TI 2012-131/III
Tinbergen Institute Discussion Paper

Managing Sales Forecasters

Bert de Bruijn
Philip Hans Franses

Econometric Institute, Erasmus School of Economics, Erasmus University Rotterdam, and Tinbergen Institute.

Tinbergen Institute is the graduate school and research institute in economics of Erasmus University Rotterdam, the University of Amsterdam and VU University Amsterdam.

More TI discussion papers can be downloaded at http://www.tinbergen.nl

Tinbergen Institute has two locations:

Tinbergen Institute Amsterdam
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 525 1600

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

Duisenberg school of finance is a collaboration of the Dutch financial sector and universities, with the ambition to support innovative research and offer top quality academic education in core areas of finance.

DSF research papers can be downloaded at: http://www.dsf.nl/

Duisenberg school of finance
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 525 8579

Managing Sales Forecasters

Bert de Bruijn[a,b,*]
Philip Hans Franses[a,b,†]
[a] Econometric Institute, Erasmus School of Economics,
Erasmus University Rotterdam, the Netherlands
[b] Tinbergen Institute, the Netherlands

November 30, 2012

Abstract

A Forecast Support System (FSS), which generates sales forecasts, is a sophisticated business analytical tool that can help to improve targeted business decisions. Many companies use such a tool, although at the same time they may allow managers to quote their own forecasts. These sales forecasters (managers) can take the FSS output as their input, but they can also fully ignore the FSS outcomes. We propose a methodology that allows to evaluate the forecast accuracy of these managers, relative to the FSS, while taking aboard latent variation across managers' behavior. We show that the results, here for a large Germany-based pharmaceutical company, can in fact be used to manage the sales forecasters by giving clear-cut recommendations for improvement.

Keywords: Forecast Support System; Sales forecasters; Forecast accuracy.
JEL classifications: M11, M31.

*Econometric Institute, Erasmus School of Economics, PO Box 1738, NL-3000 DR Rotterdam, The Netherlands; phone: +3110 4088902; email: debruijn@ese.eur.nl (corresponding author)

†Econometric Institute, Erasmus School of Economics, PO Box 1738, NL-3000 DR Rotterdam, The Netherlands; phone: +3110 4081273; fax: +3110 4089162; email: franses@ese.eur.nl

1 Introduction

Many globally operating companies rely on a forecast support system (FSS) to automatically create short- and long-horizon sales forecasts across products and countries. Often, such an FSS amounts to a sophisticated business analytical tool as monthly forecasts are automatically updated once new sales figures (and perhaps other relevant variables) become available, and also, many such systems allow for monthly updates of model creation. That is, each month the system is fed with new data, and each time another model can be used to create short- and long-horizon forecasts. The track record of recent forecast performance of such an FSS in turn provides information on structural shifts and influential data points, which can be somehow incorporated in the models that are used to create the forecasts, either by removal or by explicit inclusion using dummy variables. Models that are covered by many FSSs range from sophisticated versions of Box-Jenkins type models to basic extrapolation tools, or combinations thereof.

There is substantial literature that indicates that forecast support systems are useful tools. First, these systems take away the time and effort from managers to manually create their own forecasts. Second, an analysis of the FSS model-based forecast errors provides a useful basis to potentially improve the underlying statistical models. Third, the literature suggests that an FSS can yield quite accurate forecasts, which in turn are useful input for targeted business decisions.

Even though an FSS is recommended by academics and practitioners as a desirable business analytical tool, many companies allow the managers, who are responsible for local sales targets and for shipping and planning, to pair these forecasts from an FSS with their own forecasts. There are various reasons why managers or local sales forecasters may quote forecasts that differ from the FSS generated forecasts, see Goodwin (2000, 2002). One reason is that the sales forecaster, upon receipt of model-based forecasts, believes that somehow the expected forecast error can be reduced by including

information that could never have been included in the FSS. Foreseeable institutional changes, new tax laws, and other incidents may then be taken into account. Another reason for local sales forecasters to deviate from FSS forecasts is that they feel that the implemented model does not incorporate the correct information set or that model parameters are estimated incorrectly due to missing variables. The sales forecaster may then decide to ignore the FSS models altogether and use the forecaster's own model. Goodwin (2000, 2002) also suggests various other more psychological reasons.

Forecasts made by company-affiliated sales forecasters can also be used as a business analytical tool. Indeed, the difference between an FSS forecast and a manager's forecast can provide valuable information about what could be missing from the models in the FSS. These latter models can then be improved by incorporating these differences, see Franses and Legerstee (2013) for an illustration. Hence, an in-depth study of the factual creation of forecasts by sales forecasters (what do they do, and how?) and of their forecast error track record provides opportunities for companies to improve business decisions. In this paper we propose a methodology to carry out such an in-depth study, which we illustrate for a large database of a Germany-based pharmaceutical company. Our methodology can thus be used to manage the sales forecasters.

Literature

There is abundant evidence that sales forecasters quote different forecasts for sales data than FSS forecasts, see for example Fildes and Goodwin (2007), Bunn and Salo (1996), Sanders and Manrodt (1994), Nikolopoulos et al. (2005) and Syntetos et al. (2009). There are studies that suggest that such managerial intervention leads to improved forecasts, see Mathews and Diamantopoulos (1986) and Diamantopoulos and Mathews (1989), but on the other hand Fildes and Goodwin (2007) suggest that perhaps FSS forecasts are adjusted too often, consequently leading to a decrease in accuracy. A recent extensive study of Fildes et al. (2009) concludes that sales forecasters tend to be biased and that their over-optimism (that is, their forecasts exceed FSS forecasts)

leads to less accuracy.

Even though man-made forecasts are often available additional to FSS forecasts, and even though sales forecasters' forecasts may outperform FSS forecasts in terms of forecast accuracy, little seems to be known about what sales forecasters actually do and why they do so. Armstrong and Collopy (1998), Sanders and Ritzman (2001) and Lawrence et al. (2006) recommend that experts should keep records of their activities, but this is usually not done. As there are not many, if at all, of these records, an analyst thus needs to rely on actual data to derive what it is that sales forecasters really do. Franses and Legerstee (2009), accordingly, provide an extensive analysis of a large database for a Netherlands-based pharmaceutical company and they study the properties of (ten thousands of) sales forecasts. They show that in about 90% of all cases managers' forecasts differ from FSS forecasts. On average, there is also a slight tendency that this difference is positive. Furthermore, on average, they find that the difference between man-made forecasts and FSS forecasts is predictable.

All results documented so far in the literature provide averages across all cases. That is, usually no distinction is made for the behavior of forecasters who use the FSS forecasts as input and those who ignore these, simply because precise information on this behavior is not available. That this distinction is important however can be concluded from the survey results recently documented in Boulaksil and Franses (2009). Out of the forty-two forecasters who responded to their survey, only twenty indicated that they take the FSS forecasts as input for their own forecasts. Comparing these twenty with the remaining twenty-two, leads Boulaksil and Franses (2009) to conclude that people, who incorporate the FSS forecasts, believe that the FSS forecast is important for their own decision to adjust the FSS forecast and that they prefer to make small adjustments. The most important take-away from this survey-based study is that there are apparently distinct types of sales forecasters who display different attitudes towards the FSS and who also create their own forecasts in different ways. Any methodology that aims to analyze sales forecasts should therefore allow for such an unobserved variety

across sales forecasters, and this is what we shall do in the present paper for the first time in the literature.

The rest of this paper

In this paper we propose a methodology to analyze the quality of sales forecasts made by individuals where these man-made forecasts can deviate from available FSS forecasts. To illustrate our novel methodology, we consider a large database from a Germany-based pharmaceutical company, which has offices in a range of countries and also sells a range of products. This database has never been analyzed before, and all results in this paper are new to the literature. In the next section we provide ample details of the enormous data set, after a short review of the typical data features that are revealed in the relevant literature. In Section 3, we provide our analytical tool to link the behavior of sales forecasters with their forecast performance, while allowing for latent classes of individual forecasters with common behavior. Technical details of the relevant econometric model are relegated to an appendix. In Section 4, we discuss the main results. In Section 5, we provide some insights on how our results can be used to suggest forecasters to change some of their behavior so that they can improve their performance. Finally, in Section 6 we conclude and we provide an outlook of potentially useful further work, based on some of the limitations of our present study.

2 Data

In this section we summarize recent findings for large databases containing sales forecasts from FSSs and managers. Next, we discuss the features of our database, building on these recent findings.

Typical data and typical findings

Franses and Legerstee (2009) analyze monthly sales figures in SKUs of pharmaceutical products in seven categories. They consider data concerning 37 countries. Their

company of interest is a Netherlands-based company which uses an FSS to create forecasts, which are communicated to managers in local offices. The managers are allowed to modify the FSS forecasts and these final forecasts are recorded too. Finally, these authors also have the actual sales figures in SKUs for all months. Their sample runs from October 2004 through October 2006, implying the presence of 25 one-step-ahead forecasts for each country/category combination.

The focus in Franses and Legerstee (2009) is on the differences between model-based forecasts for SKU-level data and the managers' forecasts. Their interest lies in the frequency of the cases where this difference is not 0. Also, they address whether managers adjust model-based forecasts more upwards than downwards, that is, if there is overoptimism. Additionally, they examine whether the difference between the adjusted forecast and the FSS forecast is predictable from past data, where own past adjustment and recent model-based forecast errors could be important drivers. Finally, they examine to what extent the size of the adjustment is correlated with the model-based forecast itself. Their key findings are that in about 90% of all 30000+ cases managers adjust FSS forecasts and that in 54% of the cases this adjustment is upwards. Furthermore, the size of managers' adjustment is predictable for about 44% of the variation, and even the direction of that adjustment is found to be predictable. Forecast adjustment is found to depend mainly on habit formation.

A second recent study that considers a large database with sales forecasts and realizations is Fildes et al. (2009). These authors collected 60000+ triples of managers' forecasts, FSS forecasts and actual sales data for no less than 4 different supply-chain companies, also including a company that manufactures pharmaceuticals. In their study the authors seek to examine to what extent judgmental adjustments lead to higher forecast accuracy in terms of Mean Squared Error and in terms of the differences between the absolute percentage forecast errors across the FSS and the managers' final forecasts. Fildes et al. (2009) find that large adjustments tend to lead to more accuracy, while upward adjustments deteriorate forecasts. Further, there is evidence of optimism

amongst the managers, but this often was seen to be associated with adjustments in the wrong direction. Hence, managers (sales forecasters) tend to be optimistic at the wrong moment.

A third and related study is Franses and Legerstee (2010). These authors analyze the same data as in their 2009 study and they find that essentially managers' forecasts are not better than the FSS forecasts, in terms of Root Mean Squared Prediction Error (RMPSE). Another conclusion is that managers tend to deviate too much from the FSS forecasts. In fact, these authors claim that managers put too much emphasis on their own judgment and too little on the models in the FSS. Finally, they show that linear combinations of managers'forecasts and FSS forecasts have much higher accuracy.

Our case study

The data set that we use is provided to us by a globally operating Germany-based pharmaceutical company. Country-specific managers produce sales forecasts for a set of products, and this set is different per country. This means that there is only a single forecaster for each product in a certain country. The dataset also contains the FSS forecasts, the adjustments (which are the differences between the managers' and the FSS forecasts) and the actuals. Each pharmaceutical product can be classified to a specific category.

The full dataset concerns 11432 products with 29 monthly 3-months ahead forecasts for the period 2009-2012 (May). For many products there are only forecasts for a few of the months in the sample. Next to this, the managers are not always very precise in their reporting behavior. For example, sometimes the actuals and the forecasts are not of the same magnitude, meaning that for example an FSS forecast is reported in thousands of units, while the actual is reported in millions of units. For some cases it is clear how to bring them to the same order of magnitude, for others it is not, and these latter cases are dismissed. We have filtered the products such that only those products for which more than half of the sample period is in good condition are

selected. All three relevant series (manager forecasts, FSS forecasts and actuals) were required each to meet this criterion. Illustrative examples of our data are presented in Figures 1 and 2. After filtering, we calculate for every product/country combination the median percentage error (MPE) as compared to the median realization of sales of the corresponding product in the corresponding country. We use the median in order to robustify the accuracy measures against unnoticed badly-reported forecasts. We use the percentage error (instead of the error) to make products and countries of different sizes comparable. This is important for our econometric model below which allows for latent classes of managers with similar behavior. Similarly, we also calculate the median absolute percentage error, the median percentage adjustment and the median absolute percentage adjustment.

Data cleaning took two months, and finally, we end up with data for 2472 products across 67 forecasters. The average number of products per forecaster is 36.90. The distribution is depicted in Figure 3. Clearly, many forecasters deal with 1 to 40 products, although some care for more than 150 products. Figure 4 gives the valid forecasts per forecaster. Figure 5 gives the distribution of the median absolute percentage error, which is heavily skewed. This means that sometimes the forecast error is exceptionally large. To ensure that a few large observations alone do not dictate all the results, we will instead analyze the natural logarithm of the median absolute percentage error, of which the distribution is shown in Figure 6. More details on this database are discussed below where we deal with aspects of our methodology.

3 Methodology

Part of what forecasters do is dependent on the context. To allow for varying context to influence certain properties, we use a model with three levels. At the same time, forecasters facing similar contexts can behave similarly, and hence there may be latent (unobservable) classes of forecasters. Therefore, we propose a model in which there are

S different types of forecasters, where S ranges from 1 to a number to be estimated. A single forecaster can be entirely associated with one such class, but she can also partly be associated with different classes. The basic equations of our model are as follows, where a forecast accuracy measure (y) is on the left hand side (LHS) and aspects of forecasters behavior (X) are on the right hand side (RHS):

$$\begin{aligned} y_{p,i} &= \beta_{p,i} X_{p,i} + \varepsilon_{p,i} \\ \beta_{p,i} &= \gamma_i Z_{p,i} \\ \gamma_i &= \sum_{s=1}^{S} \psi_s \mathrm{P}[Type_i = s] \end{aligned} \tag{1}$$

with $p \in [1, ..., P]$ indicating the individual product, $i \in [1, ..., N]$ indicating the forecaster responsible for in total P_i products and $s \in [1, ..., S]$ indicating the different types of forecaster. $X_{p,i}$ contains $K_X + 1$ variables (including an intercept), and $Z_{p,i}$ contains $K_Z + 1$ variables on context. Given the discussion in Section 2 we will use the variables that form $y_{p,i}$, $X_{p,i}$ and $Z_{p,i}$ as they are denoted in Table 1.

Using just the top level, our model describes a link between the size of the percentage error on the LHS and the size and sign of the percentage adjustment on the RHS. For example, if both elements of the vector $\beta_{i,p}$ are positive, a larger adjustment tends to occur simultaneously with larger errors, and even more so for larger upward adjustments.

Adding the second level, our model describes how the contexts of the forecasters might mediate the links in the top level. For this we use three variables in our database that are assumed to be outside the direct influence of the forecaster (and are controlled by the managers of the forecasters), and these are the number of products she deals with, the number of products in the same product category as product p and the autocorrelation in the FSS forecasts, where the latter choice is based on the results in Franses and Legerstee (2009). If an element of the vector γ_i is positive, then the corresponding $\beta_{p,i}$ becomes more positive or less negative if the corresponding $Z_{p,i}$ increases. Hence, the effects of $X_{p,i}$ can be amplified or dampened.

The third level implements the different forecaster classes. Every γ_i is a linear combination of the vectors ψ_s, $s = 1, ..., S$, weighted using the probabilities of forecaster i belonging to latent class s.

To estimate the parameters in this model for a given number of forecaster types, one only needs to estimate ψ_s and the probabilities $\mathrm{P}[Type_i = s]$. The estimated values of γ_i and $\beta_{p,i}$ then follow. To estimate ψ_s and the probabilities, we use the EM-algorithm. The exact implementation of this algorithm can be found in Appendix A.

Finally, to decide on how many forecaster types we should consider, we choose to use the AIC-3 criterion (as suggested by Andrews and Currim, 2003) and the BIC criterion (which usually leads to a smaller number of classes). At the same time, we require that all classes are of sufficient size. As we only have 67 forecasters to categorize, it might be that one of the forecaster types only has 2 or 3 forecasters, especially if there is a large number of classes.

4 Results

We have estimated the model parameters in (1) for different values of S, using the estimation method as described in Appendix A. The AIC-3 criterion suggests using 6 clusters, while the BIC recommends 3 clusters. A close look at the case with 3 clusters indicates that there is one cluster with only a few observations, and this holds even more true in the first case, where four groups are very small. Because of this, we decide to limit the number of groups to 2.

As the estimated values of $\beta_{p,i}$ and $\gamma_{p,i}$ are directly dependent on the estimated values of ψ_s and on the type probabilities, we report the estimation results for ψ_1 and ψ_2 in Table 2. Both types have one characteristic in common. For both types the size of the forecast error increases if the size of the adjustment increases. For type 1 this increase is much larger (1.108 versus 0.504). Additionally, the effect of this variable for type 1 forecasters increases if the product is part of a larger category of products (0.087)

and if the autocorrelation in the FSS forecasts increases (0.173), but it decreases if there is an increase in the total number of products assigned to the forecaster (-0.179). For type 2 forecasters, these effects are not significant.

The forecast error increases for negative adjustments (compared to positive adjustments of the same size) for forecasters of type 1 (-0.145), while type 2 forecasters have a larger error in the case of positive adjustments (0.049). Also, the forecast error decreases for type 2 forecasters in the countries with more products (-0.097), but increases if the product is part of a larger product category (0.172). In the next section, we will discuss in more detail what the managerial implications are of these estimates, but for now we can conclude that there are two types of forecasters with clearly distinct behavioral characteristics. This can also be learned from the estimation results in Table 3, where we present the parameter estimates in case we assume that all forecasters constitute a single group. Due to averaging various significant effects seem to disappear.

Figure 7 shows the distribution of the estimated probabilities of a forecaster being of type 1. As can be seen, our model allows us to clearly categorize most of the forecasters into either forecasters of type 1 (the right side of the histogram) or of type 2 (the left side). Note that such a distinction would be impossible by just looking at the graphs in Figure 3 and 6. There are only a few forecasters who are a mix of both types. This shows that there really is a distinction between the two types of forecasters, and also that these classes are substantially large. We have tried to explain the categorization of our multi-level approach using available explanatory variables (for example, using a binomial probit), and we have found no significant parameters. This again indicates that one needs a multi-level or mixture model such as ours to disentangle different classes of forecast behavior.

Table 4 shows several characteristics of the forecasters per type, which are weighted averages with as weights the type probabilities. Notice both types of forecasters perform worse than the FSS forecasts, on average (which is consistent with earlier findings in the literature). There are twice as many forecasters of type 2, but type 1 forecasters

carry about twice as many products. Type 1 forecasters adjust more upwards (63.3 versus 61.4), which at first glance seems beneficial for them due to their negative sign of β_2 in Table 2 (-0.179). This effect may be countered by their number of products (indeed, $29.95 \times 0.054 = 1.6173 > 0.145$, their average parameter of X_2 is positive and higher than that of Type 2). For X_1, the reverse holds true. There, the larger number of products makes the coefficient to decrease. Taking all factors into account simultaneously, the percentage error for type 1 forecasters is larger than the percentage error of type 2, and this may be due to their actions or due to their possibility to have a more difficult forecasting task. In the next section we will highlight the prominent estimation results in a numerical experiment, which is also useful for managing the forecasters.

5 Managerial implications

Most important for forecasters, and for those who manage these forecasters, is what they can do to improve their forecasts, that is, what must change in their forecasting context and behavior to improve the forecast quality? First, for an average person of Type 1, the value for logMedAbsPercErr would be 1.069, while for Type 2 this would be 0.721, as can be seen in the bottom row of Table 5. Table 5 also shows how much this would change if certain characteristics would be adjusted by 1 or 0.1 while the other values are kept as constant. As this effect is the effect on a log-measure, this can be interpreted as a percentage change. For example, a change of 10 % in the median absolute percentage adjustment (so, more deviation from the FSS forecast) increases the median absolute percentage error of the average type 1 forecaster by 6.5 %. The effects of the two $X_{p,i}$ variables are similar across both types (top two rows in Table 5). Increasing the size of the adjustment increases the size of the error, and making more positive adjustments also increases the size of the error. This supports the empirical findings in the literature for other large datasets, see Section 2. Concerning the three

$Z_{p,i}$ variables, the effects are of opposite sign across the two types of forecasters. For example, increasing the number of products for a forecaster with 10% (while keeping the number of products per category equal, which effectively means introducing new categories to this forecaster and increasing her task) increases the absolute percentage error with 0.6% for type 1 forecasters, while type 2 forecasters will see a decrease of 0.7%. Overall, these $Z_{p,i}$ effects are smaller in absolute sense than the $X_{p,i}$ effects. One can also combine effects. For example, if both the median absolute percentage adjustment and the number of products decreases with 10 %, this will lead to values of the median absolute percentage error of 0.996 (-7.3%) for type 1 and 0.683 (-3.8%) for type 2. For type 1 this effect is larger than solely decreasing the median absolute percentage adjustment with 10 %, while for type 2 it is smaller. This difference is due to the opposite effect of a decrease in the number of products for both types.

The outcomes in Table 5 lead to the following implications for managers of the forecasters. It is more effective to change what forecasters do (the $X_{p,i}$ variables) than to change the situation they are in (the $Z_{p,i}$ variables). This is the case for two reasons. First, the effect is larger, and second, the effect is of similar sign for both types, so one does not need to distinguish between the two types. If one tries to decrease the size of the error by changing the context, one should keep in mind that different forecasters respond differently to changes in these contexts, as shown in Table 5.

6 Conclusion

In this paper we have proposed a methodology that can be used to improve forecast accuracy of sales forecasts when these are created by sales managers who can decide to quote forecasts that differ from those given by a Forecast Support System. In some cases, man-made sales forecasts improve on FSS forecasts, but in other cases these judgmental adjustments deteriorate forecast performance. As sales forecasters can have useful information that is not included in an FSS, we do not recommend to dismiss the

human touch, but instead we provide suggestions as to how one can improve forecast performance. Such an improvement can be due to changing the forecasters behavior and/or by changing the context (like number of products for which forecasts are required). To this end, we have put forward a novel methodology that links forecast performance with behavior and context, where we allow for the potential presence of distinct latent classes of forecasters. This last feature is of tantamount importance as forecasters may decide adopt the FSS forecasts, but they may also decide to ignore the FSS forecast all the way.

When we applied our methodology to a novel and large database concerning sales forecasts for pharmaceutical products, we learned that a change in the behavior of forecasters would lead to most improvement in forecast performance, more so than changing the context. More precise, smaller deviations from the FSS forecasts and less optimistic adjustments lead to better forecasts. At the same time, reducing the task of forecasters who look after not so many products by reducing that amount even further does help, while this is not the case for those forecasters who are concerned with many products already. Clearly, experienced forecasters can handle more products and can still keep performance at a constant level, while less experienced forecasters do better when they make forecasts for a smaller set of brands. At the same time, experienced forecasters tend to take FSS forecast errors more into account than less experienced one, and if they, the experienced forecasters, would pay less attention to these FSS errors, their own performance would increase. In sum, our methodology allows for clear-cut suggestions as to how sales forecasters can be managed such that their forecast performance increases.

A limitation of our study is that our analysis is merely of a descriptive nature. In turn, this immediately suggests an opportunity for further research if we were able to concretize the suggestions above and at a later moment in time observe if managerial changes had any effect. We delegate this interesting issue to further research.

A Technical details of the estimation routine

For the parameter estimation of (1) we use an Expectation-Maximization-algorithm (EM-algorithm), of which the concept was originally introduced by Dempster et al. (1977). In such an algorithm there are two steps: an Expectation-step (E-step) in which the expectation of a set of unobserved variables is taken, given current estimates of the parameters, and a Maximization-step (M-step) in which the likelihood function of the parameters is maximized, given current estimates of the unobserved variables. These steps are then repeated until convergence. In our case, the group probabilities are the unobserved variables, while the ψ_s are the parameters together with the unconditional probabilities of belonging to one group, for which we use the notation P_s. In the discussion below, y_i consists of $y_{p,i}$ for all products p that manager i is responsible for. Similar notations are used for X_i and Z_i as compared to $X_{p,i}$ and $Z_{p,i}$.

E-step

In the E-step the expectation of the group probabilities is taken, given estimates of ψ_s for $s = 1, ..., S$ and the unconditional probability P_s. This expectation is calculated by comparing for all s the individual densities $f_{i,s}(y_i; X_i, Z_i, \psi_s)$ in the case forecaster i would be fully assigned to type s, also incorporating the unconditional probability P_s. The fraction of $\mathrm{P}_s f_{i,s}(y_i; X_i, Z_i, \psi_s)$ to the sum $\sum_{s=1}^{S} \mathrm{P}_s f_{i,s}(y_i; X_i, Z_i, \psi_s)$ is the new estimate of $\mathrm{P}[Type_i = s]$.

M-step

In the M-step the likelihood function of the parameters ψ_s and P_s is maximized, given estimates of the group probabilities. It can be shown that these two variables can be estimated separately. To estimate P_s, one can simply take the averages of $\mathrm{P}[Type_i = s]$ across all i.

For estimation of ψ_s, we can show that the model reduces to a simple regression.

This can be seen by using the multilevel property of the model: $y_{p,i} = \beta_{p,i} X_{p,i} = \gamma_i Z_{p,i} X_{p,i} = \sum_{s=1}^{S} \psi_s \mathrm{P}[Type_i = s] Z_{p,i} X_{p,i}$. Define $X^*_{p,i,s} = \mathrm{P}[Type_i = s] Z_{p,i} X_{p,i}$, then the model reduces to the regression of $y_{p,i}$ on all S matrices $X^*_{p,i,s}$. This results in estimates of ψ_s, which can be used to construct estimates of γ_i and $\beta_{p,i}$.

Starting points and convergence

As the EM-algorithm might converge to a local optimum, we use several starting points for every S that we consider. The first starting points are derived from the best likelihood for the case $S-1$ as follows:

- Select a type s from the previous $S-1$ types.
- Split this type into two types by randomly assigning different proportions of $\mathrm{P}[Type_i = s]$ to types s and S, which is the new type. For each forecaster i, this is accomplished using different proportions of the original $\mathrm{P}[Type_i = s]$ to both new types s and S.
- Start the EM-algorithm and run it until convergence.

As there are $S-1$ types to split up, this results in $S-1$ outcomes. We also use R other starting points, which are constructed by randomly drawing $\mathrm{P}[Type_i = s]$ from a Dirichlet distribution with all S parameters equal to $\frac{1}{S}$, for each forecaster i. We have set R to 2500. Of these total $R+S-1$ converged estimates, the best one is chosen using the likelihood.

Following the above discussion, we end up with estimates for each S that we consider. Of course, increasing S will increase the likelihood, as in the worst case a type can always be cut into two to reduce the ideosyncratic error a bit, even though the forecasters within the types are actually from the same type. To choose the final S, one can use the AIC-3 criterion, which has been shown to perform well in a multilayered model (Anders & Currim, 2003). If a smaller number of types is desired, one can use the BIC criterion.

Further practical considerations

In the EM-algorithm, one can quickly run into numerical problems. For example, the density when a certain individual is categorized into a certain group might be so low that it is almost zero. To avoid numerical problems, natural logarithm formulations of the above are used, which can be easily adjusted to avoid such problems during estimation. If for a forecaster during the process at a certain point all densities $f_{i,s}(y;\psi_s)$ are equivalent to 0 (or the log-densities equal to $-\infty$), this forecaster is assigned to the different types using just the unconditional probabilities P_s, and then the estimation process is continued.

B Tables

	Subscript	Variable	Description
y		logMedAbsPercErr	The natural logarithm of the median of the absolute percentage error per product
X	0	Intercept	
	1	logMedAbsPercAdj	The natural logarithm of the median of the absolute percentage adjustment + 0.01
	2	medPercAdj	The median of the percentage adjustment
Z	0	Intercept	
	1	logNrProd	The natural logarithm of the amount of products the forecaster has been assigned to
	2	logNrProdCat	The natural logarithm of the number of products in the same category as the respective product and also assigned to the same forecaster
	3	corrErrModel	The autocorrelation in the errors of the FSS forecasts

Table 1: The variables in our model.

		Intercept	logMedAbsPercAdj	medPercAdj
Type 1	Intercept	**1.881** (0.165)	**1.108** (0.070)	**-0.145** (0.050)
	logNrProd	-0.083 (0.054)	**-0.179** (0.024)	**0.054** (0.016)
	logNrProdCat	-0.001 (0.054)	**0.087** (0.025)	-0.015 (0.009)
	corrErrModel	-0.178 (0.110)	**0.173** (0.050)	-0.048 (0.026)
Type 2	Intercept	**0.877** (0.133)	**0.504** (0.059)	**0.049** (0.024)
	logNrProd	**-0.097** (0.046)	-0.022 (0.021)	0.003 (0.009)
	logNrProdCat	**0.172** (0.049)	0.004 (0.022)	-0.003 (0.010)
	corrErrModel	0.178 (0.102)	0.082 (0.045)	0.034 (0.026)

Table 2: The estimates of ψ_s for both types. Standard errors in parentheses. Boldface printed estimates are significant at a 5 % significance level.

	Intercept	logMedAbsPercAdj	medPercAdj
Intercept	**1.218** (0.104)	**0.727** (0.045)	0.007 (0.021)
logNrProd	-0.051 (0.036)	**-0.043** (0.016)	-0.001 (0.007)
logNrProdCat	**0.091** (0.037)	-0.005 (0.016)	0.012 (0.006)
corrErrModel	0.108 (0.077)	**0.139** (0.034)	**-0.066** (0.017)

Table 3: The estimates of ψ if assuming that there is only one type (homogeneity). Standard errors in parentheses. Boldface printed estimates are significant at a 5 % significance level.

	Type 1	Type 2
P_s, unconditional probability	0.312	0.688
Average number of products	52.23	29.95
(SE of above)	(14.27)	(5.44)
Average of logMedAbsPercErr	1.162	0.736
(SE of above)	(0.054)	(0.040)
Average of logMedAbsPercErr for FSS	0.969	0.630
(SE of above)	(0.049)	(0.037)
Average absolute adjustment	22.50	31.74
(SE of above)	(1.088)	(4.462)
Percentage upward adjustments	63.3	61.4
(SE of above)	(0.003)	(0.002)

Table 4: Several characteristics of both forecasters types.

Variable	Change	Type 1 Outcome	Type 1 Percentage effect	Type 2 Outcome	Type 2 Percentage effect
logMedAbsPercAdj	0.1	1.134	6.5 %	0.767	4.5 %
medPercAdj	1	1.100	3.1 %	0.787	6.5 %
logNrProd	0.1	1.075	0.6 %	0.714	-0.7 %
logNrProdCat	0.1	1.063	-0.6 %	0.738	1.7 %
corrErrModel	0.1	1.038	-3.1 %	0.735	1.4 %
Average forecaster		1.069		0.721	

Table 5: The effect of marginal changes in some characteristics on logMedAbsPercErr for the average forecaster of both types.

C Figures

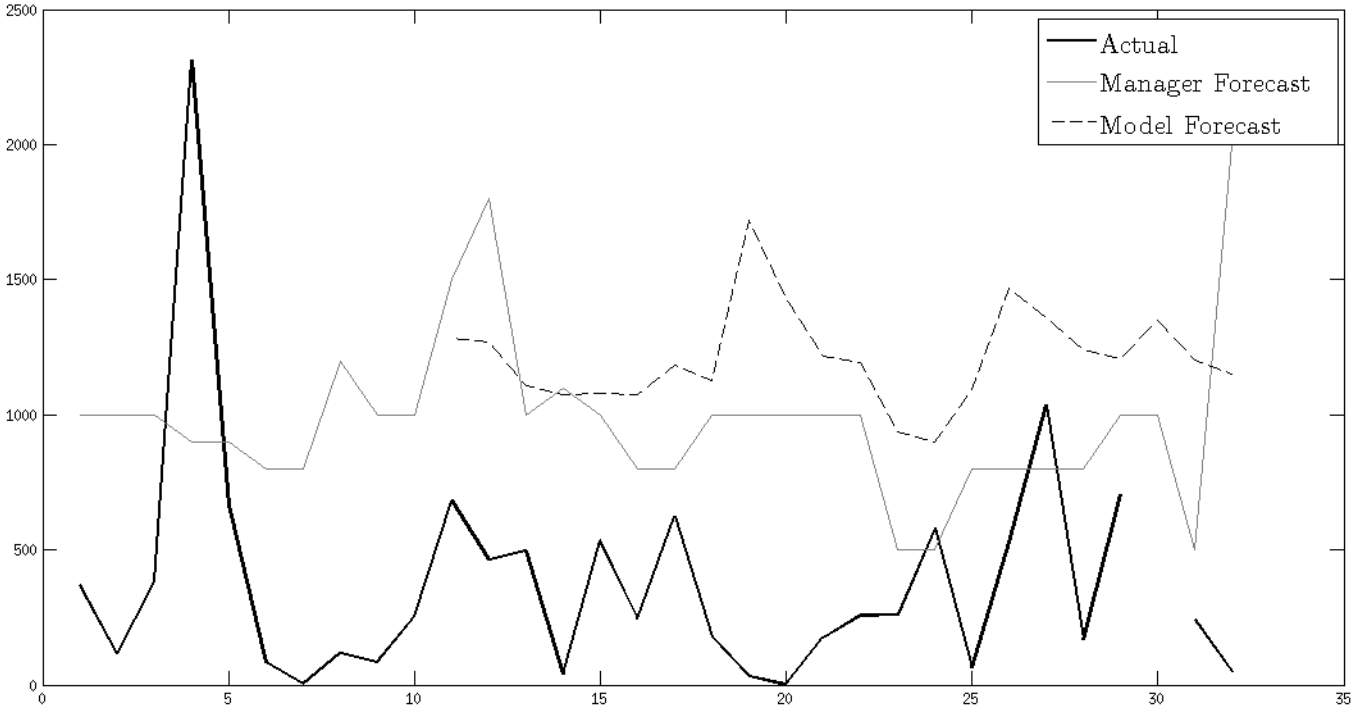

Figure 1: Illustrative sample paths of the actual, the manager forecast and the FSS forecast for one product. This path shows the manager forecast improves upon the FSS forecast.

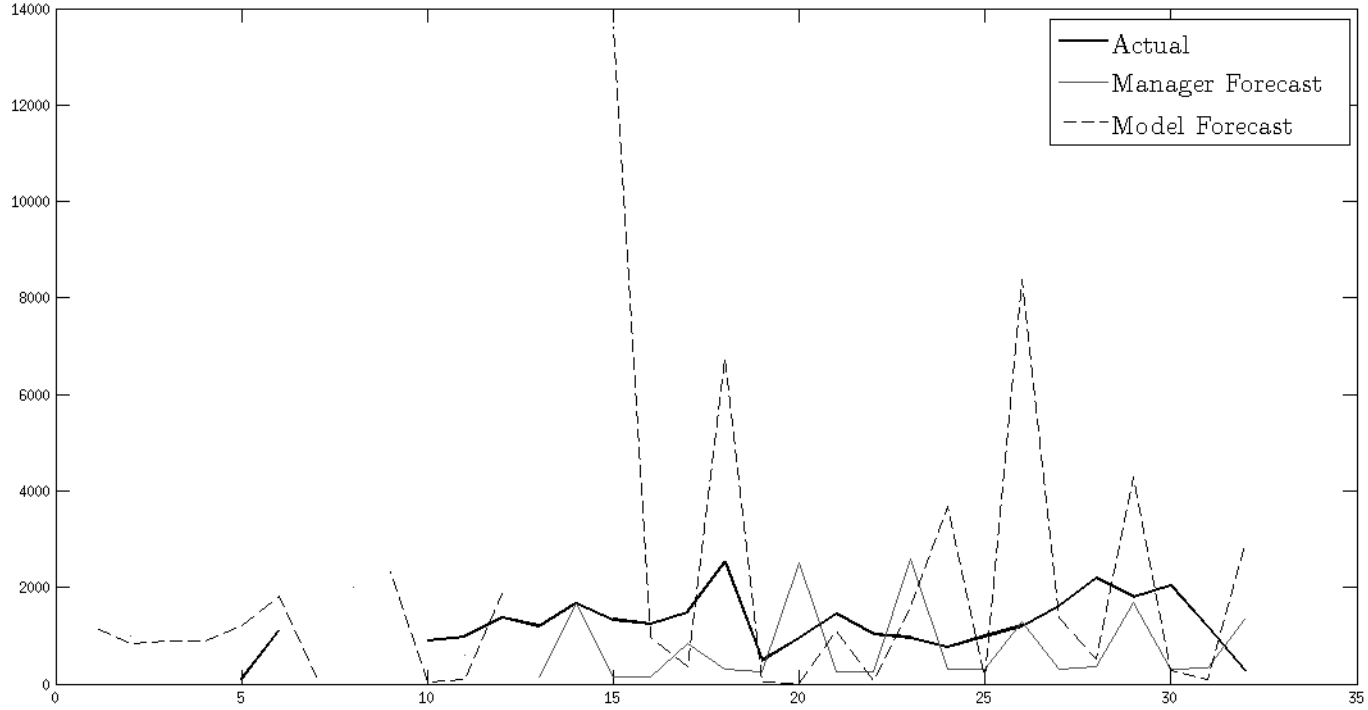

Figure 2: Another illustrative set of data of the actual, the manager forecast and the FSS forecast for one product. These paths show that there can be some missings in either the forecasts or the actuals.

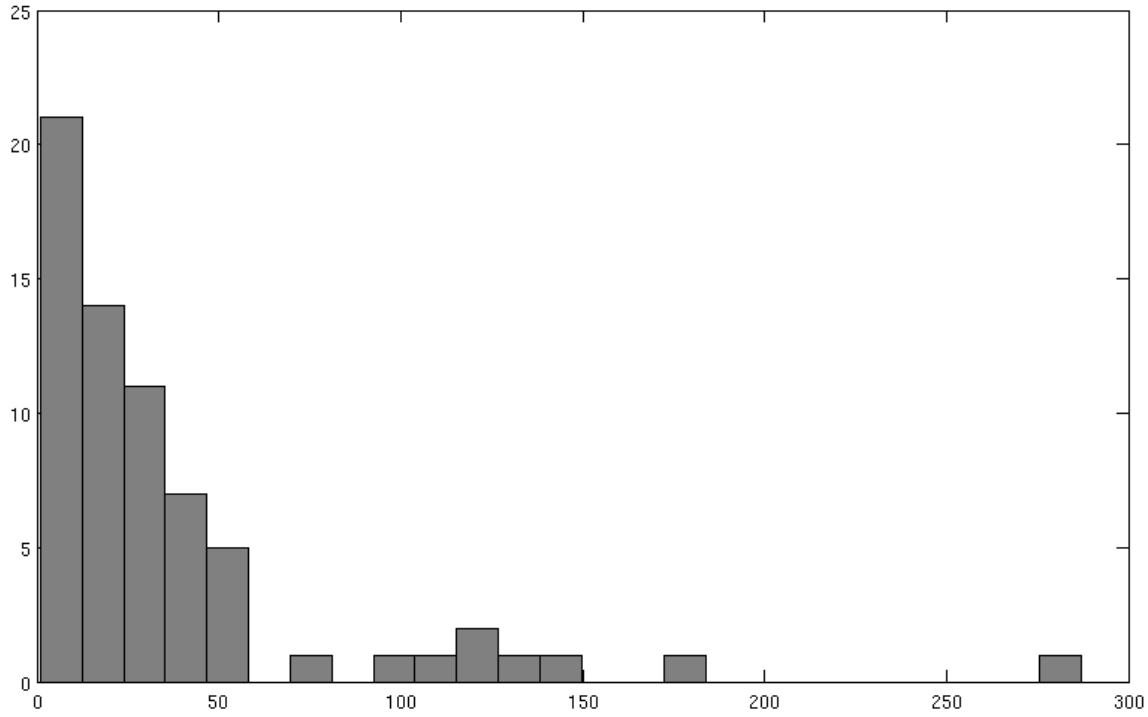

Figure 3: The distribution of the number of products per forecaster. The y-axis gives the frequency and the x-axis gives the number of products.

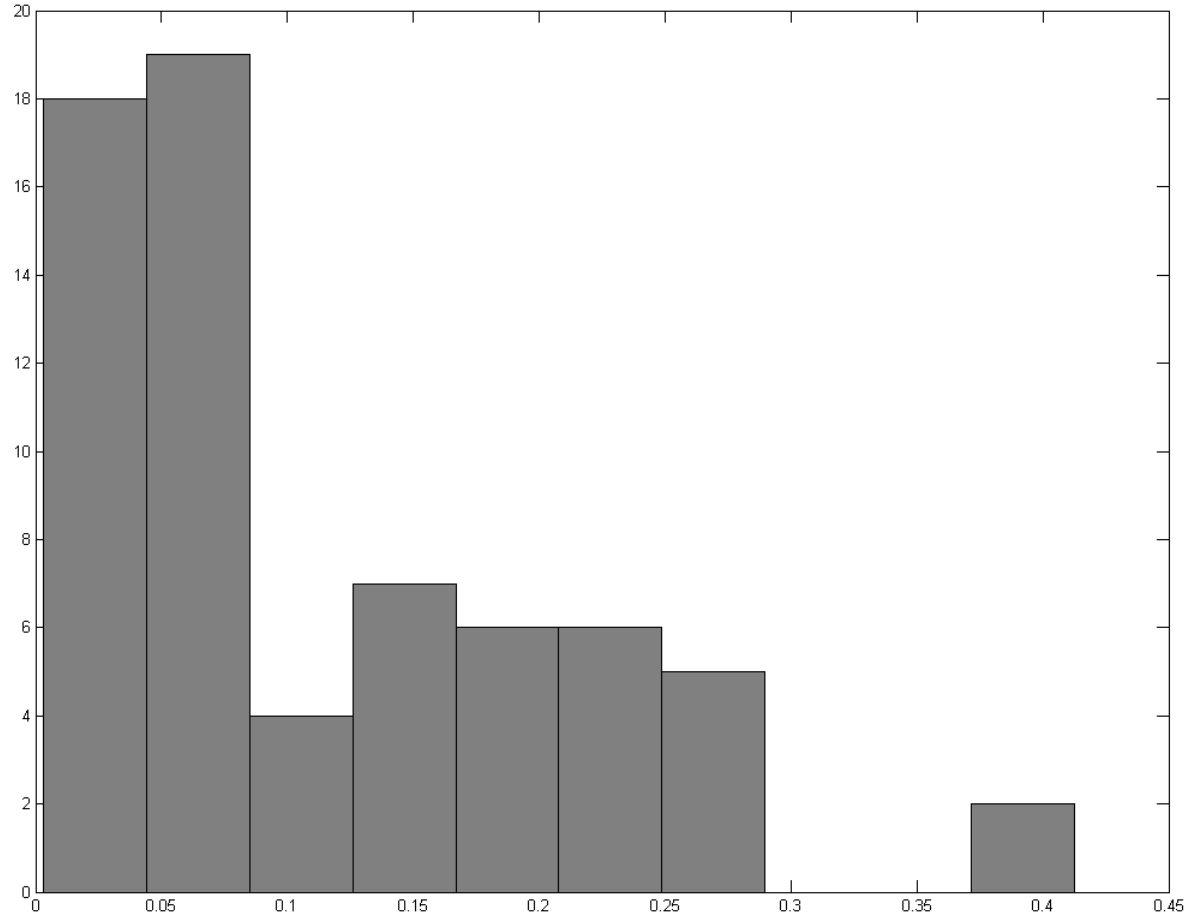

Figure 4: The distribution of the valid forecasts per forecaster, that is, the forecasts that remain after filtering. On the y-axis we give the frequency and on the x-axis we give the proportion of products that remain.

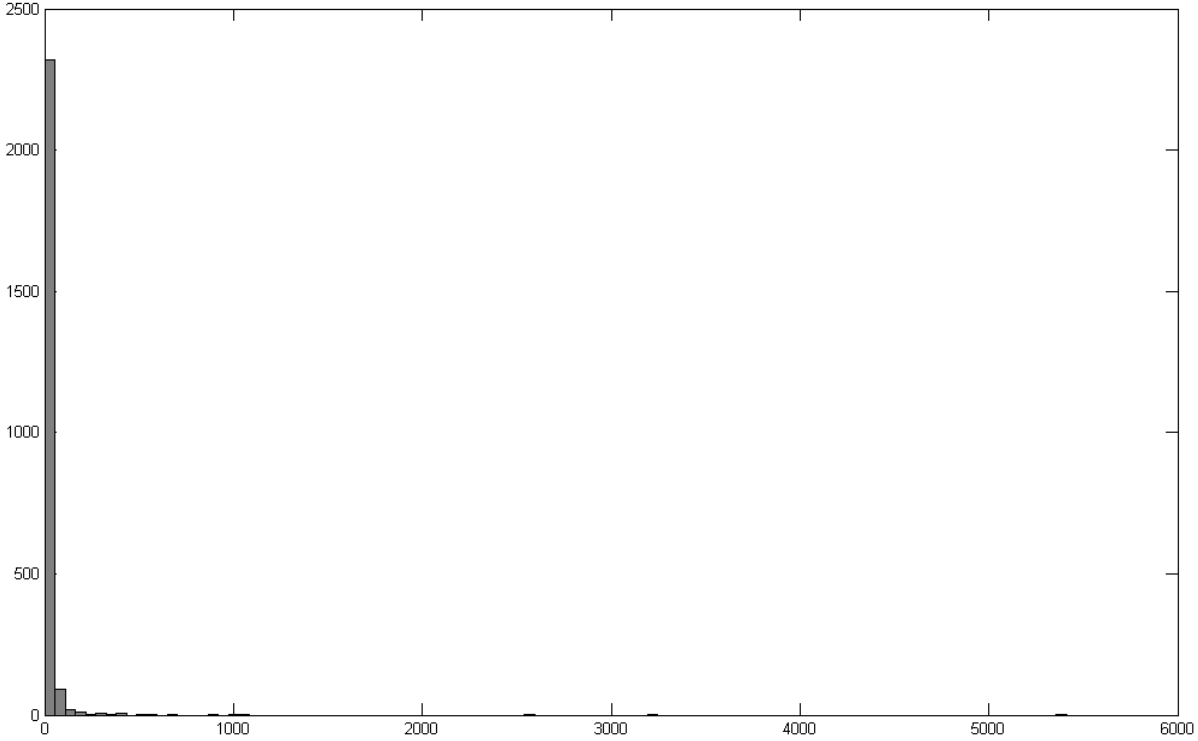

Figure 5: The distribution of the median absolute percentage error.

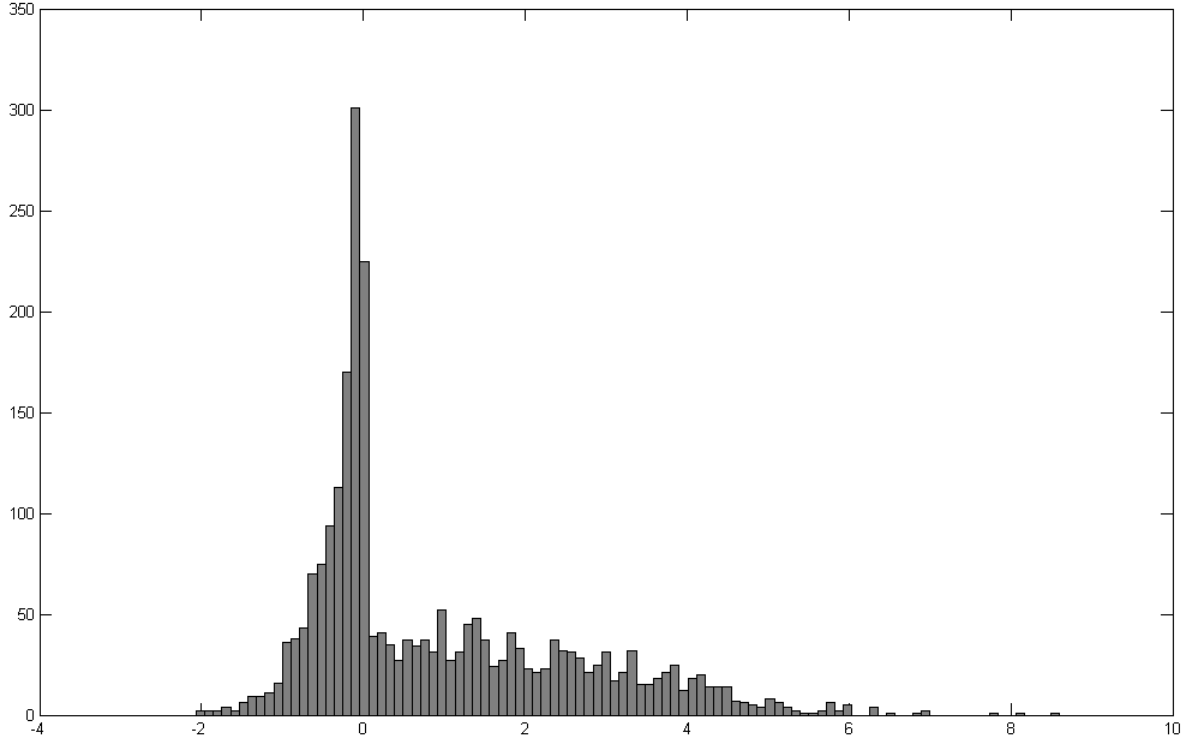

Figure 6: The distribution of the logarithm of the median absolute percentage error.

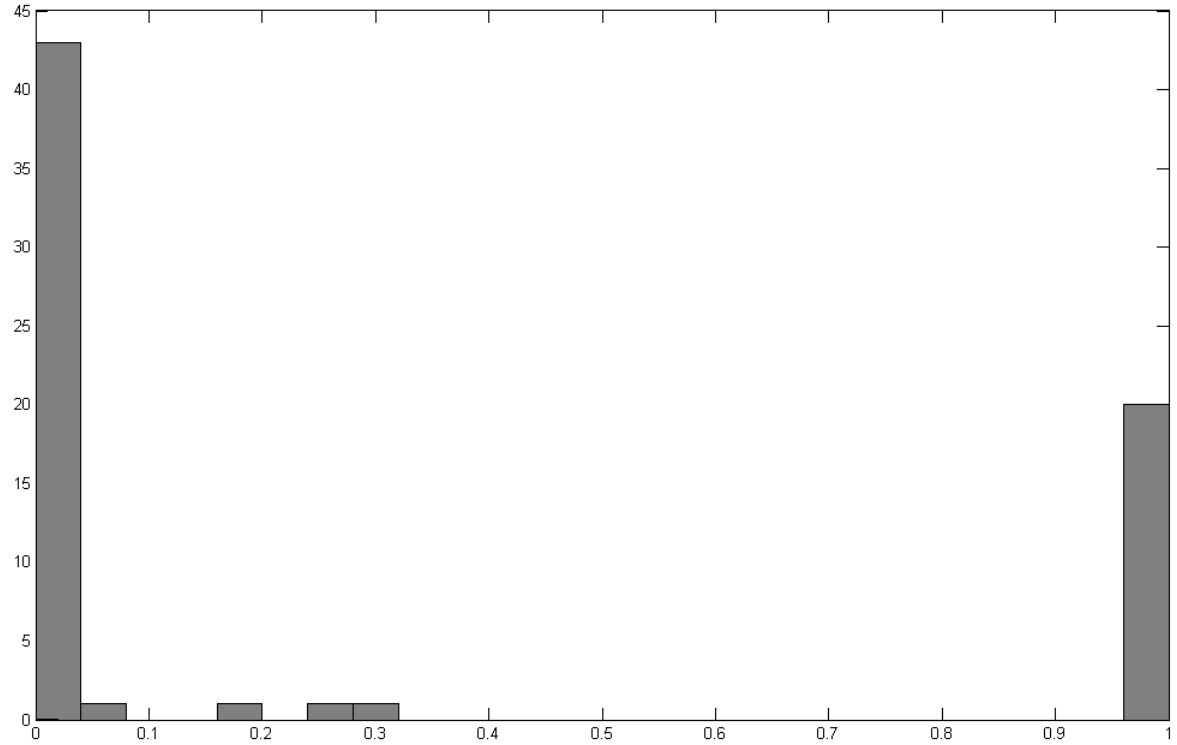

Figure 7: The distribution of the estimated probabilities of being a member of type 1 forecasters.

References

Andrews, R. L. and Currim, I. S. (2003). A comparison of segment retention criteria for finite mixture logit models. *Journal of Marketing Research*, 40(2):235–243.

Armstrong, J. and Collopy, F. (1998). Integration of statistical methods and judgement for time series forecasting: Principles for empirical research. In Wright, G. and Goodwin, P., editors, *Forecasting with Judgement*. New York: Wiley.

Boulaksil, Y. and Franses, P. H. (2009). Experts' stated behavior. *Interfaces*, 39(2):168–171.

Bunn, D. W. and Salo, A. A. (1996). Adjustment of forecasts with model consistent expectations. *International Journal of Forecasting*, 12(1):163–170.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 39(1):1–38.

Diamantopoulos, A. and Mathews, B. (1989). Factors affecting the nature and effectiveness of subjective revision in sales forecasting: An empirical study. *Managerial and Decision Economics*, 10(1):51–59.

Fildes, R. and Goodwin, P. . (2007). Against your better judgment? How organizations can improve their use of management judgment in forecasting. *Interfaces*, 37(6):570–576,596.

Fildes, R., Goodwin, P., Lawrence, M., and Nikolopoulos, K. (2009). Effective forecasting and judgmental adjustments: an empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting*, 25(1):3–23.

Franses, P. H. and Legerstee, R. (2009). Properties of expert adjustments on model-based SKU-level forecasts. *International Journal of Forecasting*, 25(1):35–47.

Franses, P. H. and Legerstee, R. (2010). Do experts' adjustments on model-based SKU-level forecasts improve forecast quality? *Journal of Forecasting*, 29(3):331–340.

Franses, P. H. and Legerstee, R. (2013). Do statistical forecasting models for SKU-level data benefit from including past expert knowledge? *International Journal of Forecasting*, 29(1):80–87.

Goodwin, P. (2000). Improving the voluntary integration of statistical forecasts and judgment. *International Journal of Forecasting*, 16(1):85–99.

Goodwin, P. (2002). Integrating management judgment and statistical methods to improve short-term forecasts. *Omega*, 30(2):127–135.

Lawrence, M., Goodwin, P., O'Connor, M., and Onkal, D. (2006). Judgmental forecasting: A review of progress over the last 25 years. *International Journal of Forecasting*, 22(3):493–518.

Mathews, B. P. and Diamantopoulos, A. (1986). Managerial intervention in forecasting. An empirical investigation of forecast manipulation. *International Journal of Research in Marketing*, 3(1):3–10.

Nikolopoulos, K., Lawrence, M., Goodwin, P., and Fildes, R. A. (2005). On the accuracy of judgmental interventions on forecasting support systems. *Working Paper. The Department of Management Science, Lancaster University.*

Sanders, N. and Ritzman, L. (2001). Judgemental adjustments of statistical forecasts. In Armstrong, J., editor, *Principles of Forecasting*. New York: Kluwer.

Sanders, N. R. and Manrodt, K. B. (1994). Forecasting practices in US corporations: Survey results. *Interfaces*, 24(2):92–100.

Syntetos, A. A., Nikolopoulos, K., Boylan, J. E., Fildes, R., and Goodwin, P. (2009).

The effects of integrating management judgement into intermittent demand forecasts. *International Journal of Production Economics*, 118(1):72–81.