



## Combination of Multi Level Forecasts

SILVIA RIEDEL

*Lufthansa Systems Berlin GmbH, Fritschstrasse 27-28, 10585, Berlin, Germany*

BOGDAN GABRYS AND SILVIA RIEDEL

*Computational Intelligence Research Group, School of Design, Engineering and Computing,  
Bournemouth University, Poole House, Talbot Campus, Poole, BH12 5BB, UK*

*Received: 1 May 2006; Revised: 3 October 2006; Accepted: 2 April 2007*

**Abstract.** This paper provides a discussion of the effects of different multi-level learning approaches on the resulting out of sample forecast errors in the case of difficult real-world forecasting problems with large noise terms in the training data, frequently occurring structural breaks and quickly changing environments. In order to benefit from the advantages of learning on different aggregation levels and to reduce the risks of high noise terms on low level predictions and overgeneralization on higher levels, various approaches of using information at different levels are analysed in relation to their effects on the bias, variance and Bayes error components proposed by James and Hastie. We provide an extension of this decomposition for the multi-level case. An extensive analysis is also carried out answering the question of why the combination of predictions using information learned at different levels constitutes a significantly better approach in comparison to using only the predictions generated at one of the levels or other multi-level approaches. Additionally we argue why multi-level combinations should be used in addition to thick modelling and the use of different function spaces. Significant forecast improvements have been obtained when using the proposed multi-level combination approaches.

**Keywords:** multi level forecasting, forecast combination, bias variance Bayes error decomposition, revenue management

### 1. Introduction

In this paper we discuss issues related to difficult real-world forecasting problems that are characterised by large noise terms in the training data, frequently occurring structural breaks, and quickly changing environments. We address applications that require multiple predictions rather than a single prediction. Each prediction represents the situation in a concrete subspace of the given target space. We illustrate our argumentation at a simple example of seasonal demand predictions for airlines. These have to be generated for different origin destination itinerary pairs (ODI) as well as different fareclasses ( $F$ ) and

different point of sales (POS). This level of forecasting, which we also call the fine/low level, is very detailed (the seasonal behaviour for a given ODI F POS combination) and therefore characterised by small numbers and very noisy data. Therefore analysts also need aggregates of the generated low level forecasts for decision making. Modern Graphical User Interfaces support this need. They offer the functionality of a data and forecast fusion to different higher levels, which represent in our example for instance the ODI level or even higher levels like country or market pairs as shown in Fig. 1.

Large noise components often lead to the decision to learn structural information or causal effects based

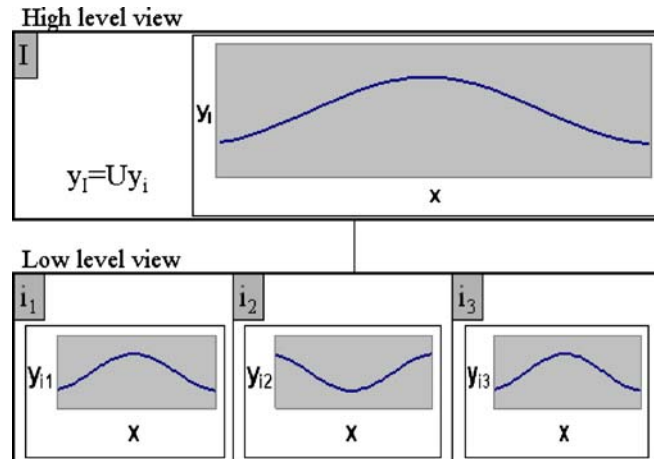


Figure 1. View from the low and the high level.

on aggregates of the input data, in other words, to carry out an input data fusion with the objective of noise reduction. There is no obvious answer to the question about the adequate level for learning. Learning at different levels is related to different types of risk. If the level is chosen too fine, relevant structural information can often not be detected properly. If on the other hand the chosen level is too general, important characteristics related to special parts in the target space may be ignored. For our example this means that if we learn seasonal factors for instance at the ODI level, we do not take into account seasonal effects in special fareclasses or point of sales properly. An overview of literature related to learning at different levels and effects of forecast aggregation is provided in [1].

In practice, the problem to find the ideal level of learning is often resolved with trial and error approaches. The choice is made only based on low level forecast errors. But if analysts make relevant decisions on the basis of a fusion of low level forecasts to a higher level, the need of high quality forecasts at higher levels should also be taken into account for the choice of the level of learning structural information.

In this paper we analyse effects of learning at two levels on the resulting forecast errors measured at these two levels. Choices that are purely made based on forecast errors measured at the low level can be unfavourable with regard to the quality of the aggregated forecasts. We base our argumentations on the error bias, variance and Bayes decomposition

proposed by James and Hastie [2]. We provide this error decomposition for the multi-level case. This enables us: (a) to analyse effects of aggregation of forecasts generated with learning at the low level to the error components at the high level, and (b) to analyse effects of using forecasts generated with learning at the high level to the error components at the low level.

As we will see that learning at both levels works well only in some cases, we also discuss the option of using forecast combination in order to make an automated choice or even to profit from knowledge at both levels. Forecast combination approaches are today a scientifically acknowledged procedures (for overviews see [3, 4, or 5]). The positive effects of forecast combination in many applications [6] have been explained in relation to different aspects and different decompositions of forecast errors and their correlation. We provide the analysis of the error components of combined multi-level forecasts at the low as well as at the high level. The analysis is based on the simplified version of the well known *optimal model* [7], the *optimal model with assumption of independence* [8], which takes into account the problem of high estimation errors of the inverse covariance matrix [9] and is purely based on past error variances. We also discuss different cases of data constellation and relation between the levels in order to show that forecasts combination works very well in cases where it represents an automatic choice of the appropriate level as well as in cases where knowledge of both levels is relevant for learning.

The paper is organised as follows:

In Section 1 we first introduce the used notation of time series data and its fusion to higher levels, the notation of forecasts with learning carried out at different levels as well as its fusion to higher levels, and the notation of forecast errors measured at different levels. We also introduce the bias–variance–Bayes error decomposition and discuss the behaviour of the different error components in case of forecast aggregation. Finally we describe an artificial example that we will use in later sections in order to illustrate the behaviour in different situations.

Then we study the effects of learning at different levels in Section 2. We first analyse learning at the low level and provide the bias–variance–Bayes decomposition of the errors if these forecasts are aggregated to the higher level. Then we discuss the case of transferring information learned at the high level to the low level and study the effects on the low level error components. We compare the two approaches and will see that both work well in some cases and have problems in others. Additionally both approaches suffer from a loss of information related to the level which has not been taken into account. This provides a motivation for the discussion of an alternative option of how to incorporate information measured at both levels, the approach of forecast combination. We finally provide a representation for the decomposed error measured at the low as well as aggregated to the high level for this approach.

Section 3 contains a detailed discussion about what happens to the error components in different concrete situations illustrated at our artificial example. We will see that the approach of forecast combination allows not only to make an intelligent automatic choice of the superior level, if it exists, but also to generate predictions that are more stable in terms of the quality of aggregates to higher levels and in case of changing environments.

In Section 4 we provide a summary of the findings and mention significant forecast improvements for a real-world application.

## 2. Problem Description and Notation

### 2.1. Notation of Multi-Level Time Series

We discuss causal models representing relationships between time series  $x_t \in \mathcal{R}^n$  and  $y_t \in \mathcal{R}$ , with  $t$  representing a time index. We further assume that  $x_t$

can be measured, that we have random noise in  $y_t$  and that an “ideal” functional relationship  $f$  exists in order to approximate  $y_t$  based on  $x_t$ . We can represent the functional relationship between input vector  $x_t \in \mathcal{R}^n$  and  $y_t \in \mathcal{R}$  by the function  $f$  and a random noise term  $\epsilon$ :

$$y_t = f(x_t) + \epsilon_{yt}, \quad (1)$$

with  $f$  the “true model” and  $\epsilon$  Gaussian with  $\epsilon_y \sim N(0, \delta_{\epsilon_y}^2)$  an independent residual component. The vector  $x$  may also contain past values or predictions of  $y$  as described in the model in [5]. In order to increase readability we will remove the parameter  $t$  in all following equations.

Let’s now assume that we do not have to predict a single time series but a whole set representing different subspaces of a target space. We will use the index  $i$  in order to indicate any given subspace (the fine/low level) for which we have to generate predictions:

$$y_i = f_i(x) + \epsilon_{yi}. \quad (2)$$

Let further index  $I$  indicate values or measurements concerning a high level view.

### 2.2. The Relation Between $y_i$ and $y_I$

It is assumed that a linear unification operator  $\cup$  over the subspaces  $i$  is defined in order to represent the aggregation from the low-level subspaces  $i$  to the higher level  $I$ . The fusion operator<sup>1</sup> carries out a weighted sum

$$z_I = \cup z_i = \frac{\sum_i \lambda_i z_i}{\sum_i \lambda_i} \quad (3)$$

over any data  $z_i$  measured at the different low levels (which could e.g. be  $f_i(x)$  or  $y_i$ ). The parameters  $\lambda_i \in \mathcal{R}$  are indicators for the relevance or size of subspace  $i$  as part of  $I$ .

Let’s assume we have given impact parameters  $\lambda_i$ . Then we get a high level representation of  $y$  following Eq. (3) with  $y_I = \cup y_i$  (high level targets are aggregates of the low level targets). As the noise component at the low level is white noise, this component is also aggregated to noise at the high level as  $\epsilon_{yI} = \cup \epsilon_{yi}$  which leads to a predictable

relationship  $f_I(x) = \bigcup f_i(x)$  fulfilling the high level relation similar to Eq. (1)

$$y_I = f_I(x) + \epsilon_{yI} = \bigcup y_i = \bigcup f_i(x) + \bigcup \epsilon_{yi}. \quad (4)$$

Let's now analyse the differences between  $f_I(x)$  and  $f_i(x)$  as these are very relevant if high level information is to be used for low level forecasting. We can expect that big differences between  $f_I$  and  $f_i$  would lead to big errors at the low level if we replace estimates of  $f_i$  by estimates of  $f_I$ . We define  $\epsilon_{fi}$  as

$$\epsilon_{fi} = f_I(x) - f_i(x). \quad (5)$$

Combining Eqs. (3), (4) and (5) it follows from

$$\begin{aligned} y_I &= \bigcup f_i(x) + \bigcup \epsilon_{yi} = \bigcup (f_i(x) - \epsilon_{fi}) + \bigcup \epsilon_{yi} \\ &= f_I(x) - \bigcup \epsilon_{fi} + \epsilon_{yI} \end{aligned} \quad (6)$$

that  $\epsilon_{fi}$  has the nice characteristics of reducing to 0 if aggregated at the high level:

$$\bigcup \epsilon_{fi} = 0. \quad (7)$$

### 2.3. Predicting $y_i$

A predefined class of functions  $h : \mathcal{R}^n \times \Phi \rightarrow \mathcal{R}$  is used in order to approximate the relationship between  $x$  and  $y$ . We first define the function space comparable to the definition given in [10]:

#### Definition (Function Space)

Let  $x_t \in \mathcal{R}^n$  be a time series and  $h : \mathcal{R}^n \times \Phi \rightarrow \mathcal{R}$  be a function with input  $x_t$  and let it depend on the parameters  $\phi \in \Phi \subset \mathcal{R}^m$ , then the function space of  $h$  is the linear space  $\mathcal{H}$  consisting of all possible functions  $h(\cdot; \phi)$  obtained by varying  $\phi$  in the domain  $\Phi$ .

We further assume that a best estimation of parameters  $\phi_i$  exists in order to approximate  $f_i$  by  $h(\cdot; \phi_i)$

$$f_i(x) \approx h(x, \phi_i) \quad (8)$$

and that the underlying distance norm is linear in a manner that for any two functions  $f_1(x) : \mathcal{R}^n \rightarrow \mathcal{R}$

and  $f_2(x) : \mathcal{R}^n \rightarrow \mathcal{R}$  with  $h(\cdot; \phi_1)$  representing the best approximation for  $f_1(x)$  and  $h(\cdot; \phi_2)$  the best approximation for  $f_2(x)$ , the best approximation for  $\lambda_1 f_1(x) + \lambda_2 f_2(x)$  is given by  $\lambda_1 h(x, \phi_1) + \lambda_2 h(x, \phi_2)$  for any  $\lambda_1, \lambda_2 \in \mathcal{R}$ .

### 2.4. The Bias–Variance–Bayes Error Decomposition

Let  $\hat{e}_i$  represent the error which will be generated in predicting  $y_i$  based on  $h(x, \hat{\phi})$  (out of sample predictions):

$$\begin{aligned} e_i &= y_i - \hat{y}_i = y_i - h(x, \hat{\phi}) \\ &= f_i(x) - h(x, \hat{\phi}_i) + \epsilon_{yi} \end{aligned} \quad (9)$$

Let's assume that we have found an estimator  $h(x, \hat{\phi}_i)$  which generates predictions without a systematic error so that  $(e_i)$  can be represented as Gaussian with  $e_i \sim N(0, \delta_{ei}^2)$ .

Then the total error variance term  $\delta_{ei}^2$  can be decomposed into different components. While different error decompositions can be found in [11] and [10], we will refer here to the decomposition of James and Hastie [2]:

$$\delta_{ei}^2 = \delta_{ehi}^2 + \delta_{e\phi i}^2 + \delta_{eyi}^2. \quad (10)$$

The first error component  $\epsilon_{hi}$  with variance  $\delta_{ehi}^2$  is called the bias. This error component is based on the fact that the class of functions  $h(x; \cdot)$  may not include the function  $f_i(x)$ . As we have assumed that an ideal parameter set  $\phi$  exists in order to estimate  $f_i(x)$  based on  $h(x; \phi)$ , the bias term of the error is defined by  $\epsilon_{hi} = f_i(x) - h(x, \phi)$ . The second term  $\epsilon_{\phi i}$  of the error with variance  $\delta_{e\phi i}^2$  is the error variance component. This term is based on the fact that the parameters  $\phi$  can not be estimated perfectly because of noise in the training data, limited number of training samples etc. The variance term of the error is defined by  $\epsilon_{\phi i} = h(x, \phi) - h(x, \hat{\phi})$ . The third term  $\epsilon_{yi}$  represents the irreducible Bayes error component in  $y_i$  which can be reduced only if more information becomes available in  $x$ .

While the third part of the error can not be reduced without including additional information (as it represents a random deviation which is not covered

by  $f$ ) the bias and variance term can be substantially influenced by the complexity of the function  $h(x; \cdot)$ . So for instance, in case of artificial neural networks (ANNs) used as our function  $h(x; \cdot)$  it depends on the choice of the architecture of an ANN or the algorithm of how to determine the parameter vector  $\phi$  based on the training data. If the function space of  $h(x; \cdot)$  is very complex, we can assume that it is able to cover  $f(x)$  very well so that we have a small bias term. But it is also difficult to estimate a complex parameter set, we have a high risk of overfitting and a large variance term. If on the other hand we use a simple class of functions  $h(x; \cdot)$  with a low dimensionality of the parameter vector  $\phi$ , we will be able to estimate the parameters well based on the training data and so have a low variance term, but we will have difficulty to cover the complexity of  $f(x)$  by  $h(x; \cdot)$  so that we have an increased bias term. For additional references and a detailed discussion of these topics see [11, 10] or [2]. The problem to find a good trade-off between error bias and variance is called the bias–variance dilemma. Different alternatives have been proposed in order to determine a good trade-off between bias and variance while learning the parameters in  $h(\cdot; \phi)$  or choosing function classes  $h(\cdot; \cdot)$  with an appropriate quality.

Let's also assume that we have a training set  $(x, y)_T$  of historical data given and that we use it in order to estimate the parameter vector  $\phi_i$  by  $\hat{\phi}_i$  so that  $h(x, \hat{\phi}_i)$  represents our best estimation for the relationship between  $x$  and  $y_i$ . If we have to produce predictions at a very fine level we risk extremely large residual terms  $\epsilon_{yi}$ . Let's therefore assume that we have only a small set of highly disturbed training data available. The reason for this small training set  $(x, y)_T$  can for instance be based on the fact that we have a quickly changing environment where only few historical data is representative for the current situation. If the function  $f(x)$  is complex it is possible that even parameters related to functions  $h(x; \cdot)$  that are clearly less complex compared to  $f(x)$  can not be estimated properly. If we want to approximate  $f(x)$  based on such a training set we have only two choices. The first option is to estimate  $f(x)$  based on structurally very poor functions  $h(x; \cdot)$  and take a large bias term into account. The second option is to choose a little bit more complex function  $h(x; \cdot)$  and risk a bias term which is a little bit lower together with a variance term which can get so big that the resulting function is completely unstable.

## 2.5. Properties of the Error Components in Relation to Forecast Aggregation

Of course the main objective is to achieve good predictions at the level of forecasting, i.e. the low level, which means a minimisation of  $\delta_{ei}^2$ . However, as in a lot of applications the generated forecasts are (also) used on an aggregated level, it is also worth to analyse the error  $\delta_{eU}^2$ . If we can find a good trade-off between the errors at different subspaces which generate more stable predictions meaning lower errors at the high level, this is certainly advantageous.

Different components of the error are related to different stability if they are aggregated. The stability depends on the correlation of an error component between different subspaces. If an error component is positively correlated between subspaces, we have to expect an error accumulation effect. If on the other hand we have no or even a negative error correlation, these errors will compensate each other well.

The error variance component is a critical component for aggregation. The values  $y_i$  are often very noisy and the noise is often highly correlated between the different subspaces. Similar deviations in the target values of the training set contain the risk of generating highly correlated residuals  $\epsilon_{yi}$ . It is therefore possible that the correlated residuals in the training set lead also to unstable (large) and highly positively correlated terms  $\delta_{e\phi_i}^2$  and therefore to very big terms  $\delta_{e\phi_U}^2$ .

The situation is different for the bias term. Because of the linearity that we have assumed for the distance norm we also know that

$$\bigcup h(x, \phi_i) = h(x, \phi_I) \quad (11)$$

is true. It follows that

$$\bigcup \epsilon_{hi} = \epsilon_{hI} \quad (12)$$

because of  $f_I(x) = \bigcup f_i(x)$ ,  $\bigcup h(x, \phi_i) = h(x, \phi_I)$  and the definition of the bias term at both levels:  $f_i(x) = h(x, \phi_i) + \epsilon_{hi}$  and  $f_I(x) = h(x, \phi_I) + \epsilon_{hI}$ . This means that all kinds of low level problems in case of more complex functions  $f_i(x)$  at the low level compared to  $f_I(x)$  compensate each other during the aggregation. If on the other hand  $f_i(x)$  is more complex in comparison to the different subspaces  $f_i(x)$ , this means

Table 1. Characteristics of example data.

Level	$\lambda_i$	$f_i(x)$	$\delta_{\text{eyi}}^2$
$i_1$	0.6	$\sin((x - 12)/(9))$	0.8
$i_2$	0.2	$-\sin((x - 12)/(9))$	2
$i_3$	0.2	$\sin((x - 12)/(9))$	2
$I$	–	$0.6\sin((x - 12)/(9))$	0.64

that we have correlations between the subspaces  $f_i(x)$  which are not extremely big. In this case we have only few compensation effects of the error bias component during the aggregation, but probably lower bias values  $\delta_{hi}^2$  because of the lower complexity of  $f_i(x)$ .

### 2.6. An Artificial Example

In O&D Revenue Management Systems [12, 13] seasonal predictions have to be carried out at a very fine level where the behaviour changes very quickly so that it is not possible to take a large number of historical data into account. Predictions have to be generated not only for different flights or origin-destination-itinerary pairs (the so called ODIs), but

also separately for different fareclasses (F) representing different prices and booking restrictions as well as different point of sales (POS). Let’s assume we have to model a seasonal dependency of the booking behaviour on the calendar week in terms of seasonal factors.<sup>2</sup> This means that in our example  $t$  represents a weekly time index, the variable  $x_t$  represents the corresponding calendar week  $x_t \in [1, 53]$   $b_t$  the number of bookings achieved in week  $t$ ,  $f(x)$  a seasonal factor representing

$$f(x) = \frac{E(b|x)}{E(b)} - 1 \tag{13}$$

as the deviation between bookings expected in the given calendar week and the total booking expectation and  $y_t$  the achieved deviation

$$y_t = \frac{b_t}{E(b)} - 1. \tag{14}$$

As the “true relationship”  $f(x)$  is not known we introduce artificial ones in order to be able to illustrate certain behaviour of different error components. We

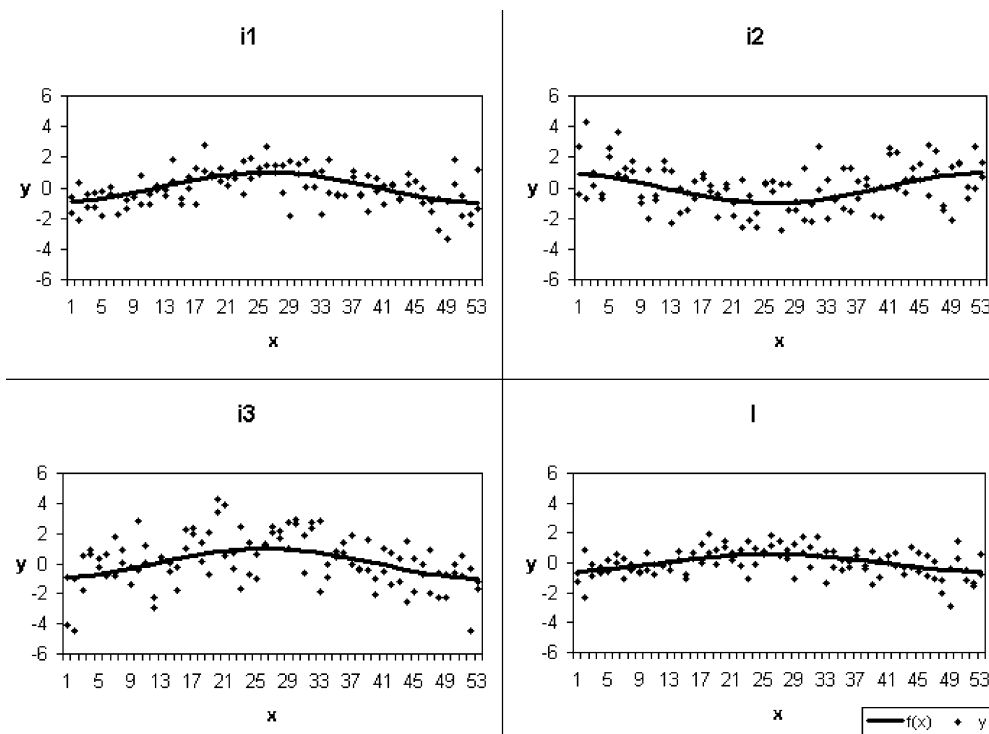


Figure 2. Artificial data generated for subspaces  $i_1$  to  $i_3$  and aggregated to the high level  $I$ .

Table 2. Error components of the forecast results.

level	learning	$\delta_{\epsilon hi}^2$	$\delta_{\epsilon \phi i}^2$	$\phi^I \delta_{\epsilon \phi i}^2$	$^{comb} \delta_{\epsilon \phi i}^2$	$\delta_{\epsilon ei}^2$	$\phi^I \delta_{\epsilon ei}^2$	$^{comb} \delta_{\epsilon ei}^2$	$w_i$
$i_1$	1	0.00	0.45	0.33	0.33	1.25	1.13	1.13	0.42
$i_2$	1	0.00	0.88	1.59	0.72	2.88	3.59	2.72	0.64
$i_3$	1	0.00	1.13	0.33	0.33	3.13	2.33	2.33	0.23
$I$	1	0.00	0.28	0.28	0.28	0.92	0.92	0.92	–
$i_1$	2	0.06	0.04	0.05	0.04	0.90	0.91	0.90	0.54
$i_2$	2	0.06	0.05	1.09	0.09	2.11	3.15	2.15	0.91
$i_3$	2	0.06	0.07	0.05	0.05	2.13	2.11	2.11	0.46
$I$	2	0.02	0.02	0.01	0.01	0.68	0.66	0.66	–

use three subspaces  $i_1$  to  $i_3$  and assume seasonal dependencies  $f_{i1}(x)$  to  $f_{i3}(x)$  and noise as described in Table 1. Figure 2 shows the assumed functions  $f_i(x)$  at the different levels together with the noisy training target values assumed for two years of training data.

Two different methods of determining/learning the parameters are used. They are both based on a function  $h(x, \phi)$

$$h(x, \phi) = \phi_x \quad (15)$$

with a parameter vector  $\phi \in \mathcal{R}^{53}$  describing the behaviour in an isolated manner for each calendar week. Because of restrictions to the possibly learned parameter sets they describe differently complex function spaces at the ODIFPOS level.

The first learning approach generating  $h(x, {}^1\hat{\phi})$  represents a very complex function space. Each parameter is only restricted to the low limit of  $-1$  which is determined by the application (a seasonal reduction of demand of more than 100% is not possible). The parameters  ${}^1\phi_1$  to  ${}^1\phi_{53}$  are learned based on historical data using the best in sample estimation corresponding to the MSE error minimisation criterion that leads to a simple average of the data related to the corresponding calendar week

$${}^1\hat{\phi}_x = \frac{1}{2} \sum_{t \text{ with } x_t=x} y_t. \quad (16)$$

The second learning approach reduces the function space by two kinds of restrictions—limits to the generated seasonal factors as well as possible differences between neighboured seasonal factors reached by smoothing the data. For the detection of each

seasonal factor neighboured values are taken into account. Additionally a lower and an upper limit of  $-0.5$  and  $0.6$  for the expected seasonal deviation are used for stabilisation purposes in order to avoid for instance a zero season assumption in case of no historical bookings measured at the ODIFPOS level for a given calendar week.

$${}^2\hat{\phi}_x = \min(\max(\frac{1}{10} \sum_{t \text{ with } x_t=x} [y_{t-2} + y_{t-1} + y_t + y_{t+1} + y_{t+2}], -0.5), 0.6). \quad (17)$$

The artificial example allows us to have a separate look at the different error components. Table 2 shows the results of different error components generated with learning method 1 and 2 as described in Eqs. (16) and (17). The high level  $I$  contains the corresponding errors of the aggregated predictions. The bias, variance and total error of the pure low level predictions (and corresponding aggregates to the higher level) can be seen for the different subspaces and the two learning methods in columns 3, 4 and 7. It can be clearly seen that learning method 2 generates better results, even if it contains a bias component bigger than zero. Learning method 1 is less stable and contains much bigger parts in the variance component. We will discuss the other columns in later sections.

The bias component generated with learning method 2 can be seen in Fig. 3 at the example of subspace  $i_1$  together with the function  $f_{i1}(x)$  and the prediction (with deviation from  $f$  because of bias plus variance error terms). The bias contains restrictions in the case of very strong seasonal effects because of the used limits of  $[-0.5, 0.6]$  as well as minimal deviations because of the smoothing.

## 2.7. Effects of Learning at Different Levels on the Error Components

We will now analyse effects of learning at the different levels on the error components. The analysis is not only focused on the low level results, we are also interested in generating high quality forecasts at the high level. This can be achieved by learning directly at the high level or by aggregating low level predictions.

## 2.8. Learning $h$ at the Low Level

Corresponding to Eq. (10) the error achieved if we learn at the low level can be decomposed into  $\delta_{ei}^2 = \delta_{ehi}^2 + \delta_{e\phi i}^2 + \delta_{eyi}^2$ .

Let's now consider the aggregated pure low level predictions

$$\hat{y}_U = \bigcup \hat{y}_i = \bigcup h(x, \hat{\phi}_i). \quad (18)$$

The aggregation leads to errors at the high level of

$$\begin{aligned} y_I - \hat{y}_U &= y_I - \bigcup h(x, \hat{\phi}_i) \\ &= y_I - \bigcup (y_i - \epsilon_{hi} - \epsilon_{\phi i} - \epsilon_{yi}) \\ &= y_I - (\bigcup y_i - \bigcup (\epsilon_{hi} + \epsilon_{\phi i}) - \bigcup \epsilon_{yi}) \\ &= \epsilon_{hl} + \bigcup \epsilon_{\phi i} + \epsilon_{yl}. \end{aligned} \quad (19)$$

As the bias–variance–Bayes decomposition holds for the high level and we have already identified  $\epsilon_{hl}$  as elements of the error bias component and  $\epsilon_{yl}$  as the Bayes we know that the elements  $\bigcup \epsilon_{\phi i}$  represent the error variance component and are so independent of the other parts of the error. We get total error variances

$$\delta_{eU}^2 = \delta_{ehl}^2 + \delta_{e\phi U}^2 + \delta_{eyl}^2. \quad (20)$$

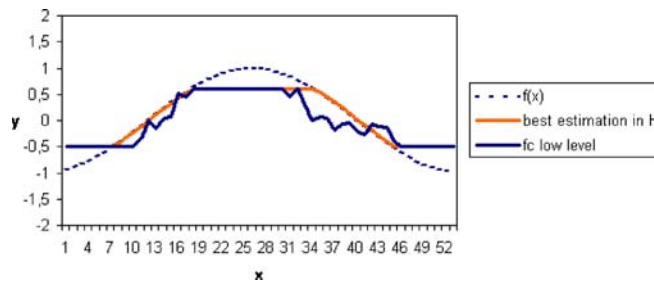


Figure 3. Function  $f_{i1}(x)$  together with the optimal and the generated prediction  $h(x, \hat{\phi}_{i1})$ .

## 2.9. Learning $h$ at the High Level

The alternative is to learn at the high level and to use the learned parameters for low level forecasts:  ${}^{\phi I} \hat{y}_i = h(x, \hat{\phi}_I)$ .

We will now analyse the composition of the resulting low level error. Combining Eqs. (2), (5) and (10) we get

$$\begin{aligned} {}^{\phi I} e_i &= y_i - {}^{\phi I} \hat{y}_i = f_i(x) + \epsilon_{yi} - h(x, \hat{\phi}_I) \\ &= f_I(x) - \epsilon_{fi} + \epsilon_{yi} - h(x, \hat{\phi}_I) \\ &= f_I(x) - \epsilon_{fi} + \epsilon_{yi} - (f_I(x) - \epsilon_{hl} - \epsilon_{\phi I}) \\ &= -\epsilon_{fi} + \epsilon_{hl} + \epsilon_{\phi I} + \epsilon_{yi}. \end{aligned} \quad (21)$$

We know that  $\epsilon_{hl}$  and  $\epsilon_{\phi I}$  are independent and that  $\epsilon_{yi}$  is pure random noise. In this case we can represent the error as

$$\begin{aligned} {}^{\phi I} \delta_{ei}^2 &= [\delta_{efi}^2 + 2Cov(\epsilon_{hl}, \epsilon_{fi}) + 2Cov(\epsilon_{\phi I}, \epsilon_{fi})] \\ &\quad + \delta_{ehl}^2 + \delta_{e\phi I}^2 + \delta_{eyi}^2. \end{aligned} \quad (22)$$

Let us now relate the above to the bias–variance–Bayes decomposition.

The series  $\epsilon_{fi}$  can again be decomposed in relation to the best approximation  $h(x; \phi_{efi}) \in \mathcal{H}$ :

$$\epsilon_{fi} = h(x; \phi_{efi}) + \epsilon_{fih} \quad (23)$$

and it follows that

$${}^{\phi I} e_i = -h(x; \phi_{efi}) - \epsilon_{fih} + \epsilon_{hl} + \epsilon_{\phi I} + \epsilon_{yi}. \quad (24)$$

The elements  $\epsilon_{fih}$  and  $\epsilon_{hl}$  belong to the bias term. Because of the linearity assumption of the approximation we know that

$$\delta_{chi}^2 = \delta_{efih}^2 + \delta_{chl}^2. \quad (25)$$

We can therefore represent the error also as

$$\begin{aligned} {}^{\phi I} \delta_{ei}^2 &= \delta_{chi}^2 + [\delta_{e\phi I}^2 + \delta_{hefi}^2 \\ &+ 2Cov(\epsilon_{\phi I}, h(x; \phi_{efi}))] + \delta_{eyi}^2 \end{aligned} \quad (26)$$

where  $\delta_{chi}^2$  belongs to the bias component,  ${}^{\phi I} \delta_{e\phi I}^2 = \delta_{e\phi I}^2 + \delta_{hefi}^2 + 2Cov(\epsilon_{\phi I}, h(x; \phi_{efi}))$  to the variance component and  $\delta_{eyi}^2$  to the residuals.

We see that learning at the high level outperforms the learning at the low level if

$$\delta_{e\phi I}^2 + \delta_{hefi}^2 + 2Cov(\epsilon_{\phi I}, h(x; \phi_{efi})) < \delta_{e\phi i}^2. \quad (27)$$

It strongly depends on the variance of  $\epsilon_{fi}$  if this relation is true, we will discuss that in more detail for different cases in the next section. While in some cases clear tendencies can be detected, the question is what level to choose for learning if the error variances are about the same:

$$\delta_{e\phi I}^2 + \delta_{hefi}^2 + 2Cov(\epsilon_{\phi I}, h(x; \phi_{efi})) \approx \delta_{e\phi i}^2. \quad (28)$$

As this decision has no relevant impact on the measured low level forecast quality, the decision should be made in relation to the high level quality as well as stability assumptions in case of changing environments.

Because of

$$\bigcup {}^{\phi I} \hat{y}_i = \bigcup h(x, \hat{\phi}_I) = h(x, \hat{\phi}_I) \quad (29)$$

we know that  $\bigcup {}^{\phi I} \hat{y}_i = \hat{y}_I$ . We profit from Eq. (7) with

$$\bigcup (-h(x; \phi_{efi}) - \epsilon_{fih}) = 0. \quad (30)$$

This indicates that in opposition to pure low level predictions we have an effect of error elimination of a part of the error variance component if the errors are aggregated to the high level. This can also have a stabilising effect in case of a changing environment

in the case when the situation does not change at the high level, i.e. shifts between the different subspaces. That's why we should always choose the higher level in these cases.

### 2.10. Using Forecast Combination

As we have already mentioned, the objective is to make choices concerning the level(s) for learning which manipulate the resulting errors concerning their correlation in a controlled manner. We have already seen that the choice of both levels for learning works well for some cases and not so well for others. The decision for one of the two approaches is difficult because the decision criterion should not depend on the pure error values at the low level. These do not take into account error variance correlations and stability effects in case of a changing environment. If we can manipulate the correlations of error variances in a manner that this is advantageous for the aggregation this should be taken into account for the choice of the level. On the other hand we want predictions at the fine level which do not only have a small error, but do also sufficiently clearly show special characteristics (features) of a given subspace if this is possible. If the data is very noisy additionally the errors can not be detected properly and, as the true function is not known, a decomposition of the error is not possible.

That's why an automated process is needed in order to make a qualified choice. Additionally it is advantageous to take not only one level into account, but to use the information present at both levels in order to generate good predictions. We need a flexible decision strategy in order to generate errors at the low level which are better or at least not much worse compared to the best choice of learning at the low or the high level, and at the same time to profit from similarities between the subspaces and levels in order to generate lower high level error variance terms. The decision process should be an automatic process which does not need to know details related to error decompositions.

Forecast combination techniques can be used in order to build complex functional approaches based on less complex ones in realising a reduction of the error bias component [4]. It can also be used in order to decrease the error variance component in following a thick modelling approach related to the setting of certain parameter values or to preprocessing [14,

15]. A similar situation compared to these tasks can be expected related to the choice of the forecast level. Each forecast level contains information based on which functional relationships, ideal parameter settings etc. can be determined, but it is likely that none of the models is optimal since it is not taking into account all the available information. Low level forecasts potentially miss general structure information. High level forecasts do not take into account the special characteristics related to the concrete part of the target space, or the representation of these characteristics is contained in the forecast model in a completely different manner than having built the model directly on the finer level. That's why it makes sense to study the approach of forecast combination as an option in order to incorporate the knowledge at the different levels and to analyse the effects on different error components at different levels.

### 2.11. Impacts of Forecast Combination on Low Level Forecasts

Using linear forecast combination on forecasts generated at the low and at the high level generates combined forecasts

$$\widehat{y}_i = w_i h(x, \widehat{\phi}_i) + (1 - w_i) h(x, \widehat{\phi}_I) \quad (31)$$

and errors

$$\begin{aligned} e_i &= w_i h(x, \widehat{\phi}_i) + (1 - w_i) h(x, \widehat{\phi}_I) - y_i \\ &= w_i (h(x, \widehat{\phi}_i) - y_i) + (1 - w_i) (h(x, \widehat{\phi}_I) - y_i) \\ &= w_i (\epsilon_{hi} + \epsilon_{\phi_i}) + (1 - w_i) (-h(x, \phi_{efi}) - \epsilon_{fih} \\ &\quad + \epsilon_{hl} + \epsilon_{\phi_I}) + \epsilon_{yi} \\ &= \epsilon_{hi} + [w_i \epsilon_{\phi_i} + (1 - w_i) (-h(x, \phi_{efi}) \\ &\quad + \epsilon_{\phi_I})] + \epsilon_{yi}. \end{aligned} \quad (32)$$

Under the assumption of independence this leads to

$$\begin{aligned} \delta_{ei}^2 &\sim \delta_{ehi}^2 + w_i^2 \delta_{\epsilon_{\phi_i}}^2 + (1 - w_i)^2 (\delta_{\epsilon_{\phi_I}}^2 + \delta_{hefi}^2 \\ &\quad + 2Cov(\epsilon_{\phi_I}, h(x; \phi_{efi})) + \delta_{\epsilon_{yi}}^2 \end{aligned} \quad (33)$$

More realistically we have to expect covariances between the different error variance components.

The difference between pure low level and pure high level forecasts is determined by the error variance component which can be approximated by

$$\begin{aligned} {}^{comb} \delta_{\epsilon_{\phi_i}}^2 &\sim w_i^2 \delta_{\epsilon_{\phi_i}}^2 + (1 - w_i)^2 (\delta_{\epsilon_{\phi_I}}^2 + \delta_{hefi}^2 \\ &\quad + 2Cov(\epsilon_{\phi_I}, h(x; \phi_{efi})). \end{aligned} \quad (34)$$

We will discuss what this means for different cases in Section 4. We will see that the weights are determined in a manner that for cases where the results generated at one level clearly outperform the other, the combination represents an automated choice of that level. For cases where both levels contain relevant information, the fusion process can even outperform the quality achieved at both levels.

This can be seen at our artificial example in comparing columns 3, 4 and 5 of Table 2. For subspaces  $i_1$  and  $i_3$  the error variance of the combined forecast is very close to the best single level results. For subspace  $i_3$  we can even outperform the results achieved at the low and the high level.

### 2.12. Impacts of Forecast Combination on Aggregated Low Level Forecasts

Forecast combination can be beneficial in order to increase the forecast quality at the low level. But the potential is still bigger if the forecasts are aggregated to the higher level as we show now in comparing combined aggregates with pure low level aggregates.

If we look at the aggregate of the combined predictions we get

$$\begin{aligned} y_I - {}^{comb} \widehat{y}_U &= y_I - \bigcup (w_i h(x, \widehat{\phi}_i) + (1 - w_i) h(x, \widehat{\phi}_I)) \\ &= y_I - \bigcup (\epsilon_{hi} + [w_i \epsilon_{\phi_i} + (1 - w_i) (-h(x, \phi_{efi}) + \epsilon_{\phi_I})] + \epsilon_{yi}) \\ &= \epsilon_{hl} + \bigcup [w_i \epsilon_{\phi_i} + (1 - w_i) (-h(x, \phi_{efi}) + \epsilon_{\phi_I})] + \epsilon_{yI} \\ &= \epsilon_{hl} + \bigcup [w_i \epsilon_{\phi_i}] + \bigcup [(1 - w_i) \epsilon_{\phi_I}] \\ &\quad - \bigcup [(1 - w_i) h(x, \phi_{efi})] + \epsilon_{yI}. \end{aligned} \quad (35)$$

We know that  $\epsilon_{hl}$  represents with  $\delta_{ehl}^2$  the bias component,  $\epsilon_{yI}$  is the Bayes, so it is clear that

$${}^{comb} \epsilon_{\phi_U} = \bigcup [w_i \epsilon_{\phi_i}] + \bigcup [(1 - w_i) \epsilon_{\phi_I}] - \bigcup [(1 - w_i) h(x, \phi_{efi})] \quad (36)$$

represents the variance error component (with variance  ${}^{comb}\delta_{\phi_{\cup}}^2$ ).

We can now write the error as

$${}^{comb}\delta_{e_{\cup}}^2 = \delta_{e_{HI}}^2 + {}^{comb}\delta_{e_{\phi_{\cup}}}^2 + \delta_{e_{YI}}^2. \quad (37)$$

Comparing the resulting error with the aggregated pure low level errors given in Eq. (20) and the high level learning error at the high level we have to again compare only the error variance terms  ${}^{comb}\delta_{e_{\phi_{\cup}}}^2$ ,  $\delta_{e_{\phi_{\cup}}}^2$  and  $\delta_{e_{\phi_I}}^2$ .

Let's now have a look what happens to the different parts of Eq. (36) during the aggregation. Compensation effects depend on the correlation of the elements at the different subspaces.

The first part is an aggregate of the weighted low level variance term  $\epsilon_{\phi_i}$ . As the low level parameter learning instabilities tend to be positively correlated, the component  $\delta_{e_{\phi_{\cup}}}^2$  can get very big and generate instabilities at the high level. This can only happen in the aggregation of the weighted elements if we have cases of big weights together with high terms  $\epsilon_{\phi_i}$ . Compared to the pure low level forecast the forecast combination represents a reduction of this component which is especially important and positive if we have big terms  $\epsilon_{\phi_i}$ .

The second part of Eq. (36) is an aggregate of weighted elements  $\bigcup[(1 - w_i)\epsilon_{\phi_I}]$ . Because of

$$\bigcup[(1 - w_i)\epsilon_{\phi_I}] = \epsilon_{\phi_I} \bigcup(1 - w_i) \quad (38)$$

this part is stable and small in case of big weights (the interesting case containing potential stability problems) and small values of  $\epsilon_{\phi_I}$  in comparison to  $\epsilon_{\phi_i}$ . In case of using only small weights this means that we generate predictions which are similar to the pure high level predictions.

The third part  $-\bigcup[(1 - w_i)h(x, \phi_{efi})]$  is determined by the function  $h(x, \phi_{efi})$ . Because  $\bigcup h(x, \phi_{efi}) = 0$  we can expect that the different elements of  $h(x, \phi_{efi})$  tend to be negatively correlated. It also follows

$$-\bigcup[(1 - w_i)h(x, \phi_{efi})] = \bigcup[w_i h(x, \phi_{efi})] \quad (39)$$

which means that we only achieve big values in cases where  $h(x, \phi_{efi})$  is relevant and  $w_i$  is big.

Comparing columns 3, 4 and 5 of Table 2 for the high level aggregate  $I$  show these positive effects of

the negative correlations for our artificial example. We can see that using forecast combination leads not only to better low level predictions, the aggregated combined predictions outperform the aggregated pure low level predictions and have the same quality as the forecasts generated directly at the high level.

We will now compare the effects of the different approaches in more detail for different cases in order to be able to make more specific statements about the expected forecast quality.

### 3. Discussion of Different Cases

#### 3.1. Case1 (*h is Too Complex to be Learned Properly Even at the High Level I*)

In this case we will have a big bias term  $\delta_{e_{\phi_I}}^2$ . The situation will probably be even worse at the low level. In any case the generated predictions will have a bad quality, but all of the other options discussed before will also have problems to reduce the error variance term. This case does not correspond to the general idea of including higher level information where the situation is more stable, we should use less complex functions  $h$  or include information generated at a higher level where the situation is more stable.

#### 3.2. Case2 (*h is Not Complex Enough*)

Geman et al. [11] argue that if we have relevant bias problems, meaning high terms  $\delta_{e_{hi}}^2$  and  $\delta_{e_{hl}}^2$  in our predictions, it is not possible to solve these problems properly without including other functions in order to approximate  $f$ . Nevertheless it can be that even with a very simple function  $h$  we get variance problems  $\delta_{e_{\phi_i}}^2$  if the training set in  $i$  is limited in sample size and characterised by high noise terms. If we get this problem we can reduce at least this part of the forecast error with the forecast combination approach.

But if we also want to reduce the bias term we have no other choice than to increase the complexity in  $h$ , which is dangerous because of the potential variance problems or to include other functions  $\tilde{h}$  that add additional information. If we also include predictions generated with  $\tilde{h}$  into the combination process, we have a chance to generate more complex functions during the fusion process and by doing so reducing the bias term.

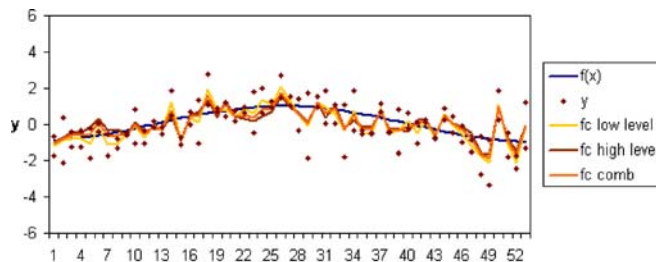


Figure 4. Predictions for subspace  $i_1$  generated with  $h(x, \hat{\phi}_{i1}^x)$ .

### 3.3. Case3 ( $i$ is Representative for $I$ )

This case means that the subspace  $i$  has a large impact  $\lambda_i$  in  $I$ . It follows that  $\delta_{ehi}^2 \approx \delta_{chl}^2$ ,  $\delta_{e\phi i}^2 \approx \delta_{e\phi l}^2$  and  $\delta_{efi}^2$  small in comparison to the other error components. The errors between the low and the high level forecasts are highly correlated and have similar size so that we will probably achieve weights near 0.5.

In this situation the best approach would be to determine the model at the low level, but choosing the high level does not make a big difference. We will not achieve any improvements using forecast combination compared to pure low level or pure high level predictions.

This case is represented in our example at subspace  $i_1$ . Figure 4 shows clearly that the predictions generated by learning at the low and the high level are strongly correlated. We have achieved combination weights of 0.42 for learning method 1 and 0.54 for learning method 2. The error of the combined prediction is in both cases very close to the best choice.

### 3.4. Case4 (Stable Situation in $i$ , But Clear Special Characteristics in $i$ )

In this case we can assume small components  $\delta_{e\phi i}^2$ ,  $\delta_{e\phi l}^2$  with  $\delta_{efi}^2$  significant. Following the strategy of forecast combination we will achieve a big weight  $w_i$

because of the high error component  $\delta_{efi}^2$  in the high level predictions (see Eq. (22)). This means that the fact, that the low level predictions should be used, can be represented by the weights very well. Also in this case it is not necessary to include higher level knowledge, but taking into account the higher level predictions with a small weight nevertheless can have a stabilizing effect at the higher level. As  $\delta_{e\phi i}^2$  and  $\delta_{e\phi l}^2$  are small and the error variance term as described in Eq. (34) is therefore strongly influenced by  $\delta_{h\phi i}^2$  we will have no problems during the aggregation (as argued in Section 3.5).

An example for this case exists in subspace  $i_2$  of our example if learned with method 2 (Fig. 5). The low level forecast has been chosen with combination weight 0.91.

### 3.5. Case5 ( $h$ is Too Complex to be Learned Properly in $i$ with $\delta_{efi}^2$ small)

In this case we have very noisy training sets with only few training data available in  $i$ . Learning only in  $i$  will lead to overfitting and big variance terms  $\delta_{e\phi i}^2$ . At the high level we have small values in all components of the error terms assuming that  $\delta_{efi}^2$  is small.

In this case the high level predictions will provide good predictions. This can also be well represented by forecast combination weights. We will achieve a

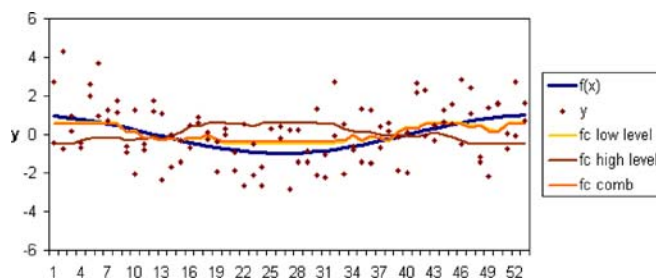


Figure 5. Predictions for subspace  $i_2$  generated with  $h(x, \hat{\phi}_{i2}^x)$ .

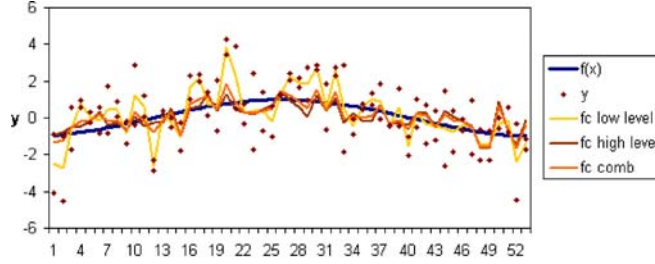


Figure 6. Predictions for subspace  $i_3$  generated with  $h(x,^1\hat{\phi}_{i3})$ .

small weight  $w_i$  and therefore no instabilities during aggregation. Forecast combination will not lead to improvements compared to the optimal choice of using only the high level predictions, but it can make this choice for us automatically.

This case is present in our example at subspace  $i_3$ . At this subspace the function  $f_{i3}(x)$  is very close to  $f_l(x)$ . Figure 6 shows that the high level predictions outperform the low level predictions. This is reflected in the combination weights of 0.22 and 0.45. The combined results even outperform slightly the high level predictions.

The higher weight in the case of the second learning method is due to a large bias error term in comparison to the error variance term. This can be seen very well in Fig. 7.

### 3.6. Case 6 ( $h$ is Too Complex to be Learned Properly in $i$ with $\delta_{efi}^2$ relevant)

This case represents the practically most relevant and also most interesting case. It means that  $\delta_{\epsilon\phi_i}^2$  is big and we have also a big error term  $\delta_{h\epsilon\phi_i}^2$ . Both predictions, pure low level and pure high level predictions will not be very good, but there is a chance that the errors are not strongly correlated. Forecast combination finds for us the best tradeoff between these two problems. We get an improve-

ment at the low level if the expectation of a low correlation between  $\delta_{\epsilon\phi_i}^2$  and  $\delta_{h\epsilon\phi_i}^2$  is true. But even if at the low level the improvement is not very big compared to the use of pure low or high level predictions, the use of forecast combination can be advantageous. Let's assume we would only choose the predictions generated at only one level.

If we would choose the pure low level predictions, we would generate error variance components  $\delta_{\epsilon\phi_i}^2$  which could cause problems during the aggregation. We would also risk instabilities in case of changing environments. In exchanging parts of  $\delta_{\epsilon\phi_i}^2$  by  $\delta_{h\epsilon\phi_i}^2$  with forecast combination we would have an increased aggregation stability (which we discussed in Section 3.5) as well as a higher stability if the situation changes. If on the other hand we would choose the high level predictions, we would generate predictions which do not represent the special characteristics in the subspace  $i$  at all which is not good for analysts or other systems which work with the generated predictions.

The situation in subspace  $i_2$  learned with method 1 in our example represents that case. The differences in the error variance term of the low and the high level learning can be clearly seen in Fig. 8. While the function learned at the low level has very high random deviations from the true function based on the noisy target data, the function learned at the high level is much smoother but has a completely

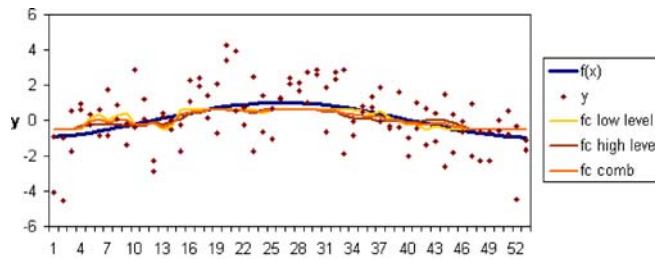


Figure 7. Predictions for subspace  $i_3$  generated with  $h(x,^2\hat{\phi}_{i3})$ .

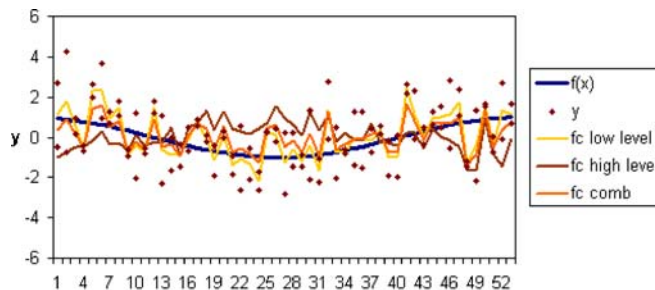


Figure 8. Predictions for subspace  $i_2$  generated with  $h(x, \hat{\phi}_{i_2}^x)$ .

different trend. It can also be seen that the combined forecast represents a good tradeoff between the two which on one hand has reduced noise and on the other hand approximates better the low level function  $f_{i_2}(x)$  than the function learned at the high level.

### 3.7. The Proposed Approach—Conclusions

As it could be seen from the previous subsections using the approach of generating multi-level forecasts and combining them seems advantageous in comparison to using pure low or high level learning. In most cases we will achieve an improved result at the low level. In cases where the low level forecast quality can only be slightly improved as compared to the best chosen individual low or high level forecast evaluated at the low level, the forecast combination process represents an automatic decision which level to choose. Additionally, in many cases we can also reach a modification in the correlation between error variance terms in a manner that the aggregate of low level forecasts gets a higher quality, which is especially important in systems where forecasts are generally aggregated in order to support decision making processes.

This can be seen by analysing the different parts in Eq. (36). We have already argued that the first component is an unstable component with elements which tend to be positively correlated. We have also mentioned that in the aggregation of the weighted elements instabilities can only happen if we have cases of large weights together with high terms  $\epsilon_{\phi_i}$ . The discussion of the different cases showed that this situation does not occur. In all cases where the elements  $\epsilon_{\phi_i}$  are big in comparison to  $\epsilon_{\phi_l}$  we do not get big combination weights  $w_i$ . We have shown that

the only cases where we do not get a small weight are the cases 3 and 6 with weights around 0.5.

While the second part of the equation is stable in any case, the third part can again contain big values in cases where  $h(x, \phi_{efi})$  is relevant and  $w_i$  is big. Again we have only the cases 3 and 6 where this can happen. We have seen that in case 3 it simply does not matter which level to choose because the low and high level are comparable and highly correlated. In case 6 we have high elements  $\epsilon_{\phi_i}$  as well as big terms  $h(x, \phi_{efi})$ . The replacement of parts of the pure low level forecast error variance terms  $\delta_{\epsilon_{\phi_l}}^2$  into  $\bigcup [w_i h(x, \phi_{efi})]$  is advantageous because of the negative correlation of the elements in  $\bigcup [w_i h(x, \phi_{efi})]$ .

Summarising we can say that in all cases where  $h(\cdot)$  is appropriate at the low or the high level (cases 3 to 6) forecast combination will generate very good results at the low as well as at the high level in comparison to pure low or high level predictions. In cases 4 and 5 forecast combination represents an automated choice of the right level. In case 6 we can even expect that the combined forecast outperforms the pure low or high level forecasts assuming the objective of generating good predictions for both levels. The most problematic cases are the cases 1 and 2 where  $h(\cdot)$  is structurally too poor or too complex for both of the levels. In these difficult cases we propose to follow the approach of “thick modelling” and the approach of using different function spaces in addition to multi-level combination.

## 4. Summary

The purpose of this paper is a theoretical analysis of the effects of multi-level forecast combination on different

error components described within the extended bias–variance–Bayes decomposition framework in comparison with the choice of a single level. We have proposed the extension of the bias–variance–Bayes decomposition to the multi-level case, analysed the effects of using multi-level information on different error components and seen that forecast combination is the best choice in comparison to the other alternatives. The proposed approach represents a completely automatic procedure that takes advantage of changes in the error components which are not only advantageous at the low level, but have also a stabilizing effect on aggregates of low level forecasts to the higher level because of changes in the correlations between the error variance components at different subspaces. We have also argued why multi-level forecast combination should ideally be connected with the use of different function spaces and/or thick modelling related to certain parameter values or preprocessing procedures. The effects of this analysis have been extensively illustrated using an artificial example that is motivated by a difficult real-world forecasting problem.

Beside the theoretical analysis we have also carried out a large number of experiments [16, 17] related to the application of revenue management seasonal forecasting. However, as the application is quite complex [18, 19], a detailed description of the experiments will be the subject of another publication. Nevertheless we would like to mention here that up to 12% error reduction has been achieved in comparison to the current optimised and tuned forecasting system of a major European carrier by using the proposed multi-level combination approach together with the different function spaces. To put this in a business context, it has to be mentioned that within this application the achieved results represent a very significant improvement bearing in mind that 2–4% of expected additional revenue is generated by airlines per 10% of reduced forecast error. As this means a lot of money, the current forecasting system is already quite well optimised and tuned and wins most of the benchmark competitions in these areas, so that these forecasts are not easy to beat.

## Notes

1. We have used this unification operator because it is a very common one. There is no restriction to this special unification

operator in the further analysis. Any other linear unification operator could be used as well.

2. Of course there are other (seasonal) impacts like e.g. day of week dependencies. We will restrict our argumentation to dependencies on the calendar week and assume no other known impacts to the booking behaviour in order to keep the example simple.

## References

1. G. Fliedner, “Hierarchical Forecasting: Issues and Use Guidelines,” *Ind. Manage. Data Syst.*, vol. 1, 2001, pp. 5–12.
2. G. James and T. Hastie, “Generalisations of the Bias/Variance Decomposition for Prediction Error”, technical report, <http://www.stat.stanford.edu/~gareth/ftp/papers/bv.ps>, 1996.
3. T. D. Russell and E. E. Adam, “An Empirical Evaluation of Alternative Forecast Combinations,” *Eur. J. Oper. Res. Econ.*, vol. 33, 1987, pp. 1267–1276.
4. L. M. De Menezes et al., “Review of Guidelines for the Use of Combined Forecasts,” *Manage. Sci.*, vol. 120, 2000, pp. 190–204.
5. A. G. Timmermann, “Forecast Combinations”, Discussion paper no. 5361, <http://www.cepr.org/pubs/dps/DP5361.asp>, 2005.
6. S. Makridakis et al., “The m2 Competition: A Real-Time Judgementally Based Forecasting Study”, *Int. J. Forecast.*, vol. 9, 1993, pp. 5–22.
7. J. M. Bates and C. W. J. Granger, “The Combination of Forecasts,” *Operations Research Quarterly*, vol. 20, 1969, pp. 451–468.
8. C. W. J. Granger and R. Ramanathan, “Improved Methods of Forecasting,” *J. Forecast.*, vol. 3, 1984, pp. 197–204.
9. E. W. Bunn, “Statistical Efficiency on the Linear Combination of Forecasts,” *Int. J. Forecast.*, vol. 1, 1985, pp. 151–163.
10. J. V. Hansen, “Combining Predictors. Meta Machine Learning Methods and Bias/Variance and Ambiguity Decompositions,” Ph.D. dissertation, 2000.
11. S. Geman, E. Bienenstock, and R. Doursat, “Neural Networks and the Bias- Variance Dilemma,” *Neural Comput.*, vol. 4, no. 1, 1992, pp. 1–58.
12. R. G. Cross, “Revenue Management,” Broadway Books, 1997.
13. McGill and van Ryzin, “Revenue Management: Research Overview and Prospects,” *Transp. Sci.*, vol. 33, no. 4, 1999.
14. C. W. J. Granger and Y. Jeon, “Thick Modelling,” *Econometric Modelling*, vol. 21, 2004, pp. 323–334.
15. M. Aiolfo and C. A. Favero, “Model Uncertainty, Thick Modelling and the Predictability of Stock Returns,” *J. Forecast.*, vol. 24, 2005, pp. 233–254.
16. S. Riedel and B. Gabrys, “Hierarchical Multilevel Approaches of Forecast Combination,” Proceedings of the OR 2004 conference, The Netherlands, 2004.
17. S. Riedel and B. Gabrys, “Evolving Multilevel Forecast Combination Models—An Experimental Study,” Proceedings of NiSIS 2005 Symposium, Albufeira, Portugal, 2005.

18. S. Riedel and B. Gabrys, "Adaptive Mechanisms in an Airline Ticket Demand Forecasting System," Proceedings of the EUNITE 2003 conference, Oulu, Finland, 2003.
19. R. Neuling, S. Riedel, and K.-U. Kalka, "New Approaches to Origin and Destination and No-show Forecasting: Excavating the Passenger Name Records Treasure," *Journal of Revenue and Pricing Management*, vol. 3, no. 1, 2004, pp. 62–72.



**Silvia Riedel** received her M.Sc. degree (german diplom) in Mathematics from the Martin- Luther University Halle Wittenberg, Germany in 1998. She has been working for Lufthansa Systems Berlin since October 1998 in the Division Revenue Management. Her current position is Operations Research Analyst. Her work focuses on the development and implementation of forecasting methods. Her responsibilities range from data analysis and development of application specific calibration and adaptation of complex mathematical models to the implementation of mathematical core components. She has presented her scientific results at different conferences, coordinates the cooperation with different universities and supervised different Masters Students. In 2003, she started a Ph.D. as external student at Bournemouth University (School of Design, Engineering & Computing) under supervision of Bogdan Gabrys.



**Bogdan Gabrys** received his M.Sc. degree in Electronics and Telecommunication (Specialization: Computer Control Systems) from the Silesian Technical University, Poland in 1994 and a Ph.D. in Computer Science from the Nottingham Trent University, UK in 1998. After many years of working at different Universities, Professor Gabrys moved to the Bournemouth University in January 2003 where he acts as a Head of the Computational Intelligence Research Group within the School of Design, Engineering & Computing. His current research interests include a wide range of machine learning and hybrid intelligent techniques encompassing data and information fusion, multiple classifier and prediction systems, processing and modelling of uncertainty in pattern recognition, diagnostic analysis and decision support systems. He published numerous research papers and is regularly invited to give talks at universities, companies and international conferences in UK and abroad. Professor Gabrys has also reviewed for various journals, edited special issues of journals, chaired international conferences, workshops and sessions and been on programme committees of a number of international conferences with the Computational Intelligence and Soft Computing theme. Professor Gabrys is a Co-Editor in Chief of the International Journal of Knowledge Based Intelligent Engineering Systems and the Chair (Academic Affairs) and a member of KES organisation Executive Advisory Board. He currently acts as a co-chair of the Natureinspired Data Technology (NiDT) focus group within the European CA project on Nature-inspired Smart Information Systems (NiSIS). In the recent years, he also acted as a corresponding person for a Key Node in the European Network on Intelligent Technologies for Smart Adaptive Systems (EUNITE) and a co-chairman of the Research Theory & Development Group on Integration of Methods. He is a senior member of the Institute of Electrical and Electronics Engineers (IEEE) and the IEEE Computational Intelligence Society and a fellow of UK's Higher Education Academy (HEA).