Dynamic Combination of Forecasts Generated by Diversification Procedures Applied to Forecasting of Airline Cancellations

Christiane Lemke*, Silvia Riedel** and Bogdan Gabrys*

Abstract— **The combination of forecasts is a well established procedure for improving forecast performance and decreasing the risk of selecting an inferior model out of an existing pool of models. Work in this area mainly focuses on combining several functionally different models, but some publications also deal with combining forecasts with the same functional approach. In the latter case, individual forecasts are generated by diversifying one or more model parameters or, if dealing with hierarchical data, by using forecasts from different levels. This work looks at multi-dimensional data from airline industry, with the aim of improving the forecast of cancellation rates for bookings. Three different methods are employed for the generation of individual forecasts.**

Forecast combinations are usually implemented in a more or less static structure, either including all available forecasts or trimming a fixed percentage of the worst performing models. For a big number of individual forecasts, this procedure can become inefficient. In this paper, a dynamic approach of pooling and trimming is applied to the generated forecasts for airline cancellation data.

I. Introduction

Forecasting demand and cancellation rates for flights is a crucial part in airline revenue management. The knowledge of the point in time when it is beneficial to restrict bookings in a lower-fare class to leave space for later booking high-fare customers is of both economical and ecological interest, producing a higher revenue for a high-demand flight and fewer unoccupied seats in a low-demand one. In previous work [1], the combination of forecasts and especially multi-level combinations for seasonality forecasts as a part of the demand prediction have proven very successful. Similar improvements are hoped for by investigating cancellation forecasts. There are two contributions in this work: Firstly, different approaches to generating individual forecasts for airline cancellation rates are investigated. Secondly, a dynamic combination approach being able to efficiently deal with the great number of individual forecasts is applied. Empirical evaluations for both of these aspects is presented. All experiments were carried out using booking and cancellation data for several flights obtained from Lufthansa Systems Berlin GmbH.

A. Time series forecasting

Time series forecasting is a very active area of research. As far as individual methods are concerned, approaches based on exponential smoothing are popular, yet simple to use methods ([2], [3]). Statisticians and econometricians tend to rely on complex ARIMA models and their derivatives ([4]). The machine learning community mainly looks at neural networks, either using Multi-Layer-Perceptrons with time-lagged time series observations as inputs as, for example, in [5] and [6], or recurrent networks with a memory, see, for example, [7]. Extensive empirical studies have been carried out, a very recent one being the M3 competition, which investigated 3003 time series. Results have been published in [3]. It seems, however, as if no method has ever proven successful across various studies and time series. This is mainly due to the fact that time series can have very diverse characteristics and underlying data generation processes, which makes it impossible to design a method working well for all of them.

Traditionally, several functionally different approaches are being applied to a problem before picking the one that is most suitable. More recently however, other approaches than just using different methods have been pursued in order to try a number of different individual forecasts. One method was proposed in [8] under the name of thick modelling. The general idea here is to generate different models using the same functional approach by varying one or more parameters used in the building or forecasting process of the model. This has shown to decrease model risk and improve forecasting performance.

A second method can be used in applications where data is available in a hierarchical structure. In airline industry, booking and cancellation numbers can be aggregated on different levels, for example for each or all points of sale or days of the week. A recent example of a publication investigating approaches to hierarchical forecasting is [9].

B. Forecast combination

Research in the area of combination of time series forecasts has a long track record, with the first related publication dating back to 1969. The motivation comes from the fact that all models of a real world data generation process are likely to be misspecified and picking only one of the available models is risky, especially if the data and consequently the performance of models change over time. Forecast combination is a reliable method of decreasing model risk and aims at improving forecast accuracy by exploiting the different strengths of various models while also compensating their weaknesses.

Usually, weighted linear combinations of forecasts with equal weights or weights that are in some way based on past

* Smart Technologies Research Centre, School of Design, Engineering and Computing, Bournemouth University, Poole House, Talbot Campus, Poole, BH12 5BB, UK, Phone: +44 1202 595298, Fax: +44 1202 595314, email: bgabrys@bournemouth.ac.uk

** Lufthansa Systems Berlin GmbH, Salzufer 8, 10587 Berlin, Germany, Phone: +49 30 34007141, Fax: +49 30 34007100, email: silvia.riedel@lhsystems.com

978-1-4244-2774-1/09/$25.00 ©2009 IEEE

performance are employed ([10]), furthermore, regression models can be used ([11], [10]). Literature in the area of nonlinear forecast combination is quite sparse, which is probably due to the lack of evidence of success as stated in [10]. Quite recently, pooling algorithms have been investigated, providing the possibility of dynamically trimming bad forecasts from the ensemble and generating a more efficient combination. Aiolfi and Timmermann proposed variance-based pooling in [12], which was used in investigations done in this work.

This paper looks at an approach to generate an extensive pool of diverse individual forecasts by using parameter, level and functional diversification. Both traditional combination methods and a pooling approach extended to multi-level forecasts are compared. Section II looks at the diversifications using the example of airline cancellation data, section III describes the traditional combination algorithms and the pooling approach that are compared in empirical experiments in section IV. The last section concludes.

II. Generating diverse individual forecasts

It is a common agreement that individual forecasts in a combination should differ from each other to produce a combination result that improves upon individual forecasts. In general, it is desirable that the forecasts to be combined are as accurate as possible, while weaknesses one forecast may have should preferably be compensated by the others. The concept of encompassing is investigated in [13], stating that one important characteristic of a superior individual forecast is its encompassing of rival forecasts, i.e. it includes all the information other models give. Forecasts that are encompassed by others are redundant in a combination. A recent application of this concept is given in [14].

This leads to the question of how to generate a pool of diverse individual forecasts. The most natural way is using methods with different functional approaches. Examinations presented in [15] and [16] come to the conclusion that both linear and nonlinear models should be present in a forecast combination. In previous work published in [17], a functionally diverse method pool consisting of methods like moving average or exponential smoothing, ARIMA models and neural networks has been investigated. However, the airline booking and cancellation time series used in this work can be characterised as being very short, prone to problems related to predicting small numbers and subject to strong time restrictions as many forecasts have to be generated in a very short time. For these reasons, only simple individual methods can be applied here; namely, three different forecasts are used: single exponential smoothing, Brown's exponential smoothing and a regression approach. These methods have proven to be successful in both forecast accuracy and compliance with time restrictions.

Individual forecasts can furthermore be diversified by thick modelling as presented in [8]. Airline cancellation forecasting at Lufthansa Systems Berlin is based on rates, restricting actual rates to certain confidence limits in both the history building and the forecasting process. Preliminary experiments have shown that manipulating some parameters for confidence limit calculation, making them slimmer or wider, has a positive effect if the resulting individual forecasts are combined. Four different sets of parameter values are used to add a second dimension to the set of individual forecasts.

The third and last diversification method can only be applied to hierarchical data sets, who are however quite common in service and retail industry. The history for airline cancellation rate forecasts in this application is usually built on the finest possible level, which means using data collected per fareclass, day of week, point of sale and origin-destination-itinerary. On this finest level, important characteristics that are only visible when looking at the bigger picture, i.e. a higher aggregation level, might be lost. On the other hand, using a high level might omit characteristics specific to certain parts of the data and lead to inferior forecasts as well. Generation of forecasts based on different levels and combining the resulting individual forecast saves the forecaster from having to choose one single aggregation level for the forecasting process. In this example, forecasts are generated using data from the finest level as well as data aggregated over days of week, fareclass and compartment.

III. Combining and pooling forecasts

Five different traditional forecast combination methods have been evaluated in the empirical experiments. Most of them have been introduced a long time ago, though still many researchers rely on these basic methods. Traditional combination methods include:

- Simple average : The available forecasts are averaged.
- Simple average with trimming: The forecasts are averaged as well, but only the best 80% are taken into account.
- Outperformance model: Weights for a linear combination are assigned based on the number of times a forecast performed best in the past [18].
- Variance-based model: Weights are assigned in relation to past error variance [19].
- Optimal model: Weights are calculated according to [20], taking covariance information into account.

All of the methods have strengths and weaknesses as reviewed in [21]. The simple average with and without trimming has the reputation of being notoriously hard to beat. The outperformance model, only rewarding methods performing best at a given point of time, omits all relative performances in its weight calculation. Only the optimal model takes covariance information of the individual forecasts into account, which is regarded as unstable especially if the number of forecasts is high in relation to the available forecasts as stated in many publications, for example in [20] and [12].

Relatively recently, Aiolfi and Timmermann introduced conditional combination strategies in [12]. Following empirical results saying that a good or bad performing forecast is more likely to keep performing well or badly instead of changing its performance, they group a number of forecasts that are diversified in functional approach and model parameters in two or three clusters using a k-means algorithm on their past error variance. Forecasts are then pooled within the groups before combining them with one of the following strategies:

- selecting the previously best performing cluster and averaging the forecasts contained in it,
- excluding the cluster that performs worse and averaging forecasts from the other clusters,
- combining forecast averages of each of the clusters using least squares regression or
- doing the same as in the previous approach but shrink weights towards equal weights.

In this work, the second approach has been investigated using three and four clusters. The combination method used for obtaining one forecast per pool is the simple average with trimming. Past experiments have shown that if the number of forecasts in a cluster exceeds five, it is useful to dismiss the worst performing ones.

IV. Empirical investigation

The data set investigated comprises 63 weeks of booking and cancellation data for 16 Origin-Destination-Opportunities (ODOs), which represent the routing between an origin and destination airport. Eleven of the flights take place within Europe, five of them are intercontinental ones. From the 63 available weeks, the first 28 are used as an initial period for history building for the individual models. The first 54 weeks are then used for learning forecast combination weights. Yearly seasonality is not accounted for in the cancellation rate forecasting process, as it is included in the booking forecasts. Out-of-sample error calculation takes place over the last ten weeks. The error measure used is the mean absolute deviation of the forecasted and the actual net booking numbers. The forecasted net bookings are calculated by the difference of the booking forecast and the absolute forecasted cancellation numbers in their unconstrained version, which means that all influences resulting from booking classes being closed for various reasons have been removed. The booking forecast is obtained using a single established method to assess the impact of the cancellation rates only. Forecasts are made for final booking and cancellation numbers on the finest possible level at 22 precedent data collection points (DCPs) of pre-defined time spans before departure. Errors are aggregated over point of sale and fareclass by simply adding them up to obtain a high level view, which the given result numbers are based on. Result tables in this section give numbers that represent the relative improvement compared to the best individual forecast for each of the DCPs the best individual forecast being determined dynamically for each DCP. Three sets of experiments are carried out, which are described in more detail in the next sections.

A. Ordinary combinations

As mentioned in the introduction, the most widely used combination types are combinations of functionally different forecasts. The three different individual models, single exponential smoothing, Brown's exponential smoothing and regression, have been combined with the five traditional combination models presented in section III.

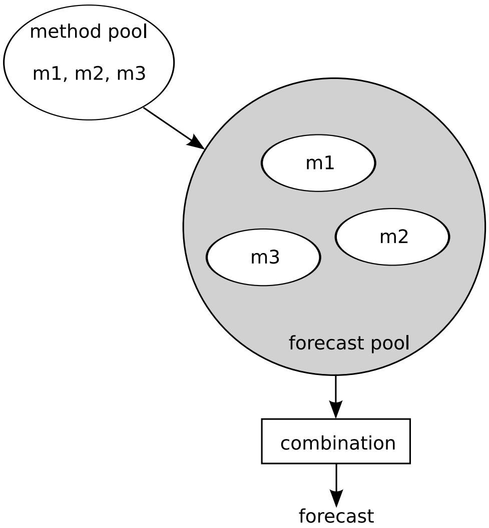

Fig. 1. Sketch of a combination of three forecasting methods.

A sketch of this procedure can be found in Figure 1. The pooling approach has been omitted for this experiment as well as the next one, as the number of three forecasts is too small to be organised in three or four clusters.

As shown in Table I, each forecast combination fails to outperform the best individual method on average. The most stable method, the outperformance model, only improves upon the best individual method at six out of the 22 data collection points. The best case here is an improvement of 2.5%, while the worst case deteriorates performance by 4.2%.

The basic experiments show that the simple combination of three functionally different individual forecasts does not improve performance when applied to airline cancellation forecasting. Reasons for that can be the small number of individual forecasts and their correlation, as combinations do generally not improve performance if input forecasts are too similar. The next two experiments investigate other methods of extending the pool and the diversity of available input forecasts.

B. Adding parameter diversified forecasts

In this experiment, the pool of forecasts is generated as depicted in Figure 2. Each of the three forecasting methods is built in four different ways, by diversifying a parameter set for calculating confidence limits in each of the forecasting processes.

Table II shows small improvements compared to the previous experiment, especially for the simple average with trimming, the outperformance model and the variance-based

TABLE I

Relative improvement in % of five forecast combination models (simple average (AVG), simple average with trimming (SAT), outperformance model (OUTP), variance-based (VAR), optimal (OPT)) compared to the best of 3 individual forecasts. Performance is given for 22 data collection points (DCP).

DCP	AVG	SAT	OUTP	VAR	OPT
0	-12.413	-12.413	-0.964	-11.157	-8.447
1	-0.164	-0.164	-1.281	-0.297	-4.835
2	1.951	1.951	0.093	1.858	4.348
3	3.493	3.493	2.502	3.566	3.346
4	1.823	1.823	2.112	2.022	1.637
5	0.233	0.233	1.119	0.886	-0.265
6	-0.718	-0.718	0.311	-0.005	-4.002
7	-1.141	-1.141	0.025	-0.271	-6.185
8	-1.904	-1.904	-0.261	-0.873	-5.307
9	-3.049	-3.049	-1.146	-1.852	-9.838
10	-4.433	-4.433	-1.795	-2.912	-7.265
11	-4.954	-4.954	-1.954	-2.870	-6.559
12	-5.139	-5.139	-2.207	-2.960	-5.535
13	-5.073	-5.073	-2.214	-3.253	-4.577
14	-6.836	-6.836	-3.183	-4.952	-4.730
15	-7.601	-7.601	-3.213	-5.596	-5.364
16	-7.857	-7.857	-4.228	-6.740	-9.290
17	-6.550	-6.550	-3.808	-5.549	-8.606
18	-5.517	-5.517	-3.776	-4.972	-7.916
19	-4.660	-4.660	-3.548	-4.331	-8.733
20	-3.587	-3.587	-2.534	-3.264	-6.241
21	-1.977	-1.977	-1.621	-1.800	-8.783
average	-3.458	-3.458	-1.435	-2.515	-5.143
minimum	-12.413	-12.413	-4.228	-11.157	-9.838
maximum	3.493	3.493	2.502	3.566	4.348

approach. The first two of them are now able to outperform the best individual method at 13 of the 22 data collection points. Mainly due to a performance outlier at the first data collection point however, the overall improvement is still negative, if only slightly.

The results of this experiment show that all of the combination methods were able to improve their average performance and increase the number of data collection points at which individual methods were outperformed. However, methods still suffer from negative performance outliers, especially at the early DCPs and there are still many cases where performance gets worse.

C. Adding the level dimension

Encouraged by the small improvements using parameter diversification, two more diversification dimensions are now added, exploiting the hierarchical nature of the airline data in this application. In addition to the functionally different forecasts using diversified parameters for their forecasting processes, forecasts are now also generated on the basis of the data aggregated over day of week and compartment as indicated in Figure 3. The sets of different parameters have been reduced to two for this experiment to reduce computational effort.

Because of the larger number of individual forecasts generated, the variance-based pooling approach described in section III has been employed in addition to the five traditional combination methods. Because of the non-deterministic nature of the clustering algorithm, eight structures are generated, picking the best one based on their pseudo-out-of-sample performance.

Amazing improvements can be observed in the result table III. The best traditional models, the outperformance and the variance-based model, outperform the best individual method at 21 out of 22 data collection points with improvements of up to 33.5%. The outstanding method however is the variance-based pooling, with three and with four clusters.

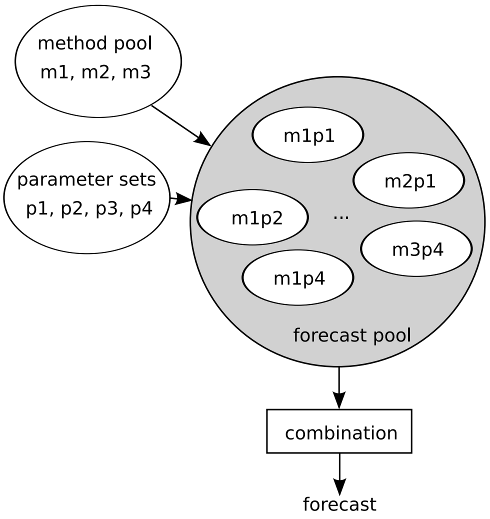

Fig. 2. Sketch of a combination of three forecasting methods, with additionally diversifying model parameter values.

TABLE II

Relative improvement in % of five traditional forecast combination models compared to the best individual method. Additional individual forecasts were generated using parameter diversification.

DCP	AVG	SAT	OUTP	VAR	OPT
0	-19.943	-12.200	-5.202	-18.501	-8.867
1	-2.050	-1.721	-3.086	-2.127	-4.441
2	-0.127	0.775	-1.877	-0.157	2.227
3	1.599	2.923	0.902	1.668	2.123
4	0.397	2.080	0.866	0.641	1.090
5	-0.360	1.923	0.859	0.497	-1.157
6	-0.834	1.321	0.334	-0.002	-1.975
7	-0.216	2.107	0.846	0.759	-4.042
8	0.250	2.656	2.214	1.239	-4.019
9	-0.853	1.503	1.290	0.264	-8.757
10	-1.528	1.313	1.253	0.079	-6.135
11	-1.392	2.012	1.699	0.719	-4.095
12	-1.284	1.257	1.244	0.421	-3.213
13	-1.588	1.083	1.034	0.117	-4.393
14	-2.496	0.401	0.893	-0.763	-3.788
15	-3.544	-1.014	0.689	-1.665	-4.459
16	-3.688	-2.422	-0.237	-2.549	-8.057
17	-3.338	-2.320	-0.663	-2.395	-7.725
18	-3.229	-2.044	-1.390	-2.622	-6.032
19	-2.957	-2.666	-1.757	-2.625	-14.953
20	-2.401	-1.594	-1.251	-2.088	-4.261
21	-1.326	-1.048	-0.944	-1.161	-7.593
average	-2.314	-0.258	-0.104	-1.375	-4.660
minimum	-19.943	-12.200	-5.202	-18.501	-14.953
maximum	1.599	2.923	2.214	1.668	2.227

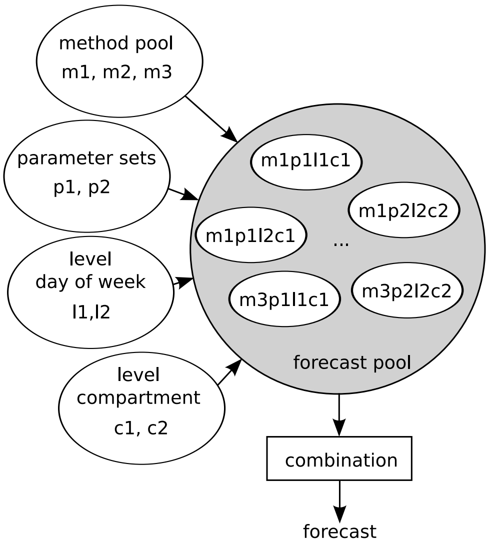

Fig. 3. Sketch of a combination of three forecasting methods, with additionally diversifying model parameters and learning levels.

TABLE III

RELATIVE IMPROVEMENT IN % OF FIVE TRADITIONAL FORECAST COMBINATION MODELS (SIMPLE AVERAGE (AVG), SIMPLE AVERAGE WITH TRIMMING (SAT), OUTPERFORMANCE MODEL (OUTP), VARIANCE-BASED (VAR), OPTIMAL (OPT)) AND THE VARIANCE-BASED POOLING APPROACH USING 3 AND 4 CLUSTERS (VBP3/VBP4) COMPARED TO THE BEST INDIVIDUAL METHOD. ADDITIONAL INDIVIDUAL FORECASTS WERE GENERATED USING PARAMETER AND LEVEL DIVERSIFICATION.

DCP	**AVG**	**SAT**	**OUTP**	**VAR**	**OPT**	**VBP3**	**VBP4**
0	-11.8	-7.4	-0.5	-10.8	-45.9	4.7	2.0
1	1.3	0.1	1.0	0.7	-18.3	6.8	6.3
2	4.8	4.1	4.8	3.7	-6.1	4.6	6.0
3	5.8	5.3	6.1	5.1	-5.1	7.7	5.8
4	2.7	2.2	2.8	2.3	-11.4	7.9	9.1
5	1.3	2.9	1.4	2.0	-9.7	5.0	5.3
6	3.0	4.3	2.1	3.4	-16.8	4.7	5.8
7	6.8	8.3	6.5	7.6	-16.2	9.3	8.9
8	10.8	12.8	10.9	12.5	-7.6	14.7	14.8
9	11.5	13.7	11.0	13.4	-13.4	14.6	15.1
10	13.5	16.5	15.3	16.8	4.7	19.3	19.5
11	17.3	20.4	19.9	21.6	8.4	23.9	24.2
12	17.7	21.0	21.0	23.5	8.1	24.9	25.0
13	17.6	21.0	22.0	24.8	12.0	26.6	26.9
14	16.1	20.3	23.7	28.0	17.7	32.4	32.4
15	13.0	18.2	24.7	28.5	-3.2	34.9	34.6
16	7.5	15.8	22.5	28.8	20.0	36.9	36.9
17	7.0	16.3	23.2	30.8	26.6	35.9	35.9
18	3.5	14.1	23.6	31.4	14.2	37.6	37.5
19	-4.2	8.6	22.7	31.4	26.3	37.5	36.9
20	-13.1	1.4	21.2	33.5	30.6	37.0	37.3
21	-49.3	-26.7	12.6	29.3	18.9	31.0	34.8
avg	3.8	8.8	13.6	16.7	1.5	20.8	21.0
min	-49.3	-26.7	-0.5	-10.8	-45.9	4.6	2.0
max	17.7	21.0	24.7	33.5	30.6	37.6	37.5

TABLE IV

PERCENTAGE OF TIMES A FORECAST GOT EXCLUDED FROM A COMBINATION.

	l1c1	**l1c2**	**l2c1**	**l2c2**
p1m1	48.2	5.5	11.2	1.5
p1m2	72.5	27.0	26.4	12.9
p1m3	56.4	13.1	17.3	5.7
p2m1	64.0	8.4	16.7	2.1
p2m2	83.1	27.9	28.8	12.3
p2m3	71.8	15.5	22.3	5.6

Both methods achieve improvements of up to 37%, outperforming all other combination methods at every single data collection point. Even in the worst case, improvements for this method amount to 2%. On average, the performance of both pooling methods is quite similar.

Exploiting information from higher data aggregation levels to generate a bigger pool of individual models leads to significant improvements in the cancellation forecast delivered by traditional combining models, however, at few DCPs, performance is still worse than that of the best individual model. The extremely good results of the dynamic pooling approach, which always outperforms individual forecasts in this experiment, shows that the extension of this approach to hierarchical data was successful.

V. ANALYSING VARIANCE-BASED POOLING

The variance-based pooling approach excluded a dynamic number of forecasts from each generated structure. With regards to the application, it is interesting to analyse how often which individual methods got excluded. This is shown in Table IV. For example, the first cell to the top left means, that a forecast using exponential smoothing (m1), parameter set one (p1) and a low aggregation level for both day of week (l1) and compartment (c1) was not present in 48.2% of the best structures.

One thing that becomes clear in the table is that forecasts being calculated on the low levels are being excluded most frequently (first column) while forecasts from the high levels (fourth column) are included most of the times. This strongly indicates that higher level forecasts perform better than the low level ones. Furthermore, Brown's exponential smoothing (m2, second and fifth row) is excluded more frequently than forecasts based on the other two methods, making it the seemingly weakest individual method of the three.

Having a closer look at the generated result structures, one example structure for pooling with three clusters is shown in Figure 4. The best performing cluster contains four methods, which were combined using a simple average to obtain the first input for the final combination. The second cluster contains eight methods. Since this number exceeds the maximum allowed number of five individual forecasts per combination, the worst performing 3 ones are trimmed before averaging the remaining five to obtain the second input for the final combination. The third cluster contains the twelve worst performing forecast of the method pool and is discarded completely; the remaining two pooled forecasts are averaged.

In the experimental results presented in the previous section, it can clearly be seen that adding forecasts diversified by the level of their calculation to the pool of individual

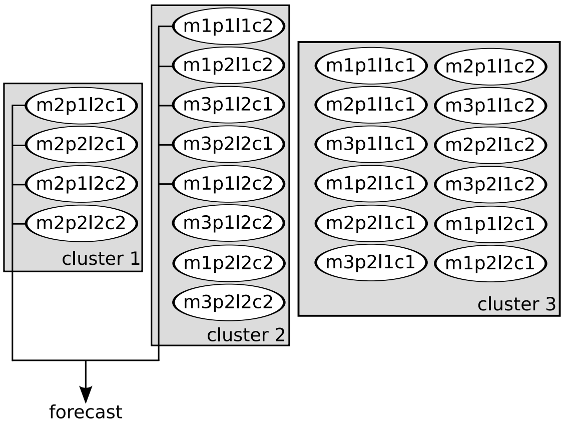

Fig. 4. One generated pooling structure, using three clusters. The worst performing cluster is discarded.

forecasts is very beneficial. However, it can also be seen that the variance-based pooling significantly outperforms the other combination approaches. Two reasons for this success can be given. Firstly, as opposed to all other methods, the number of forecasts included in the combination is determined dynamically for each problem by generating clusters and omitting the worst one. If using the simple average with trimming, the cutoff point is arbitrary and could exclude a method with a performance very similar to one retained for the combination. The other methods try to make use of all available forecasts, regardless of how bad they might be. The average number of forecasts included in a combination based on variance-based pooling is 17.4, ranging from 15.5 to 19.0 for different flights. Simple average with 80% trimming would statically include 19.2 methods in a combination.

The second reason for the superior performance can be attributed to inner cluster weights. In the three-cluster-scenario, the pooled forecasts of the two best performing clusters would get weights of $\frac{1}{2}$ each, in the four-cluster-scenario, the three best performing clusters would get weights of $\frac{1}{3}$. This is not related to the number of methods in each cluster, which can range from one to five in the algorithm used here. The weight assigned to a cluster is then distributed evenly, meaning it has to be divided by the number of forecasts in the cluster. Considering the example above, each of the nine forecasts included in the combination would get a weight of $\frac{1}{9}$ in a simple average combination. With the variance-based pooling however, the forecasts of the first cluster would get increased weights of $\frac{1}{8}$ each, while the five forecasts of the second cluster would decrease their weights to $\frac{1}{10}$ each as depicted in Figure 5.

Keeping in mind that one cluster contains forecasts with similar performance and information, a bigger group of forecasts agreeing with each other is consequently punished and looses weight in favour of forecasts grouped in smaller clusters. In this way, the pooling approach can ensure a better balance in the combination by taking interaction and similarity of forecasts into account when calculating weights.

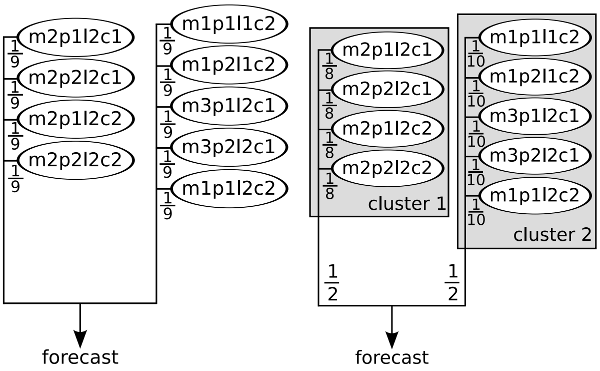

Fig. 5. Combination weights for a generated pooling structure using three clusters, left: simple average, right: variance-based pooling.

VI. Conclusion

This work investigated time series forecasting and forecast combinations applied to airline cancellation data. Experiments combining three different methods did not improve forecasting performance, on the contrary, performance gets worse compared to the best individual forecast. Therefore, diversification procedures for generating a larger pool of input forecasts have been investigated. Generating more forecasts by parameter diversification slightly improved results, still without a robust performance gain. However, additional generation of forecasts on different aggregation levels of the data before combining resulted in a significant reliable performance gain; in the best case improvements of up to 37% compared to the best individual forecast could be achieved.

In general, it was shown that parameter and level diversification procedures can make forecast combinations successful, even if the basic combination of functionally different forecasts is not very promising due to high correlation values and a small number of applicable methods. It was furthermore shown that the variance-based pooling approach proposed by Aiolfi and Timmermann can successfully be extended to using input forecasts from multiple levels and thus is an astonishingly useful method when dealing with hierarchical data. Reasons for the superior performance have been discussed and are related to better exploiting disagreement of individual forecasts and the dynamic number of forecasts included in a combination structure.

The investigation of the variance-based pooling approach was not exhaustive, more numbers of clusters as well as other outer- and inner-cluster combination methods could still be evaluated. Future work will also look at dynamically evolving combination structures for cancellation forecasting, as previously done for seasonality forecasts ([22]).

References

[1] S. Riedel, "Forecast combination in revenue management demand forecasting," Ph.D. dissertation, Bournemouth University in collaboration with Lufthansa Systems Berlin GmbH, 2007.

[2] E. S. Gardner, "Exponential smoothing: The state of the art–Part II," *International Journal of Forecasting*, vol. 22, pp. 637–666, October-December 2006.

[3] S. Makridakis and M. Hibon, “The M3-Competition: Results, Conclusions and Implications,” *International Journal of Forecasting*, vol. 16, no. 4, pp. 451–476, 2000.

[4] G. Box and G. Jenkins, *Time Series Analysis*. Holden-Day San Francisco, 1970.

[5] G. Zhang, B. Patuwo, and M. Hu, “Forecasting with Artificial Neural Networks: The State of the Art,” *International Journal of Forecasting*, vol. 14, pp. 35–62, 1998.

[6] J. Faraway and C. Chatfield, “Time series forecasting with neural networks: a comparative study using the air line data,” *Applied Statistics*, vol. 47, no. 2, pp. 231–250, 1998.

[7] T. Koskela, M. Lehtokangas, J. Saarinen, and K. Kaski, “Time series prediction with multilayer perceptron, fir and elman neural networks,” in *Proceedingsof the World Congress on Neural Networks*, 1996, pp. 491–496.

[8] C. Granger and Y. Jeon, “Thick modeling,” *Economic Modelling*, vol. 21, no. 2, pp. 323–343, 2004.

[9] R. J. Hyndman, R. A. Ahmed, and G. Athanasopoulos, “Optimal combination forecasts for hierarchical time series,” Monash University, Department of Econometrics and Business Statistics, Monash Econometrics and Business Statistics Working Papers 9/07, Jul. 2007. [Online]. Available: http://ideas.repec.org/p/msh/ebswps/2007-9.html

[10] A. Timmermann, “Forecast Combinations,” in *Handbook of Economic Forecasting*, G. Elliott, C. Granger, and A. Timmermann, Eds. Elsevier, 2006, pp. 135–196.

[11] C. Granger and R. Ramanathan, “Improved methods of combining forecasts,” *Journal of Forecasting*, vol. 3, no. 2, pp. 197–204, 1984.

[12] M. Aiolfi and A. Timmermann, “Persistence in Forecasting Performance and Conditional Combination Strategies,” *Journal of Econometrics*, vol. 127, no. 1-2, pp. 31–53, 2006.

[13] Y. Y. Chong and D. F. Hendry, “Econometric evaluation of linear macro-economic models,” *Review of Economic Studies*, vol. 53, no. 4, pp. 671–90, August 1986, available at http://ideas.repec.org/a/bla/restud/v53y1986i4p671-90.html.

[14] Y. Fang, “Forecasting combination and encompassing tests,” *International Journal of Forecasting*, vol. 19, no. 1, pp. 87–94, 2003.

[15] G. P. Zhang, “A combined arima and neural network approach for time series forecasting,” *Neural Networks in Business Forecasting, Hershey, PA: Idea Group Publishing:*, pp. 213–225, 2004.

[16] N. Terui and H. K. van Dijk, “Combined forecasts from linear and nonlinear time series models,” *International Journal of Forecasting, Volume 18, Issue 3*, pp. 421–438, July-September 2002.

[17] C. Lemke and B. Gabrys, “Do we need experts for time series forecasting?” in *Proceedings of the 16th European Symposium on Artificial Neural Networks, Bruges*, 2008, pp. 253–258.

[18] D. Bunn, “A Bayesian Approach to the Linear Combination of Forecasts,” *Operational Research Quarterly*, vol. 26, no. 2, pp. 325–329, June 1975.

[19] P. Newbold and C. Granger, “Experience with Forecasting Univariate Time Series and the Combination of Forecasts,” *Journal of the Royal Statistical Society. Series A (General)*, vol. 137, no. 2, pp. 131–165, 1974.

[20] J. Bates and C. Granger, “The combination of forecasts,” *Operational Research*, vol. 20, no. 4, pp. 451–468, 1969.

[21] L. M. de Menezes, D. W. Bunn, and J. W. Taylor, “Review of guidelines for the use of combined forecasts,” *European Journal of Operational Research*, vol. 120, no. 1, pp. 190–204, 2000.

[22] S. Riedel and B. Gabrys, “Dynamic pooling for the combination of forecasts generated using multi level learning,” in *Proceedings of the International Joint Conference on Neural Networks*, 2007.