Chapter 20
Economic Measures of Forecast Accuracy for Demand Planning: A Case-Based Discussion

Thomas Ott, Stefan Glüge, Richard Bödi, and Peter Kauf

Abstract Successful demand planning relies on accurate demand forecasts. Existing demand planning software typically employs (univariate) time series models for this purpose. These methods work well if the demand of a product follows regular patterns. Their power and accuracy are, however, limited if the patterns are disturbed and the demand is driven by irregular external factors such as promotions, events, or weather conditions. Hence, modern machine-learning-based approaches take into account external drivers for improved forecasting and combine various forecasting approaches with situation-dependent strengths. Yet, to substantiate the strength and the impact of single or new methodologies, one is left with the question how to measure and compare the performance or accuracy of different forecasting methods. Standard measures such as root mean square error (RMSE) and mean absolute percentage error (MAPE) may allow for ranking the methods according to their accuracy, but in many cases these measures are difficult to interpret or the rankings are incoherent among different measures. Moreover, the impact of forecasting inaccuracies is usually not reflected by standard measures. In this chapter, we discuss this issue using the example of forecasting the demand of food products. Furthermore, we define alternative measures that provide intuitive guidance for decision makers and users of demand forecasting.

1 Introduction

1.1 Sales Forecasting and Food Demand Planning

Accurate demand forecasts are the backbone of successful demand planning. In particular, for food products with short life cycles the choice of the most suitable

T. Ott (✉) · S. Glüge · R. Bödi
ZHAW Zurich University of Applied Sciences, Winterthur, Switzerland
e-mail: thomas.ott@zhaw.ch

P. Kauf
PROGNOSIX AG, Richterswil, Switzerland

M. Braschler et al. (eds.), *Applied Data Science*,
https://doi.org/10.1007/978-3-030-11821-1_20

forecasting method is of central concern for business and hence the question is a driver for applied research activities. It does not come as a surprise that a plethora of different forecasting methods have been developed and suggested for food demand planning (e.g., Da Veiga et al. 2014; Žliobaitė et al. 2012; Taylor 2007). The most prevalent methods are based on time series models or state space models, notably, exponential smoothing, Holt-Winters method, ARIMA models, Kalman filters or regression models [see, e.g., De Gooijer and Hyndman (2006) for an overview of the most common methods]. Furthermore, artificial neural networks have been used for demand planning for a long time (Doganis et al. 2006), while only more recently other classes of machine learning techniques such as regression trees or random forests (Bajari et al. 2014) have been utilized. Common commercial software solutions for demand planning, such as Inform add*ONE or SAP APO (Vasumathi and Shanmuga Ranjani 2013), typically employ one or more of these methods.

Demand planning takes more than good forecasts. For the actual planning, a number of boundary conditions such as inventory constraints have to be considered. Sales forecasting should focus on the demand of a product irrespective of these constraints as they often distort the figures about the actual demand. In an operational setting we often face the problem of one-step-ahead forecasting, that is, for a product we want to predict the demand at time step t based on the demand observations from times $t-1, t-2, \ldots, t-n$. In the following, we use X_i for the actual demand and F_i for the respective forecast. In order to estimate the past demand values $X_i(i = t-1, t-2, \ldots, t-n)$, the actual sales data is used. Special care has to be taken in stock-out situations, as sales data underestimates the real demand of the product. At the same time, the real demand of some substitute product might be overestimated. Hence, the availability of accurate past demand data is nontrivial. For the following considerations we will ignore this problem and assume that X_i closely reflects the actual demand.

1.2 *Successful Demand Forecasting: The Past and the Future Inside*

Statistical forecasting algorithms try to capture past sales patterns and project them into the future. However, these patterns can be disturbed or can even undergo disruptive changes. An experienced procurement manager has some intuition and beliefs about the driving factors of structural interruptions and their impact on sales quantities. Hence, she or he may adjust the forecasts manually, in accordance with the assumed impact of the driving factors that she or he considers relevant in advance. In practice, a manual intervention is often made when promotions are planned or when an upcoming event or specific weather conditions are supposed to influence sales. Clearly, human intuition can be an important source to incorporate the impact of information about the future, and yet, human intuition is limited. For instance, for humans it is often difficult to grasp cross-effects of many factors and, as

a consequence, humans often tend to overestimate the influence of a single factor. Hence, especially when dealing with a large product portfolio, a need for supporting software solutions arises that evaluate and employ external drivers for enhanced sales forecasting.

An example of such a software solution is PROGNOSIX Demand, which combines various forecasting approaches and additionally incorporates the experience of human experts in cases where not enough (or unreliable) data is available. The methodology is based on the common experience that there is not a single best forecasting method for everything. Depending on the product, available data and the current sales situation, different methods are more or less suitable. Hence, it is important to evaluate the methods in terms of performance, where the performance is usually put into relation with the forecast error, or forecast accuracy, respectively, evaluated over a certain period of time. Subsequently, there is a need for suitable error or accuracy measures. In the following, we will thus first discuss common error measures. However, in practice, one has to decide for one measure in order to judge the performance of different methods and to select the best one. Does it matter which error measure is used? What is the economic significance of the error? The answer is not always clear when using conventional measures, as we will illustrate in the subsequent sections.

1.3 Traditional Measures of Forecast Accuracy

The goal of good forecasting is to minimize the forecasting error(s).

$$e_t = F_t - X_t \tag{20.1}$$

The error is positive, if the forecast is too high, and negative, if the forecast is too low. Usually, the error is defined with opposite signs. Here, in the context of sales forecasting, we prefer the convention in Eq. (20.1), as a positive error means that we have some unsold products left (oversupply).

Traditional measures of forecast accuracy, also referred to as forecast error metrics, can be subdivided into four categories (Hyndman 2006). We will quickly review each by providing the most popular metrics for one-step ahead forecasts.

1. **Scale-dependent metrics** are directly based on the forecast errors e_t
   The most popular measures are the mean absolute error (MAE):

$$\mathrm{MAE}(n) = \frac{1}{n}\sum_{t=1}^{n} |e_t| \tag{20.2}$$

and the root mean square error (RMSE):

$$\mathrm{RMSE}(n) = \sqrt{\frac{1}{n}\sum_{t=1}^{n} {e_t}^2} \tag{20.3}$$

Here and in the following we assume that the forecasting series is evaluated over a period $t = 1, \ldots, n$.

2. **Percentage error metrics** aim at scale-independence, such as the widely used mean absolute percentage error MAPE:

$$\mathrm{MAPE}(n) = \frac{1}{n}\sum_{t=1}^{n} \left| \frac{e_t}{X_t} \right| \tag{20.4}$$

MAPE has the disadvantage of being asymmetric, as for a given forecast value F_t and $|e_t|$, the penalty is heavier if $e_t < 0$. Therefore, a symmetric form of the MAPE is used sometimes, where the denominator is replaced by $\frac{(X_t+F_t)}{2}$, or alternative measures have been suggested (e.g., Kolassa and Schütz 2007).

3. **Relative error metrics** compare the error of the forecasting with the error of some benchmark method. Usually, the naïve forecast (i.e., X_{t-1} for F_t) is used as benchmark, where the forecast value for a one-step ahead forecast is simply the last observed value. One of the measures used in this context is the relative mean absolute error (RelMAE), defined as

$$\mathrm{RelMAE}(n) = \frac{1}{n}\sum_{t=1}^{n} \frac{|\, e_t \,|}{|\, X_t - X_{t-1} \,|} \tag{20.5}$$

Here we assume that X_{t-1} is also available. Similarly, we can define the relative RMSE, also known as Theil's U (De Gooijer and Hyndman 2006).

4. **Scale-free error metrics** have been introduced to counteract the problem that percentage error metrics and relative error metrics are not applicable if zeros occur in the denominators. The mean absolute scaled error MASE (Hyndman 2006) introduces a scaling by means of the MAE from the naïve forecast, where the last value is used as forecast:

$$\mathrm{MASE}(n) = \frac{1}{n}\sum_{t=1}^{n} \left(\frac{|\, e_t \,|}{\frac{1}{n-1}\sum_{i=2}^{n} |X_i - X_{i-1}|} \right) \tag{20.6}$$

All these measures come along with certain advantages and disadvantages. For example, percentage error metrics are often recommended for comparing forecast performance across different time series. Drawbacks are the inapplicability if a demand value is zero and the vagueness of percentage values regarding the interpretation of the economic impact. For sales forecasting, a MAPE of 1% may be economically significant or insignificant, depending on the sales volume. The next

sections will address the issue of the economic significance of errors on the basis of concrete examples.

2 Cost Error Metrics

2.1 Which Metric Is Best: A Toy Example

We first study a prototypical situation in demand forecasting by means of a slightly caricatured toy example. For this, we created a random sequence of $n = 100$ samples from a Gaussian distribution with mean $\mu = 10$ and standard deviation $\sigma = 1$ (arbitrary units). This sequence is interpreted as the sales baseline. We then added five random peaks with height of $\Delta h = 4\mu$, which represent the increased demand due to external factors. Real-world examples of such factors are sales promotions/special offers, holidays, special weather conditions etc. The generated sequence is shown in Fig. 20.1. Furthermore, the output of two different forecasting models is depicted. The first model is a perfect baseline model that, however, cannot anticipate the peaks. The second model is able to perfectly predict the peaks, but is always slightly overestimating the sales otherwise. We modeled this situation by a slight upshift of the original time series by 1 unit.

Imagine a planner that has to decide which model to choose for future predictions. She or he has to resolve the trade-off between hitting the peaks while being slightly

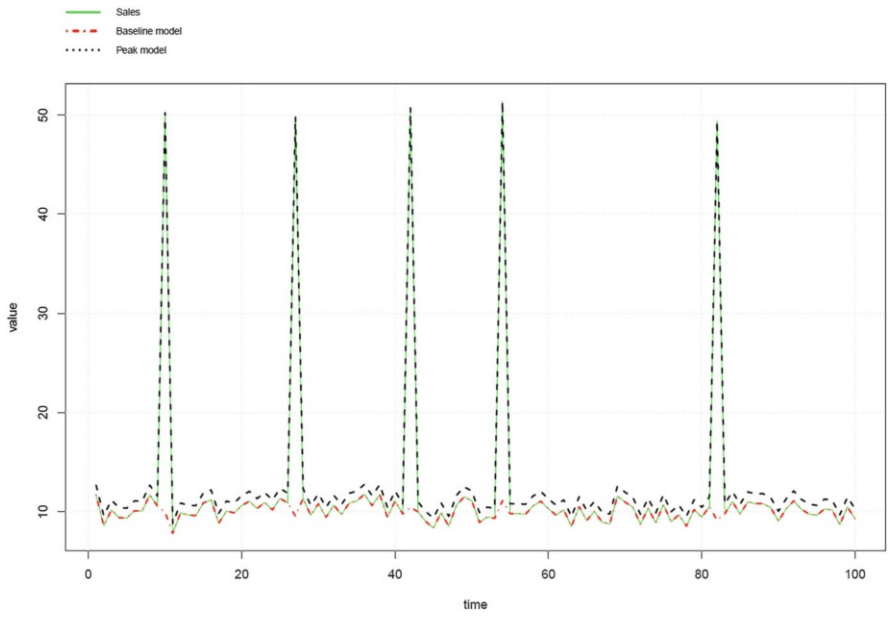

Fig. 20.1 Sales sequence (green) with five disruptive peaks, a perfect baseline model (red) that misses the peaks and a perfect peak model (black) which is slightly shifted in between peaks

Table 20.1 Results of four different error metrics

	MAE	RMSE	MAPE	relMAE
Baseline model	2.0	8.9	0.04	0.05
Peak model	0.95	0.97	0.10	2.74

The best results are highlighted

wrong in the meantime and being accurate most of the time but missing the peaks. For her or his decision, the planner evaluates the observed sequence using MAE, RMSE, MAPE, and relMAE. The results of the evaluation are shown in Table 20.1.

The result is ambiguous. MAE and RMSE speak in favor of the peak model, while MAPE and relMAE favor the baseline model. According to MAE, the time series model seems to be about "twice as bad" as the peak model; according to the RMSE, it seems to be even about "nine times as bad." Similar arguments can be produced for the comparison between MAPE and relMAE in favor of the baseline model.

The example illustrates the limitation of forecast error metrics for decision making. How can we resolve this issue? At the end of the day an economically relevant metric is defined by cost, that is, the financial consequences that come along with the prediction errors. Costs, however, can be highly product-specific and market-specific. Moreover, they depend on stock-keeping processes, an aspect we will discuss later.

2.2 Constructing Cost-Based Error Metrics

For now, let us assume that forecasting errors and costs are in direct relation. This is typically the case for fresh food products that cannot be stored and for which cost are directly related to sales. In consequence, a forecast that is too high results in costs for food waste and a forecast that is too low yields costs for stock-out. For goods that can be stored for an (un)limited time, there are storage costs instead of waste costs. In most practical cases, there will be a mixture of these types of costs. In any case, we assume that the forecast error e_t can be directly translated into costs $c(\cdot)$ and the costs do not depend on the history, that is,

$$c((X_t, F_t), (X_{t-1}, F_{t-1}), (X_{t-2}, F_{t-2}), \ldots) = c(e_t). \tag{20.7}$$

In the following, we will explain to what extend the metrics discussed above are able to reflect these costs and what kind of adaption would be needed to better account for real costs. We propose a generalized Mean Cost Error (MCE) of the following form:

$$\mathrm{MCE}(n) = s\left(\frac{1}{n}\sum_{t=1}^{n} c(e_t)\right), \tag{20.8}$$

where $c(\cdot)$ is a cost function and $s(\cdot)$ is a scaling function. Obviously, MCE defines a general form of a scale-dependent metric; MAE and RMSE can be considered special instances of MCE (see Fig. 20.2a, b).

If MAE and RMSE are interpreted in the framework of MCE, then it becomes apparent that these metrics impose some specific assumptions on the costs that may not be very natural in practice.

From the perspective of cost, a natural first approach is to neglect economies of scale and assume proportionality. Hence, excess stock cost or food waste cost ($e_t > 0$) increase proportional to the volume of the leftovers, that is, proportional to the forecasting error. For instance, costs may increase proportional to the manufacturing cost per unsold item or to the storage cost per unsold item (a: costs per item for $e_t > 0$). Similarly, stock-out cost increases proportional to the stock-out, for example, proportional to the unrealized profit or margin per item (b: costs per item for $e_t < 0$). Consequently, a first model is a piecewise linear cost model.

As a special class of MCE, we thus define the linear MCE (linMCE) as

$$\mathrm{linMCE} = \frac{1}{n}\sum_{t=1}^{n} c_{ab}(e_t) \quad \text{with} \quad c_{ab}(e_t) = \begin{cases} ae_t & \text{if } e_t \geq 0 \\ -be_t & \text{if } e_t < 0 \end{cases} \tag{20.9}$$

The measure is usually asymmetric as $a \neq b$ in general. In this setting, MAE is a special symmetric instance of linMCE (c.f. Fig. 20.2a, c).

Furthermore, we define a generalized class of scale-independent metrics that we call Mean Cost Percentage Error MCPE, as follows:

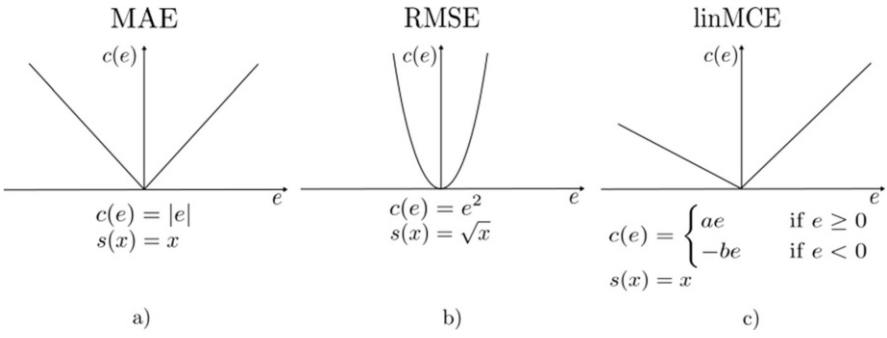

Fig. 20.2 Cost function of MAE, RMSE, and linMCE as special instances of MCE

$$\mathrm{MCPE}(n) = s\left(\frac{1}{n}\sum_{t=1}^{n}\frac{c(e_t)}{p(X_t)}\right), \tag{20.10}$$

where $p(\cdot)$ is given by some reference costs that are in connection with the real demand at time t. In the linear approach, we may for instance assume that $p(X_t)$ is proportional to the sales price of a product and to the number of items sold, that is, $p(X_t) = p \cdot X_t$. Hence, we define the linear MCPE as

$$\mathrm{linMCPE} = \frac{1}{n}\sum_{t=1}^{n}\frac{c_{ab}(e_t)}{p \cdot X_t} \quad \text{with} \quad c_{ab}(e_t) = \begin{cases} ae_t \text{ if } e_t \geq 0 \\ -be_t \text{ if } e_t < 0 \end{cases} \tag{20.11}$$

The measure can be interpreted as the mean of the costs due to the forecasting error in relation to the sales volume per forecasting period. MAPE is a special case of linMCPE with $a = b = p = 1$.

2.3 Sensitivity Analysis for linMCE

In order to calculate linMPE, we need to specify the parameters a (oversupply cost) and b (stock-out cost) for each product. Therefore, the costs per item for oversupply and for stock-out have to be made explicit. In practice, the parameters may be difficult to quantify exactly as, for instance, the oversupply cost can consist of a variable mixture of costs for food waste and storage. Thus, we may be interested in a more general comparison of forecasting methods or models with respect to the parameters a and b. For this sensitivity analysis we dissect the linMCE in an a-part and a b-part [i.e., using the Heaviside step function $h(\cdot)$]:

$$\mathrm{linMCE} = \frac{1}{n}\sum_{t=1}^{n} c_{ab}(e_t) = a \cdot \underbrace{\left(\frac{1}{n}\sum_{t=1}^{n} e_t h(e_t)\right)}_{\mathrm{linMC}\ \ E_a \geq 0} - b \cdot \underbrace{\left(\frac{1}{n}\sum_{t=1}^{n} e_t (1 - h(e_t))\right)}_{\mathrm{linMC}\ \ E_b \leq 0} \tag{20.12}$$

We can then study the relative performance of two forecasting models, $M1$ and $M2$, in dependence on the ratio x of a and b as follows:

$$f\left(x=\frac{a}{b}\right) = \frac{\text{linMCE}^{M1}}{\text{linMCE}^{M2}} = \frac{a\cdot\text{linMCE}_a^{M1} - b\cdot\text{linMCE}_b^{M1}}{a\cdot\text{linMCE}_a^{M2} - b\cdot\text{linMCE}_b^{M2}} = \frac{x\cdot\text{linMCE}_a^{M1} - \text{linMCE}_b^{M1}}{x\cdot\text{linMCE}_a^{M2} - \text{linMCE}_b^{M2}} \tag{20.13}$$

with the restriction that $x \geq 0$. Model 1 outperforms model 2 if $f(x) < 1$. Hence, as a critical condition for x we obtain

$$x_{\text{crit}} = \frac{\text{linMCE}_b^{M1} - \text{linMCE}_b^{M2}}{\text{linMCE}_a^{M1} - \text{linMCE}_a^{M2}} \tag{20.14}$$

In practice, one has to perform a case-by-case analysis to decide whether the critical point is in the relevant range $x \geq 0$ and to determine the values of $f(x)$. Hence, it is more convenient to plot this function, as we will discuss in the next section.

3 Evaluation

3.1 Calculating the Linear MPE: Toy Example Revisited

For our toy example we calculate the function $f(x)$ in a straightforward manner. The time series consists of $n = 100$ observations and 5 peaks with peak height $\Delta h = 4\mu = 40$. The peak model is shifted by $\Delta v = 1$ off the peaks. Hence, for the comparison of the baseline model ($M1$) and the peak model ($M2$), we get

$$f\left(x=\frac{a}{b}\right) = \frac{\text{linMCE}^{\text{baseline}}}{\text{linMCE}^{\text{peak}}} = \frac{2b}{0.95a} = \frac{2.11}{x} \tag{20.15}$$

which is derived from the following analysis of the linMCE (Table 20.2).

As a cross-check we see that the values for the MAE in Table 20.1 are retrieved for $a = b = 1$. The function $f(x)$ in Eq. (20.15) is continuously decreasing and the critical point is at $x = 2.11$. Therefore, the baseline model should be preferred if $a > 2.11 \cdot b$ ($x > 2.11$) and the peak model is the right choice if $a < 2.11 \cdot b$. That is, the peak model performs better if the oversupply (food waste/storage) costs per item are smaller than about two times the stock-out costs per item. This is due to the fact that a larger b in comparison to a, and hence a smaller x, puts a heavier penalty on stock-out situations that occur for the baseline model during peaks.

Table 20.2 Dissection of linMCE for the toy example (cf. Fig. 20.1)

	$a\cdot\text{linMCE}_a$	$-b\cdot\text{linMCE}_b$	linMCE
Baseline model ($M1$)	0	$5\cdot\Delta h\cdot\frac{b}{n} = 2b$	$2b$
Peak model ($M2$)	$(n-5)\cdot\Delta v\cdot\frac{a}{n} = 0.95a$	0	$0.95a$

3.2 Real World Example

In this section, we turn the focus on a real world example. Figure 20.3 depicts the demand data for a food product (weekly sales of a fruit juice) from a retail supplier. The sales sequence (blue curve) comprises of $n = 146$ values with mean $\underline{x} = 18{,}266$ and standard deviation $\sigma = 3783$ (units). The time series shows some characteristics that are typical for many food products, that is, there are characteristic peaks and dents due to promotions and the series exhibits a falling trend and, hence, is not stationary.

Following the toy example introduced above, we chose and fitted two models for one-step-ahead predictions. Both models are based on regression trees. However, they show a rather complementary behavior comparable to the models in the toy example before (cf. Fig. 20.3). One model can be considered as baseline model (red curve, model 1). It is able to predict the general development of the time series, but misses the peaks. In contrast, the second model (peak model; black dotted curve, model 2) takes into account additional external information and hence, is able to predict most peak demands. The price to pay is a reduced reliability between peaks. The model even predicts peaks that do not occur at all in the actual sales sequence.

With regard to the traditional error measures we observe the same picture as for the toy example (cf. Table 20.1). MAE and RMSE favor the peak model, whereas

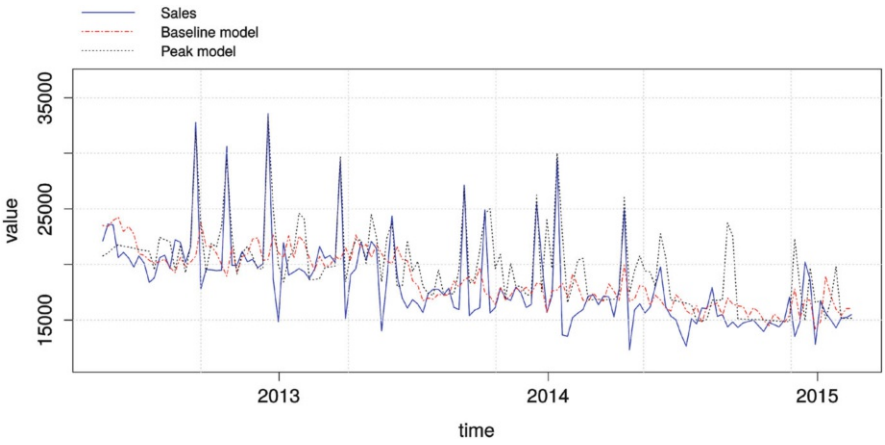

Fig. 20.3 Typical sales sequence and two different forecasts for a food product

Table 20.3 Results of four different error metrics for the real world example

	MAE	RMSE	MAPE	relMAE
Baseline model	2134.85	3234.45	0.1125	3.7248
Peak model	2080.34	2951.00	0.1253	6.2250

The best results are highlighted

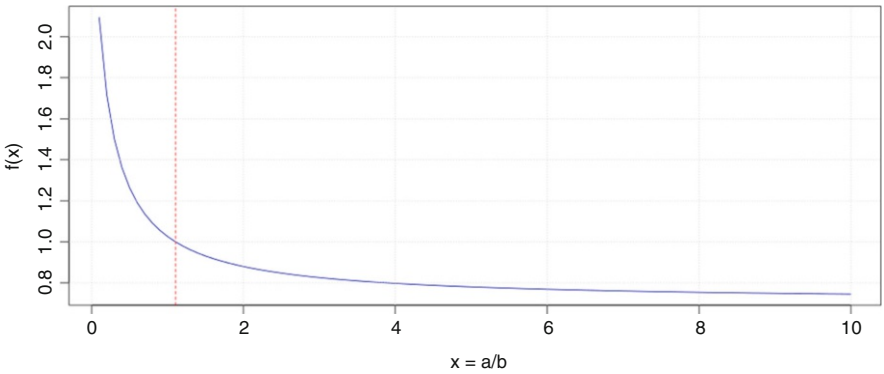

Fig. 20.4 Comparison of baseline model versus peak model as a function of the ratio a/b. On the right side of the critical point (red line), the baseline model should be preferred

MAPE and relMAE suggest using the baseline model (Table 20.3). In fact, relMAE even indicates that the naïve model should be used. The error measures were computed over the whole sequence.

The sensitivity analysis based on linMCE again allows for a clearer picture. In Fig. 20.4, the function f according to Eq. (20.13) is depicted. The critical point, highlighted by a red vertical line, is at $x_{\text{crit}} = 1.105$. We can conclude that the baseline model should be used in case $a/b > 1.105$, that is, if the stock-out cost per item is clearly smaller than the oversupply cost per item. In case $a/b < 1.105$, the peak model performs better since the stock-out costs per item are almost equal or larger than the oversupply costs per item. Again, for ratios $a/b < 1.105$, stock-out situations that occur for the baseline model during a peak are penalized more heavily and the costs for the baseline model are increased accordingly.

For the comparison of more than two models we suggest pairwise comparisons of each model with a benchmark, for example, the naïve prediction (last value is used as predicted value), which allows for a ranking of the models for each value of x by comparing the functions:

$$b_{\text{model}}\left(x = \frac{a}{b}\right) = \frac{\text{linMCE}^{\text{model}}}{\text{linMCE}^{\text{benchmark}}} \tag{20.16}$$

The result of this comparison for the baseline model and the peak model is shown in Fig. 20.5. We can identify three different regimes:

1. $0 < x < 1.105$, that is, if the oversupply costs per item are less than 1.105 times the stock-out costs, the peak model outperforms the baseline model and the benchmark model, the benchmark model is the worst choice.
2. $1.105 < x < 2.050$, the baseline model outperforms the peak model and the benchmark model, and the benchmark model is the worst choice.

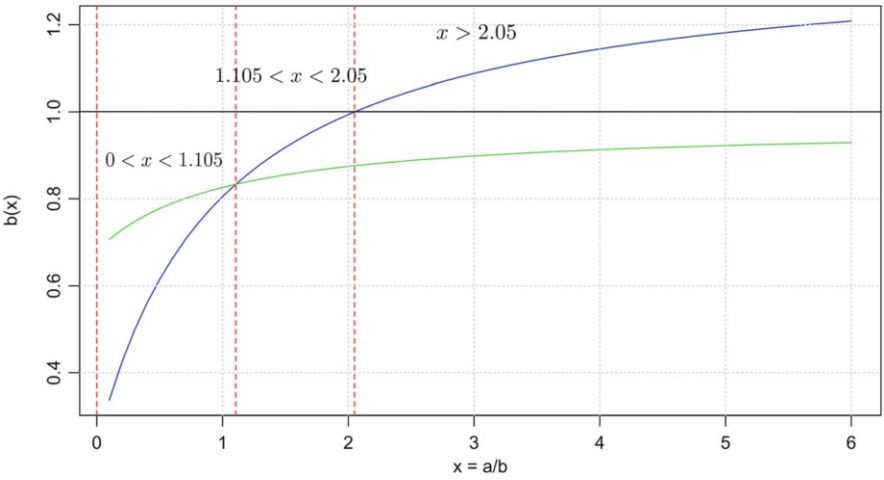

Fig. 20.5 Comparison of models versus naive model: green line: b_{baseline}, blue line: b_{peak}, black line: benchmark. The red lines indicate preference regimes of different models

3. $x > 2.050$, that is, if the oversupply costs are more than 2.050 times larger than the stock-out costs, the baseline model is best, the peak model is worst.

If we finally want to decide which model to use for our product, we need to make assumptions about the parameters a and b. For our example product (fruit juice), a per unit price of 1 CHF has to be paid to the producer and the product is sold for 1.2 CHF. Hence, the margin per unit is 0.2 CHF and this value is used to estimate the stock-out cost ($b \approx 0.2$). The oversupply parameter a is a bit harder to estimate in this case. As the time of permanency of this product is relatively large in comparison to the supply circle, an oversupply leads to an increase in stock rather than to food waste. The stock-keeping costs are estimated to be 10% of the average stock value, that is, 0.1 CHF per unsold unit per cycle. Hence, according to a first approximation, we choose $a \approx 0.1$.

In conclusion, we have $x = a/b \approx 0.5$ and hence the sensitivity analysis suggests to use the peak model.

3.3 Stock-Keeping Models: Beyond Simple Cost Measures

The measures for forecast costs presented so far were functions of demands/sales X_t and forecasts F_t. These measures are applicable in straightforward manner for goods with short expiration times as in this case, the parameters a and b are relatively easy to estimate (a corresponds to the production/base prize and b corresponds to the margin per unit). The estimations become more complicated for products with a longer time of permanency. In this case, as we have seen in the example above, we

have to make further assumptions about the stock-keeping process as the proposed measures do not take into account storage capacities. In general, for goods that can be stored, storage capacities, storage cost, and logistics strategies should be taken into consideration for a more reliable evaluation of the economic impact of forecasting algorithms. In the following, we present a simple stock-keeping model including stock capacities, service level, storage cost, ordering times, and product margins.

An important key figure in logistics is the *beta* service level, defined as the probability that an arbitrary demand unit is delivered without delay. Typical target values for the *beta* service level are at 0.99 or even higher. From the service level, a safety stock level can be derived, depending on assumptions about the demand distribution and the forecast accuracy. Intuitively, the more reliable the forecasts are, the lower the safety stock level can be, given a certain *beta* service level. Typically, the demand distribution is not known, but has to be estimated indirectly through the sales distribution (not yielding information about, e.g., potential stock-outs), as we pointed out earlier.

To compute the safety stock level in practice, normally distributed forecast errors $e_t = F_t - X_t$ are usually assumed (Brown 1967). From these errors, the root mean square error RMSE(e_t) can be computed. Defining

$$t_{\text{beta}} = \frac{((1 - \text{beta})\ D)}{\left(\text{beta}\ \sqrt{\text{Lt}}\ \text{RMSE}(e_t)\right)}, \tag{20.17}$$

where D is the average demand and Lt the average lead time, we compute $w = \sqrt{\log\left(\frac{25}{t_{\text{beta}}{}^2}\right)}$ and approximate the safety stock factor k_{beta} as (according to Brown 1967)

$$k_{\text{beta}} = \frac{-5.3926 + 5.6611 \times w - 3.8837 \times w^2 + 1.0897 \times w^3}{1 - 0.725 \times w + 0.5073 \times w^2 + 0.06914 \times w^3 - 0.0032 \times w^4}. \tag{20.18}$$

From k_{beta}, the safety stock level is computed as safety stock level $= k_{\text{beta}}$ sigma $\sqrt{\text{Lt}}$, with sigma denoting the standard deviation of the forecast errors e_t. Details on the derivation of the safety stock level can be found in Brown (1967).

With these foundations (simplifying Lt = 1), a stock-keeping model can be defined through

$$\text{Orders for time } t+1 = F_{t+1} + \text{safety stock level} - \text{stock at time } t.$$

To evaluate different forecasting strategies, X_{t+1} can be used as simulated demand and costs for stock-keeping and lost sales can be simulated for each forecasting model.

Table 20.4 Stock-keeping model results for the example presented in Fig. 20.3

Quantity	Baseline model	Peak model
Average stock level (units)	22,544	23,164
Safety stock level (units)	3886	3415
Effective *beta* service level (%)	98.08	99.92
Stock-keeping costs (CHF)	6329	6504
Opportunity costs (CHF)	10,266	385
Stock-keeping + opportunity costs (CHF)	16,595	6889

At a margin of 20% (0.2 CHF), stock-keeping cost differences are by magnitudes lower than the differences in opportunity costs for the two models

Applied to the example presented in Fig. 20.3, assuming again a per unit price of 1 CHF paid to the producer, a per unit price of 1.2 CHF paid by the customer, an average lead time Lt = 1 time period, a safety stock factor $k_{\text{beta}} = 0.99$, and annual stock-keeping costs of 10% of the average stock value, Table 20.4 shows a comparison between the baseline model and the peak model (146 periods). Note that more complex inventory models would allow for further parametrizations of expiration times for a product and correspondingly for estimations of waste cost.

As expected from Fig. 20.3, the peak model is more valuable in terms of opportunity costs than the baseline model. For stock-keeping cost, the baseline model is slightly more profitable. The effective beta service level "effective *beta* service level" is close to 99% for both models, indicating that forecast error distributions are in accordance with the assumptions stated above. The decision upon which model should be used can now be based on total costs. In this example, the peak model is to be preferred. This finding is in line with our result based on the linMCE analysis, where we found $\frac{a}{b} \tilde{} 0.5$. The stock-keeping model, however, allows for a more robust estimate of the economic significance of the two forecasting models. From Table 20.4 we see that choosing the right model helps reduce the costs by almost 60%, when changing from the baseline model to the peak model. Or in other words, if the decision would have been based on either MAPE or relMAE, the cost due to forecasting errors of the chosen model would have been at least 2.4 times as high as necessary in the case of the optimal decision.

4 Conclusions and Lessons Learned

Error metrics are used to evaluate and compare the performance of different forecasting models. The traditional, most widely used metrics such as MAE, RMSE, MAPE, and relMAE come along with certain disadvantages. As our examples from food demand forecasting illustrated, their values are often difficult to interpret regarding the economic significance and they may yield incoherent accuracy rankings. In practice, economically relevant metrics are linked to the costs that are caused by the prediction errors. We introduced a class of such measures that allow for

considering different weights for oversupply/excess stock costs and stock-out costs. It turns out that traditional measures can be interpreted as special cases in this class with specific cost configurations. In practice, costs for oversupply or stock-out might be difficult to determine. In order to cope with this issue, we introduced a method that enables a straightforward sensitivity analysis. It allows for choosing the optimal forecasting method on the basis of a relatively rough estimate of cost ratios.

The proposed cost error metrics, however, have no memory. That is, they assume that there is no stock available at the beginning of each step and the demand is equal to the supply of goods. This assumption is reasonable for the approximate evaluation of a forecasting method. However, real costs may not always directly reflect this performance, for example, for stocked goods a too low demand forecast does not necessarily lead to stock-out cost. It might even help reducing stocks and hence a bad forecast can have a positive effect on the costs. In order to better approximate real costs, simplified stock-keeping models can be used.

We illustrated the discussed aspects by means of a toy and a real world example. From these case studies we learned the following:

- The choice of the best forecasting model depends on the ratio of oversupply costs and stock-out costs.
- In particular, a baseline model should be preferred over a peak model if the oversupply costs are much higher than the stock-out costs and vice versa.
- Common error metrics do not account for this observation and can lead to bad model decisions.
- A bad model decision can easily result in an increase of the cost or the nonrealized earning potential by a factor of 2.4 for a single product.

An important aspect regarding the choice of optimal models that has not been discussed is the length of the evaluation time window. On the one hand, if the evaluation window is too short, random deviations without any significance can be predominant. On the other hand, if this window is too long, the good performance of a model in the distant past might masque structural disruptions that can cause a poor performance in the near future. For the model selection process, we thus generally suggest introducing an additional optimization loop that regularly adjusts the optimal length of the evaluation window. There is clearly not a unique optimally performing forecasting algorithm for everything. Similarly, to assess the qualities and economic values of forecasts, there is not a unique best forecast error measure. Different aspects, mainly involving costs of stock-keeping, stock-out, and waste, but also supply chain and marketing strategies (customer satisfaction, ecologic reputation, transport optimization, etc.) should be considered when evaluating forecasting procedures. The strategies presented here may provide a contribution to the goal of creating more economic value from demand forecasting.

References

Bajari, P., Nekipelov, D., Ryan, S., & Yang, M. (2014). Machine learning methods for demand estimation. *American Economic Review, Papers and Proceedings, 105*(5), 481–485.

Brown, R. (1967). *Decision rules for inventory management*. New York: Reinhart and Winston.

Da Veiga, C. P., Da Veiga, C. R. P., Catapan, A., Tortato, U., & Da Silva, W. V. (2014). Demand forecasting in food retail: A comparison between the Holt-Winters and ARIMA models. *WSEAS Transactions on Business and Economics, 11*, 608–614.

De Gooijer, J. G., & Hyndman, R. J. (2006). 25 years of time series forecasting. *International Journal of Forecasting, 22*, 443–473.

Doganis, P., Alexandridis, A., Patrinos, P., & Sarimveis, H. (2006). 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.

Hyndman, R. J. (2006). Another look at forecast-accuracy metrics for intermittent demand. *Foresight, 4*, 43–46.

Kolassa, S., & Schütz, W. (2007). Advantages of the MAD/Mean ratio over the MAPE. *Foresight: The International Journal of Applied Forecasting, 6*(6), 40–43.

Taylor, J. W. (2007). Forecasting daily supermarket sales using exponentially weighted quantile regression. *European Journal of Operational Research, 178*(1), 154–167.

Vasumathi, B., & Shanmuga Ranjani, S. P. (2013). Forecasting in SAP-SCM (Supply Chain Management). *International Journal of Computer Science and Mobile Computing, 2*(7), 114–119.

Žliobaitė, I., Bakker, J., & Pechenizkiy, M. (2012). Beating the baseline prediction in food sales: How intelligent an intelligent predictor is? *Expert Systems with Applications, 39*(1), 806–815.