Quantile forecast optimal combination to enhance safety stock estimation

Juan R. Trapero[a,*], Manuel Cardós[b], Nikolaos Kourentzes[c]

[a] *University of Castilla-La Mancha*
Department of Business Administration, Ciudad Real 13071, Spain
[b] *Universitat Politècnica de València*
Department of Business Organization, Camino de Vera s/n, Valencia 46022, Spain
[c] *Lancaster University Management School*
Department of Management Science, Lancaster, LA1 4Y X, UK

Abstract

The safety stock calculation requires a measure of the forecast error uncertainty. Such errors are usually assumed Gaussian iid (independent, identically distributed). However, deviations from iid deteriorate the supply chain performance. Recent research has shown that, alternatively to theoretical approaches, empirical techniques that do not rely on the aforementioned assumptions, can enhance the safety stock calculation. Particularly, GARCH models cope with time-varying heterocedastic forecast error, and Kernel Density Estimation do not need to rely on a determined distribution. However, if forecast errors are both time-varying heterocedastic and do not follow a determined distribution, the previous approaches are inadequate. To overcome this, we propose an optimal combination of the empirical methods that minimizes the asymmetric piecewise linear loss function, also known as tick loss. The results show that combining quantile forecasts yields safety stocks with a lower cost. The methodology is illustrated with simulations and real data experiments for different lead times.

Keywords: Quantile forecasting, Safety stock, Risk, Supply chain, Kernel Density Estimation, GARCH, combination, tick loss.

*Corresponding author.
Email addresses: juanramon.trapero@uclm.es (Juan R. Trapero), mcardos@doe.upv.es (Manuel Cardós), nikolaos@kourentzes.com (Nikolaos Kourentzes)

1. Introduction

Traditionally, the supply chain forecasting literature has mainly focused on point forecasts, with these being a measure of central tendency of the forecast density function, such as the mean or median. Point forecasts are vital for driving production systems based on Material Requirements Planning (Silver et al., 1998). However, little attention has been paid to measure the uncertainty around those forecasts, although important applications such as determining the safety stock and reorder point in many replenishment policies depend on estimating the uncertainty.

Exponential smoothing has been one of the most popular forecasting methods to construct point forecasts for supply chain purposes (Gardner, 2006). Hyndman et al. (2008) embedded exponential smoothing within a statistical framework, based on the State Space approach, capable of yielding the whole forecast distribution for different forecasting horizons. The popularity of exponential smoothing has been due to its relative simplicity and accuracy in practice. However, often, how well it approximates the demand process is not sufficiently tested, mainly due to the limited repertoire of forecasting model alternatives in software.

Ideally, if the demand generating process is correctly identified, the forecasts obtained from that identified model provide errors that are iid (independent, identically distributed) and the statistical forecast distribution is perfectly defined as a function of the forecasting model parameters and the forecasting horizon. This holds regardless of the forecasting model, which could be, for example, State Space Exponential Smoothing (Hyndman et al., 2008), ARIMA (Box et al., 1994) or Unobserved Components (Harvey, 1989).

However, given the complex relationships that drive the demand, it is reasonable to question whether correct identification of the process is possible. Thus, the iid assumptions about the error should be tested. However, typically in supply chain forecasting we assume that these assumptions hold and instead, we focus on comparing different forecast error metrics, such as Mean Absolute Percentage Error (MAPE) or Mean Squared Errors (MSE), without analyzing residual autocorrelation and deviations from the assumed statistical distribution (Barrow and Kourentzes, 2016). Syntetos et al. (2010) investigate the relationship between forecasting accuracy improvements and stock control metrics, like cycle service level and inventory investments, and

suggest that potential autocorrelation of the forecast errors (over the forecast horizon) may have an effect on cycle service level under-performance. Trapero et al. (2016) show that if the residuals are not iid, the supply chain performance can deteriorate through different service levels than targeted, higher volume of backorders or greater inventory investment.

An alternative to using theoretical models is their empirical counterparts (Chatfield, 2000). Data driven approaches do not rely on iid assumptions and they have provided promising results in calculating prediction intervals (Williams and Goodman, 1971). When the iid assumptions are not satisfied, other techniques like GARCH (Bollerslev, 1986) and Kernel Density Estimation (KDE; Silverman, 1986) can be useful to forecast the size of uncertainty.

Within a supply chain case study, Syntetos and Boylan (2008) analyzed the empirical performance of alternative forecasting methods for slow moving items. They identified the estimation of the variability of the lead time forecast errors as an area, amongst others, with scope for improving the overall performance of the system. Zhang and Kline (2007) pointed out that, although temporal demand heteroscedasticity has not been included into the inventory management literature, it is present in industrial time series. In addition, computational results show that ignoring temporal heteroscedasticity can increase company's inventory costs up to 30%, when demand autocorrelation is highly positive. Trapero et al. (2016) applied parametric models, such as GARCH, to the problem of determining the variability of the lead time forecast error. They showed with simulated and real data that for high lead times, forecast error standard deviation presented temporal autocorrelation and that GARCH models yielded promising results.

Another miss-specification typically found in a supply chain context is normality deviation (one of the other areas of improvement identified by Syntetos and Boylan, 2008). For intermittent demand data, bootstrapping is a non-parametric alternative (Willemain et al., 2004). Syntetos et al. (2015) compared the performance of bootstrapping with respect to parametric methods. For continuous demand data, non-parametric approaches like KDE has been succesfully reported by Strijbosch and Heuts (1992) and Manary et al. (2009). Manary et al. (2009) employed KDE at the Intel company for correcting forecast bias, nonnormal forecast errors and heterogeneous forecast errors resulting in safety stock reductions of approximately 15 percent. Trapero et al. (2016) also analyzed KDE and showed that worked reasonably well for lower lead times, where nonnormality issues were a common situation. However, if forecast errors possess both standard deviation autocorrelated

and an unknown density function, which one should be used?

In this work we propose a quantile combination scheme that provides the quantiles of the lead time forecast density function required to determine the safety stock. By combining, we circumvent the need for choosing a specific empirical approach. Furthermore, the combination proportions for each method are data driven. The proposed quantile combination is optimal in the sense that it minimizes the tick loss function (see section 3.2). Initially, such a combination was defined by Giacomini and Komunjer (2005) and applied to value-at-risk models. In that same reference, authors point out the little empirical work that has been done in the context of conditional quantile forecasting, despite the extensive body of literature in economics and management science attesting the usefulness of forecast combination for point forecasts (Clements and Harvey, 2011). In fact, to the best of the author's knowledge, this is the first time quantile combination is applied to safety stock estimation.

The classical assumption of normality and independence are evaluated for different lead times by means of the Jarque-Bera and Engle tests, respectively, in a real data case study. Our results show that non-normality mainly happens for low lead times. In that case, the proposed combination achieves a closer cycle service level with respect to the target one and it also reduces the inventory investment and backorders, these improvements are summarized in a tick loss reduction. For high lead times, the main issue is temporal heteroscedasticity rather than non-normality. Here, the combination scheme reduces the tick loss for most of the quantiles under study, except for the extreme quantile (99%), where it does not improve over GARCH method.

The rest of the paper is organized as follows: section 2 provides relevant background research. Section 3 discusses why a quantile combination of forecasts is required instead of a point or density forecast. Furthermore, this section describes the proposed combination scheme. Section 4 discusses implementation aspects of the experiments, defining the criteria to measure the performance of forecasts, as well as, the point forecasting algorithm employed. Section 5 carries out different Monte Carlo simulations to show the influence of residual autocorrelation and asymmetric demand distributions on the proposed combination approach. Section 6 presents the case study that is utilized to assess the proposed approach. Finally, the last section summarizes the main conclusions.

2. Background research

If the demand forecasting error is Gaussian iid with zero mean and constant variance, the safety stock (SS), for a target Cycle Service Level (CSL), expressed as the target probability of no stockout over the lead time, can be computed as:

$$SS = k\sigma_L \tag{1}$$

where $k = \Phi^{-1}(CSL)$ is the safety factor; $\Phi(\cdot)$ denotes the standard normal cumulative distribution function; and σ_L stands for the standard deviation of the forecast error for a certain lead time L that it is assumed to be constant and known.

The main challenge in (1) is estimating σ_L. There are two alternatives: a theoretical and an empirical approach. Regarding the theoretical option, first, an estimation of σ_1 (one-step ahead standard deviation of the forecast error) is provided and, since the updating forecast step is usually smaller than the lead time, subsequently, an analytic expression that relates σ_L and σ_1 is employed. For instance, Hyndman et al. (2008) show that if the demand can be modeled as a local level model, i.e., an ETS(A,N,N) with parameter α, the conditional variance for the lead-time demand is:

$$\sigma_L = \sigma_1\sqrt{L}\sqrt{1 + \alpha(L-1) + \frac{1}{6}\alpha^2(L-1)(2L-1)}. \tag{2}$$

Note that Wecker (1979), Johnston and Harrison (1986) and Graves (1999) also arrived at the same expression assuming a demand process that follows an ARIMA(0,1,1), which is equivalent to ETS(A,N,N). In addition, σ_1 can be estimated based on applying a single exponential smoothing on the Mean Squared Error (MSE), such as $\sigma_1 = \sqrt{MSE_{t+1}}$ and MSE is updated as new observations become available as follows:

$$MSE_{t+1} = \alpha'\epsilon_t^2 + (1-\alpha')MSE_t, \tag{3}$$

where $\epsilon_t = y_t - F_t$, y_t is the actual value at time t and F_t is the forecast value for the same period. Within this first alternative, Prak et al. (2017) pointed out that if demand distribution parameters are not known and should be estimated, as it usually happens, safety stocks should include a correction factor.

Alternatively to the theoretical approach, an empirical parametric approach can be employed (Syntetos and Boylan, 2006, 2008), in which case σ_L is estimated directly from the lead time forecast error such as:

$$\sigma_L = \sqrt{MSE_{L,t}}, \tag{4}$$

and

$$MSE_{L,t} = \gamma\epsilon_{L,t}^2 + (1-\gamma)MSE_{L,t-1}, \tag{5}$$

where $\epsilon_{L,t} = \sum_{i=t-L+1}^{t}(y_i - F_i)$ is the cumulative lead time forecast error. For this heuristic method we need neither the forecasting model/method nor its parameters, i.e., it does not rely on an ARIMA(0,1,1) demand process. In practice, this is particularly advantageous when Forecasting Support Systems do not provide such information to users. Nonetheless, expressions (4)-(5) in conjunction with (1) retain the assumption that lead time forecast errors are normally distributed.

When applying an exponential smoothing to forecast errors, as in (5), implicitly we assume that σ_L can be time-varying, thus, the independence assumption is also relaxed. In this sense, to cope with time-varying volatility, we can use the generalized autoregressive conditional (GARCH) models (Bollerslev, 1986), that represent a more parsimonious and less restrictive version of the ARCH(p) models (Engle, 1982). GARCH(p,q) models express the conditional variance of the forecast error at time $t+1$, as a linear function of both q lagged squared error terms (ϵ_t^2) and p lagged conditional variance terms. For example, GARCH(1,1) model is given by:

$$\sigma_{t+1}^2 = \omega + a_1\epsilon_t^2 + \beta_1\sigma_t^2. \tag{6}$$

Note that SES in (3) can be seen as a particularization of a GARCH model: the integrated GARCH model (IGARCH; Nelson, 1990), with $\beta_1 = 1 - a_1$ and $\omega = 0$.

Following the same intuition of (4)-(5), we can apply the GARCH(1,1) over the cumulative lead time forecast error instead the one-step ahead forecasting error, thus, equation (6) can be rewritten as:

$$\sigma_{L,t+1}^2 = \omega' + a_1'\epsilon_{L,t}^2 + \beta_1'\sigma_{L,t}^2. \tag{7}$$

In this work, we focus our analysis on the GARCH(1,1) model using an overlapping approach to estimate $\sigma_{L,t}$, as new sample becomes available (Boylan and Babai, 2016). To avoid any confusion with (6), the GARCH(1,1) over the cumulative lead time forecasting error in (7) is called CGARCH(1,1).

It is likely that some demand distributions present important asymmetries, particularly when they are subject to promotions or special events. In

these cases, the typical normality assumption for the forecast errors may not hold and empirical non-parametric approaches can be useful. In this work, as a non-parametric approach we use the Kernel Density Estimation (KDE). To consider non-parametric methods, the safety stock calculation should be reformulated as:

$$SS = Q_L(CSL), \tag{8}$$

where $Q_L(CSL)$ is the lead time forecast error quantile at the probability defined by CSL. This quantile can be estimated non-parametrically from the empirical distribution of the generated lead time forecast errors (e_L). According to Silverman (1986), if $f(x)$ represents the probability density function of the lead time forecast errors, its formula for a series X at a point x is given by:

$$f(x) = \frac{1}{Nh} \sum_{j=1}^{N} K\left(\frac{x - X_j}{h}\right), \tag{9}$$

where N is the sample size, $K(\cdot)$ is a kernel smoothing function that integrates to one and h is the bandwidth (Silverman, 1986).

3. Combining quantile forecasts

3.1. Point, quantile and density forecasts, what should we combine?

Ideally, if we could model the underlying demand process for each SKU, we would not require to use any combination scheme. However, that assumption is unrealistic and, although we could identify the process, in practice, it is not always possible to implement it in supply chain companies for several reasons: i) many companies judgmentally adjust statistical forecasts, (Fildes et al., 2009; Trapero et al., 2013) which may bias the residuals; ii) companies rely on forecasting software vendors with limited forecasting techniques (Fildes, 2017); iii) the choice of demand forecasting model may not be under the control of the operations/inventory planning department that is responsible for setting the safety stock (Manary and Willems, 2008). Therefore, if we have doubts about the validity of the point forecasting models our company have available, the iid assumptions should be questioned and a combination of the different forecasting errors obtained is a reasonable approach for enhancing the calculation of the safety stock.

Combining forecast literature can be divided in two streams: combining point forecasts and combining density forecasts. Combining point forecasts,

i.e. the mean or median of the predictive distribution, has been studied substantially since Granger and Ramanathan (1984). However, its analysis for a supply chain forecasting context has been scarcer (Barrow and Kourentzes, 2016). On the other hand, research on combining density forecast is more recent and mainly applied to financial time series (Clements and Harvey, 2011; Hall and Mitchell, 2007; Geweke and Amisano, 2011). A thorough analysis regarding its utility in a supply chain forecasting context has not been done yet. Between these two areas, we can place combining quantile forecasts, which is very close to combining prediction intervals (Granger et al., 1989). Again, there is a lack of research about the potential benefits of quantile combination in supply chain applications.

Despite the obvious advantages of obtaining the whole predictive distribution, as it provides more information than a point or quantile forecast, it has the issue that it is difficult to compare alternative forecasting algorithms and, particularly, determine the precise quantiles that predictive distributions may differ (Boylan and Syntetos, 2006). In other words, a forecasting algorithm, in overall terms, may provide a closer predictive distribution to the true one, however for certain quantiles that we may be mostly interested for, the results may be worse. Another potential limitation of forecasting the whole distribution is that to evaluate it, we need to specify/estimate the unknown true density of the variable to be forecasted, although this limitation can be avoided by using scoring rules (Hall and Mitchell, 2007).

In this work, since we focus on safety stocks, we do not require the whole forecast distribution. Here, we consider several quantiles that are related to the cycle service level that companies aim for, typically between 85 % and 99 %. We follow the recommendation given by Boylan and Syntetos (2006), where for inventory calculations attention should be restricted to the upper end of the cumulative distribution.

3.2. Proposed combination scheme

Christoffersen (1998) pointed out that combining non-parametric methods with time-varying variance estimators is likely to enhance prediction interval estimates. In this work, we follow that suggestion to determine the safety stock. Note that prediction intervals can be obtained from quantile forecasts (Granger et al., 1989), while the safety stock is the difference between the upper prediction interval limit and the mean.

Trapero et al. (2016) compared different theoretical and empirical methods to determine the safety stock. Among then, KDE and GARCH over the

cumulative lead time forecasting error obtained better results. Thus, both methods have been chosen in this work to combine them by minimizing the tick loss function (Giacomini and Komunjer, 2005). In this sense, the tick loss function, which is also known as the *linlin*, *hinge*, *pinball* or *newsvendor* loss (Gneiting, 2011), is an asymmetric piecewise linear loss function,

$$TL_{\alpha}(y_t, F_t) = \begin{cases} \alpha|y_t - F_t| & if F_t \leq y_t \\ (1-\alpha)|y_t - F_t| & if F_t \geq y_t \end{cases}, \tag{10}$$

of order $\alpha \in (0, 1)$, where any α-quantile of the predictive distribution is an optimal point forecast (Gneiting, 2011). In this work, the target quantile is given by the CSL. In addition, the quantile α reflects the asymmetry in cost terms. From a newsvendor point of view, the cost of underforecasting (C_a) it is not the same cost than overforecasting (C_b). In that sense, α-quantile can be related to costs such as, $\alpha = \frac{C_a}{C_a + C_b}$ (Gneiting, 2011). For example, if the CSL=0.9, it means that the cost of underforecasting is 9 times the cost of overforecasting. Note that in supply chain applications it is common to use the CSL interpretation based on quantiles rather than costs, given the difficulty in estimating the cost C_a.

Given the KDE and GARCH quantiles $Q^1_{L,t}(CSL)$ and $Q^2_{L,t}(CSL)$ respectively, we can combine them to obtain the safety stock (SS_t) in the following form (given that the quantiles are based on the same point forecast, Hall and Mitchell, 2007):

$$SS_t = w_1 \cdot Q^1_{L,t}(CSL) + w_2 \cdot Q^2_{L,t}(CSL), \tag{11}$$

where (w_1, w_2) lies in some compact subset of $\mathbb{R}^2$. Following indications by Giacomini and Komunjer (2005) we do not impose the restriction $w_1 + w_2 = 1$. More details about the restrictions on the combination weights, see (Granger and Ramanathan, 1984). A way to estimate (w_1, w_2) is to minimize the expected tick loss:

$$(w_1^* w_2^*) = \underset{(w_1, w_2)}{\arg\min}\, E_t \left[TL_{CSL}(y_t, F_t + SS_t) \right] \tag{12}$$

Another alternative to determine the optimal combination weight is to maximize the conditional coverage Christoffersen test p-value. Christoffersen (1998) developed statistical tests to assess the unconditional coverage on the basis of the following indicator:

$$I_t = \begin{cases} 1 & \text{if } y_t \in [0, F_t + Q_{L,t}(CSL)] \\ 0 & \text{if } y_t \notin [0, F_t + Q_{L,t}(CSL)] \end{cases}, \tag{13}$$

where $Q_{L,t}(CSL)$ is the quantile for a certain target CSL and y_t is the actual value. Note that coverage is the percentage of times that the actual value is lower than $Q_{L,t}(CSL)$ and it is highly related to the CSL concept in stock control. The limitation of the unconditional coverage test is that it does not measure whether the ones and zeros in (13) come clustered together in a time-dependent fashion (Christoffersen, 1998). In order to address this, Christoffersen proposes the conditional coverage test, which is a combination of the tests for unconditional coverage and independence. The idea behind such a combination is that the resulting quantile forecasts would be robust to potential autocorrelation in the forecast error variability and further deviations from assumed statistical distributions.

Interestingly, both optimization alternatives were connected by Giacomini and Komunjer (2005, Lemma 1, page 419). In that reference is shown that the correct conditional coverage condition is equivalent to requiring optimality of an interval forecast with respect to tick loss function. In this work we have preferred to determine the optimal weights by using (12), since it is a relative evaluation, unlike the Christoffersen test that is an absolute evaluation (Giacomini and Komunjer, 2005). Note that the combination process can be easily automated, which is especially recommended in a supply chain context, where a vast number of products have to be forecasted.

4. Experimental setup

To assess the performance of the proposed combination we need to define point-forecasts for the demand distribution, appropriate evaluation criteria and benchmark models. These considerations are described here and are relevant to the simulations and the case study that follow.

4.1. Point forecast

In order to compute the safety stock the first step is to calculate the demand point forecast. The respective forecast error is used for the variance/quantile forecasting methods described previously. This two-step approach was also employed by (Granger et al., 1989).

In this work we use the single exponential smoothing (Gardner, 1985, 2006) to obtain the point forecasts:

$$F_{t+1} = \alpha y_t + (1 - \alpha) F_t, \tag{14}$$

where $0 < \alpha < 1$. Given the recursive nature of exponential smoothing, it is necessary to initialize the method. We optimize the initial value together with the smoothing parameter α by minimizing the in-sample mean squared error. The lead time forecast is $F_L = \sum_{h=1}^{L} F_{t+h} = L \cdot F_{t+1}$, which is required to compute the lead time forecast error $\epsilon_{L,t}$ in (5) and (7). The choice of the method is appropriate given the data considered, as detailed in the next sections.

4.2. Evaluation of the alternative approaches

4.2.1. Tradeoff curves

The different methods to compute the safety stock are evaluated by its direct effects on stock control by means of tradeoff curves (Gardner, 1990; Syntetos et al., 2015). These curves allow to compare different methods in a realistic fashion, given that from a practitioner view, they are the most meaningful. The three variables considered in the curves are the achieved cycle service level, the inventory investment and backorders.

We assume a newsvendor framework (Beutel and Minner, 2012; Lee, 2014). The achieved cycle service level is calculated as the percentage of times that the real lead time demand falls within the prediction intervals for a certain SKU, where the prediction interval is the sum of the point forecast plus the safety stock. Then, the average of that percentage across SKUs is computed. The inventory investment is the average of the upper bound of the prediction interval per SKU and across SKUs. Note that the safety stock depends on the CSL target, and since the point forecast is the same for every method considered, the main differences found between methods are due to different safety stocks calculations. Additionally, we calculate the backorders by summing the units out of the quantile per each SKU on the hold out sample and then calculating the average of these sums across SKUs. Note that, as we do not have any information available about both, the backorder cost and holding cost per SKU, performance metrics as backorders and inventory investment are function of physical units.

4.2.2. Tick loss function

Although tradeoff curves provide rich information, a potential problem with such curves is that they may be difficult to interpret when comparing several methods. Basically, the practitioner needs to handle three metrics at the same time, that is, achieved cycle service level, inventory investment and backorders. It may occur that, when comparing different methods, some of them may provide a good service level, although at the expense of higher inventory investments. Therefore, unless a method is better at the three metrics, the choice of the "best" method is not totally clear.

A possible solution is to utilize the tick loss function given its economic interpretation. In this sense, the tick loss function averages the assymetric cost of underforecasting and overforecasting, resulting in only one metric. Therefore, the "best" method will be the one with the lowest loss value. To complete the information provided by tradeoff curves and to facilitate the comparison between methods, we will include another graph with the tick loss value obtained for each method.

4.3. Implementation details and benchmark models

In this work the target cycle service levels are set to: 85%, 90%, 95% and 99%. The data is split in four parts of equal length. The first part (25% of the data) is used to compute both the exponential smoothing parameter and its initial value, so as to determine the point demand forecast. The second part is employed to estimate the KDE, CGARCH and Naïve (Defined below) methods for estimating the safety stock. The third part is utilized to calculate the weights in (11). Finally, the last part (25% of the data) is devoted to test the quantile forecasts of the considered methods. Such a sample size distribution is employed for both simulation and real data experiments.

Regarding estimation algorithms, for the KDE, we use the Epanechnikov kernel smoothing function and the bandwidth is set to the appropriate value that is optimal for normal distribution densities (Bowman and Azzalini, 1997). Estimation of CGARCH(1,1) parameters are based on the econometric toolbox from MATLAB, selecting an interior-point optimization algorithm.

Furthermore, two benchmarks are considered. Firstly, a naïve benchmark to assess the KDE and CGARCH, based on the constant empirical lead time standard deviation calculated on the hold-in sample and (1). Secondly, we also benchmark the proposed combination by implementing a combination

of KDE and CGARCH intervals, where each is assigned a weight of 50 %, instead of the optimal set (w_1^*, w_2^*).

5. Simulation results.

To evaluate the performance of the combination scheme proposed, several Monte Carlo simulations with 100 repetitions were carried out, where the length of each simulation was set to 500 observations. The simulation exercise will deal with the common situation where the point forecasting method does not precisely follow the demand generating process. That experiment has been implemented by simulating an AR(1) demand process, whereas the point forecasting method is an exponential smoothing. Such a demand process has been chosen because it is frequently found in empirical datasets. For instance, Ali et al. (2012) find AR(1) to be the most frequent process (30.3 % of the series). Similarly, Trapero et al. (2014) find AR(1) to describe 56.25 % of the series in their dataset.

Let D_t be the demand at time t that follows an AR(1) process:

$$D_t = \mu + \phi D_{t-1} + \epsilon_t, \tag{15}$$

where μ is a positive constant, ϕ is the autoregressive parameter and ϵ_t is i.i.d. normally distributed with zero mean and variance σ^2. The values chosen for the simulation were $\mu = 100$, $\sigma^2 = 50$, and ϕ is allowed to vary between -0.9 and 0.9.

Figure 1 shows the tick loss function on the hold-out sample averaged across repetitions (100) and then, across quantiles (85%, 90%, 95% and 99%) against the autoregressive parameter ϕ. The lead times analyzed in this simulation were 1 and 4 weeks, plotted in the upper and lower panel, respectively. Considering the upper panel with a lead time of 1 week, the parametric CGARCH approach provides a lower average tick loss for every ϕ parameter, closely followed by the proposed Optimal Quantile Combination (OQC). That indicates some autocorrelation on the forecasting errors variability, which may be caused by the fact that the forecasting model did not exactly match with the demand generating process. Since the error term in (15) is normally distributed, non-parametric KDE achieves a higher loss. The 50%-50% combination approach lies between the loss obtained by KDE and CGARCH, very similar to the Naïve. Interestingly, the average tick loss increases for negative values of ϕ. Regarding the lead time of 4 weeks, the

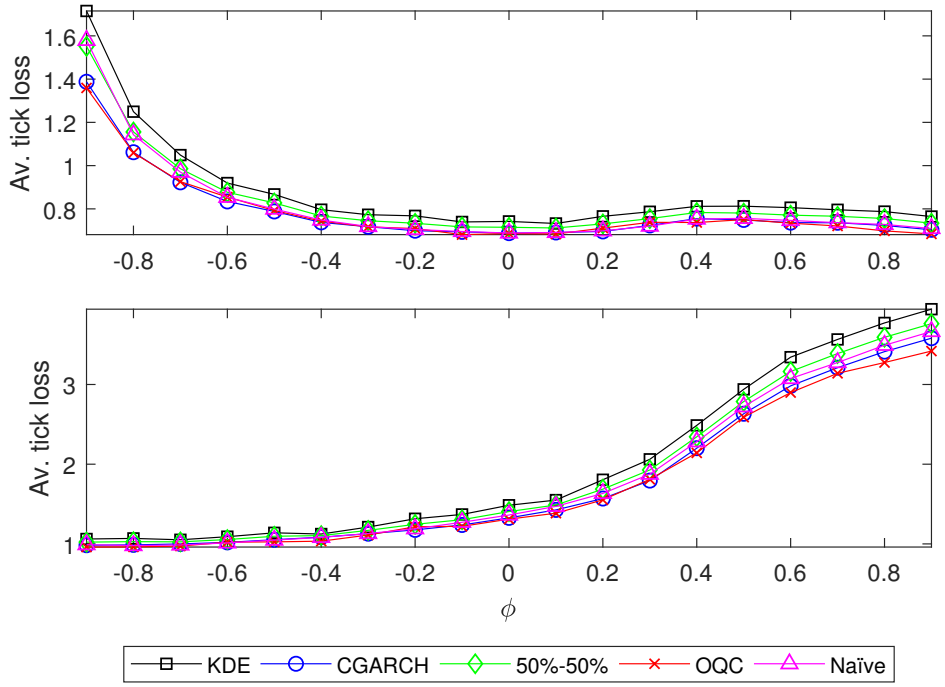

Figure 1: Upper panel: average tick loss for a lead time 1 and normal distribution. Lower panel: average tick loss for a lead time 4 and normal distribution.

combination approach OQC and CGARCH yield the lowest loss, where OQC slightly improves CGARCH for higher values of positive ϕ.

Figure 2 depicts the average tick loss against ϕ when the error distribution in demand is not normally distributed. In this case, we have added to expression (15) a log-normal noise with mean 0.9 and variance 1.4. Again, upper panel corresponds to lead time 1 and lower panel to lead time 4. Considering the upper panel, since the error is not normal, KDE can capture such deviation from normality due to its non-parametric nature providing a low value of tick loss only improved by OQC for most of ϕ values. For a lead time 4 (lower panel), as a consequence of the central limit theorem, the non-normality is mitigated and CGARCH outperforms Kernel, although again, the combination proposed yields the best results.

Figures 1 and 2 show the overall results. Nonetheless, to see the performance of the proposed combination scheme more disaggregated, we can choose a particular value of ϕ, and plot the tradeoff curves and the tick loss value for each quantile of interest. Figure 3 shows the trade-off curves for

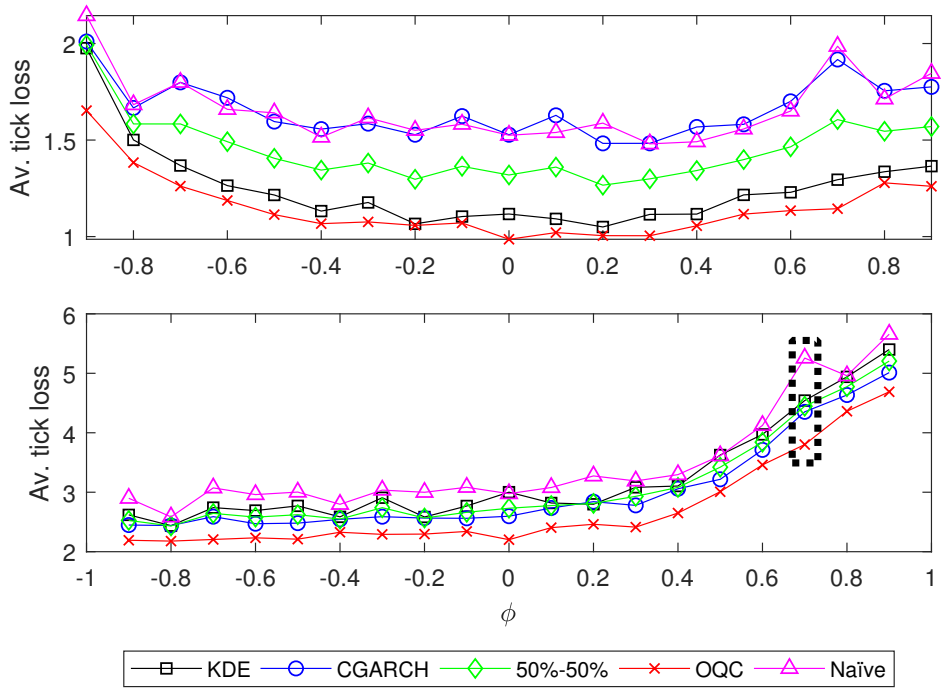

Figure 2: Upper plot: average tick loss values for a lead time 1 and log-normal distribution. Lower plot: average tick loss values for a lead time 4 and log-normal distribution. The marked region is expanded in figures 3 and 4 and table 1.

Table 1: Tick loss values corresponding to Figure 4. Minimal values are highlighted in bold.

Method	CSL			
	0.85	0.95	0.9	0.99
KDE	11.65	9.29	6.31	2.93
CGARCH	10.15	7.68	4.70	**1.51**
50%-50%	10.56	8.0	4.94	1.75
OQC	**10.08**	**7.56**	**4.63**	1.62
Naïve	12.83	10.21	6.68	2.61

$\phi = 0.7$, lead time 4 and log-normal distribution, that corresponds to the points included in the marked rectangle of Figure 2. The tradeoff curves consist of representing achieved CSL and backorders against inventory investment. Each curve has four plotting symbols corresponding to the four CSL targets. The tradeoff curves show how OQC provides good results, since it almost achieve the target CSL but a lower inventory investment. However, it is difficult to indicate which method is the "best", since OQC provides lower inventory investment, although it does not achieve the target cycle service levels and provides a higher level of backorders with regards to CGARCH. An additional plot that can help with the interpretation is shown in Figure 4. In that figure, the value of the tick loss function is calculated for each CSL and method. For that particular case, we can see that KDE, Naïve, 50%-50% obtain higher tick loss values, whereas OQC obtains slightly lower values than CGARCH for most of CSL targets (0.85, 0.9, 0.95). For the quantile 0.99, CGARCH provided the lowest loss, followed by OQC. The tick loss values of such a figure are also found in Table 1, where the minimal values for each CSL are highlighted in bold.

6. Experimental results. A case study.

6.1. Dataset

The dataset employed in this paper has been previously used by Barrow and Kourentzes (2016) and originates from a major UK fast moving consumer goods manufacturer specialized in the production of household and personal care products. In total 229 products with 173 weekly observations per product are available. According to Barrow and Kourentzes (2016) there are no seasonal SKUs and only a minority (21 %) exhibits a small trend.

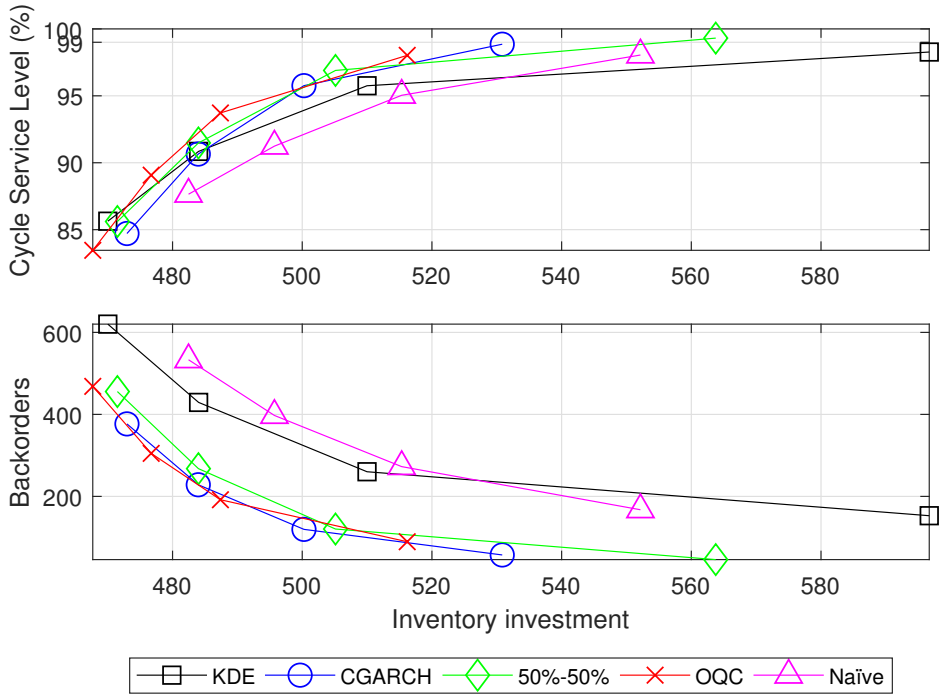

Figure 3: Tradeoff curves for AR(1) with $\phi = 0.7$, lead time 4 and lognormal distribution. The target cycle service levels are 85%, 90%, 95% and 99%

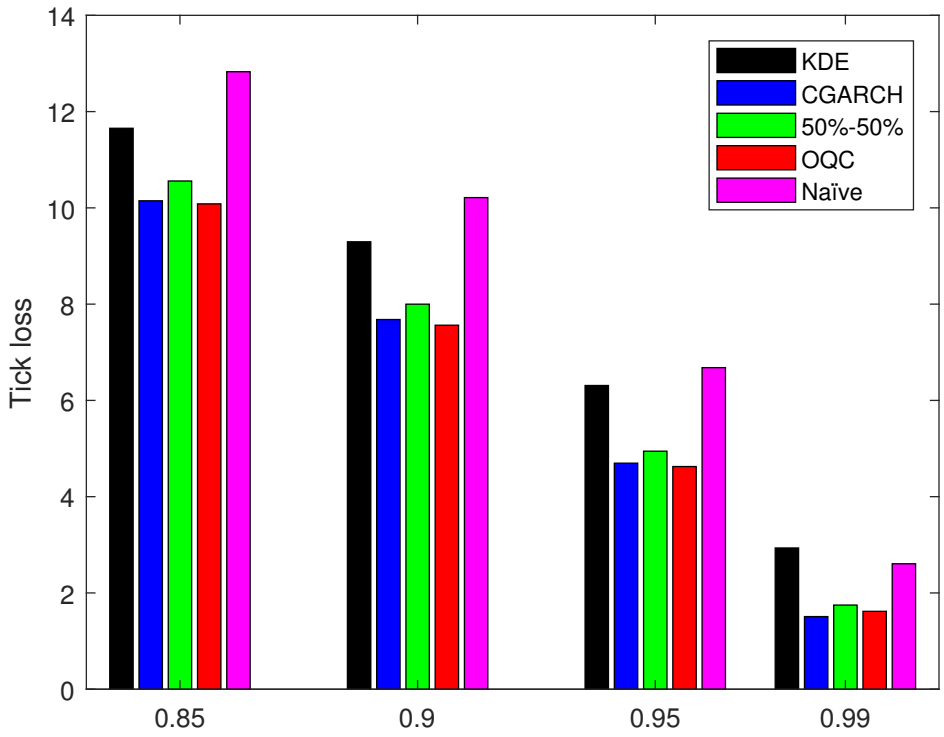

Figure 4: Tick loss values for the same case analyzed in Figure 3. The target cycle service levels are 85%, 90%, 95% and 99%

We use single exponential smoothing to produce the point forecasts for all SKUs, even if some exhibit trend. Although exponential smoothing may not always be the best option to produce point forecasts, this will allow us to assess the performance and robustness of the proposed combination approach when the forecasting model is not the underlying data generating process. Note that this is a common problem both in research and practice. The true process is typically unknown and its identification is not trivial. Furthermore, in an industrial context it is commonplace to use a simple univariate statistical forecasts that are subsequently adjusted judgmentally to encompass additional information (Fildes et al., 2009), or promoted sales (Trapero et al., 2013), to fit better the historical sales, since company forecasting support systems are often not equipped with an adequate repertoire of forecasting models.

6.2. Results

We have carried out two simulations with the real demand data for lead times equal to 1 and 4 weeks. In doing so, this experiment will also shed some light on the influence of the lead time over the combination approach. In order to compare the different inventory investments and backorders across SKUs, sales have been normalized with respect to the in-sample mean.

Figure 5 shows the tradeoff curves of the manufacturer data for a lead time of 1 week. Regarding the achieved CSL, all the methods underachieve the target CSL. However, for quantiles 95% and 99%, OQC outperforms the rest of methods. Potential deviations of normality in the error distribution, which are common for low lead times, make Naïve and CGARCH underachieve the highest target CSL (99%). KDE slightly improves it, but at the expense of higher inventory investments. Focusing on backorders, the combination approach OQC achieves the best results with the lowest level of backorders and for CSL targets 85%, 90% and 95% with the lowest inventory investment too.

Figure 6 shows the tick loss value obtained from each technique. This graph shows the benefits of OQC with respect to the rest of methods for every CSL target.

Figure 7 shows the trade-off curves of the manufacturer data for a lead time of 4 weeks. In terms of achieved CSL, CGARCH provides the best results with the lower deviation from the target. KDE and Naïve yield poor results. Among the combination approaches, OQC and 50%-50% achieves a similar CSL, although OQC at a substantial lower inventory investment.

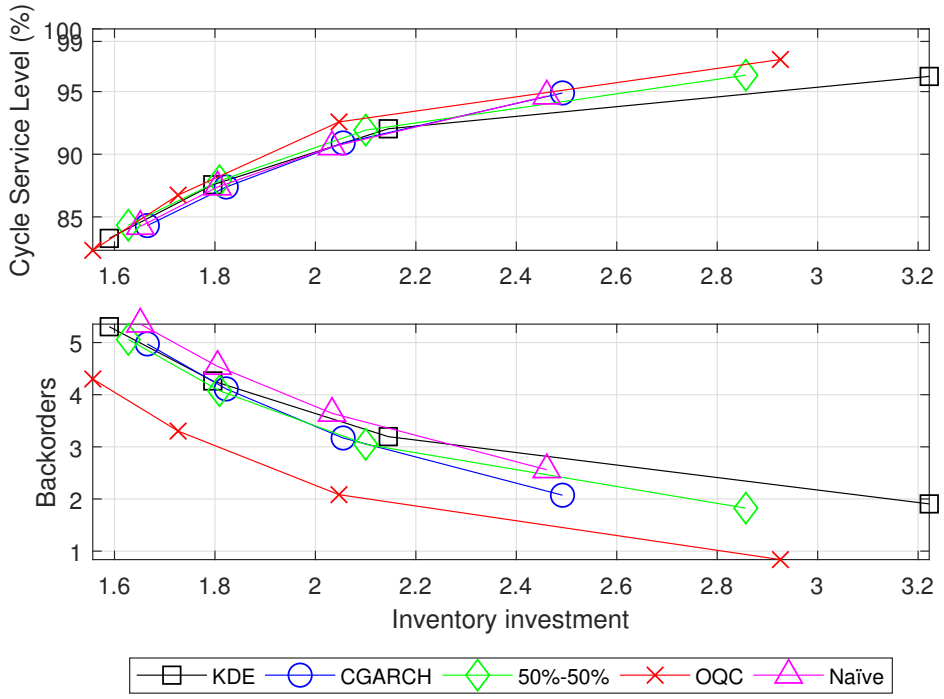

Figure 5: Tradeoff curves for the manufacturer data assuming a lead time equal to 1 week. The cycle service level target are 85%, 90%, 95% and 99%

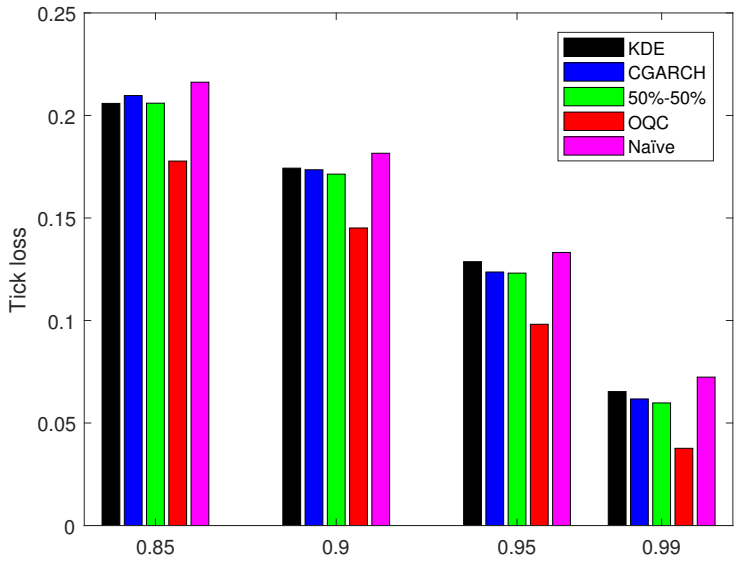

Figure 6: Tick loss values for each CSL target obtained from the manufacturer data assuming a lead time equal to 1 week.

Regarding backorders, OQC outperforms the rest of techniques for all CSL targets, except 99% in favor of CGARCH. These results can be summarized by the tick loss function, whose values are shown in Figure 8. That figure shows that OQC approach reduces the loss for most of CST targets, except for CSL=99%, where CGARCH is shown to be better. Note that these experimental results coincide with those obtained by simulated data in the previous section. In general terms, OQC shows a superior performance with respect to the other methods for every quantile when lead time is 1. For higher lead times, that improvement remains except for the 99% quantile, where CGARCH works better.

A superior performance of KDE for lower lead times is the result of normality deviations on the forecast errors. In the same sense, when CGARCH provides good results for higher lead times is as a consequence of forecast errors conditional heterocedasticity. To provide evidence of these we calculated the Engle test for residual heteroscedasticity and the Jarque-Bera test for Gaussian distribution, for lead times ranging from 1 to 4 weeks. These results are shown in Table 2. The second column in that table shows the percentage of SKUs that reject the null hypothesis of no conditional heteroscedasticity. The third columns shows the percentage of SKUs that reject

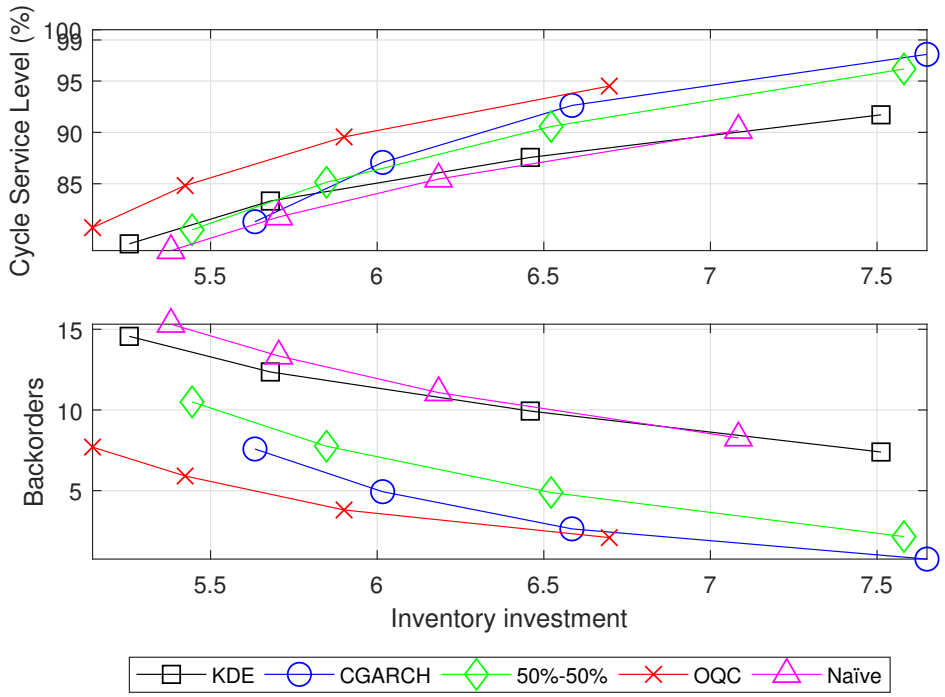

Figure 7: Tradeoff curves for the manufacturer data assuming a lead time equal to 4 weeks. The target cycle service level are 85%, 90%, 95% and 99%

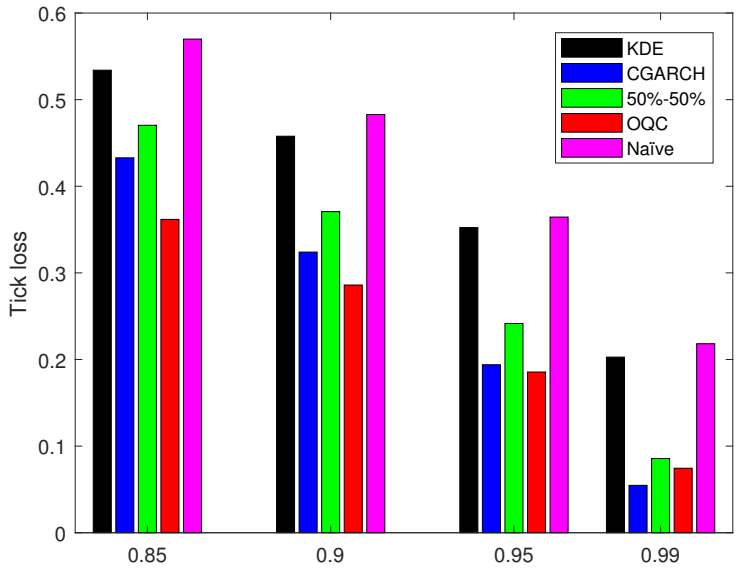

Figure 8: Tick loss values for each target CSL obtained from the manufacturer data assuming a lead time equal to 4 weeks.

Table 2: Engle test for residual heteroscedasticity applied to our dataset assuming different lead times

Lead time	Engle test(% SKUs)	Jarque-Bera test(% SKUs)
1	30.2	85.6
2	91.7	81.2
3	98.2	74.6
4	99.1	69.9

the null hypothesis of Gaussian distribution. To compute those statistical tests, we have employed the forecasting errors, where the sample part devoted to optimize the point forecasting method has been removed.

That table shows the relationship that exists between the lead time and both, the conditional heterocedasticity and normality deviations on the forecasting errors. In summary, as the lead time increases, the percentage of SKUs that do not pass the null hypothesis of no conditional heteroscedasticity is higher. In the same sense, for higher lead times, the percentage of SKUs that reject the null hypothesis of normality is lower due to the central limit theorem.

7. Conclusions

Supply chain management requires forecasts of the demand mean and variance. In particular, safety stocks are based on the estimation of quantiles of the demand forecast error distribution. Such quantiles are related to the cycle service levels that are important for achieving company goals. We propose to combine different empirical approaches to determine the safety stock in a more robust fashion with respect to traditional iid assumptions. The main idea is to combine alternative empirical approaches, which can be parametric and non-parametric, each with its own merits, with optimal weights. These weights are optimal in the sense that they minimize the tick loss function (Giacomini and Komunjer, 2005). To the best of authors' knowledge, this is the first time quantile combination has been applied to compute the safety stock. The results show that the combination reduces the loss function for the different CSL targets. If the lead time is high, such a combination still provides promising results for most of the quantiles, although for extreme quantiles as 99%, CGARCH as a single method is better suited. This conclusion is supported by both simulated and real data.

As a byproduct, the use of the tick loss function provides another assessment tool to complement the tradeoff curves, which are difficult to interpret when handling different methods and any of them show a clear superiority.

The results have direct implications for practice. Even when the demand point forecasts are acceptably accurate, the iid assumptions in the safety stock calculation can harm the inventory performance. Although for certain forecasting models there are analytical expressions for the variance over lead time, again retaining the same assumptions, these are typically not considered in practice and are furthermore invalidated when the final forecasts are adjusted to include managerial judgment (Trapero et al., 2013). This creates a problem for practice, as the calculated safety stocks have several weaknesses, with apparent effects on inventory. On the other hand, using empirical approaches overcomes these limitations, as the assumptions are relaxed and the considered error distribution originates from the final forecast, which can include any further adjustments or be based on any model/method, irrespective of the existence of analytical expressions. The proposed combination improves the empirical approaches to achieve superior inventory performance. Crucially, this is done in a way that it is easy to implement with existing forecasting systems, since the only required input is the historical forecast errors. The proposed approach can benefit existing forecasting systems and

forecasting methods directly. A further advantage of the proposed approach that is relevant to practice is that it is fully automatic and data driven, and therefore implementable in the context of supply chain forecasting, where it is typical to require predictions for a very large number of SKUs.

This work has presented the combination scheme on the basis of two quantile forecasts, KDE and CGARCH, however, the inclusion of more techniques in the combination is straightforward.

According to Giacomini and Komunjer (2005), little empirical work has been done in the context of combining conditional quantile forecasting. This work presents a first attempt to show the merits of such a combination in a supply chain context to determine the safety stocks. However, further research should verify these findings in other industrial datasets, for example, on slow-moving items, where bootstrapping non-parametric methods have been successfully applied (Willemain et al., 2004). Furthermore, this work was limited to a newsvendor framework, however, other stock control policies of the order-up-to level type should also be investigated, as well as, the impact of the method to determine the safety stock on the demand variability of other upwards supply chain members through the bullwhip effect.

Acknowledgment

This work was supported by the European Regional Development Fund and Spanish Government (MINECO/FEDER, UE) under the project with reference DPI2015-64133-R. The authors would like to acknowledge the useful comments and references of three anonymous referees that led to a considerably improved version of the article.

References

Ali, M. M., Boylan, J. E., Syntetos, A. A., 2012. Forecast errors and inventory performance under forecast information sharing. International Journal of Forecasting 28 (4), 830 – 841, special Section: Election Forecasting in Neglected Democracies.

Barrow, D., Kourentzes, N., 2016. Distributions of forecasting errors of forecast combinations: implications for inventory management. International Journal of Production Economics 177, 24–33.

Beutel, A.-L., Minner, S., 2012. Safety stock planning under causal demand forecasting. International Journal of Production Economics 140 (2), 637 – 645.

Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics 31 (3), 307 – 327.

Bowman, A. W., Azzalini, A., 1997. Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations. Vol. 18. OUP Oxford.

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., 1994. Time series analysis: Forecasting and Control. (3rd. Edition). Upper Saddle River, New Jersey: Prentice Hall.

Boylan, J., Syntetos, A., June 2006. Accuracy and Accuracy Implication Metrics for Intermittent Demand. Foresight: The International Journal of Applied Forecasting (4), 39–42.

Boylan, J. E., Babai, M. Z., 2016. On the performance of overlapping and non-overlapping temporal demand aggregation approaches. International Journal of Production Economics 181, Part A, 136 – 144, sI: {ISIR} 2014.

Chatfield, C., 2000. Time-series forecasting. CRC Press.

Christoffersen, P. F., 1998. Evaluating interval forecasts. International economic review 39 (4), 841–862.

Clements, M. P., Harvey, D. I., 2011. Combining probability forecasts. International Journal of Forecasting 27 (2), 208 – 223.

Engle, R. F., 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. Econometrica: Journal of the Econometric Society 50, 987–1007.

Fildes, R., Winter 2017. Research into Forecasting Practice. Foresight: The International Journal of Applied Forecasting (44), 39–46.

Fildes, R., Goodwin, P., Lawrence, M., 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.

Gardner, E. S., 1985. Exponential smooothing: The state of the art. Journal of Forecasting 4 (1), 1–28.

Gardner, E. S., 2006. Exponential smoothing: The state of the art, Part II. International Journal of Forecasting 22, 637–666.

Gardner, Jr., E. S., 1990. Evaluating forecast performance in an inventory control system. Management Science 36 (4), 490 – 499.

Geweke, J., Amisano, G., 2011. Optimal prediction pools. Journal of Econometrics 164 (1), 130–141.

Giacomini, R., Komunjer, I., 2005. Evaluation and combination of conditional quantile forecasts. Journal of Business & Economic Statistics 23 (4), 416–431.

Gneiting, T., 2011. Quantiles as optimal point forecasts. International Journal of Forecasting 27 (2), 197 – 207.

Granger, C., White, H., Kamstra, M., 1989. Interval forecasting. Journal of Econometrics 40 (1), 87 – 96.

Granger, C. W., Ramanathan, R., 1984. Improved methods of combining forecasts. Journal of Forecasting 3 (2), 197 – 204.

Graves, S. C., 1999. A single-item inventory model for a nonstationary demand process. Manufacturing & Service Operations Management 1 (1), 50–61.

Hall, S. G., Mitchell, J., 2007. Combining density forecasts. International Journal of Forecasting 23 (1), 1 – 13.

Harvey, A., 1989. Forecasting Structural Time Series Models and the Kalman Filter. Cambridge University Press.

Hyndman, R. J., Koehler, A. B., Ord, J. K., Snyder, R. D., 2008. Forecasting with Exponential Smoothing: The State Space Approach. Springer-Verlag, Berlin.

Johnston, F. R., Harrison, P. J., Mar 1986. The variance of lead-time demand. Journal of the Operational Research Society 37 (3), 303–308.

Lee, Y. S., 2014. A semi-parametric approach for estimating critical fractiles under autocorrelated demand. European Journal of Operational Research 234 (1), 163 – 173.

Manary, M. P., Willems, S. P., Mar 2008. Setting safety-stock targets at Intel in the presence of forecast bias. Interfaces 38 (2), 112–122,158–159.

Manary, M. P., Willems, S. P., Shihata, A. F., Sep. 2009. Correcting heterogeneous and biased forecast error at intel for supply chain optimization. Interfaces 39 (5), 415–427.

Nelson, D. B., 1990. Stationarity and persistence in the garch(1,1) model. Econometric Theory 6 (3), 318–334.

Prak, D., Teunter, R., Syntetos, A., 2017. On the calculation of safety stocks when demand is forecasted. European Journal of Operational Research 256 (2), 454 – 461.

Silver, E., Pyke, D., Peterson, R., 1998. Inventory Management and Production Planning and Scheduling. Wiley.

Silverman, B., 1986. Density Estimation for Statistics and Data Analysis. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, Bristol.

Strijbosch, L., Heuts, R., 1992. Modelling (s, Q) inventory systems: Parametric versus non-parametric approximations for the lead time demand distribution. European Journal of Operational Research 63 (1), 86 – 101.

Syntetos, A. A., Babai, M. Z., Jr., E. S. G., 2015. Forecasting intermittent inventory demands: simple parametric methods vs. bootstrapping. Journal of Business Research 68 (8), 1746 – 1752, special Issue on Simple Versus Complex Forecasting.

Syntetos, A. A., Boylan, J. E., 2006. On the stock control performance of intermittent demand estimators. International Journal of Production Economics 103 (1), 36 – 47.

Syntetos, A. A., Boylan, J. E., 2008. Demand forecasting adjustments for service-level achievement. IMA Journal of Management Mathematics 19 (2), 175–192.

Syntetos, A. A., Nikolopoulos, K., Boylan, J. E., 2010. Judging the judges through accuracy-implication metrics: The case of inventory forecasting. International Journal of Forecasting 26 (1), 134 – 143, special Section: European Election Forecasting.

Trapero, J. R., Cardos, M., Kourentzes, N., 10 2016. Empirical safety stock estimation based on kernel and GARCH models. Tech. rep., Lancaster University, Department of Management Science.

Trapero, J. R., García, F. P., Kourentzes, N., 2014. Impact of Demand Nature on the Bullwhip Effect. Bridging the Gap between Theoretical and Empirical Research. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1127–1137.

Trapero, J. R., Pedregal, D. J., Fildes, R., Kourentzes, N., 2013. Analysis of judgmental adjustments in the presence of promotions. International Journal of Forecasting 29 (2), 234 – 243.

Wecker, W. E., January 1979. The variance of cumulative demand forecasts. Working Paper. Graduate School of Business. University of Chicaco 5.

Willemain, T. R., Smart, C. N., Schwarz, H. F., 2004. A new approach to forecasting intermittent demand for service parts inventories. International Journal of Forecasting 20 (3), 375 – 387.

Williams, W. H., Goodman, M. L., 1971. A simple method for the construction of empirical confidence limits for economic forecasts. Journal of the American Statistical Association 66 (336), 752–754.

Zhang, G. P., Kline, D. M., 2007. Quarterly time-series forecasting with neural networks. IEEE Trans. Neural Netw. 18 (6), 1800–1814–.