Food Wholesales Prediction: What is Your Baseline?

Jorn Bakker and Mykola Pechenizkiy

Eindhoven University of Technology
P.O. Box 513, NL-5600 MB, Eindhoven, The Netherlands
{j.bakker, m.pechenizkiy}@tue.nl

Abstract. Sales prediction is an important problem for different companies involved in manufacturing, logistics, marketing, wholesaling and retailing. Different approaches have been suggested for food sales forecasting. Several researchers, including the authors of this paper, reported on the advantage of one type of technique over the others for a particular set of products. In this paper we demonstrate that besides an already recognized challenge of building accurate predictive models, the evaluation procedures themselves should be considered more carefully. We give illustrative examples to show that e.g. popular *MAE* and *MSE* estimates can be intuitive with one type of product and rather misleading with the others. Furthermore, averaging errors across differently behaving products can be also counter intuitive. We introduce new ways to evaluate the performance of wholesales prediction and discuss their biases with respect to different error types.

1 Introduction

The success of different companies depends today on their ability to adapt quickly to the changes of their business environment. An accurate and timely sales prediction is particularly important for the companies involved in manufacturing, logistics, marketing, wholesaling, and retailing.

In the food and beverages market, food service companies often have to deal with short shelf-life products, and uncertainty and fluctuations in consumer demands. These variations in consumer demand may be impacted by the high number of factors including e.g. price change, promotions, changing consumer preferences, or weather changes [5]. Furthermore, a large share of the products sold in that market is sensitive to some form of seasonal change due to the different cultural habits, religious holidays, fasting, and alike. All these factors imply that some types of products are sold mostly during the limited period(s) of time.

Although it is known that some seasonal pattern is expected, the predictive features that define these season are not always directly observed. Therefore, drops and rises in sales which are accommodated by the changing seasons are often difficult to predict. Regarding inventory management, this results often in a stock-out at the start of the season and perishable or obsolete goods at the

end of a seasonal period. Thus, both shortage and surplus of goods can lead to loss of income for the company.

Time-series research has been traditionally suggesting ARIMA (autoregressive moving average) and ANN (artificial neural networks) approaches to address the problem of sales prediction. Despite of the continuous efforts devoted to come up with a right algorithm, and a number of comparative studies focused on identifying the strongest one, researchers are not clearly in favor of one particular method. Nonlinearity prevents the success of simple linear models, while rather short lengths of the time series are insufficient to learn more complex models [1]. It is rather intuitive that no single method is best in every situation and that combining different models might be an effective way to improve accuracy of (sales) prediction. Interestingly, both data mining and time series forecasting research pointed out into this promising direction [6] [4].

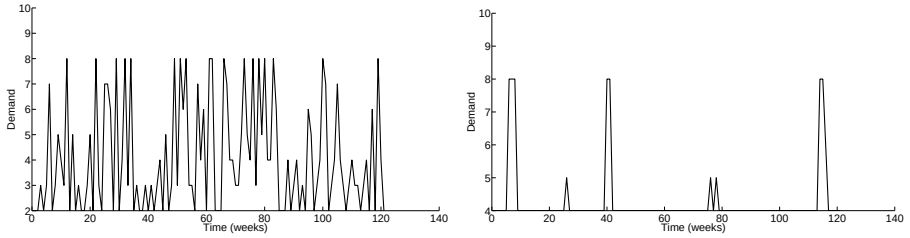
Anyhow, the challenge of building accurate predictive models has been already recognized among both researchers and practitioners. In this paper we reconsider the problem of evaluating the performance of time series forecasting and, particularly, food wholesales prediction and emphasize that this issue is also far from being trivial. Sales data typically comprises of many different products that exhibit very different types of behavior. Standard error measures like Mean Absolute Error (*MAE*) and Mean Squared Error (*MSE*) yield biased results when applied on the different types of time series. They can be intuitive with one type of product and rather misleading with the others. Naturally, we often want to compare the performance of several methods across a number of products (time series). This requires an error or accuracy performance measure to remain intuitive when aggregated over several datasets. Due to imbalances and structural differences this is not always possible (or not advisable). We compare the intuition behind the different popular error measures, discuss their limitations, and introduce new approaches and measures that may allow to get a better insight on the prediction performance.

2 Food sales prediction evaluation

In this section we illustrate that not just the wholesales prediction but also the evaluation and comparison of different prediction techniques across various products (datasets) is not trivial and requires careful considerations. Let us illustrate first that traditional *MAE* or *MSE* like measures can be rather unintuitive because sales data typically comprises of many different products that exhibit very different types of behavior.

Consider wholesales figures for two products given in Fig. 1; *Product 1* has a lot of variation of (and no constant) demand whereas *Product 2* is periodic and shows constant demand between the peaks. Taking an error measure like the *MSE*, it would be easy to achieve a good performance on the highly periodic series (like with *Product 2*) by taking a naive predictor that just chooses the last observed value as the prediction for the next point, or always outputs the most popular value, i.e. the value corresponding to the constant demand in this case,

or computes a moving average. Thus, *MSE* of an optimal predictor will be close to *MSE* of a naive predictor that makes the comparison of *MSE*'s of different predictors meaningless. (Since from the domain perspective the peak demand is more important to predict than long lasting flat areas, we can see here also additional connections to the issues of class imbalance and one-class classification that are well-known in machine learning). It is not difficult to notice that *MSE* of the same naive predictor for the *Product 1* would lead to very bad results but e.g. a moving average approach would perform reasonably well, i.e. likely not worse that performance of any *learnable* predictor. Thus, if we try to aggregate the *MSE*'s over the two products, the average *MSE*'s will be misleading. Therefore, using the *MSE* is not preferable if we want the performance measure to yield a result that is intuitively comparable over all the time series in the data set.



(a) Product 1 (b) Product 2

Fig. 1. The structural difference between two representative products.

This claim with respect to the *MSE* can be generalized to any error measure that uses the unscaled prediction error. In order to address this issue, a scaled measure and a baseline that provides the reference scale for the performance measurement is needed. We will consider a couple of corresponding possibilities.

2.1 Error Measures

Error measures that have been proposed in the literature [2] and were commonly applied for evaluation of time series forecast include:

- Mean Squared Error: $MSE = \frac{1}{n}\sum_{t=1}^{n} e_t^2$,
- Root Mean Squared Error: $RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2}$,
- Mean Absolute Error: $MAE = \frac{1}{n}\sum_{t=1}^{n} |e_t|$,
- Mean Absolute Percentage Error: $MAPE = \frac{1}{n}\sum_{t=1}^{n} |p_t|$,
- Mean Absolute Scaled Error: $MASE = \frac{1}{n}\sum_{t=1}^{n} \left|\frac{e_t}{MAE(Baseline)}\right|$,

where the error $e_t = y_t - f(t)$, y_t is the actual value, $f(t)$ is the predicted value by model f at time t, and the percentage error $p_t = \frac{e_t}{y_t}$.

Both *MSE* and *RMSE* are well known and widely used to validate the accuracy of a model. In the machine learning field these measures are used to evaluate the performance of a given algorithm. In the forecasting of time series, however, they are deemed not so suitable because of the aforementioned scaling differences and the sensitivity to outliers. We will present and discuss a scaled version of the *MSE*.

The *MAPE* has been recommended for measuring accuracy among many different time series. However, it should be noticed that in cases where y_t is very close to zero, the resulting *MAPE* will become infinite or invalid. The same holds for the *MASE*, i.e. in case the *MAE(Baseline)* is close to zero, but this case is special in the following sense. The *MASE* uses, in contrast with the other error measures, explicit scaling with respect to some baseline. Notice that if *MAE(Baseline)* is close to zero, the baseline itself is a good predictor. The advantage of *MASE* is that the accuracy of a given model can directly be related to the baseline regardless of scale.

2.2 Baseline Predictors

Selection of suitable baselines is important for identifying reference points which would allow comparing among different alternative techniques, but also to have a better understanding of how much worse (or better) a particular technique performs with respect to known optimal (or simply reasonably good) and worst (or clearly bad) cases.

Naive prediction baseline. The naive prediction baseline ("choose the last observed value as the prediction for the next point") is a widely used baseline in forecasting methods. The intuition behind using this baseline is, that regardless of the accuracy of a given predictor it should always perform better than the naive prediction. Scaling towards the naive predictor does not have an upper bound. In our investigation we consider the *MAE* applied to the naive predictor (f_{naive}):

$$MAE(f_{naive}) = \frac{1}{n-1}\sum_{t=1}^{n}|y_t - y_{t-1}|. \tag{1}$$

Worst case prediction baseline. The worst case scenario gives us an upper bound of poor performance and can be used to scale the error of different predictors between 0 and 1 and directly compare the predictive performance of different algorithms. This approach can only be used in cases where the maximal value can be computed. But, a priori, this baseline also suffers from a bias with respect to structurally different time series. In the new error measure that we introduce in the next subsection, the *MSE* of the worst case (f_{WC}) baseline is used:

$$f_{WC} = \frac{1}{n}\sum_{t=1}^{n}(y_t - \max_{i=1}^{\alpha}|i - y_t|)^2, \tag{2}$$

where α is the number of levels to which time series is approximated.

Sample biased evaluation. The bias of the f_{WC} prediction can be decreased by selecting "interesting" data points from test data. The "interesting" parts of the time serie in Fig. 1 are not the long stretches of constant values, but the peaks. If only the peaks are taken into account in the accuracy calculation, the error estimate becomes more adequate from the domain point of view.

The selection of test data points to be considered in the accuracy calculation should be handled with care. The only points eligible to be deselected are points for which the following two properties hold:

i) the last actual value y_{t-1} is equal to y_t, and
ii) the error $e_t = 0$.

All other points are in the test data. In other words, this selection procedure selects everything except for the points that did not change in the recent past and have been predicted correctly by all the considered approaches. This approach is similar to the $MASE$ with the important difference that it is scaled to an interval between 0 and 1:

$$f_{WCscaled} = \frac{MSE(f(\bar{t}))}{MSE(f_{WC}(\bar{t}))}, \tag{3}$$

where $\bar{t}$ is the vector of selected points, $f(\bar{t})$ the output of the prediction model, and $f_{WC}(\bar{t})$ the worst case prediction. This approach has some similarities to computing misclassification error separately for the true positive class.

3 Experiment design and results

In order to demonstrate the characteristics of the aforementioned error measures we conducted experiments on a real wholesales data. In this section we present an overview of the experiment design, the results with respect to the error measures, and some additional tradeoffs in the evaluation of food sales prediction algorithms.

3.1 Experiment design

For our study we selected several products provided by Sligro Food Group N.V., which encompasses food retail and food service companies selling directly and indirectly to the entire Dutch food and beverages market and has about 60.000 products in stock. The products are selected in such a way that they represent different type of behavior (more seasonal vs. more chaotic, cf. Fig. 1) to demonstrate and investigate the bias of different accuracy measures in different types of time series.

Data preprocessing. The data warehouse consists of all the transactions made in a period of over two years. For weekly predictions (which are most important for wholesales), the resulting time series of aggregated transactional data contains 120 data points, from which the first 77 instances are used as the training set and the last 43 instances are used for progressive evaluation of the predictors.

Accumulated and aggregated transactional data was transformed with piece wise approximation to 8 levels that reflect the variation in sales from very low (1) to very high (8). Thus the data has a predefined upper bound and we can compute the error of the worst case.

Besides the standard time-series features like history of sales, moving averages, and slopes each data (time) point contain information about promotions, (school and public) holidays and weather which are known to impact the wholesales for certain types of products. A simple filter-based individual feature selection is used to address the problem of high dimensionality.

Learning techniques. We experimented with three predictors: a moving average over a window of size 6 (*MA6*), a logistic regressor (*LR*), and an ensemble learning algorithm (*ENS*). The moving average, is a very basic predictor that is being used in practice to aid prediction of demand. The logistic regression, is a method that is commonly used in prediction problems. The ensemble learning, is known to be a promising approach for prediction in changing environments [3], and recent studies in time series forecasting and data mining have shown that combining different classifiers for sales prediction can lead to better results [6] [4].

3.2 Results

The results of the experiments are displayed in Table 1. For each prediction method and product we present the error estimates computed over the test (i.e. out-of-sample) data with different considered error measures. For all of the error measures in the table the smaller the value the accurate the predictor is.

The first thing to be observed is the difference between the *MSE*, *RMSE*, *MAE*, and *MAPE* and the results of *MASE* and $MSE(f_{WCscaled})$. For the first group of error measures it holds that the forecasting results on *Product 1* are worse than for *Product 2*. This is due to the fact that predicting the contant demand is easy for every considered technique. Not surprisingly, the *MASE* and $MSE(f_{WCscaled})$ send an opposite message. In the case of *MASE* all three predictors perform worse than the naive predictor in case of *Product 2* and better than the naive predictor in case of *Product 1*.

The second thing to note is the difference between *MASE* and $MSE(f_{WCscaled})$. While in the *MASE* case *MA6* performs worse than the *ENS*, the $MSE(f_{WCscaled})$ shows that the *MA6* performs better. Please notice that these two measures are on completely different scale, so a direct comparison is hard. In Fig. 2 we can see what kind of errors (i.e. difference between the true labels and predictions) different predictors make.

Apart from the $MSE(f_{WCscaled})$, each of the errors shown in Table 1 are unbounded. Since the $MSE(f_{WCscaled})$ is scaled between 0 and 1 with respect to

Table 1. Performance of *MA6*, *LR*, and *ENS* on *Product 1* (P1) and *Product 2* (P2)

.

		MA6		LR		ENS	
	Range	**P1**	**P2**	**P1**	**P2**	**P1**	**P2**
MSE	$0 \,..\, \infty$	3.79	1.09	6.47	1.05	5.88	1.23
RMSE	$0 \,..\, \infty$	1.95	1.05	2.54	1.02	2.43	1.11
MAE	$0 \,..\, \infty$	1.51	0.49	2.05	0.40	1.84	0.35
MAPE	$0 \,..\, 1$	0.45	0.10	0.60	0.07	0.50	0.59
MASE(f_{naive})	$0 \,..\, \infty$	0.93	2.28	0.93	2.17	0.90	1.63
MSE$(f_{WCscaled})$	$0 \,..\, 1$	0.14	0.16	0.22	0.05	0.20	0.18

the worst case policy, the values shown here can be considered as traditional misclassification errors.

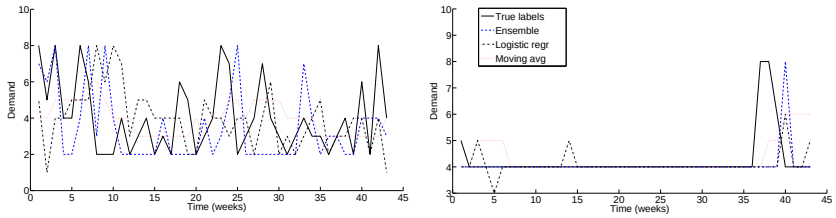


Fig. 2. The true and predicted sales for products from Fig. 1.

Summary on the considered measures. Let us remind that our aim is to find a performance measure that enables aggregating the performance results over all data sets in a given database regardless of their structural differences. Any error measure that is not scaled cannot be used for this purpose. Compare, e.g., the outcome of the *MSE* (or *RMSE*) results between the two products in Table 1. For all algorithms that were tested, it holds that $MSE_{P1} > MSE_{P2}$, whereas the same is not true for the *MASE* and *MSE*$(f_{WCscaled})$ measure. The question is whether the *MASE* and *MSE*$(f_{WCscaled})$ measures are reliable enough in order to allow for cross product validation.

The *MASE* measure provides a direct way to compare the predictor to a meaningful baseline. When comparing the outcomes of the unscaled measures in Table 1 to the Fig. 2 for *Product 2*, it becomes immediately clear that something is wrong. Where the unscaled measures report a very low error, the *MASE* indicates that all of the algorithms perform worse than the naive predictor. However, since *MASE* is unbounded, it does not give a relative and normalized accuracy measure.

The $MSE(f_{WCscaled})$ measure provides an accuracy measure that is bounded. Since the measure is scaled towards the worst case predictor, it is only usable if the upper bound of the time series is known. In principle, this measure can be aggregated over all the different products in the database. What remains to be seen is whether the selection procedure is fair enough to provide an unbiased accuracy measure. If the amount of selected points in the test set is relatively low the measure can become biased because of the underlying *MSE* measure.

Both *MASE* and $MSE(f_{WCscaled})$ give a more accurate and intuitive performance measure than the traditional evaluation measures. The *MASE* is particularly useful in cases where ranking is used between different predictors because of its comparing nature. The $MSE(f_{WCscaled})$ gives a direct error measure on the prediction, but it assumes that a maximum value is known for the time series. This last assumption is not trivial in the context of data streams.

Other biases in predictions. In the domain of food sales predictions there are actually different *types* of errors with different impact on the performance. An overestimation of demand will have different impact on the outcome (application of) the prediction than an underestimation. Therefore, performance measures that take this into account seem natural in this context.

Comparing how often predictors forecast either too low or too high, might indicate a bias of each predictor towards certain type of error. In Fig. 3 the number of "misses" (estimated too low, i.e. points for which $y_t - f(t) < 0$) and "false alarms" (estimated too high, i.e. points for which $y_t - f(t) > 0$) are shown for a selection of products. Each of these points corresponds to a predictor, the *MA6* (dashed red circles) or *ENS* (solid black circles). Each pair of points corresponding to a certain product is connected via a line. We can observe that the products that have many flat parts are in the lower left corner, whereas the products having more chaotic behavior are in the upper right. We can also see that *MA6* always has higher number of misses, i.e. under predictions, but *ENS* for the majority of product has higher false alarm rates. Similar comparison of different predictors can be performed with respect to "too late" vs. "too early" or other types of errors.

In the field of food sales prediction an error might be less grave if the predicted amount is needed within some safety boundary. If a company overstocks at time t, it might be at some time $t + n$ demand is rising. If the time difference n is then small enough the stock might still get sold, resulting in a lower cost than predicted by the algorithm.

It is often important to know if the demand curve is close to the structural shape of the predicted curve. Dynamic Time Warping (DTW) can be applied to the demand curve and the predicted curve to find the distance reflecting the performance of the predictor (see Fig. 4)*ENS* has, apart from a few examples, a clear advantage over *MA6* when it comes to structural differences. Fig. 5 shows actual DTW mappings for *Product 1* and *Product 2*. Please notice that assigning different costs to each of the aligning directions in DTW we can also introduce a desired bias.

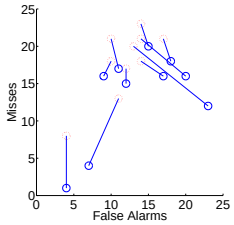


Fig. 3. Number of misses against the number of false alarms.

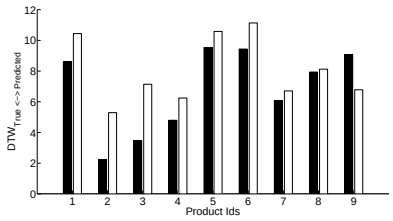


Fig. 4. DTW as accuracy measure; ENS (black), MA6 (white).

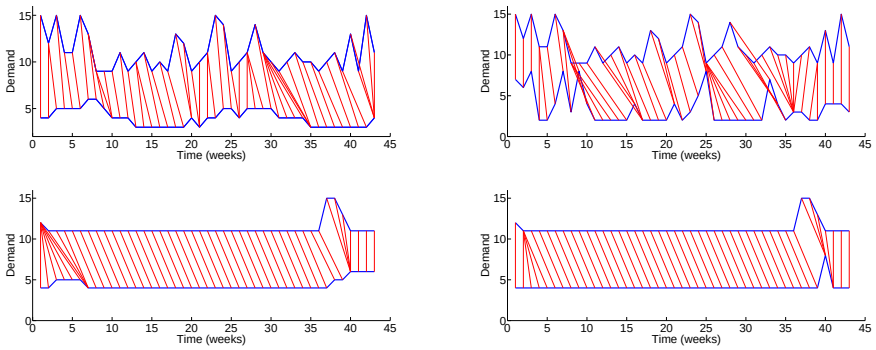


Fig. 5. Tolerating 'wrong time' prediction with nonlinear alignment of the true to the predicted labels for *Product 1* (top) and *Product 2* (bottom) for *MA6* (left) and *ENS* (right). For visualization purposes, the true label values are increased by 7.

4 Conclusions

Food sales prediction is an important and challenging problem having some connections to the problem of predicting in changing environments. In this paper we emphasized that besides this already recognized challenge, the problem of performance evaluation is also far from being trivial. Our previous experience showed that it was not always appropriate to use any of the suggested in the literature measures across different products within a business as a result of rather different behavior in sales, volume and supply characteristics. Here, we considered the different traditional ways of measuring the sales prediction accuracy (that is essential for monitoring and comparing the performance of employed approaches) and discussed and illustrated their limitation with real food sales data. Instead of averaging error estimates across products, someone may try to compare averaged ranks. However, with increasing number of learners to compare and yet questionable appropriateness of a error measure, the averaged rank can be also rather unstable and thus not informative.

In this paper we introduced and experimentally analyzed one new measure that does allow comparing performance of different predictors across different products with different types of time series structure. Beside this, we introduced and demonstrated the use of a generic approaches to measure other biases like optimistic vs. pessimistic and early vs. late prediction biases. We considered the use of the dynamic time warping distance as accuracy measure which may allow to prevent or to tolerate the certain types of errors.

Ultimately, the performance measure would be expressed in the form of a cost function (e.g. on the amount of money the company saves or loses by choosing a particular prediction strategy) that allows directly optimize various parameters with a cost-sensitive learning approach or multi-objective optimization. Our further work in this direction includes development of more generic cost-sensitive approach for evaluating foodsales prediction performance.

Acknowledgements. This research is supported by The Netherlands Organisation for Scientific Research NWO HaCDAIS project. We are thankful to Sligro Food Group BV for providing us with the data and domain knowledge.

References

1. P. Doganis, A. Alexandridis, P. Patrinos, and H. Sarimveis. Time series sales forecasting for short shelf-life food products based on artificial neural networks and evolutionary computing. *Journal of Food Engineering*, 75(2):196–204, 2006.
2. R. Hyndman and A. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688, 2006.
3. L. I. Kuncheva. Classifier ensembles for changing environments. In F. Roli, J. Kittler, and T. Windeatt, editors, *Multiple Classifier Systems*, volume 3077 of *Lecture Notes in Computer Science*, pages 1–15. Springer, 2004.
4. P. Meulstee and M. Pechenizkiy. Food sales prediction: "if only it knew what we know". In *ICDM Workshops*, pages 134–143. IEEE Computer Society, 2008.
5. J. van der Vorst, A. Beulens, W. de Wit, and P. van Beek. Supply chain management in food chains: improving performance by reducing uncertainty. *International Transactions in Operational Research*, 5(6):487–499, 1998.
6. P. G. Zhang. Time series forecasting using a hybrid arima and neural network model. *Neurocomputing*, 50:159–175, 2003.