Davydenko A., a.davydenko@lancaster.ac.uk

LUMS' Inaugural Research Conference, June 30, 2010, Lancaster, UK

How to Measure the Quality of Demand Forecasts Efficiently:
a New Class of Forecasting Performance Metrics

Demand data is typically characterised by a large number of time series with a high variation of actual values. In case of intermittent demand the available data also contains a high proportion of zero or relatively low (compared to forecast errors) actual values. In these conditions well-known error measures cannot be efficiently applied since they become vulnerable to outliers and biases induced by corresponding averaging techniques. By using an enhanced calculation scheme, the proposed class of metrics is aimed to overcome the limitations of existing approaches and to ensure a reliable and comprehensive comparison of demand forecasts.

The features shown on Fig. 1 render well-known error measures unsuitable when measuring the quality of demand forecasts (see the table below). The proposed enhanced scheme is based on aggregating benchmark ratios representing various relative characteristics of errors across time series. To ensure the correct properties of a benchmark summary indicator (see Fleming and Wallace 1986) the aggregation of benchmark ratios is performed using the weighted geometric mean.

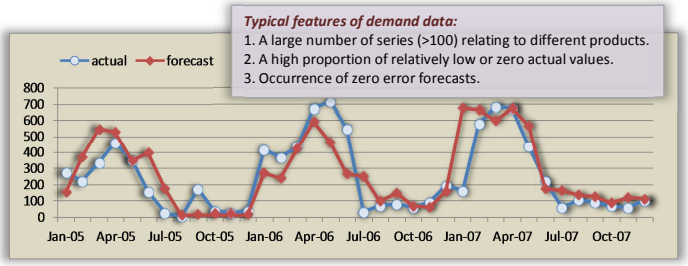

Fig. 1. A real-world demand time series with one-step-ahead managerial forecasts.

Measure	Calculation scheme	Applicability/Limitations
All measures based on percentage errors (PEs)	Let the forecasting error for a given time period t and product i be $e_{i,t} = Y_{i,t} - F_{i,t}$, where $Y_{i,t}$ is a demand value for product i observed at time t, $F_{i,t}$ is the forecast of $Y_{i,t}$. The percentage error (PE) is found as $p_{i,t} = 100 \times e_{i,t}/Y_{i,t}$.	PEs can be aggregated across periods and across series, but PE-based measures have the following general limitations: - observations with zero actual values cannot be processed; - dividing by low actuals results in extreme percentage values that do not allow for a useful interpretation (since they are not necessarily harmful or damaging); - therefore the evaluation of intermittent demand forecasts becomes intractable due to a large proportion of zero and close to zero actual values; - all PE-based measures can be misleading when the improvement in accuracy correlates with actual value on the original scale (Davydenko, Fildes and Trapero 2010).
Mean Absolute Percentage Error (*MAPE*) (and other possible measures based on the arithmetic mean of percentage values)	$\mathrm{MAPE} = \mathrm{mean}(\lvert p_{i,t}\rvert)$, where $\mathrm{mean}(\lvert p_{i,t}\rvert)$ is the sample mean of $\lvert p_{i,t}\rvert$ over all available values.	The sample mean of percentage variables gives a highly inefficient estimate and is severely affected by extreme cases (see Hyndman and Koehler 2006).
Median Absolute Percentage Error (*MdAPE*) (and other possible measures based on the median of percentage values)	$\mathrm{MdAPE} = \mathrm{median}(\lvert p_{i,t}\rvert)$, where $\mathrm{median}(\lvert p_{i,t}\rvert)$ is the sample median of $\lvert p_{i,t}\rvert$ over all available values.	The sample median of percentage variables is resistant to the influence of extreme cases, but at the same time it is insensitive to large errors even if they are not outliers or extreme percentage values. Comparing accuracy using MdAPE shows the improvements related to 50% of lowest absolute percentage errors. However, the improvement of MdAPE can be accompanied by more damaging remaining errors lying above the median if the shapes of error distributions differ. Therefore MdAPE is not easily interpretable and is not sufficiently indicative of changes in forecasting performance when methods have different shapes of error distributions. If the dispersion of errors remains the same, MdAPE makes a comparison in favour of a method producing errors with a heavier tailed distribution.
Measures based on relative errors	Let the relative error be $\mathrm{RE}_{i,t} = e_{i,t}/e^{\mathrm{b}}_{i,t}$, where $e^{\mathrm{b}}_{i,t}$ is the forecast error obtained from a benchmark method. Then $\mathrm{Mean\ Relative\ Absolute\ Error\ (MRAE)} = \mathrm{mean}(\lvert \mathrm{RE}_{i,t}\rvert)$, $\mathrm{Median\ Relative\ Absolute\ Error\ (MdRAE)} = \mathrm{median}(\lvert \mathrm{RE}_{i,t}\rvert)$, etc.	Averaging ratios of absolute errors across individual observations overcomes the problems related to dividing by actual values, but has the following limitations: - REs cannot be obtained in cases of zero forecasting errors (when the actual and forecasted demands coincide); - REs are prone to outliers arising when original errors are close to 0 (see Hyndman and Koehler 2006).
Mean Absolute Scaled Error (*MASE*) proposed in (Hyndman and Koehler 2006)	For the scenario when forecasts are produced from varying origins but with a constant horizon the MASE is found as $\mathrm{MASE} = \mathrm{mean}(\lvert q_{i,t}\rvert), \quad q_{i,t} = \frac{e_{i,t}}{\mathrm{MAE}^{\mathrm{b}}_i}$, where $\mathrm{MAE}^{\mathrm{b}}_i$ – mean absolute error (MAE) of a benchmark (naïve) forecast for series i.	MASE overcomes some of the disadvantages of the previous schemes. However, it was shown in (Davydenko, Fildes and Trapero 2010) that MASE is equivalent to the weighted arithmetic mean of MAEs. As a result of using the arithmetic mean to average benchmark ratios across series the MASE scheme - induces a bias towards overrating benchmark; - is affected by extreme cases when MAE of benchmark forecast is relatively low. MASE is also vulnerable to outliers or structural breaks in time series history (Kolassa and Schütz 2007).
The proposed metrics to indicate average relative characteristics of errors (*AvgRelMAE*, *AvgRelMSE*, etc.)	The proposed statistic representing average relative performance is constructed as $L = \left(\prod_{i=1}^{m} r_i^{n_i}\right)^{1/\sum_{i=1}^{m} n_i}, \quad r_i = \frac{c_i}{c_i^{\mathrm{B}}}$, where m – total number of time series, c_i^{B} – characteristic of forecasting errors of a benchmark method for series i, c_i – characteristic of forecasting errors of the method being evaluated against the benchmark for series i, n_i – number of observations used to find r_i. For example, $\mathrm{AvgRelMAE} = \left(\prod_{i=1}^{m} \mathrm{RelMAE}_i^{n_i}\right)^{1/\sum_{i=1}^{m} n_i}, \mathrm{RelMAE}_i = \frac{\mathrm{MAE}_i}{\mathrm{MAE}_i^{\mathrm{B}}}$.	The advantages of this approach is that it - gives easily interpretable and informative (useful) results; - efficiently uses all available information; - ensures objective comparison without introduction of biases or outliers induced by the calculation procedure itself; - is suitable for intermittent demand or low actual demand data, as well as for data containing zero errors or negative observations; - enables a comparison of forecasts according to various possible criteria (such as average relative improvements in terms of a specified loss function); - can be directly extended to robust schemes.

Application to Real-World Data

The presented statistic was applied to evaluate the effectiveness of managerial adjustments to model-based forecasts (Davydenko, Fildes and Trapero 2010). From Fig. 2 it can be seen that the analysis based on PEs or scaled errors is unreliable due to the highly diffuse and skewed distributions. AvgRelMAE is less affected by extreme cases and ensures a more objective comparison of forecasts. Using the new metric it was possible to describe the accuracy of final managerial forecasts in terms of an average relative improvement of MAE of model-based forecast (see the table below).

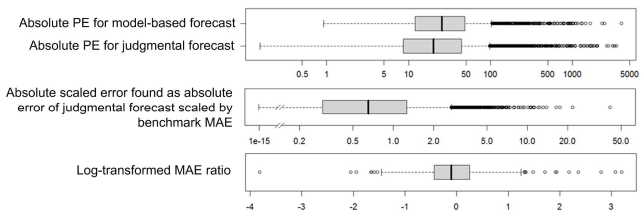

Fig. 2. Boxplots for absolute percentage errors, absolute scaled errors, and log-transformed MAE ratios.

Accuracy of Judgmental Adjustments According to Different Error Measures				
	MAPE (2% trim)	MdAPE	MASE	AvgRelMAE
Model-based (statistical) forecast	**34.51 %**	24.98 %	**1.00**	1.00
Judgmentally adjusted forecast	37.22 %	**19.98 %**	1.02	**0.90**

Conclusions

The proposed scheme overcomes many limitations and disadvantages of well-known error measures. It can be efficiently used to compare the quality of demand forecasts across time series with minimal assumptions about the features of the data. The proposed general metric ensures informative and objective comparison with the benchmark method. Using the general scheme it is possible to construct aggregated indicators of relative performance in accordance with different criteria such as improvements in MAE or MSE. The example of application to real-world data has shown that these metrics can be used to perform a comprehensive and reliable analysis of forecasting performance in practical settings.

References

Davydenko, A., R. Fildes, and J. Trapero. *Measuring the Accuracy of Judgmental Adjustments to SKU-level Demand Forecasts*. Lancaster University Management School Working Paper 2010/26, 2010.

Fleming, P.J., and J.J. Wallace. "How not to lie with statistics: the correct way to summarize benchmark results." *Communications of the ACM*, 29, no. 3 (March 1986): 218-221.

Kolassa, S., and W. Schütz. "Advantages of the MAD/Mean Ratio over the MAPE." *Foresight: The International Journal of Applied Forecasting*, no. 6 (2007): 40-43.

Hyndman, R.J., and A.B. Koehler. "Another look at measures of forecast accuracy." *International Journal of Forecasting* 22, no. 4 (2006): 679-688.