Forecasting seasonal demand: a serious limitation of Winters' forecasting procedure and the added value of product-aggregation
van Donselaar, K.H.

Published: 01/01/2003

Document Version
Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

• A submitted manuscript is the author's version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

Citation for published version (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 21. Dec. 2018
Forecasting seasonal demand:
a serious limitation of Winters’ forecasting procedure
and the added value of product-aggregation

Karel van Donselaar
Eindhoven University of Technology,
PO BOX 513, Paviljoen E3
5600 MB EINDHOVEN
The Netherlands
k.h.v.donselaar@tm.tue.nl
tel. 040 – 247 2869
fax 040 – 246 4531
Abstract
The well-known method for forecasting seasonal demand, Winters’ procedure, has a serious drawback: if the relative demand uncertainty increases (e.g. due to larger product assortments) or if the amount of historical demand data decreases (e.g. due to smaller product life cycles), the quality of the forecasts deteriorates. In this paper mathematical models are created to quantify these effects. These models help researchers, teachers and practitioners to better understand why and when Winters’ forecasting procedure may deteriorate.

One way to improve again the performance of Winters’ procedure may be to use the concept of product-aggregation: if different products have a similar seasonal pattern, the seasonal indices from Winters’ method can be determined from the product family’s aggregate demand. Mathematical modelling as well as simulation is used to assess the added value of product-aggregation. It turns out that impressive improvements can be achieved, especially in case demand uncertainty is high or (when forecasting is applied in inventory systems) in case the lead times are large.

Keywords:
Forecasting, exponential smoothing, seasonality, aggregate data, forecast error, gamma distribution, inventory systems

About the author:
Karel van Donselaar is assistant professor at the subdepartment of Logistics at Eindhoven University of Technology in The Netherlands. He wrote his dissertation on alternatives for MRP in stochastic multi-echelon inventory systems and the determination of safety stock norms in supply chains. The results of his dissertation are being applied in the automotive and the pharmaceutical industry. His main research interests include inventory control, forecasting, the value of demand information in supply chains and logistics in the retail sector.
1. Introduction

Makridakis and Wheelwright (1989) were right in their forecast, when they said ‘...one of the primary needs of management in the 1990s is to deal with an increasingly uncertain environment, one very dissimilar to that of the 1960s, the decade when most of the widely used statistical methods of forecasting were developed. During the 1960s most of the industrial world witnessed unprecedented economic stability.....’. Indeed, in the last decade assortments have grown rapidly and at the same time product life cycles have been reduced substantially. As a result, forecasting short-term demand for an individual item has become more difficult. This is especially true for seasonal demand. The standard approach, i.e. modeling the seasonal demand for individual products independently, no longer works or may no longer be optimal. For example Winters’ forecasting method, which will be explained later on, needs 2 years historical data as a minimum just to do the initialization\(^1\). Many products nowadays already have a life cycle which is less than two years. Next to this, due to the growing assortments, there are simply too many products with too few data for each individual product to construct reliable forecasting models.

A concept, which might help in dealing with these developments, is aggregation. Dalhart (1974) suggested for example to construct product families, consisting of different individual products with similar seasonal patterns. The aggregated demand per product family is used to find the seasonal indices. These indices are then considered to be known when a forecast is made at the individual product level. Based on the idea that often the demand at the product family level is relatively less erratic compared with demand at the product level, it is assumed that separating the seasonal pattern from the randomness will be easier at the product family level.

In paragraph 2 the literature on (dis)aggregation in demand forecasting is reviewed. In paragraph 3 the forecasting method based on product-aggregation is described in detail. In paragraph 4 the impact of two dominant variables on the performance of Winters’ procedure is quantified, resulting in the following two conclusions:

1. The higher the demand uncertainty, the more difficult it is for Winters’ procedure to distinguish the seasonal pattern from randomness
2. The shorter the initialization and the more uncertain demand per review period, the higher the risk that huge forecast errors arise.

Next, the concept of forecasting based on product-aggregation is evaluated by means of mathematical modelling and by means of a theoretical simulation experiment. In the simulation experiment forecasting is applied to an inventory control system. Paragraph 5 describes the simulation model. It turns out that product-aggregation works very well and may lead to inventory reductions, which in some situations exceed 50%. The inventory reduction depends mainly on the coefficient of variation of deseasonalized demand\(^2\). These and other simulation results are presented in paragraph 6. The article ends with the main conclusions of the research.

\(^1\) in case the seasonal pattern is a yearly pattern.

\(^2\) In practice many different circumstances are present, which are not incorporated in these theoretical models. Therefore in a separate paper (Dekker and Van Donselaar, 2002) real sales data from two companies are used to test the concept of forecasting based on product-aggregation in practice.
2. Literature review on aggregation in demand forecasting

The literature on aggregation in demand forecasting is sparse. In the bibliography of Kwong et al. (1995), which covers more than 4200 articles on forecasting, written in the period 1979-1993, only three articles are devoted to some aspect of aggregation. The famous book of Box, Jenkins and Reinsel (1994) deals with many aspects of forecasting, but does not address the issue of aggregation.

Essentially, two small streams of research were found, which are dealing with (dis)aggregation in relation to forecasting. The first stream aims to operationalize the concept of demand forecasting based on product-aggregation and to test the added value of the concept based on empirical data. Dalhart (1974) introduced the concept of combining products into a product family to obtain a better estimate of the seasonal pattern. Withycombe (1989) tested the concept with (4-weekly or) monthly data for 29 products from a computer peripherals supplier, which he divided into 6 product families. He reports an average 11.5% reduction in a forecast error called MSE. Bunn and Vassilopoulos (1993 and 1999) extended the work of Dalhart and Withycombe by providing a broader comparison of methods, addressing the issue of forming the product families and testing various approaches on a sample of 54 series with 4-weekly sales data from a chain of department stores. Their empirical tests again revealed the superiority of the product-aggregation concept, although sometimes the reduction in forecast errors was not overwhelming. Dekker and Van Donselaar (2002) compared several forecasting methods, which are based on either product-aggregation, time-aggregation (i.e. aggregating demand over time, for example by consolidating weekly data into 4-weekly data) or correlation. They used weekly sales data for plastic tubes from a electro-technical wholesaler and for beers from a food-wholesaler. Their results show that the added value of product-aggregation may differ per industry (since seasonality may have different characteristics in different industries) and that the concept of correlation is a very powerful element in forecasting in case seasons are not constant every year (which typically is true for products with sales depending on the weather).

The second stream of research deals with the issue of top-down versus bottom-up forecasting. This issue is often encountered in Hierarchical Planning Systems. For example, in production planning two planning levels may be distinguished: Aggregate Production Planning and Detailed Production Planning. Aggregate Production Planning typically looks ahead 6 to 24 months and uses time-buckets like months or quarters. Input for this planning is an aggregate forecast. This forecast may reflect the requirements for a group of products, which all share a critical capacity or a critical component. The result of the Aggregate Production Planning can be translated to the detailed short-term planning of the individual products by using a planning bill-of-material.

Giesberts (1993) discusses two ways to perform this translation. The first is to disaggregate the Aggregate Production Plan into production plans per product (see also Ling and Sari, 1987). The second way is to disaggregate the aggregate sales forecast into a sales forecast per product (see Ling and Sari, 1987 and Everdell, 1984). Giesberts compares two ways to derive the sales forecast at the product level: top-down from the aggregate forecast or directly from the historical data at the product level (this second method is called direct detailed forecasting). Giesberts proves that direct detailed forecasting is optimal in case the demand for different products is independent and top-down forecasting is optimal in case the products

---

Footnote:

A planning bill-of-material specifies for each individual product its percentage of total demand for the product group.
have identical demand processes with full positive correlation of the changes in the systematic pattern.

Just like Giesberts, we will make a comparison between top-down forecasting and direct detailed forecasting. In contrast to Giesberts, we will assume that demand follows a seasonal pattern. Furthermore, the top-down forecasting procedure evaluated in this paper is different from the procedure evaluated by Giesberts. In this paper the aggregate forecast is not translated to the product level via a planning bill-of-material. Instead, only the seasonal indices are passed on from the aggregate level to the product level. The demand level is still determined at the individual product’s level. This may be especially beneficial if the mix within the family changes over time.

In this paper we focus on short-term demand forecasting for demand processes which consist of both a level and a seasonal pattern. For reasons of analytical tractability we restrict ourselves to demand processes without a trend. A well-known method for forecasting seasonal demand is Winters’ exponential smoothing procedure (1960). Winters’ procedure has a good track record for forecasting seasonal demand, when applied at a high aggregation level (e.g. in Aggregate Production Planning), using monthly or quarterly time-buckets and using many years of historical data (see e.g. De Leeuw et al. (1998)). In this paper we investigate the performance of Winters’ procedure if it is applied at the individual product level with a limited amount of historical data and when applied for short-term forecasting. Short-term forecasting implies that we are dealing with weekly or even daily time-buckets. We will compare the performance of Winters’ procedure with the forecasting method based on product-aggregation (abbreviated as AGG), which is explained next.

3. The forecasting method based on product-aggregation (AGG)

The forecasting method based on product-aggregation starts with an one-time initialization of variables like the seasonal indices and/or the demand level and then continues with a periodic update of these variables. The initialization procedure first estimates for every period $t$ the deseasonalized demand level for this period by calculating the centered 1-year Moving Average demand (roughly speaking this is a summation of the total demand during half a year before this period and half a year after this period, divided by the number of periods per year; see Silver et al., 1998 for details on this calculation). Next an estimator for the seasonal index in period $t$ is determined by dividing the actual demand in period $t$ by the estimated deseasonalized demand.

Note that, because of the moving average procedure described above, we cannot obtain estimates for the first half-season nor for the last half-season. So in case the initialization period is two years, only one demand observation is available for every period in the season to estimate its seasonal index. In case more than two years of data are available, the estimates for the seasonal indices for similar periods in different years are averaged in order to dampen random effects.

When Silver et al. describe the initialization phase for Winters’ procedure they ‘suggest a minimum of 4 complete seasons, using 5 or 6 (with care) if that much is available’. In case less seasons are available they already advise to use information

---

4 In case there is a trend, alternative methods can be applied, see e.g. Holt (1957).
on the season from other sources: an existing individual item or group of items. Basically they also suggest here, in case insufficient data exist, to use the product-aggregation procedure to find the seasonal pattern. In this paper we will compare Winters’ procedure (based on the initialization procedure for situations with sufficient data) with the product-aggregation procedure in order

- to quantify the consequences if Winters’ standard procedure is still used although there are less than 4 complete seasons available and
- to show which parameters in the system have the biggest impact on the error we make when we ignore the fact that the number of data is limited.
- to show the effect of decreased demand variation due to aggregation.

Before the forecasting method can be applied, a product family has to be defined and selected. A product family is defined here as a set of different products all having a similar seasonal pattern. Once it is decided which products are part of the product family, the demand for these products is aggregated. Next, Winters’ method is applied to this aggregated demand in order to determine the seasonal indices. With these indices the demand at the individual item level is deseasonalised. Then simple exponential smoothing is applied to this deseasonalised demand, resulting in an estimator for the level of demand. The final step is to multiply this estimator with the seasonal index found at the product family level.

In order to describe the procedure above in terms of mathematical formulas, we define:

\[ x_t \] = actual demand of the product family in period \( t \)

\[ x_{jt} \] = actual demand of product \( j \) in period \( t \)

\[ \hat{a}_{jt} \] = estimated level of demand for product \( j \) in period \( t \)

\[ \hat{f}_{jt}^\mathcal{N} \] = normalised seasonal index for the product family in period \( t \)

\[ \hat{x}_{jt, t+i} \] = demand forecast for product \( j \) in period \( t+i \), made at the end of period \( t \)

\[ \hat{f}_t \] = seasonal index for the product family in period \( t \), before normalisation to 1

\[ \hat{a}_t \] = estimated level of aggregate demand for product family in period \( t \)

\( P \) = number of periods in one season

\( N_{pro} \) = number of products in the product family

The first step is to aggregate the sales data for all items belonging to the product family:

\[ x_t = \sum_{j=1}^{N_{pro}} x_{jt} \]

Every period a new forecast is made, the first step will be to determine new seasonal indices at the family level. This is done by applying Winters’ method on the aggregated demand data (in the formulas below the smoothing parameters at the family level are denoted by \( \alpha_{Fam}^\mathcal{W} \) and \( \gamma_{Fam}^\mathcal{W} \)).

---

5 The seasonal indices for \( P \) periods are normalized to make sure that they will always add up to \( P \) (see Silver et al., 1998, for details on this).
\[ \hat{a}_t = \alpha_{W} \hat{x}_t / \hat{f}_{t-P} + (1 - \alpha_{W}) \hat{a}_{t-1} \]
\[ \hat{f}_t = \gamma_{W} \hat{x}_t / \hat{a}_t + (1 - \gamma_{W}) \hat{f}_{t-P} \]

The estimates for the seasonal indices \( \hat{f}_t \) are normalised, to make sure they add up to P. This yields:

\[ \hat{f}_t^N = \hat{f}_t \cdot \hat{c}_t^N \]

with \( \hat{c}_t^N \) the correction factor needed to normalize the seasonal indices\(^6\).

Next, on the item level, the demand level is determined. This is done by applying simple exponential smoothing on item demand after this item demand is deseasonalized using the seasonal indices of the product family.

\[ \hat{a}_{j,t} = \alpha_w \hat{x}_{j,t} / \hat{f}_{t-P} + (1 - \alpha_w) \hat{a}_{j,t-1} \]

Finally the forecast at item level is calculated by simply multiplying the level with the seasonal index.

\[ \hat{\xi}_{j,t+1} = \hat{a}_{j,t} \hat{f}_{t+1-P} \]


The performance of Winters’ procedure depends on the demand uncertainty, the length of the initialization period and whether product-aggregation is applied or not. To demonstrate and quantify this, a theoretical analysis is performed in this paragraph. More specifically, in this paragraph we will prove that:

1. The higher the demand uncertainty, the more difficult it is for Winters’ procedure to distinguish the seasonal pattern from randomness.
2. The shorter the initialization and the more uncertain demand per review period, the higher the risk that occasionally huge forecast errors arise.

The impact of demand uncertainty on the ability to recognize seasonal patterns will be explained in subparagraph 4.1. The next two subparagraphs analyze the impact of a small initialization period and uncertain demand on the size of the forecast error. Subparagraph 4.2 deals with this issue in case Winters’ procedure (IND) is applied. The analysis in subparagraph 4.3 shows to what extent demand forecasting based on product-aggregation (AGG) helps to prevent huge forecast errors.

\(^6\) The interested reader is referred to Silver et al. for a description of the exact procedure to determine this correction factor.
4.1 The impact of demand uncertainty on the ability to recognize seasonal patterns

To see the impact of increasing demand uncertainty, imagine a system with one product having deterministic demand. For many years the demand for this product was equal to \( a \) (the demand level) times a seasonal index \( f_{t-p} \). Assume the demand pattern for this system was known in the past, so \( f_{t-p}, \hat{a}_t \), and \( \hat{x}_t \) used to be:

\[
\begin{align*}
\hat{f}_{t-p} &= \hat{f}_{t-p} = f_{t-p} \\
\hat{a}_t &= a \\
\hat{x}_t &= x_t = a.f_{t-p}
\end{align*}
\]

for all \( t < \tau \)

Now imagine that in period \( \tau \) demand for this product suddenly becomes stochastic. The expected demand and the seasonal indices remain the same, but this is unknown to the forecaster.

Assume demand for this product in period \( \tau \) turns out to be

\[ x_\tau = (1+z).f_{\tau-p}.a \]

The variable \( z \) denotes the extent to which the actual demand deviates from the expected demand. If demand uncertainty is very high, very high or very low values of \( z \) are more likely. If we apply Winters’ procedure, we see that this change in demand has the following impact on the estimator for the demand level and the seasonal index for period \( \tau \):

\[
\begin{align*}
\hat{a}_\tau &= \alpha \left[ \frac{(1+z).f_{\tau-p}.a}{f_{\tau-p}} \right] + (1 - \alpha).a = (1 + \alpha z).a \\
\hat{f}_\tau &= \gamma \left[ \frac{(1+z)f_{\tau-p}.a}{(1 + \alpha z).a} \right] + (1 - \gamma).f_{\tau-p} = \left[ 1 + \gamma \frac{(1-\alpha)z}{l + \alpha z} \right] f_{\tau-p}
\end{align*}
\]

To see the impact of \( z \) on the estimator of the seasonal index, we determine the derivative of \( \hat{f}_\tau \):

\[
\frac{df_\tau}{dz} = \frac{(1-\alpha)\gamma}{(1+\alpha z)^2} f_{\tau-p}
\]

This derivative is larger than zero for any \( \alpha < 1 \). Recall that \( z \) represented a random change in the demand; the seasonal pattern remained unchanged. Yet, we see that if Winters’ procedure is applied, the change in demand results in a change of the seasonal index as well. Moreover, we found that the higher the random deviation in demand, the higher the adaptation in the seasonal index. This implies that the higher the demand uncertainty, the more difficult it will be for Winters’ procedure to find the right seasonal indices.
4.2 The risk for large forecast errors when Winters’ procedure (IND) is applied

In case the initialization period is equal to two years, only one demand observation is available for every period in the season to estimate its seasonal index. This means that for a particular product \( j \) the estimate for the seasonal index in period \( t-P \) is simply equal to the actual demand in period \( t-P \) divided by an estimator for the level of demand (which is equal to the centered 1-year Moving Average demand; see paragraph 3):

\[
\hat{f}_{j,t-P} = \frac{x_{j,t-P}}{\hat{a}_{j,t-P}}
\]

After normalization we get:

\[
\hat{f}^N_{j,t-P} = \hat{f}_{j,t-P} \cdot \hat{c}^N_{j,t-P}
\]

with \( \hat{c}^N_{j,t-P} \) the correction factor needed to normalize the seasonal indices.

In case Winters’ procedure is applied at the individual product level, then in the first year after the initialization at the end of period \( t \) the estimator for the level of demand is equal to:

\[
\hat{a}_{j,t} = \alpha \left( \frac{x_{j,t}}{\hat{f}^N_{j,t-P}} \right) + (1-\alpha)\hat{a}_{j,t-1}
\]

The forecast for \( i \) periods ahead is equal to

\[
\hat{x}_{j,t+i} = \hat{a}_{j,t} \cdot \hat{f}^N_{j,t+i-P}
\]

So in the first year after initialization the forecast error, measured here as the Mean Absolute Deviation (MAD), is equal to:

\[
MAD_{j,t+i} = \left| \frac{x_{j,t+i} - \hat{x}_{j,t+i}}{\hat{f}^N_{j,t+i-P}} \right|
\]

\[
= \left| \frac{x_{j,t+i} - \left[ \alpha \left( \frac{x_{j,t}}{\hat{f}^N_{j,t-P}} \right) + (1-\alpha)\hat{a}_{j,t-1} \right] \cdot \hat{f}^N_{j,t+i-P}}{\hat{f}^N_{j,t+i-P}} \right|
\]

\[
= \left| \frac{x_{j,t+i} - \left[ \alpha \frac{x_{j,t}}{\hat{f}^N_{j,t-P}} \cdot \hat{a}_{j,t-P} + (1-\alpha)\hat{a}_{j,t-1} \right] \cdot \hat{c}^N_{j,t-P} \cdot \hat{f}^N_{j,t+i-P}}{\hat{f}^N_{j,t+i-P}} \right|
\]

The ratio \( x_{j,t} / x_{j,t-P} \) plays a crucial role in this formula.

To be able to further quantify the impact of this ratio on the MAD, we assume that:

- \( x_{j,t-P} \) and \( \hat{a}_{j,t-P} \) are independent and that
- deseasonalised demand is gamma distributed.

Strictly speaking, \( x_{j,t-P} \) and \( \hat{a}_{j,t-P} \) are dependent, but this dependency is very limited as \( \hat{a}_{j,t-P} \) is a 1-year centered Moving Average demand, including not only \( x_{j,t-P} \) but also the demand in \( P-1 \) other periods. So, if \( P \) is large, this dependency is very limited.
From Stuart and Ord (1994) we know: if \( X_1 \) and \( X_2 \) are gamma distributed with parameters \( \lambda \) and \( n_1 \) resp. \( n_2 \), then \( X_1 / X_2 \) is distributed according to the beta distribution function of the second type with parameters \( n_1 \) and \( n_2 \):

\[
X_1 / X_2 \sim B_2(n_1, n_2).
\]

The expectation of \( X_1 / X_2 \) exists for \( n_2 > 1 \) and equals \( n_1 / (n_2 - 1) \). Note that this expectation is independent of the scaling parameter \( \lambda \). In our case both \( X_{j,t} \) and \( X_{j,t-p} \) are gamma distributed with parameter \( n \), where \( 1/n \) is equal to the squared coefficient of variation of the deseasonalised demand per review period\(^7\): \( 1/n = CV_p^2 \).

So in case \( X_{j,t} \) is gamma distributed, we have

\[
E(X_{j,t} / X_{j,t-p}) = \frac{n}{n-1} = \frac{1}{1-CV_p^2} \quad \text{if } n > 1 \quad \text{(i.e. if } CV_p < 1)\]

We call the expectation of \( X_{j,t} / X_{j,t-p} \) the 'inflation factor', since it is the main cause that the MAD is being inflated. Figure 1 shows the inflation factor as a function of \( CV_p \). The inflation factor increases rapidly as \( CV_p \) gets close to 1.

Beyond \( CV_p = 1 \), the inflation factor no longer exists (i.e. it is infinitely high). So in case \( CV_p > 1 \) huge forecast errors may arise, when Winters’ procedure is applied.

![Figure 1](image.png)

**Figure 1** The inflation factor as a function of \( CV_p \) when demand is gamma\(^8\) distributed and Winters’ procedure is applied.

The results in Figure 1 suggest that application of Winters’ procedure is acceptable as long as \( CV_p < 0.3 \) (since then the inflation is less than 10%), and is no longer advisable if \( CV_p > 0.5 \) (since then the inflation is already more than 33%). For intermediate values Winters’ procedure should be applied with great care, i.e.

---

\(^7\) Note that in our simulations we had \( CV_p^2 = (P/48) \times CV^2 \)

\(^8\) The fact that the inflation factor may be very high does not depend on the assumption that demand is gamma distributed. If for example \( X_i \sim U[0,1] \) for \( i=1 \) and \( i=2 \), then \( E(X_1/X_2) \) already no longer exists. See Mood et al. (1974, p. 188).
preferably with very low values for the smoothing factor $\alpha^9$ and with a human intervention to check whether forecasts are still reasonable if they become very large.

The reason for the high inflation factor is the fact that every now and then $X_{t-P}$ gets very close to zero. If e.g. $X_{t-P}$ equals 0.001 times its expected value, $X_t / X_{t-P}$ will be in the order of magnitude of 1,000, leading to an extremely high $MAD$ (see formula (1)).

The reasoning above clearly highlights the parameters which have a devastating effect on the determination of the reorder point in the system: both $CV$ and $P$ have a direct impact on the probability of a low demand in any period during the initialization. $P$ also has a direct negative impact on the probability that any such low demand may occur: in case $P = 48$, there are 48 periods during the initialization in which demand may be relatively low, while if $P = 12$, there are only 12 probabilities for such low demand. Furthermore $\alpha$ and $L$ have a clear impact: the smaller $\alpha$ and $L$, the less the reorder point is inflated in case $X_t / X_{t-P}$ is very large. This explains why the optimal $\alpha$ was so low in our simulations, as will be discussed in paragraph 6.

4.3 The risk for large forecast errors when AGG is applied

As mentioned in paragraph 4.2, in case the initialization period is equal to two years, only one demand observation is available for every period in the season to estimate its seasonal index. In case AGG is applied, this implies that for the product family to which a particular product $j$ belongs, the estimate for the seasonal index in period $t-P$ is simply equal to the actual aggregate demand in period $t-P$ divided by an estimator for the level of the aggregate demand (which is equal to the centered 1-year Moving Average aggregate demand):

$$\hat{f}_{t-P} = x_{t-P} / \hat{a}_{t-P}$$

After normalization we get:

$$\hat{f}_{t-P} = \hat{f}_{t-P} \cdot \hat{c}_N$$

with $\hat{c}_N$ the correction factor needed to normalize the seasonal indices.

In case AGG is applied, then in the first year after the initialization at the end of period $t$ the estimator for the level of demand for product $j$ is equal to:

$$\hat{a}_{j,t} = \alpha \cdot x_{j,t} / \hat{f}^{N}_{t-P} + (1-\alpha) \hat{a}_{j,t-1}$$

The forecast for $i$ periods ahead is equal to

$$\hat{x}_{j,t+i} = \hat{a}_{j,t} \cdot \hat{f}^{N}_{t+i-P}$$

So, if AGG is applied, in the first year after initialization the Mean Absolute Deviation ($MAD$) for product $j$ is equal to:

---

9 this very low value of $\alpha$ casts a serious doubt on whether Winters’ procedure should be applied at all in these situations, since the procedure then hardly reacts to any structural change in the demand level.
In order to make this expression comparable to formula (1), we can rewrite it as:

\[
MAD_{j,t+1}^{AGG} = \left| x_{j,t+1} - \left[ \alpha \frac{x_{j,t}}{\hat{\alpha}_{t-p}^N} + (1 - \alpha)\hat{\alpha}_{t-1} \right] \hat{f}_{t+1-p}^N \right|
\]

(2)

The inflation factor here is \( E\left( x_{j,t} / (x_{t-p,\hat{\alpha}_{t-p} / \hat{\alpha}_{t-p}) \right) \).

Again we note that \( \hat{\alpha}_{j,t-p} / \hat{\alpha}_{t-p} \) is hardly dependent on \( x_{t-p} \) if \( P \) is large.

If we assume independence, we can rewrite the inflation factor as:

\[
E\left( x_{j,t} / x_{t-p} \right) / E\left( \hat{\alpha}_{j,t-p} / \hat{\alpha}_{t-p} \right)
\]

In case of \( Npro \) independently identically distributed products, we have

\[
E\left( \hat{\alpha}_{j,t-p} / \hat{\alpha}_{t-p} \right) = 1 / Npro
\]

So, in case AGG is applied and there are \( Npro \) products with i.i.d. demand, the inflation factor in the \( MAD \) is approximately equal to

\[
E\left( Npro x_{j,t} / x_{t-p} \right)
\]

In case AGG is applied and in case demand for all products is independently identically gamma distributed, we have

\[
X_{j,t} \sim G(\lambda, n) \quad \text{and} \quad X_{t-p} = \sum_{j=1}^{Npro} X_{j,t-p} \sim G(\lambda, Npro \cdot n).
\]

So, again using the results in Stuart and Ord (1994), the inflation factor in case AGG is applied is equal to:

\[
E\left( X_{j,t} / (X_{t-p} / Npro) \right) = \frac{v}{v-1} = \frac{Npro}{Npro - CV_p^2}
\]

(3)

since

\[
v = Npro \cdot n = Npro / CV_p^2
\]

(4)
To interpret $v$, note that the deseasonalised aggregated demand in period $t$ for $N_{pro}$ products is gamma distributed with parameters $a_i$ and $v$.

Formulas (3) and (4) can also be written as:

$$N_{pro} = \frac{I}{I-1} \cdot CV_p^2$$

(5)

with $I$ the inflation factor.

Table 1 shows the minimal size of the product-family ($N_{pro}^{10}$) which is needed to reach an inflation factor $I$, for several values of $CV_p$.

<table>
<thead>
<tr>
<th>$I$</th>
<th>$CV_p$</th>
<th>0.1</th>
<th>0.25</th>
<th>0.5</th>
<th>1.0</th>
<th>1.5</th>
<th>2.0</th>
<th>3.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.01</td>
<td>2</td>
<td>7</td>
<td>26</td>
<td>101</td>
<td>228</td>
<td>404</td>
<td>909</td>
<td></td>
</tr>
<tr>
<td>1.02</td>
<td>1</td>
<td>4</td>
<td>13</td>
<td>51</td>
<td>115</td>
<td>204</td>
<td>459</td>
<td></td>
</tr>
<tr>
<td>1.05</td>
<td>1</td>
<td>2</td>
<td>6</td>
<td>21</td>
<td>48</td>
<td>84</td>
<td>189</td>
<td></td>
</tr>
<tr>
<td>1.1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>11</td>
<td>25</td>
<td>44</td>
<td>99</td>
<td></td>
</tr>
<tr>
<td>1.2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>6</td>
<td>14</td>
<td>24</td>
<td>54</td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>7</td>
<td>12</td>
<td>27</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>8</td>
<td>18</td>
<td></td>
</tr>
</tbody>
</table>

Table 1 The minimal size of the product-family, as a function of the coefficient of variation of deseasonalized demand and the inflation factor.

One of the added values of the product-aggregation procedure (AGG) is that the probability of a large value for $X_t / X_{t-P}$ is reduced. This is due to the fact that the seasonal indices in AGG are derived from a more stable demand pattern. This helps to explain that AGG performs particularly well when $N_{pro}$ is large, when $CV$ is large and when $P$ is large.

From the analysis above it is clear that if huge forecast errors arise, they are mainly due to the error made during the initialisation of the seasonal index. One solution was proposed and evaluated here: using demand aggregation for a product-family to estimate the seasonal indices. Other potential directions to prevent these errors are: - to set a limit to the change from period to period in either the seasonal index or the reorder point or - to smooth seasonal indices based on the values in periods directly preceding or following the period in question or (similarly) - to use a transcendental model (that is a mix of sine waves) to model seasonal demand (Brown, 1981, Chapter 9) or - to use models which are specifically designed for (very) short product life cycles (see e.g. Kurawarwala (1996))

---

10 Here $N_{pro}$ is derived from equation (5) and then rounded to the first integer, which is higher than or equal to $N_{pro}$.
5. The simulation model

In addition to the theoretical analysis in paragraph 4, we also used simulation as a tool to study the performance of Winters’ forecasting procedure as well as the added value of product-aggregation. There are three reasons for this. First, the analytical results derived in paragraph 4 are based on a number of assumptions (e.g. an initialization period of two years, focussing on the \( \text{MAD} \) in the first year after initialization and assuming independence between \( x_{j,t-p} \) and \( \hat{a}_{j,t-p} \)). Secondly, in paragraph 4 the magnitude of the inflation factor was determined, but formula (1) and (2) show that the forecast error (\( \text{MAD} \)) is a function of the inflation factor and some other variables. Thirdly, a poor forecast in one particular period may have longer lasting effects, which are not included in the \( \text{MAD} \). An inventory system, for example, may order too much if the forecast was too high and as a result the inventory may well be above target for many periods after the mistake was made.

Therefore we tested the forecasting procedures in a simulation model for a single-echelon inventory system facing seasonal demand. The performance of the forecasting procedures was measured by comparing the amount of inventory on hand needed in these systems. As a benchmark we used the amount of inventory needed in case the traditional Winters’ forecasting procedure was applied at the individual product level (IND). This was compared with the amount of inventory needed when demand forecasting was based on product-aggregation (AGG).

Note that for systems with stationary demand, there is a direct link between the amount of inventory on hand needed and MSE. For non-stationary demand, however, there is no such direct link.

In the simulation model a one-product periodic review inventory system with positive leadtime and lot-sizing was simulated during 500,000 periods per simulation run. A period was either 1 week or 4 weeks (depending on the parameter ‘review period’). A year consisted of 48 weeks. Each simulation run consisted of thousands of subruns. Every subrun started with an empty system and with two years of initialization, needed for Winters’ forecasting procedure. Then during a period, which was equal to the leadtime plus the review period, the inventory system was filled with scheduled receipts and inventory on hand. From that moment on, the performance of the forecasting procedure was measured every time period (i.e. every week or every four weeks, depending on the review period) during another two years. The length of this period was deliberately restricted to two years to take into account (at least to some extent) the fact that in practice the product life cycle is limited.

The demand was modelled as a gamma distributed stochastic variable with no trend and with average demand per week equal to 100 units times a seasonal index. The gamma distribution is known to be very suitable to model demand (see Burgin (1975). The seasonal pattern is depicted in Figure 2 (the maximum expected demand is at the end of the first quarter and equals 175, whereas the minimum expected demand is at the end of the third quarter and equals 25). The coefficient of variation of demand per week was equal to a constant CV in all weeks (resulting in a relatively high (low) standard deviation of demand during the high (low) season).

In the simulations it was assumed that all products which belong to the product family which was used in AGG, have identical demand distributions.
The service level was defined as the fill rate, i.e. the percentage of demand which could be delivered from stock immediately. In all simulations the target service level was equal to 95%.

The inventory reorder strategy was \((R,s_i,nQ)\): if the inventory position was below \(s_i\) at a review moment, the inventory was raised with a quantity equal to (a multiple of) the lot-size \(Q\) in such a way that after ordering the inventory position was at or above the reorder point \(s_i\) again.

In systems with non-stationary demand both the reorder point and the safety stock norm (which is part of the reorderpoint) should be dynamic: during the high season the average demand and the standard deviation of demand are relatively high and as a result the forecast errors during those periods are also high, leading to the need of a high safety stock norm in those periods (see also Silver et al., par. 8.7, 1998).

In the simulation model, the reorder point \(s_i\) for a system with leadtime \(L\) was set equal to\(^{11}\):

\[
 s_i = (1 + m) \sum_{j=1}^{L+1} \hat{x}_{j,i+1} \\
\]

with \(m\) being a multiplier, which depends on the target service level. Typically, the reorder point is set equal to the demand forecast for the next \(L+1\) periods plus an additional safety stock, which depends on the standard deviation of the forecast error during those periods. There are three reasons why we chose to express the reorder point differently, i.e. only in terms of the demand forecasts for the next \(L+1\) periods:

1. Since the forecasting in every subrun is only done for two years, hardly any historical data are available to estimate the standard deviation of the forecast error for every individual week.
2. In the simulations the multiplier \(m\) which corresponded with a 95% service level, was often slightly negative (often somewhere between \(-0.5\) and \(0\)). The reason for this is the fact that in many simulations the expectation of (forecast minus actual demand) is substantially larger than zero. Standard theory on determination of stocknorms assumes the average forecast error is equal to zero.

\(^{11}\)In the simulations a maximum was set to this reorder point (this maximum is equal to \(100\times(L+1)\times\text{Average demand in the first period}\)) to avoid really excessive reorder points.
3. Finally, according to Silver et al., it is reasonable to assume that the standard deviation of the forecast error depends on the level of the demand forecast.

Since there are no mathematical expressions available in the literature, which link the multiplier $m$ used in this formula with the target service level, we determined this multiplier via a simple search method: the inventory system was simulated for a particular value of the multiplier. Based on the resulting service level, the multiplier was adapted. This process was repeated until the multiplier was found which yielded the target service level. In each system the multiplier $m$ was optimised for each forecasting procedure.

The parameters used in the models, i.e. $\alpha$ (the smoothing factor for the level) and $\gamma$ (the smoothing factor for the season), have been optimized by repeatedly simulating the system with every possible combination of $\alpha$ and $\gamma$, with $\alpha$ equal to 0.01, 0.05, 0.1, 0.2 or 0.4 and $\gamma$ equal to 0.05, 0.1, 0.2 or 0.5.

6. Results

To get an impression of the advantage to be obtained by forecasting based on product-aggregation, we simulated 36 different inventory systems with the following parameters:

- Coefficient of variation of deseasonalised demand per week ($CV$) = 0.1, 0.5 or 1.0
- Number of (review) periods per season ($P$) = 48 or 12. This also determines the time-bucket used in the forecasting method: if $P = 48$, the time-bucket is 1 week, while if $P = 12$, the time-bucket is 4 weeks (later on also referred to as 1 month)
- Lead time ($L$) = 4 or 16 weeks,
- Lot-size ($Q$) = 1, 400 or 1600 units (with average demand per week equal to 100 units in all systems).

The number of products in the product family ($N_{pro}$) was equal to 10.

The forecasting method based on product-aggregation (AGG) as well as the traditional forecasting method at the individual product level (IND) both use smoothing factors $\alpha$ and $\gamma$. So ideally we have to optimise these smoothing factors for every individual inventory system which is being simulated. As Silver et al. indicate, it is not very handy in practice to determine the optimal smoothing factors individually. An alternative would be to find the smoothing factors which minimise the total inventory on hand for all 36 systems mentioned above, under the restriction that in all systems the same smoothing factors are applied.

For each of the 36 inventory systems, the average inventory on hand is determined. Table 2 shows the sum of these inventories if we use IND (row 1) or AGG (row 2) and if we use individually optimized (column 1) or group-optimized (column 2) smoothing factors. This Table demonstrates that the product-aggregation procedure is very powerful. If row 1 is compared with row 2, we see that the total inventory on hand can be more than halved, while still yielding the same service level. If column 1 is compared with column 2, we see that the added value of optimising the smoothing factors for every individual system is limited: it reduces the total inventory on hand with slightly more than 1%. Apparently it suffices to do the optimisation for the group as a whole.
Table 2. The inventory on hand in case AGG or IND is applied and in case the smoothing factors are optimised per scenario or for the group as a whole. The group-optimal smoothing factors (both in AGG and IND) were: $\alpha$=0.01 and $\gamma$=0.1.

To understand the impact of optimisation over $\alpha$ and $\gamma$, we compare the average total inventory on hand for all 36 scenario’s for different values of $\alpha$ and $\gamma$. Table 3 summarises the results (the results for $\gamma$=0.05 and $\gamma$=0.1 are omitted since they provide no additional insight). It appears that optimisation over $\gamma$ has very little impact. However, choosing the wrong value for $\alpha$ may lead to quite some additional inventory.

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>$\gamma$=0.1</th>
<th>$\gamma$=0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>28,245</td>
<td>28,467</td>
</tr>
<tr>
<td>0.05</td>
<td>28,944</td>
<td>29,278</td>
</tr>
<tr>
<td>0.1</td>
<td>30,235</td>
<td>30,559</td>
</tr>
<tr>
<td>0.2</td>
<td>32,490</td>
<td>32,384</td>
</tr>
<tr>
<td>0.4</td>
<td>36,732</td>
<td>35,885</td>
</tr>
</tbody>
</table>

Table 3. The total inventory on hand for all 36 systems in case AGG is applied and the smoothing factors are ‘group-optimal’.

To get a more detailed understanding in which situation optimisation over $\alpha$ is important, we compare the average inventory on hand for several scenario’s in which AGG is applied. For each scenario we compare the situation with ($\alpha$=0.01, $\gamma$=0.1) and ($\alpha$=0.2, $\gamma$=0.1). The latter combination is a reasonable parameter setting, according to Silver et al.$^{12}$ Among the scenario’s we vary the review period, the coefficient of variation of demand per week and the leadtime (see Table 4). In all scenario’s we choose the lot-size to be equal to 400 units, since the impact of the lot-size is limited in this comparison.

From Table 4 it is clear that optimisation over the smoothing factors $\alpha$ and $\gamma$ is especially useful if the review period is small and if at the same time the standard deviation of the demand during the leadtime is very large. This is not surprising, since we know from paragraph 4 that the inflation factor is large if $CV_p$ is large, i.e. if both $CV$ and $P$ are large. Formula (1) shows that a small $\alpha$ helps to dampen the effect of the inflation factor. The worst case shows that the inventory on hand is approximately 70% too high if the smoothing factors are not optimised.$^{13}$

$^{12}$ Silver et al. also suggest to optimise $\alpha$ and $\gamma$ per product group rather than using these reasonable values for the entire assortment. In addition to Silver et al.’s suggestion we indicate the size of the possible error and in which type of environment this error is largest.

$^{13}$ The results in Table 3 hold for the situation in which AGG is applied. In case the traditional forecasting at the individual product level (IND) is applied, the inventory on hand in the same worst case scenario ($CV$=1.0, review period=1 week and L=16 weeks) is equal to 8,502 if $\alpha$=0.01 and $\gamma$=0.1 and equal to 20,752 if $\alpha$=0.2 and $\gamma$=0.1: an increase of more than 140%!
Table 4. The inventory on hand for each individual product for two different values of $\alpha$ in situations where we used AGG.

<table>
<thead>
<tr>
<th>P</th>
<th>CV</th>
<th>L</th>
<th>$\alpha = 0.01$</th>
<th>$\alpha = 0.2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\gamma = 0.1$</td>
<td>$\gamma = 0.1$</td>
</tr>
<tr>
<td>week (P=48)</td>
<td>0.1</td>
<td>4</td>
<td>247</td>
<td>248</td>
</tr>
<tr>
<td>week</td>
<td>0.1</td>
<td>16</td>
<td>276</td>
<td>280</td>
</tr>
<tr>
<td>week</td>
<td>0.5</td>
<td>4</td>
<td>356</td>
<td>395</td>
</tr>
<tr>
<td>week</td>
<td>0.5</td>
<td>16</td>
<td>553</td>
<td>864 (!)</td>
</tr>
<tr>
<td>week</td>
<td>1.0</td>
<td>4</td>
<td>635</td>
<td>837 (!)</td>
</tr>
<tr>
<td>week</td>
<td>1.0</td>
<td>16</td>
<td>1178</td>
<td>2021 (!)</td>
</tr>
<tr>
<td>month (P=12)</td>
<td>0.1</td>
<td>4</td>
<td>487</td>
<td>489</td>
</tr>
<tr>
<td>month</td>
<td>0.1</td>
<td>16</td>
<td>511</td>
<td>521</td>
</tr>
<tr>
<td>month</td>
<td>0.5</td>
<td>4</td>
<td>600</td>
<td>600</td>
</tr>
<tr>
<td>month</td>
<td>0.5</td>
<td>16</td>
<td>800</td>
<td>823</td>
</tr>
<tr>
<td>month</td>
<td>1.0</td>
<td>4</td>
<td>915</td>
<td>918</td>
</tr>
<tr>
<td>month</td>
<td>1.0</td>
<td>16</td>
<td>1481</td>
<td>1541</td>
</tr>
</tbody>
</table>

All these results indicate that in case Winters’ procedure is applied in a situation with a season, no trend and high demand uncertainty, the smoothing parameter $\alpha$ should be chosen very low.

Based on the insights we obtained above, we will assume in the remaining part of this paper that $\alpha=0.01$ and $\gamma=0.1$ are reasonable values for the environment we are studying.

Next, we are interested in the root causes for the success of AGG. Here to we will perform a sensitivity analysis for the main parameters in our model: CV, P, L, Q and Npro. We start from a base scenario with CV=1.0, P=48, L=4 weeks, Q=400 units and Npro=10 products (in case AGG is applied). In Figure 3 we first vary the demand uncertainty\(^{14}\) and demonstrate the impact on the average inventory in this base system. The aggregation procedure turns out to be especially worthwhile if demand is very uncertain. But even with moderate demand uncertainty, the aggregation procedure may already yield 10% inventory reduction.

---

\(^{14}\) In practice, demand uncertainty may be well above 1.0. Snyder et al. (2002) e.g. used a data set for 345 jewelry lines sold by The Monet Group. The coefficient of variation of weekly sales in this recent data set typically varied between 1.0 and 3.5 during low season and around 2.0 during high season.
Figure 3  The average inventory on hand for several values of CV in case the review period equals 1 week (P=48)

The same experiment with forecasting per month (P=12) results in Figure 4. If forecasting is done per month, the results are less dramatic: the aggregation procedure adds less value in such a situation. Of course this is mainly due to the fact that the coefficient of variation of demand per month is equal to only half times CV (assuming a month consists of 4 weeks). Still, if demand is relatively uncertain (CV>0.5), the aggregation procedure yields inventory reductions ranging here from 10% till 60%.

Figure 4  The average inventory on hand for several values of CV in case the review period equals 1 month (P=12)

Figure 5 shows how the size of the product-family (Npro) helps to improve the performance of AGG. It shows that already in a situation with 2 products in a product family the savings can be huge. The added value of having more products in the family decreases rapidly.
7. Conclusions

The classic forecasting method for seasonal demand, Winters’ procedure, has two main disadvantages:
- The higher the demand uncertainty, the more difficult it is for Winters’ procedure to distinguish the seasonal pattern from randomness.
- The shorter the initialization and the more uncertain demand per review period, the higher the risk that huge forecast errors arise.
Winters’ procedure was published in 1960, when demand was still relatively predictable. In the second half of the 20th century however, product assortments have grown and product life cycles have become shorter. As a result of these and other developments, the demand uncertainty has increased. Still, many textbooks on Operations Management refer to Winters’ procedure when demand is seasonal, without stressing enough these fundamental changes or without being specific on the environments in which it can be applied. The results in this paper clearly show that, if only two years of sales data are available for initialization, application of Winters’ procedure for forecasting weekly demand may result in huge forecast errors in case the coefficient of variation of the deseasonalized weekly demand is large. In case of a small initialization period and if this coefficient of variation is larger than 0.3, Winters’ procedure should already be applied with great care (e.g. by using a very small smoothing factor $\alpha$ in order to avoid large forecast errors). In case this coefficient of variation is larger than 0.5 in combination with a small initialization period, it is no longer advised to use Winters’ forecasting procedure in its original form.

For situations with high demand uncertainty, short-term forecasting can be improved by first aggregating demand for a product family, consisting of products having similar seasonal patterns. Winters’ method is applied on this aggregated demand to find the seasonal indices, and next these indices are used as an exogenous variable, when forecasts are made at the individual product level. The aggregation concept in this forecasting procedure helps to reduce the (consequences of) forecast errors substantially, particularly when the coefficient of variation of deseasonalized demand is large, when the time-bucket in forecasting is small and when the leadtime is large. The improvements can already be very substantial if the product family consists of only 2 products. The added value of larger product-families depends on the coefficient of variation of the deseasonalised demand. In this paper formulae are derived, which give an indication of the appropriate size of the product-family as a function of the demand uncertainty and the desired forecasting accuracy.

**Literature**


Everdell, R. (1984), Master Scheduling. APICS publication.


