E-06
Naïve2	0.057	-5.9	3.10E-09
Robust-Trend	0.036	-9.3	0

4.2 Complexity

This section investigates the effect of forecast method complexity on ROC performance, measured by PAUC. Previous research ranked the M3 forecast methods for complexity (Koning et al., 2005). We eliminated Rule-Based Forecasting from 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 comparing forecasts to last historical value. We rescaled the published rankings accordingly and averaged them to yield the complexity variable in Exhibit 6. We created a dummy variable for Box-Jenkins methods, expecting such methods to behave differently than the other methods. Exhibit 4 also includes the PAUC for $FPR \leq 0.20$. We expected the relationship between complexity and partial AUC to be positive.

Exhibit 4.

Data set for PAUC as influenced by forecast method complexity.

<i>Forecasting Method</i>	<i>BJ</i>	<i>Complexity</i>	<i>PAUC</i>
Single	0	2	0.100
Holt	0	3	0.102
Robust-Trend	0	4.33	0.082
Winter	0	5.33	0.100
Dampen	0	6.17	0.114
PP Autocast	0	7.5	0.118
Theta SM	0	8	0.102
Theta	0	11	0.128
Comb SHD	0	11.33	0.108
Box-Jenkins	1	11.67	0.114
Autobox1	1	12.33	0.086
Autobox3	1	14.67	0.104
Autobox2	1	14.67	0.101
ARARMA	1	15.17	0.085
Smart FCS	0	16.83	0.121
Flores-Pearce2	0	16.83	0.124
Flores-Pearce1	0	16.83	0.113
Forecast X	0	18.83	0.110
Forecast Pro	0	21.17	0.123
AutomatANN	0	21.5	0.126

We regressed PAUC complexity and the Box-Jenkins dummy variable, with results reported in Exhibit 5. Results were significant for both variables at a p-value < 0.05 and in the expected direction for complexity. Box-Jenkins methods as a group are significantly worse than the other

methods, in regard to partial AUC performance. Thus we have evidence that forecast method complexity positively affects ROC performance of forecast methods.

Exhibit 5.

Regression results for effect of complexity on ROC performance.

<i>Variable</i>	<i>Coefficients</i>	<i>Standard Error</i>	<i>t-stat</i>	<i>p-value</i>
Intercept	0.09765	0.00538	18.144	0
BJ	-0.01614	0.00548	-2.946	0.00904
Rank	0.00121	0.00041	2.946	0.00904

Regression Statistics

Multiple R	0.68210
R Square	0.46526
Adjusted R Square	0.40235
Standard Error	0.01045
Observations	20

4.3 Dominant Methods

For any given false positive rate, the forecasting method with the highest true positive rate is dominant. Exhibit 6 shows those forecasting methods whose ROC curves are dominant within specified FPR ranges. Exhibit 7 has ROC curves for a subset of top-performing forecast methods as well as some dominated methods for comparison. Although Flores Pearce 2 and Smart FCS were not significantly worse than Theta in Exhibit 5, they were not included in Exhibit 7 because they were not dominant over any range.

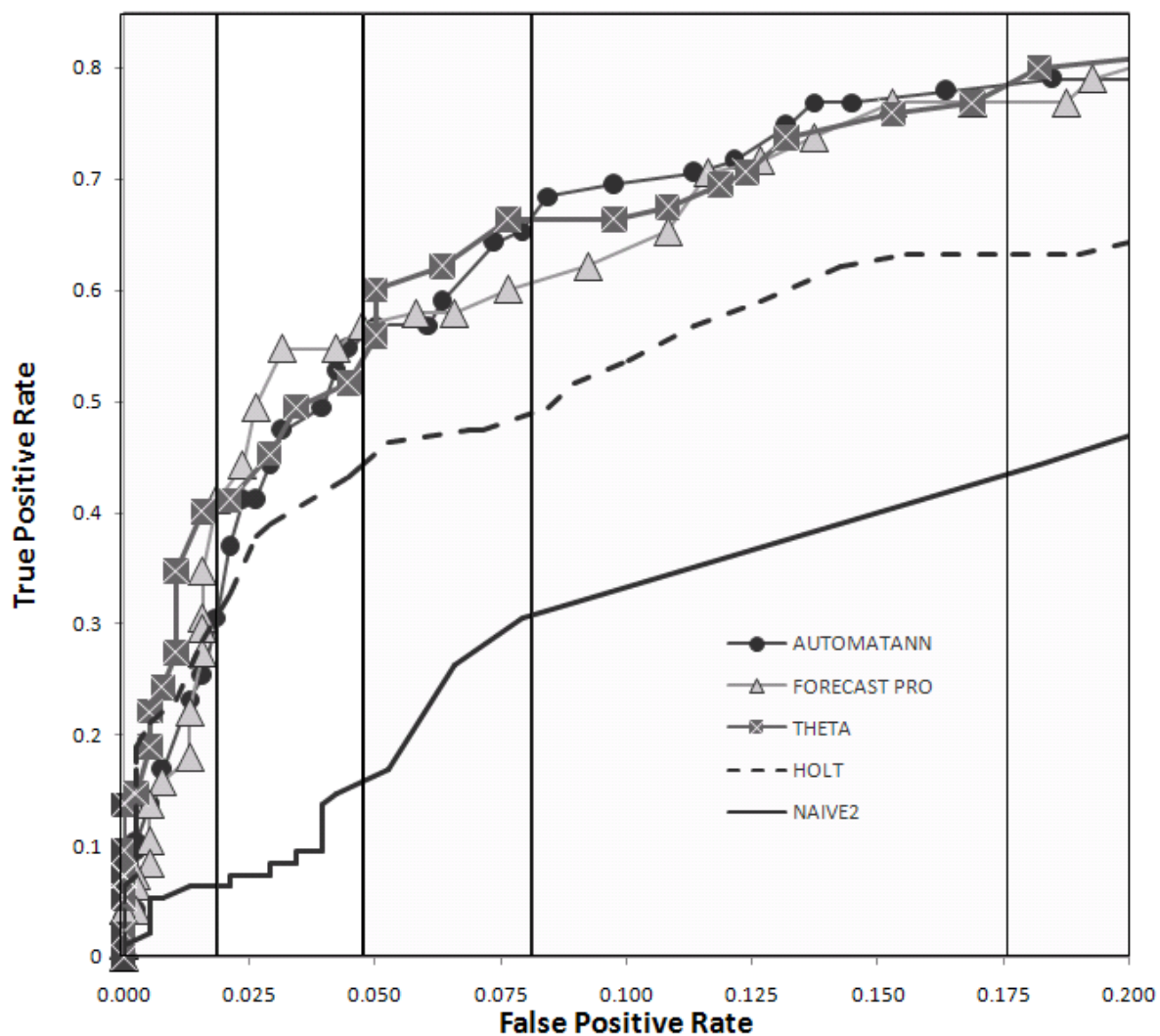
Exhibit 6. Dominant forecast methods by FPR range.

<i>Forecasting Method</i>	<i>Lower FPR</i>	<i>Upper FPR</i>	<i>Lower TPR</i>	<i>Upper TPR</i>	<i>Partial AUC</i>
Theta	0.00	0.02	0.00	0.40	0.0058
Forecast Pro	0.02	0.05	0.41	0.57	0.0158
Theta	0.05	0.08	0.55	0.66	0.0190
AutomattANN	0.08	0.17	0.67	0.78	0.0658
Theta	0.17	0.21	0.77	0.82	0.0321
Forecast Pro	0.21	0.26	0.82	0.83	0.0415
Forecast X	0.26	0.29	0.84	0.86	0.0257

Flores Pearce 2	0.29	0.40	0.86	0.92	0.0986
Forecast Pro	0.40	0.50	0.93	0.94	0.0935
Flores Pearce 2	0.50	0.58	0.94	0.96	0.0762
Theta	0.58	0.68	0.95	0.97	0.0966
Forecast Pro	0.68	0.83	0.97	0.99	0.1482
Box Jenkins	0.83	0.90	0.99	1.00	0.0700
Various	0.90	1.00	1.00	1.00	0.0990

Exhibit 7.

ROC curves for a selection of forecast methods: three top performers for $FPR \leq 0.20$ and two dominated methods.



4.4 Combination forecasts

It is well-known that a simple average combination of methods often forecasts most accurately (e.g., Clement, 1989). We created such a combination forecast using the average of the top three forecast methods (Theta, Forecast Pro, and AutomatANN). We also created a maximum combination forecast that had signaled positives whenever any of the top three methods had a signaled positive. Additionally, a median combination forecast was created that had signaled positives whenever two of the top three methods had a signaled positive. Exhibit 8 shows the average combination forecast to have a statistically greater PAUC for one out of three of its three component methods. Exhibit 9 shows the ROC curves for the average combination forecast method and its components.

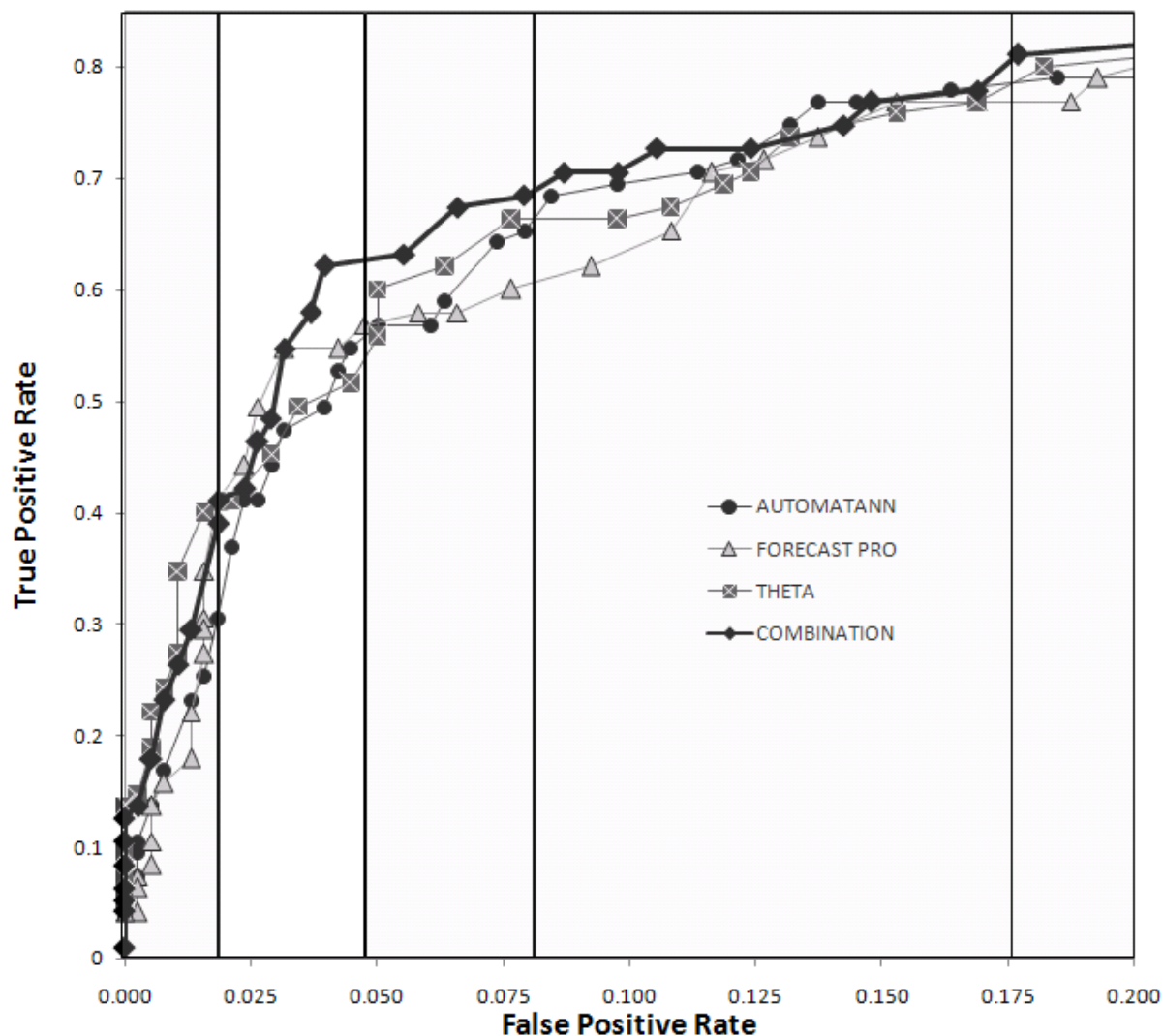
Exhibit 8.

Paired comparisons of the combination forecasts partial AUC with its components via bootstrapping.

Method	Partial AUC	Std Error	Delta Std Error	z-value	p-value
Average Combo	.132	.0095			
Max Combo	.131	.0095	.0096	-0.3	.38
Median Combo	.130	.0095	.0031	-0.9	.18
Theta	.128	.0094	.0041	-1.1	.13
AutomattANN	.126	.0094	.0048	-1.3	.09
Forecast Pro	.123	.0093	.0042	-2.1	.02

Exhibit 9.

ROC curves for the average combination forecast method and its component methods.



5. Conclusion

This paper has applied ROC analysis to M3 Competition micro monthly time series for one-month-ahead forecasts. Using the partial-area-under-the-ROC-curve (PAUC) criterion and paired comparison testing via bootstrapping, 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. A regression model of PAUC on a judgmental index for forecast method complexity and a dummy variable for Box-Jenkins methods provides further confirming evidence.

We also found that a combination forecast, consisting of the simple arithmetic average of the top three methods, to perform better than the component methods as well as combinations constructed from “OR” rules with any one or pair of methods signaling exceptions.

References

- Cohen, J., S. Garman, & W. L. Gorr (2006). Empirical calibration of time series monitoring methods using receiver operating characteristic curves, Heinz School, Carnegie Mellon University Working Paper 2007-27.
- DeLong, E. R., D. M. DeLong, and D. L. Clarke-Pearson. 1988. Comparing the areas under two or more correlated receiver operating curves: A nonparametric approach. *Biometrics* 44: 837-845.
- Dorfman, D.D. and Alf, E. Jr. (1969) Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals -Rating method data. *J. Math. Psychol.* 6, 487-496.
- Gorr, W. L (2007). Forecast accuracy measures for exception reporting, Heinz School, Carnegie Mellon University Working Paper 2007-27.
- Hanley, J. A., and B. J. McNeil. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143: 839-843.
- 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, p. 839-843.
- Hibon M. and Makridakis S. [2000]. The M3 Competition: results, conclusions and implications. *International Journal of Forecasting* 16, 451-476.
- Janes, H., Longton, G. M., & M. Pepe, "Accommodating Covariates in ROC Analysis" (January 22, 2008). UW Biostatistics Working Paper Series. Working Paper 322. <http://www.bepress.com/uwbiostat/paper322>
- Koning, A. J., P. K. Franses, M. Hibon, & H. O. Stekler, H.O. (2005). The M3 competition: Statistical tests of the results, *International Journal of Forecasting* 21, 397-409
- Pepe MS, Longton GL. 2005. Standardizing markers to evaluate and compare their performances. *Epidemiology*. 16 (5): 598–603.