International Journal of Production Research, 2016
Vol. 54, No. 17, 5271–5282, http://dx.doi.org/10.1080/00207543.2016.1167983

Taylor & Francis
Taylor & Francis Group

To select or to combine? The inventory performance of model and expert forecasts

Xun Wang and Fotios Petropoulos*

Logistics & Operations Management Section, Cardiff Business School, Cardiff University, Cardiff, UK

(*Received 20 October 2015; accepted 10 March 2016*)

Demand forecasting is a crucial input of any inventory system. The quality of the forecasts should be evaluated not only in terms of forecast accuracy or bias but also with regards to their inventory implications, which include the impact on the total inventory cost, the achieved service levels and the variance of orders and inventory. Forecast selection and combination are two very widely applied forecasting strategies that have shown repeatedly to increase the forecasting performance. However, the inventory performance of these strategies remains unexplored. We empirically examine the effects of forecast selection and combination on inventory when two sources of forecasts are available. We employ a large data-set that contains demands and (statistical and judgmental) forecasts for multiple pharmaceutical stock keeping units. We show that forecast selection and simple combination increase simultaneously the forecasting and inventory performance.

Keywords: forecasting; judgement; inventory; bullwhip effect; selection; combination

1. Introduction

Demand forecasting is of critical importance for effective operations and planning. Forecast information as the input of any inventory system should not only be evaluated in terms of their accuracy but also for the inventory control implications. Forecast selection and combination, which have been long established as more accurate and beneficial techniques in the forecasting literature compared to a single forecasting method, are yet to be evaluated from an inventory perspective.

In the field of inventory management and production smoothing, there has been ample discussion on the impact of forecasting methods on inventory-related costs. This research often assumes a known demand type and focus on a specific forecasting method, e.g. moving average (Chen et al. 2000), simple exponential smoothing (Dejonckheere et al. 2002), trended and seasonal exponential smoothing (Wright and Yuan 2008) and damped trend forecasting (Li, Disney, and Gaalman 2014). Figure 1(a) presents a simple flowchart of the case where a set of model forecasts, together with the actual demand, is the input for the inventory system. While the forecasting performance of the model forecasts can be directly measured by comparing them to actual demands, the inventory performance can be measured in terms of orders variance (bullwhip effect), inventory variance, inventory-related cost and service levels.

In many cases, model forecasts are judgmentally revised before they are fed into the inventory system (Figure 1(b)). Previous studies have compared the performance of model vs. expert forecasts on both accuracy and inventory implications (Syntetos et al. 2009; Syntetos, Nikolopoulos, and Boylan 2010). They used empirical demand and forecast data as input to an inventory system model and observed via simulation the performance of the inventory system in terms of stock levels and service levels. However, an important characteristic of the inventory system is the cost brought by changes in ordering and production, which has been overlooked.

Sometimes, more than one set of forecasts are available. Usually, these are the outputs of different statistical methods. However, it could also be the model forecasts and their judgmentally revised counterparts (Figure 1(c)) when the latter are not considered as the direct input of the inventory system (i.e. the final forecast). In the cases where multiple sets of forecasts are available, selection of the most suitable set of forecasts or combination of them are considered as good practices in the forecasting literature.

Automatic selection between forecasts typically assesses the forecasting performance of a set of potential estimators on the in-sample or a hold-out set of data (Fildes and Petropoulos 2015). However, concerning the relationship between forecast performance and inventory performance, it has been discovered that accurate forecast does not always result in high efficiency of the inventory system, hence a cost metric rather than an error metric should be prioritised when selecting the forecasting strategy (Flores, Olson, and Pearce 1993). This fact calls a need to explore how inventory-focused selection rules perform in practice.

*Corresponding author. Email: PetropoulosF@cardiff.ac.uk

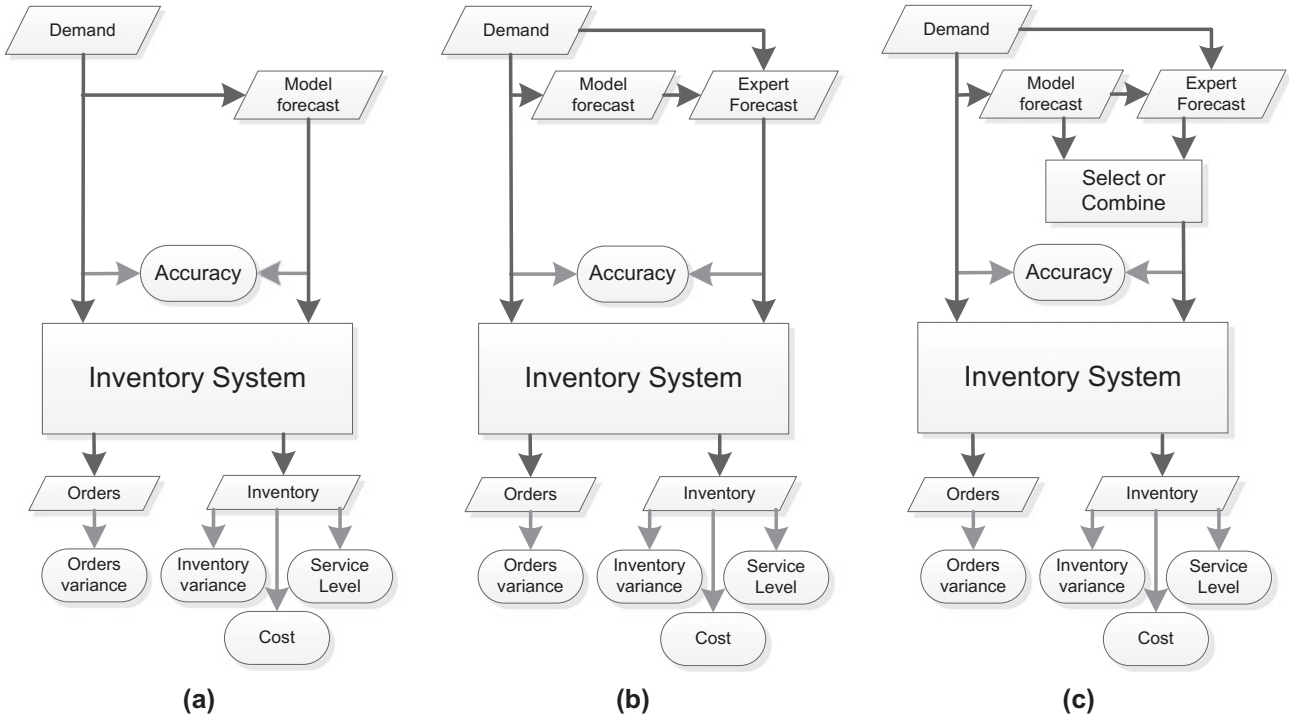

Figure 1. The forecasts fed into the inventory system could be: (a) model forecasts, (b) judgmentally revised (expert) forecasts, or (c) forecasts derived from selection or combination.

Forecast combinations have long been considered beneficial for improving the forecasting performance, with simple equally weighted combinations being proved to work equally well compared to more sophisticated combination schemes (Timmermann 2006; Franses and Legerstee 2011; Genre et al. 2013). Nevertheless, to the best of our knowledge, the performance of forecast combinations has not been previously evaluated in terms of their inventory implications.

The current work focuses on shedding more light on the under-explored links between forecasting and inventory performance. We examine for the very first time the inventory implications of forecast selection and combination. To achieve this, we use a large data-set from a pharmaceutical company which contains model and expert forecasts along with the actual values. This data-set has been analysed before from a forecasting point of view (e.g. see: Franses and Legerstee 2009, 2010, 2013; Petropoulos, Fildes, and Goodwin 2016), neglecting the comparison of model vs. expert forecasts and their combination on an inventory level.

The remaining of the paper is organised as follows. Section 2 provides the theoretical formulations and assumptions of the inventory system model. Section 3 presents the empirical data to be used and elaborates on the experimental design. Section 4 provides the empirically obtained results and discussion. Section 5 concludes.

2. Theoretical formulation

Firstly, the notations will be introduced. We have d for demand, f for forecast, o for order, i for inventory on-hand and w for the work-in-process (WIP). IP is the inventory position which is the sum of net inventory (inventory on-hand minus backorders) plus WIP. S is the order-up-to level which is variable over time. t is the index for discrete time periods which are positive integers. Specifically, $f_{t,t+1}$ denotes the demand forecast made at period t for period $t+1$. Moreover, $E(x)$ is the mean or expectation of a random variable x; Σ_{xy} is the sample covariance between x and y; Σ_{xx} is the sample variance of x; and $\sigma_x = \Sigma_{xx}^{1/2}$.

2.1 *The inventory system*

We assume the sequence of events in the inventory system as follows: in each period, the firm receives the products it ordered l periods ago, then actual demand is observed and inventory is consumed. The forecast accuracy in the last period can then

be evaluated. Based on the information (demand, forecasts and forecast errors) the forecast for the demand in the next period is then produced (we will introduce how forecasts are generated in Section 3). We further assume an order-up-to policy is adopted in the inventory system, where the order quantity is decided in the following manner:

$$o_t = S_t - IP_t$$

In other words, the ordering quantity in period t equals to order-up-to level at t minus the inventory position at time t. This ordering policy has been adopted in a lot of inventory control literature such as Lee, Padmanabhan, and Whang (1997). S_t can be calculated as

$$S_t = \sum_{k=1}^{l} f_{t,t+k} + ss_t$$

where the first term represents the forecast for lead-time demand, and ss_t is the safety stock, calculated as $ss_t = \Phi^{-1}(\alpha^*)\sigma_e$, where α^* is the target service level, $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution function of the standard normal distribution and σ_e is the standard deviation of forecast errors. In the long term, ss_t can be seen as a constant under the assumption of stationary demand input. The inventory position is updated in each period as follows:

$$IP_t = IP_{t-1} + o_{t-1} - d_t$$

This is simply a balance equation of the level of inventory position with flow-in (the orders) and flow-out (the demand). The inventory position can be further broken down to inventory on-hand and WIP:

$$i_t = i_{t-1} + o_{t-l} - d_t$$
$$w_t = w_{t-1} + o_{t-1} - o_{t-l}$$

where l is defined as the replenishment lead-time, which is the time delay between issuing an order and receiving the products. When we assume that $l = 1$ and the safety stock is constant, the system model can be further simplified:

$$o_t = f_{t,t+1} + ss - i_t \tag{1}$$

and

$$i_t = i_{t-1} + o_{t-1} - d_t \tag{2}$$

The noticeable differences brought by the $l = 1$ assumption are that (a) we only need one-step-ahead forecasts and (b) there is no WIP when $l = 1$. In the following analysis, we will keep this assumption due to the limited data points available to us for each SKU. In practice, this means the intervals between order issuance and shipment delay are kept the same.

2.2 *Performance measures*

We measure both the accuracy of the forecasts and the performance of the inventory system. For the inventory system, we investigate both the first- and second-order performance measures. Firstly, we are interested in the second-order moments of the inventory system variables, i.e. variances of order quantity and inventory level over time. The variance of order quantity can also be understood as the bullwhip effect (Chen and Disney 2007; Potter et al. 2009). The rationale behind this measure is that both order and inventory fluctuations are costly in supply chain operations. Unnecessary inventory fluctuation induces both inventory holding and stockout cost; whereas fluctuation in orders increases the demand variability that the suppliers are facing and subsequently brings inefficiency in supply chains. For manufacturers in the supply chain, fluctuating production is detrimental to the production system which leads to repeated idling and overtime of machines and staff.

The first-order measure, i.e. the average cost linearly associated with system variables, is more significant and relevant in a practical way. The average total cost TC in T periods consists of two parts, holding cost and backlog cost, calculated as

$$TC = hE(i^+) + bE(i^-) = \frac{h}{T}\sum_{t=1}^{T} i_t^+ + \frac{b}{T}\sum_{t=1}^{T} i_t^- \tag{3}$$

where h and b are unit-holding and backlog costs and $i^+ = \max(i, 0)$ and $i^- = \max(-i, 0)$. So i_t^+ represents the positive inventory on-hand and i_t^- the backlogs. The assumption behind this cost measure is that the cost grows proportionately with system variables at a rate which is defined as the unit cost per item. h and b also define the target service-level α^* in the Newsvendor setting such that $\alpha^* = b/(h + b)$. TC can be reduced by bringing down inventory variance – when inventory

variance decreases, both average inventory holding and average stockout decreases as well. Low inventory variance also leads to lower safety stock requirement under a given service level, which is defined as the probability of demand satisfaction.

Forecasting accuracy has long deemed to be a positive contributor to the inventory system performance. However, this has been challenged in recent years by several researchers, either from an empirical perspective (Flores, Olson, and Pearce 1993; Syntetos, Nikolopoulos, and Boylan 2010) or from a mathematical perspective (Gaalman 2006), especially for the cost relevant to fluctuation in orders and production. This leads to the dilemma between forecasting accuracy and inventory system cost.

From the above, we conclude that the variances of the variables, especially orders and inventory, are central in the cost analysis. Since the aim of this paper was to empirically verify the impact of forecasting on inventory performance, in the next subsection we shall establish some results on forecast and inventory performance.

2.3 *Theoretical results on inventory performance*

Define e_t as the forecast error at time t, i.e. $e_t = d_t - f_{t-1,t}$, for the inventory variance, we have the following result:

PROPOSITION 1 *For the inventory system stated above, the following relationship holds:*

$$\Sigma_{ii} = \Sigma_{ee} \tag{4}$$

Proof See Vassian (1955). □

This result, that the variance of inventory equals to the variance of forecast errors, is quite well known. Consequently, as long as the order-up-to policy is in use, one can simply use the forecast errors instead of the inventory to determine the safety stock.

For the order variance, we have the following:

PROPOSITION 2 *For the inventory system stated above, the following relationship holds:*

$$\Sigma_{oo} = \Sigma_{ff} + \Sigma_{ee} + 2\Sigma_{fe} \tag{5}$$

Proof Substitute (2) into (1), we have $o_t = f_{t,t+1} + ss - o_{t-1} - i_{t-1} + d_t$. Also, lag (1) for one period we have $o_{t-1} = f_{t-1,t} + ss - i_{t-1}$. Therefore, $o_t = f_{t,t+1} - f_{t-1,t} + d_t = f_{t,t+1} + e_t$. Calculating the variance on both sides gives us (5). □

The terms in (4) and (5) are all measures for the forecast. They can be interpreted in the following way. Firstly, if forecasts are unbiased, then smaller Σ_{ee} indicates higher smoothness of the forecast errors. Note that Σ_{ee} is asymptotically equal to the mean squared error (MSE) as the sample size grows. Secondly, Σ_{ff} is the variance of the forecasts which measures how volatile the forecast sequence is. Lastly, Σ_{fe} gives the estimate of the relationship between the observed forecast error and subsequent forecast made in one period. It estimates the response of a certain forecasting method under the stimulation of an error.

The above analysis shows that the inventory variance is solely determined by the variance of forecast errors; however, the order variance is also influenced by the forecast variability and error-forecast response. Therefore, the forecasts with the smallest MSE do not always lead to the best performance of the inventory system, in terms of the smoothness of order sequence. We should note that in the analysis presented in this subsection, there is no need to specify the pattern of the demand process, except the assumption of stationarity. However, the interpretation of Σ_{fe} only makes sense when $l = 1$, as for when $l > 1$, f_t becomes the forecast for the demand over the lead-time and is not comparable with d_t, the demand in a single period.

3. Experimental design

3.1 *Data*

The data-set contains 1101 sets of time series, each including one series of real (demand) data, one forecast series generated by statistical forecasting models and one forecast series produced by experts after judgmentally revising the model forecasts. Both real data series and forecast series have 25 data points. Model and expert forecasts refer to one-step-ahead predictions.

The model forecasts are the output of a commercial statistical forecasting software which takes into account the statistical forecasts of several widely used methods (e.g. Holt-Winters and Box-Jenkins) and optimally selects the most appropriate for each period and each SKU.

Some of the time series contain invalid data points, while some others are non-stationary. We have run stationarity tests (ADF and KPSS) on all the 1101 sets of data, in which 580 series have passed at least one test, and all the data are valid.

Two further series have been excluded from our analysis as it was found that the model forecast contained several outliers. In these outliers, the model forecast shows severe and unexplainable fluctuations in the presence of relatively smooth real demand data. So, the empirical results presented in Section 4 are based on 578 time series.

In order to be able to directly compare and summarise the variances of the orders and the inventory, we standardise the data. Let f_m and f_e be the model forecast and expert forecast, respectively. We make the following transformations:

$$d = \frac{d - E(d)}{\sigma_d}$$
$$f_m = \frac{f_m - E(f_m)}{\sigma_d}$$
$$f_e = \frac{f_e - E(f_e)}{\sigma_d}$$

All the series will therefore have means of zero; the standard deviation of demand is one; and the comparative relationship in the variance and covariance will be preserved.

3.2 *Candidate strategies*

Model selection for forecasting is a research area that has attracted much attention over the last 40 years (Petropoulos et al. 2014). Selection between forecasts produced from different models may be applied in either an individual (for each SKU) or aggregate (cluster of SKUs) fashion.

Aggregate selection is the strategy of choosing a single source of forecasts. Such sources could include specific statistical forecast models, forecasts judgmentally produced by experts or even judgmentally adjusted model forecasts. Then, forecasts for all time series are produced using the uniquely selected forecast generation process. Aggregate selection is very easy to apply and could be of value when dealing with homogeneous time series (Fildes 1989). However, the global application of a specific source of forecasts does not take into account the individual characteristics of each time series; nor can it distinguish the cases where a particular model fails.

The simplest implementation of aggregate selection for the data in hand would be the universal (across all series and without updating over origins) application of either the model or the expert forecasts. So, the first two strategies are as follows:

S1 Unconditionally use the model forecasts (MF).
S2 Unconditionally use the expert forecasts (EF).

Individual selection refers to separately (per series) selecting an 'optimal' source of forecast. This strategy has the obvious advantage of taking into account specific time series features (such as trend, seasonality and volatility) appearing in each series individually. The performance of individual selection generally outperforms that of aggregate selection (Fildes and Petropoulos 2015) with the expense of the additional complexity and computational cost.

Different criteria for selecting the 'optimal' per-series forecasts have been investigated in the literature. Selection on information criteria (such as Akakie's Information Criterion, AIC, or Bayesian information criterion, BIC) is considered one of the main approaches (Hyndman et al. 2002; Billah et al. 2006; Hyndman and Khandakar 2008). Information criteria select the model with the best fit, after penalising based on the number of parameters (model complexity). Selecting based on minimising AIC is asymptotically equivalent to selecting based on minimising the one-step-ahead forecast MSE. However, this can be regarded as a practical disadvantage as in many real applications forecasts are produced for a horizon greater than one.

Apart from selecting forecasts based on information criteria, other researchers have focused on linking the selection of specific models to the time series features present in each series. For example, Collopy and Armstrong (1992) present a set of 99 rules for selecting and combining between methods based on 18 times series features. Also, Petropoulos et al. (2014) perform a regression analysis to identify the factors affecting forecasting accuracy. Using seven time series features and one decision variable (the forecast horizon), they analyse the effect of each factor on the accuracy of widely used forecasting methods. This analysis is then used for the proposition of a model selection protocol. Focusing on count data, Syntetos, Boylan, and Croston (2005) proposed a framework to select between estimators for intermittent demand by measuring the average inter-demand interval and the squared coefficient of variation of the non-zero demand occurrences.

Another possibility for selecting between forecasts would be to select based on their past forecasting performance, assuming that the top performer in the previous period(s) will be the top performer in the next period(s). The approach of identifying the best candidate based on past performance means that the available in-sample information has to be divided

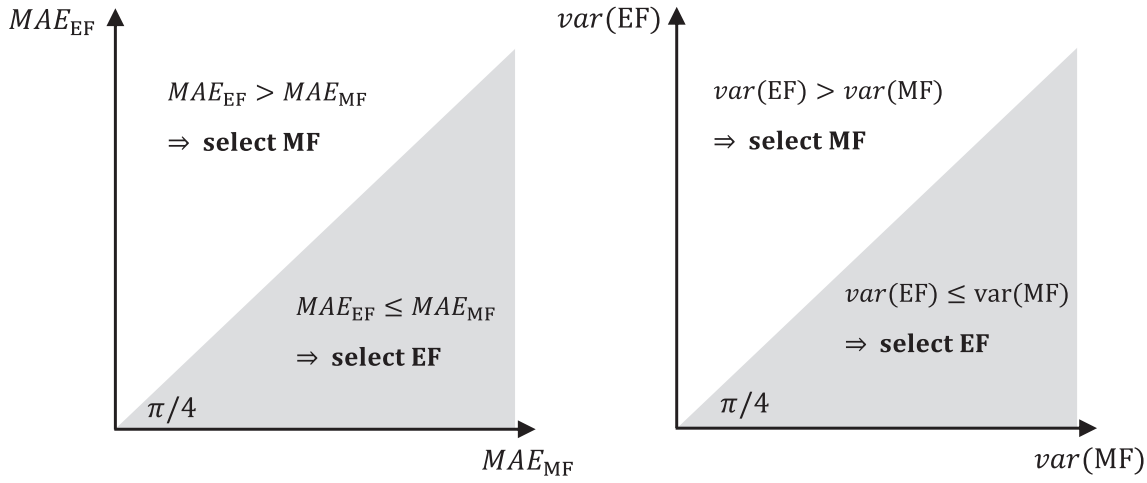

Figure 2. Forecast selection based on MAE (left) and variance (right).

in a training set, which is used to produce the forecasts based on which the selection will be made, and a validation set, where the past performance will be measured. Using a single validation window is also known as selecting based on the performance from a static origin, whereas multiple validation windows (cross-validation) involve the production of forecasts from multiple origins, usually through a rolling-origin process (Tashman 2000).

As more new data become available and are incorporated into the in-sample, forecast selection based on any approach (information criteria, past performance, etc.) may select the same or a different source of forecasts. In this sense, forecast selection can be regarded as dynamic over time: the selected source of forecast at t may be different than the one selected at $t-1$. The dynamic element is expected to be more apparent in an individual selection set-up compared to an aggregate one.

The implementation of a dynamic individual forecast selection for the current research would imply the per series selection of the 'optimal' one-step-ahead forecast, choosing between the model forecast or the expert-adjusted one. After dividing the total sample of triplets of actuals, model and expert forecasts into an initial validation sample (of size t) and a test sample (of size $T-t$), we measure the past performance of the model and the expert forecast over the validation sample and select the most appropriate forecast for period $t+1$. This procedure is repeated for every $t \leq T-1$, as to achieve an one-step-ahead out-of-sample rolling-origin evaluation.

To measure the past forecasting performance, we choose a relatively simple error measure, the mean absolute error (MAE). MAE is suitable for measuring the accuracy of the estimates.[1] However, having as the objective function the minimisation of absolute forecast error (maximisation of the forecast accuracy) is not always linked with maximum inventory performance. Equation (5) suggests that order variance is positively correlated with forecast variance. So, another possibility for forecast selection that would be directly linked with inventory performance would be to select the forecasts with the lower variance. While the selection based on past forecasting performance can occur only for the periods that we have already observed the actual values (i.e. periods 1 up to t), selection based on the variance of the forecasts can also include the point forecasts calculated for period $t+1$.

So, the next two strategies that we consider in this study are:

S3 Dynamic individual forecast selection based on the one-step-ahead past forecasting performance, as measured by the MAE.
S4 Dynamic individual forecast selection based on the variances of the one-step-ahead forecasts.

A graphical representation of S3 and S4 is depicted in Figure 2. Hereafter, these strategies will be referred to as forecast selection on accuracy (FSA) and forecast selection on variance (FSV), respectively.

An alternative to forecast selection, when multiple forecasts are available, is combination. Forecast combinations have been long considered very promising (Makridakis and Winkler 1983; Clemen 1989) and have been shown to improve the accuracy even under the simplest set-ups (Genre et al. 2013). Also, Fildes and Petropoulos (2015) showed that combination may be a better strategy to either aggregate or individual selection when dealing with non-trended series. Additionally, and what is very important for inventory purposes, forecast combinations are expected to lead to forecast errors with lower

variances (Hibon and Evgeniou 2005) meaning lower safety stocks, while a combination of two forecasts will always have variance that is at most equal to the maximum of the variances of either of the forecasts.

Unconditional (50–50%) combination between the database information and the managers expertise has been proposed by Blattberg and Hoch (1990). In the special case of model forecasts and their judgmentally adjusted counterparts, Fildes et al. (2009) and Franses and Legerstee (2011) showed that a simple combination (which is equivalent with damping the judgmental adjustment towards the statistical output) results in superior performance from either model or expert adjusted forecasts. So, the combination strategy considered in the current research is:

S5 Unconditional (50–50%) combination of model and expert (adjusted) forecasts.

All strategies are benchmarked against a 'no change' (Naive method) forecasting approach, where the forecast for period $t+1$ is the actual demand at period t.

3.3 *Experiment procedure*

In this research, the simulation experiment is carried out as follows. For the standardised time series for each SKU that contains 25 data points, we use the first 10 data points as the initial in-sample data which corresponds to the initial validation sample.[2] The initial value of order is set to be the mean demand in the in-sample, and the initial value of inventory is set as the safety stock, plus the mean of expert forecast, minus the mean demand. The reason is that we assume the expert forecast is regarded as the operational forecasts during the initial in-sample, where $t \in [1, 10]$. We also calculate the forecast accuracy and forecast variance, based on which we produce the forecast at the 11th period using FSA or FSV strategies.

The last 15 periods act as the out-sample data which are inputted into the inventory system described by (1) and (2). For FSA and FSV strategies, a dynamic selection process is adopted. That is, to select the 'optimal' forecast for period $t+1$, the forecast accuracy and variability of both model and expert forecasts will be compared and the preferred forecast will be chosen based on the respective strategy. As discussed in the previous subsection, forecast accuracy and variability are calculated over the periods 1 to t and 1 to $t+1$, respectively. Such dynamic selection is not needed for MF, EF and combination strategies. After the forecasts are generated, the output variables, i.e. the orders and inventory, are updated.

Each strategy is then evaluated in terms of inventory performance by calculating order and inventory variance, total (holding and backlog) cost, and achieved service level. The inventory performance is contrasted to the forecasting performance which is calculated by three measures, the mean error ME $= E(e)$ for measuring forecast bias, the MAE $= E(|e|)$ for measuring forecast accuracy and the MSE $= E(e^2)$. The arithmetic mean across the 578 stationary series under investigation is presented for each of the performance measures.

One technical detail needs to be noted here. To make the ordering decision at the 25th period, one needs the demand forecast for the 26th period, which is not available in this data-set. Therefore the order variance is derived from 14 periods' data, from 11th to 24th; whereas the other measures, including inventory variance, inventory-related cost and accuracy, are based on 15 periods' data, from 11th to 25th. This is also a reason why we have to assume $l = 1$, because longer lead-times would reduce the amount of data available in our experiment.

For the measure of cost, we set $h = 1$ and investigate three scenarios for unit backlog cost: $b = 9$, $b = 19$ and $b = 99$. These correspond to optimal target service levels α^* to be 90%, 95% and 99%, respectively. Note that in this experiment the safety stock level is updated every period with σ_e being calculated from the first period up to the current period. Also we argue that in the pharmaceutical supply chain, it is reasonable to assume $b \gg h$ due to the profound (care and cost) implication of drug shortage (Saedi, Kundakcioglu, and Henry 2016).

4. Empirical results

4.1 *Comparison of different forecasting strategies*

Table 1 presents the summarised performance of the different forecasting strategies as discussed in Section 3.2. We also present the performance of a simple benchmark, the Naive method.

The different forecasting strategies are also contrasted for statistically significant differences between them, by performing paired t-tests. Strategies that are significantly different at level 0.05 are marked in Table 2 for order (♠) and inventory (♡) variances, total cost (♣) and forecast accuracy based on MAE (◇). The differences for achieved service level were statistically insignificant in all cases. In all cases, the results for 99% targeted service level have been considered ($b = 99$) in line with the $b \gg h$ assumption.

The value of the ME is positive across all strategies, suggesting negatively biased forecasts (under-forecasting). This is especially true in the cases of Naive method and model forecasts. Franses and Legerstee (2009) found that experts tend to adjust the forecasts more often upwards than downwards. In this paper, we find that the ME of model forecasts is higher than

Table 1. Summary of the results of the different forecasting strategies.

	Metric	Naive	Model forecasts	Expert forecasts	FSA	FSV	Combination
Forecasting performance	ME	0.049	0.022	0.008	0.005	0.014	0.015
	MAE	1.025	0.786	0.788	0.746	0.734	0.725
	MSE	1.845	1.138	1.207	0.976	0.934	0.950
$\alpha^* = 90\%$	Σ_{oo}	4.815	1.594	1.592	1.334	1.171	1.246
$(h = 1, b = 9)$	Σ_{ii}	1.975	1.204	1.301	1.059	1.013	1.030
	TC	2.471	1.997	1.993	1.924	1.920	1.865
	α	0.898	0.906	0.907	0.912	0.911	0.907
$\alpha^* = 95\%$	Σ_{oo}	4.853	1.619	1.613	1.358	1.195	1.265
$(h = 1, b = 19)$	Σ_{ii}	2.018	1.233	1.338	1.084	1.035	1.053
	TC	3.092	2.534	2.531	2.452	2.448	2.375
	α	0.941	0.945	0.944	0.949	0.946	0.945
$\alpha^* = 99\%$	Σ_{oo}	4.944	1.681	1.667	1.415	1.253	1.310
$(h = 1, b = 99)$	Σ_{ii}	2.121	1.304	1.424	1.143	1.089	1.106
	TC	4.999	4.252	4.261	4.154	4.178	4.079
	α	0.977	0.979	0.977	0.979	0.979	0.977

ME: mean error (bias measure); MAE: mean absolute error (accuracy measure); MSE: mean squared error; Σ_{oo}: order variance; Σ_{ii}: inventory variance; TC: total cost; α: achieved service level.

Table 2. Statistically significant differences (99% service level).

	Naive	Model forecasts	Expert forecasts	FSA	FSV	Combination
Naive	–	♠♡♣◇	♠♡♣◇	♠♡♣◇	♠♡♣◇	♠♡♣◇
Model forecasts	♠♡♣◇	–		♠♡◇	♠♡◇	♠♡♣◇
Expert forecasts	♠♡♣◇		–	♠♡◇	♠♡◇	♠♡♣◇
FSA	♠♡♣◇	♠♡◇	♠♡◇	–	♠♡◇	♠◇
FSV	♠♡♣◇	♠♡◇	♠♡◇	♠♡◇	–	
Combination	♠♡♣◇	♠♡♣◇	♠♡♣◇	♠◇		–

♠ denotes statistically significant differences (at 0.05 level) for order variances.
♡ denotes statistically significant differences (at 0.05 level) for inventory variances.
♣ denotes statistically significant differences (at 0.05 level) for total cost.
◇ denotes statistically significant differences (at 0.05 level) for accuracy (MAE).
The differences of achieved service levels were statistically insignificant across the various approaches.

the ME of expert forecasts which explains the upwards adjusting tendency of experts (remember that $e = d - f$). At the same time the ME of expert forecasts is closer to zero, meaning that expert forecasts are less biased than model forecasts. However, the difference is not statistically significant.

Further comparing the performance of the two aggregate approaches (model forecasts or expert forecasts applied unconditionally), we can observe that there are practically no differences in the inventory nor the forecasting performance. Any small differences are statistically insignificant. While the indifferent performance in terms of forecasting accuracy comes to verify previous studies that employed the same data (Franses and Legerstee 2010; Petropoulos, Fildes, and Goodwin 2016), this is the first time that this database is used to evaluate the inventory performance of the judgmentally adjusted model forecasts. When unconditionally applied, expert adjustments on model forecasts do not add value in terms of inventory performance. At the same time, the insignificant differences in forecast accuracy remain insignificant in terms of inventory performance. The comparison of both aggregate strategies with the Naive method reveals statistically significant differences for all metrics considered: both model forecasts and expert forecasts perform much better than a no-change method.

Selection of the most accurate forecast based on the observed past forecasting performance (FSA) generally outperforms aggregate selection of either model or expert forecasts. It reduces order variance and inventory variance on average by 16%, with smaller percentage gains observed for the total (inventory and backlog) cost (up to 3.7% for a targeted customer service level of 90%) and the forecast accuracy (up to 5.3%). It is worth mentioning that the relative decrease in the cost is negatively

Table 3. Statistical properties of various forecasting strategies.

		Naive	Model forecasts	Expert forecasts	FSA	FSV	Combination
Σ_{ff}	Variance of forecasts	0.907	0.539	0.659	0.399	0.305	0.380
Σ_{ee}	Variance of forecast errors	1.964	1.165	1.228	1.008	0.963	0.977
Σ_{fe}	Covariance of forecasts and forecast errors	0.979	−0.065	−0.152	−0.049	−0.065	−0.066
Σ_{oo}	Theoretical order variance	4.829	1.574	1.583	1.309	1.138	1.225
	Empirical order variance	4.944	1.682	1.667	1.415	1.253	1.310
Σ_{ii}	Theoretical inventory variance	1.833	1.165	1.228	1.008	0.963	0.977
	Empirical inventory variance	2.121	1.304	1.424	1.143	1.089	1.106

correlated with α^*, meaning higher gains from FSA for lower customer service levels. Also, although insignificant, the FSA strategy gives the best performance in maintaining the target service level.

Forecast selection based on variances (FSV) performs interestingly well in all measures considered, even outperforming FSA. It is ranked first across the candidate forecasting strategies in terms of order and inventory variance and second in terms of cost, accuracy and achieved service level. Reduction in the bullwhip effect (order variance) is as high as 27% compared to aggregate strategies. Surprisingly, selecting on variances does not only improve the inventory performance compared to FSA, but also further increases the forecast accuracy. In any case, FSV is certainly a very promising strategy for maximising inventory performance, compared to the application of either model or expert forecasts.

Previous studies (Fildes et al. 2009; Franses and Legerstee 2011; Petropoulos, Fildes, and Goodwin 2016) showed that a simple equal-weighted combination strategy between the model and expert forecasts outperforms either of them in forecast accuracy. This result is not just confirmed here but also reinforced, as improvements are significant in terms of inventory performance as well. The 50–50% combination not only gives the best cost performance across the strategies considered in this study but also offers the best accuracy results.

The differences among all strategies in the achieved service levels are not significant at a 0.05 level. However, all strategies can only meet a low target service level and fail on a high target. Generally, the FSA and FSV strategies outperforms the others when considering achieved service levels.

4.2 *Discussion of the results*

The performance of the forecasting strategies on system variances can easily be explained by Propositions 1 and 2. We have broken down the relevant forecast performance in Table 3 and have compared the theoretical and empirical values of variances. The theoretical values are derived from (4) and (5). We can see that the theoretical and empirical results fit very well. The discrepancies are attributed primarily to the variable safety stock and secondarily to the fact that the forecasts in the data-set are not completely unbiased. For brevity we have only included empirical data for $\alpha = 99\%$. It can be easily derived that when target service level decreases, both variance and cost will fall.

Firstly, since the variance of inventory is completely dominated by the variance of forecast errors, we know the forecasting strategy with the least Σ_{ee} should produce the smoothest inventory flow, which is the FSV. The order variance is determined by three factors: the variance of forecast, the variance of the forecast errors and the error-forecast response. Since FSV generates both the smoothest forecasts and the smoothest forecast errors, it is expected that it also leads to the lowest order variance. The combination method on the other hand performs better than FSA in terms of variance of both forecasts and errors; however, neither of them is as good as FSV.

Interestingly, FSV outperforms FSA in terms of both Σ_{ff} and Σ_{ee}; the former is anticipated, since FSV always chooses the forecast that leads to lower forecast variance; whereas the latter is quite counter-intuitive. A possible explanation is as follows: consider that the firm needs to select the forecast at period t. At this instant, the most up-to-date information are the forecasts for $t + 1$ and demand for t. In other words, the variance information is up to $t + 1$ but the error information is only up to t. This means that FSV is always able to utilise one period's extra information than FSA irrespective of the error measure chosen for FSA (MAE or MSE). We believe this is the reason of FSV's better performance.

Looking at the term Σ_{fe}, we see that in most strategies there exists a mild negative relationship between the forecast error and the subsequent forecast (except the Naive method which in turn generates a much greater order variance). Statistics show that such behaviour is much stronger in expert forecasts than in other strategies. This suggests that experts tend to adjust the forecast in the opposite direction to the previous error. This is consistent with Petropoulos, Fildes, and Goodwin (2016). From (5) we see that such negative relationship is beneficial in reducing order variance. However, this does not make

the expert forecasts a more preferable strategy than the model forecasts. The reason is that the adjustments also increase the variance of forecasts and the variance of forecast errors. As such, the benefit brought by such adjustments is not enough to compensate the downsides.

When the inventory distribution is unimodal, the expected holding and backlog costs increase with the inventory variance which increases with the variance of forecast errors. Specifically, when the inventory is normally distributed, the total cost can be calculated as $TC = (h + b)\Phi^{-1}(\alpha)\sigma_i$ (Hadley and Whitin 1963). Therefore, we can establish that lower forecast error variance leads to lower total cost, as can be seen from Tables 1 and 2. This is generally true when we compare the three groups of strategies, i.e. the Naive forecast, the aggregate selection approaches (MF and EF) and model selection/combination approaches (FSA, FSV and combination). Due to the amplification effect of h and b, the differences within the group become less significant.

From a more general perspective, different cost implications of order and inventory variance suggest that supply chain participants should consider the trade-off between order and inventory when selecting the suitable forecasting strategy. In the downstream of the supply chain, the inventory-related cost is usually directly charged to the company. So retailers have a strong incentive to reduce the inventory variance. On the other hand, the danger of high-order variability is always hidden and deemed unimportant. However, volatile orders increase the uncertainty at the supplier and manufacturer's echelons, forcing the supplier to bring up the safety stock and eventually harm the efficiency of the entire supply chain.

By adjusting the weights on inventory and order variances, the firm gradually changes its role in the spectrum of being selfish and altruistic (Hosoda and Disney 2006). Although previous studies have investigated other adjusting methods such as proportional inventory control, the above analysis has shown that it can also be achieved by selecting an appropriate forecasting method. The variance of forecast errors is still dominant in both order and inventory variance, and should be the imperative factor when the objective is to reduce local inventory cost. But if the whole supply chain is taken into account, firms need to carefully choose the forecasting strategy so that variance of forecasts and forecast errors are properly leveraged.

5. Conclusions

In this paper, we have conducted an empirical analysis to investigate the inventory performance of model and expert forecasts. The analysis is carried out based on a widely used inventory system model and fairly general assumptions. A real data-set containing data for actual demands, model and expert forecasts for many pharmaceutical products is used as input to the model. We have also incorporated selection and combination strategies from model and expert forecast data.

When comparing different forecasting methods, we found no significant differences in the performance of model and expert forecasts. This is true both in terms of forecast accuracy – which confirms previous studies – and inventory performance. Forecast selection and combination improve both inventory and forecasting performance compared to either model or expert forecasts. The improvements are also in most cases statistically significant. Interestingly, under the stationarity assumption, forecast selection based on forecast variability generates more accurate forecasts than that based on accuracy.

Forecast selection based on forecast variability results in the lowest order and inventory variances. On the other hand, combination is the best strategy for minimising the total cost and maximising forecast accuracy (measured by MAE). From a managerial viewpoint where simplicity matters, the application of equal-weights combination of the model and expert forecasts brings balanced improvements in terms of minimising the variances and the inventory-associated costs while increasing forecast accuracy compared to aggregate approaches.

We have provided theoretical formulation to breakdown the causes of simulation output. The analysis suggests that apart from the already well-known property that inventory variance is positively affected by the variance of the forecast errors, the same relationship exists between orders and forecast errors. In addition, the variance of the forecasts and the covariance between forecasts and lagged forecast errors also have a positive impact on the order variance. However, in this data-set, the variance of forecast errors and the variance of forecasts are dominant in determining the performance of order smoothing.

Future research should focus on how forecast selection based on forecast variability performs when the stationarity assumption is relaxed. Also, a natural way forward is to test the inventory and forecasting performance of the various strategies when the candidate pool of forecasting strategies is increased, e.g. by investigating different combinations of statistical methods. Lastly, a limitation of this study is the reliance on a single database from the pharmaceutical industry. It would be useful if the results of this study (and other similar studies comparing statistical and judgmental estimates) were replicated using data from a wider range of companies/industries.

Acknowledgements

We would like to thank the Editor, Professor Alexandre Dolgui, the Associate Editor and the two anonymous reviewers for their very constructive comments. We are grateful to Professor Philip Hans Franses for sharing with us this very interesting database. Also, we would like to thank Professor Stephen Disney and Professor Aris Syntetos for their very helpful feedback on earlier versions of this study.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes

1. We have also considered measuring the past forecasting performance using the MSE, a proxy to the variance of forecast errors and we found no noteworthy differences compared to selecting based on MAE.
2. Note that in this research no training sample is required as both model and expert forecasts are provided.

References

Billah, B., M. L. King, R. D. Snyder, and A. B. Koehler. 2006. "Exponential Smoothing Model Selection for Forecasting." *International Journal of Forecasting* 22 (2): 239–247.

Blattberg, R. C., and S. J. Hoch. 1990. "Database Models and Managerial Intuition: 50% model + 50% Manager." *Management Science* 36: 887–899.

Chen, Y. F., and S. M. Disney. 2007. "The Myopic Order-up-to Policy with a Proportional Feedback Controller." *International Journal of Production Research* 45 (2): 351–368.

Chen, F., Z. Dresner, J. K. Ryan, and D. Simchi-Levi. 2000. "Quantifying the Bullwhip Effect in a Simple Supply Chain: The Impact of Forecasting, Lead Times, and Information." *Management Science* 46: 436–443.

Clemen, Robert T. 1989. "Combining Forecasts: A Review and Annotated Bibliography." *International Journal of Forecasting* 5: 559–583.

Collopy, F., and J. S. Armstrong. 1992. "Rule-based Forecasting: Development and Validation of an Expert Systems Approach to Combining Time Series Extrapolations." *Management Science* 38: 1394–1414.

Dejonckheere, J., S. M. Disney, M. R. Lambrecht, and D. R. Towill. 2002. "Transfer Function Analysis of Forecasting Induced Bullwhip in Supply Chains." *International Journal of Production Economics* 78: 133–144.

Fildes, R. 1989. "Evaluation of Aggregate and Individual Forecast Method Selection Rules." *Management Science* 39: 1056–1065.

Fildes, R., P. Goodwin, M. Lawrence, and K. Nikolopoulos. 2009. "Effective Forecasting and Judgmental Adjustments: An Empirical Evaluation and Strategies for Improvement in Supply-chain Planning." *International Journal of Forecasting* 25: 3–23.

Fildes, R., and F. Petropoulos. 2015. "Simple versus Complex Selection Rules for Forecasting Many Time Series." *Journal of Business Research* 68 (8): 1692–1701.

Flores, B., D. Olson, and S. Pearce. 1993. "Use of Cost and Accuracy Measures in Forecasting Method Selection: A Physical Distribution Example." *International Journal of Production Research* 31: 139–160.

Franses, P. H., and R. Legerstee. 2009. "Properties of Expert Adjustments on Model-based SKU-level Forecasts." *International Journal of Forecasting* 25: 35–47.

Franses, P. H., and R. Legerstee. 2010. "Do Experts' Adjustments on Model-based SKU-level Forecasts Improve Forecast Quality?" *Journal of Forecasting* 29: 331–340.

Franses, P. H., and R. Legerstee. 2011. "Combining SKU-level Sales Forecasts from Models and Experts." *Expert Systems with Applications* 38: 2365–2370.

Franses, P. H., and R. Legerstee. 2013. "Do Statistical Forecasting Models for SKU-level Data Benefit from Including Past Expert Knowledge?" *International Journal of Forecasting* 29: 80–87.

Gaalman, G. 2006. "Bullwhip Reduction for ARMA Demand: The Proportional Order-up-to Policy versus the Full-state-feedback Policy." *Automatica* 42: 1283–1290.

Genre, V., G. Kenny, A. Meyler, and A. Timmermann. 2013. "Combining Expert Forecasts: Can Anything Beat the Simple Average?" *International Journal of Forecasting* 29 (1): 108–121.

Hadley, G., and T. Whitin. 1963. *Analysis of Inventory Systems*. Englewood Cliffs, NJ: Prentice Hall.

Hibon, M., and T. Evgeniou. 2005. "To Combine or Not to Combine: Selecting among Forecasts and their Combinations." *International Journal of Forecasting* 21: 15–24.

Hosoda, T., and S. M. Disney. 2006. "The Governing Dynamics of Supply Chains: The Impact of Altruistic Behaviour." *Automatica* 42 (4): 1301–1309.

Hyndman, Rob J., and Yeasmin Khandakar. 2008. "Automatic Time Series Forecasting: The Forecast Package for R." *Journal of Statistical Software* 27 (3): 1–22.

Hyndman, Rob J., Anne B. Koehler, Ralph D. Snyder, and Simone Grose. 2002. "A State Space Framework for Automatic Forecasting using Exponential Smoothing Methods." *International Journal of Forecasting* 18 (3): 439–454.

Lee, H. L., V. Padmanabhan, and S. Whang. 1997. "Information Distortion in a Supply Chain: The Bullwhip Effect." *Management Science* 43: 546–558.

Li, Q., S. M. Disney, and G. Gaalman. 2014. "Avoiding the Bullwhip Effect using Damped Trend Forecasting and the Order-up-to Replenishment Policy." *International Journal of Production Economics* 149: 3–16.

Makridakis, S., and R. Winkler. 1983. "Average of Forecasts: Some Empirical Results." *Management Science* 29: 987–996.

Petropoulos, F., R. Fildes, and P. Goodwin. 2016. "Do 'big losses' in Judgmental Adjustments to Statistical Forecasts Affect Experts' Behaviour?" *European Journal of Operational Research* 249: 842–852.

Petropoulos, F., S. Makridakis, V. Assimakopoulos, and K. Nikolopoulos. 2014. "Horses for Courses' in Demand Forecasting." *European Journal of Operational Research* 237: 152–163.

Potter, A. T., D. R. Towill, T. Böhme, and S. M. Disney. 2009. "The Influence of Multi-product Production Strategy on Factory Induced Bullwhip." *International Journal of Production Research* 47 (20): 5739–5759.

Saedi, S., O. E. Kundakcioglu, and A. C. Henry. 2016. "Mitigating the Impact of Drug Shortages for a Healthcare Facility: An Inventory Management Approach." *European Journal of Operational Research* 251: 107–123.

Syntetos, A. A., J. E. Boylan, and J. D. Croston. 2005. "On the Categorization of Demand Patterns." *Journal of the Operational Research Society* 56 (5): 495–503.

Syntetos, A. A., K. Nikolopoulos, and J. E. Boylan. 2010. "Judging the Judges through Accuracy-implication Metrics: The Case of Inventory Forecasting." *International Journal of Forecasting* 26: 134–143.

Syntetos, A. A., K. Nikolopoulos, J. E. Boylan, R. Fildes, and P. Goodwin. 2009. "The Effects of Integrating Management Judgement into Intermittent Demand Forecasts." *Journal of the Operational Research Society* 60: 611–618.

Tashman, Leonard J. 2000. "Out-of-sample Tests of Forecasting Accuracy: An Analysis and Review." *International Journal of Forecasting* 16 (4): 437–450.

Timmermann, Allan. 2006. "Forecast Combinations." In *Handbook of Economic Forecasting*, Vol. 1, edited by C. W. J. Granger, G. Elliott and A. Timmermann, 35–196. Amsterdam: Elsevier.

Vassian, H. J. 1955. "Application of Discrete Variable Servo Theory to Inventory Control." *Journal of the Operations Research Society of America* 3 (3): 272–282.

Wright, D., and X. Yuan. 2008. "Mitigating the Bullwhip Effect by Ordering Policies and Forecasting Methods." *International Journal of Production Economics* 113 (2): 587–597.