State Space Models for Seasonal Aggregation in Sales Forecasting

Steve Djajasaputra* Tom Heskes
Department of Computer Science, University of Nijmegen
Toernooiveld 1, 6525 ED, Nijmegen, The Netherlands

Pim Ouwehand
Department of Technology Management, TU Eindhoven
P. O. Box 513, 5600 MB, Eindhoven, The Netherlands

June 9, 2005

Keywords:
Seasonality; State-space models; Sales forecasting; Seasonal aggregation; Combining seasonality; Kalman filter; Holt-Winters method

Abstract

This paper describes a way to improve forecasts by simultaneously forecasting a group of products that exhibit a similar seasonal pattern. There have already been several previous publications that demonstrated forecast improvements using seasonal aggregation. However, these papers focused on various *ad hoc* methods to combine seasonal indices from aggregated time series. Instead, we develop state space models in which aggregation is naturally incorporated. Our primary contribution is the seasonal aggregation extension of the Harvey's dummy seasonal model and the trigonometric seasonal model. Using sales data, we show the possible improvement of forecasting accuracy using these aggregation models, compared with forecasting individual time series. The empirical results suggest that the *truncated* harmonic trigonometric seasonal aggregation model (a trigonometric seasonal model with a reduced number of harmonics) is the most promising approach to perform seasonal aggregation forecasting in terms of forecast accuracy and computational cost.

1 Introduction

Sales forecasting is a crucial issue in business, reliable forecasts may save a lot of money on production and logistic planning. Seasonality is one of the distinctive characteristics of sales data; a typical example is the summer peaks in soft drinks sales. Therefore, having reliable estimates of seasonality is important in sales forecasting. With product life-cycles becoming shorter, it is getting more difficult to obtain reliable estimates of seasonal effects due to the smaller amount of data available. By aggregating a group of products, more data become available to improve the estimates of seasonal effects.

This paper describes a way to improve forecasts by simultaneously forecasting a group of time series which exhibit a similar seasonal pattern. The potential benefit of aggregation, however, may not only be restricted to sales applications but may also be relevant to many other forecasting problems such as econometric modeling (e.g. in subsection 3.2 we discuss aggregation on US industrial production data).

*A draft submitted to the International Journal of Forecasting 2005. Corresponding author. Tel.: +31-24-3652632; fax: +31-24-3653356; *e-mail address:* steved@cs.ru.nl (Steve Djajasaputra)

The possibility of forecast improvement using aggregated sales data has already been reported in several previous publications, for example Dalhart (1974), Withycombe (1989), Bunn (1999) and Dekker et al. (2004). However, these papers focused on various *ad hoc methods* to combine seasonal indices from aggregated time series. Instead, we develop state space *models* in which aggregation is naturally incorporated. First, we collect several time series which share a common seasonal pattern into a group. In our aggregation model, these time series share the same *joint seasonal states*. In this way, we aggregate the seasonality of these time series. This modeling approach also offers several advantages: it makes the underlying assumptions more explicit and easily extendable.

The aggregation models may yield better predictions since they have lower *overfitting* risks compared with non-aggregation models. Overfitting is the phenomenon that an over-complex model captures too many details of the data (including noise that it is fitted on), which is not relevant for prediction. Consequently, the model performs well on the training data used for parameter estimation, but performs badly in out-of-sample prediction. On the other hand, a simple model (but with enough complexity to capture the underlying process generating the data) does not capture all of the details in the training data, thus will be less likely to capture irregularities due to noise. Hence, a simple model has less risk of overfitting and might perform better in out-of-sample prediction. Seasonal aggregation leads to a simpler model since a group of time series shares common seasonal states, while a non-aggregation model has seasonal states for each time series. So we can expect that an aggregation model has less risk of overfitting and thus might give better predictions.

This paper is organized as follows. In section 2, we start with a review of two seasonal state space representations based on a basic structural model (BSM) for a single time series as described by Harvey (1989), i.e. the dummy seasonal model (HS henceforth) and the trigonometric/Fourier-harmonics seasonal model (FS henceforth). We also consider the *truncated* harmonics trigonometric seasonal models (TS henceforth), i.e. trigonometric seasonal model with a reduced number of harmonics. Our primary contribution is the seasonal-aggregation extension of these models which is described in subsection 2.2: the dummy seasonal aggregation (HA), the trigonometric/Fourier-harmonic seasonal aggregation (FA) and the truncated harmonic seasonal aggregation (TA). We show the possible improvement in forecasting accuracy using HA (section 3), FA and TA (section 4). In section 5, we compare our probability approach with other aggregation methods. Finally, after we discuss the conclusions (section 6), we discuss possible extensions of our models as future research in section 6. Throughout this paper bold letters denote vectors and capital letters denote noise parameters (Q, R) and matrices.

2 Models and methods

In this section, we specify the *non aggregation* seasonal models (HS, FS, TS) and the *aggregation* seasonal models (HA, FA, TA) using a general framework: the linear state-space matrix form

$$y_t = C\mathbf{x}_t + \epsilon_t^{(y)}, \tag{1}$$

$$\mathbf{x}_t = A\mathbf{x}_{t-1} + \boldsymbol{\epsilon}_t^{(x)}, \tag{2}$$

where $\mathbf{x}_t$ is a state vector that comprises the level and the seasonal effects and $\boldsymbol{\epsilon}_t^{(x)}$ is the noise vector for this state. The transition matrix A governs how this state vector evolves (e.g. according to equation (3) in case of the dummy seasonal model). C is an observation matrix, which maps the state vector to the observation y_t. In this section, different seasonal models (HS, FS, TS, HA, FA, TA) are specified using different specifications for their transition matrix A, the state space transition equation for $\mathbf{x}_t$, and the observation matrix C.

A detailed discussion about these structural time series models are given by Harvey (1989, p. 40-42) and Durbin and Koopman (2001, p. 38-42). Proietti (2000) provides an insightful comparison of different seasonal representations, including the dummy seasonal model and trigonometric seasonal model.

In the standard BSM by Harvey (1989), one decomposes time series into local levels, seasonalities and trends. In this paper, we do not incorporate trends in our models since we do not observe any trends in our sales data and we want to have simple models that focus on the seasonality. Nevertheless, a generalization to models with trend is straightforward.

2.1 Modeling seasonality using state-space models for a single time series

2.1.1 Dummy seasonal model

Harvey and Todd (1983) describe a seasonal time series as follows:

$$\begin{aligned} \text{Observation equation:} \quad & y_t = l_t + s_t + \epsilon_t^{(y)}\,, \qquad \epsilon_t^{(y)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}[0, R]\,, \\ \text{Transition equations:} \quad & l_t = l_{t-1} + \epsilon_t^{(l)}\,, \qquad \epsilon_t^{(l)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}[0, Q^{(l)}]\,, \\ & s_t = -\sum_{j=1}^{p-1} s_{t-j} + \epsilon_t^{(s)}\,, \quad \epsilon_t^{(s)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}[0, Q^{(s)}]\,. \end{aligned} \tag{3}$$

Here t denotes the time (in our sales data typically measured in weeks), l_t denotes the level, s_t denotes "dummy" seasonal effects which sum to zero when added over the entire period p. y_t represents the variable that we observe (e.g. the logarithm of sales figures). The $\epsilon_t^{(y)}, \epsilon_t^{(l)}, \epsilon_t^{(s)}$ are noise terms independently drawn from Gaussian distributions with zero mean and stationary variances R, $Q^{(l)}$ and $Q^{(s)}$ respectively. Rewriting (3) in the matrix form (2), (1), the $p \times 1$ state vector is $\mathbf{x}_t = [l_t, s_t, \ldots, s_{t-p+2}]^T$ and its $p \times 1$ noise state vector is $\boldsymbol{\epsilon}_t^{(x)} = [\epsilon_t^{(l)}, \epsilon_t^{(s)}, 0, \ldots, 0]^T$. The $p \times p$ transition matrix A is given by

$$A = \left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & A_s \end{array}\right], \text{ with } A_s = \begin{bmatrix} -1 & \cdots & -1 & -1 \\ 1 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 \end{bmatrix}. \tag{4}$$

The $p-1 \times p-1$ transition matrix A_s deals with the seasonal part and the "1" in the upper left of A deals with the level part. The transition equation (2) of HS then obeys

$$\mathbf{x}_t = \begin{bmatrix} l_t \\ \hline s_t \\ s_{t-1} \\ \vdots \\ s_{t-p+2} \end{bmatrix} = \left[\begin{array}{c|ccc} 1 & 0 & \cdots & 0 \\ \hline 0 & -1 & \cdots & -1 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{array}\right] \begin{bmatrix} l_{t-1} \\ \hline s_{t-1} \\ s_{t-2} \\ \vdots \\ s_{t-p+1} \end{bmatrix} + \begin{bmatrix} \epsilon_t^{(l)} \\ \hline \epsilon_t^{(s)} \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \tag{5}$$

which is equivalent to (3) above.

The $1 \times p$ observation matrix is $C = [1, C_s]$, where $C_s = [1, 0, \ldots, 0]$ is the $1 \times p-1$ observation matrix for the seasonal part and the "1" in the C is for the level part.

2.1.2 Trigonometric/Fourier-based seasonal model

Another way to express seasonality is by using a trigonometric/Fourier expansion to represent seasonal effects, as described in West and Harrison (1997, p. 247-253) and Harvey (1989, p. 41-42). The state vectors in the trigonometric seasonal model (FS) consist of the level l_t and Fourier harmonic component pairs $\mathbf{f}_{r,t}$: $\mathbf{x}_t = [l_t, \mathbf{f}_{1,t}, \ldots, \mathbf{f}_{p/2,t}]^T$ where

$$\mathbf{f}_{r,t} = \begin{cases} \begin{bmatrix} c_{r,t} & c_{r,t}^* \end{bmatrix}^T & \text{if } r \neq p/2 \\ c_{p/2,t} & \text{if } r = p/2\,. \end{cases}$$

$c_{r,t}$ and $c_{r,t}^*$) are the r^{th} harmonic components at time t. In FS, the $p \times p$ transition matrix is given by

$$A = \left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & A_s \end{array}\right], \text{ with } A_s = \begin{bmatrix} G_1 & 0 & \cdots & 0 \\ 0 & \ddots & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & G_{p/2} \end{bmatrix}, \tag{6}$$

where

$$G_r = \begin{cases} \begin{bmatrix} \cos(\frac{2\pi}{p}r) & \sin(\frac{2\pi}{p}r) \\ -\sin(\frac{2\pi}{p}r) & \cos(\frac{2\pi}{p}r) \end{bmatrix} & \text{if } r \neq p/2\,, \\ 1 & \text{if } r = p/2\,, \end{cases}$$

is the transition matrix for the harmonic pairs $\mathbf{f}_{r,t}$. Hence, the transition equation (2) of FS reads

$$\mathbf{x}_t = \begin{bmatrix} l_t \\ \mathbf{f}_{1,t} \\ \vdots \\ \mathbf{f}_{p/2,t} \end{bmatrix} = \left[\begin{array}{c|ccc} 1 & 0 & \cdots & 0 \\ \hline 0 & G_1 & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & G_{p/2} \end{array}\right] \begin{bmatrix} l_{t-1} \\ \mathbf{f}_{1,t-1} \\ \vdots \\ \mathbf{f}_{p/2,t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_t^{(l)} \\ \boldsymbol{\epsilon}_{1,t}^{(f)} \\ \vdots \\ \boldsymbol{\epsilon}_{p/2,t}^{(f)} \end{bmatrix}, \tag{7}$$

where

$$\boldsymbol{\epsilon}_{r,t}^{(f)} = \begin{cases} \begin{bmatrix} \epsilon_{r,t}^{(c)} & \epsilon_{r,t}^{(c)*} \end{bmatrix}^T & \text{if } r \neq p/2\,, \\ \epsilon_{p/2,t}^{(c)} & \text{if } r = p/2\,. \end{cases} \tag{8}$$

$\epsilon_{r,t}^{(c)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}[0, Q^{(f)}]$ and $\epsilon_{r,t}^{(c)*} \overset{\text{i.i.d.}}{\sim} \mathcal{N}[0, Q^{(f)}]$) are the noise in the harmonic components.

In FS, the $1 \times p$ observation matrix is $C = [1, C_s]$, where "1" in the C is for the level part and $C_s = [\mathbf{c}_1 \ \ldots \ \mathbf{c}_{p/2}]$ is for the seasonal part. The $\mathbf{c}_r$ in the C_s, which maps the r^{th} harmonic to the observation y_t, is:

$$\mathbf{c}_r = \begin{cases} \begin{bmatrix} 1 & 0 \end{bmatrix} & \text{if } r \neq p/2, \\ 1 & \text{if } r = p/2\,. \end{cases}$$

It is also possible not to use all the $p/2$ harmonic pairs $\mathbf{f}_{r,t}$, we name this variant of FS: the *truncated* harmonics trigonometric seasonal model (TS). This means that we can omit some harmonic pairs $\mathbf{f}_{r,t}$ from (7) and also their corresponding $\mathbf{c}_r$ in C, G_r in A, and $\boldsymbol{\epsilon}_{r,t}^{(f)}$ in (7). Leaving out harmonics reduces the dimensionality of the seasonal state, which can be advantageous since it is computationally less demanding and it has a lower risk of overfitting (which might lead to a better forecast accuracy). In section (4), we use TS (and TA) to investigate the forecast accuracy as we vary the number of harmonics.

2.2 State-space models for seasonal aggregation

We assume that as several time series within a group share a common seasonality characteristic, we can apply a structural model where these n individual time series $\mathbf{y}_i = [y_{i,1}, \ldots, y_{i,L}]$ *share* the *same* seasonal states. (Here, L is the length of the time series and i is the time series index.) For example, we propose the following seasonal aggregation model that is based on the dummy seasonal model (HA):

$$\begin{aligned} \text{Observation equation:} \quad & y_{i,t} = l_{i,t} + s_t + \epsilon_{i,t}^{(y)}\,, \quad \epsilon_{i,t}^{(y)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}[0, R]\,, \\ \text{Transition equations:} \quad & l_{i,t} = l_{i,t-1} + \epsilon_{i,t}^{(l)}\,, \quad \epsilon_{i,t}^{(l)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}[0, Q^{(l)}]\,, \\ & s_t = \sum_{j=1}^{p-1} s_{t-j} + \epsilon_t^{(s)}\,, \quad \epsilon_t^{(s)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}[0, Q^{(s)}]\,. \end{aligned} \tag{9}$$

Hence, each time series has its own level $l_{i,t}$ but shares the same seasonal effects s_t with other time series within a group. In the same spirit, we can apply this concept for the seasonal aggregation using the trigonometric model (FA) and the truncated harmonics trigonometric model (TA).

In the general state space form, the state vector $\mathbf{x}_t$ in these aggregation models consists of the level states $\mathbf{x}_t^{(l)} = [l_{i,t}, \ldots, l_{n,t}]^T$ of the time series $y_{i,t}$ (for $i = 1...n$) and the *shared* seasonal states, which is $\mathbf{x}_t^{(s)} =$

$[s_t, \ldots, s_{t-p+2}]^T$ in HA or $\mathbf{x}_t^{(s)} = [\mathbf{f}_{1,t}, \ldots, \mathbf{f}_{p/2,t}]^T$ in FA (and TA with omission of some harmonics $\mathbf{f}_{r,t}$). All of these seasonal aggregation models (HA, FA, TA) have the transition matrix

$$A = \left(\begin{array}{c|c} I_n & 0 \\ \hline 0 & A_s \end{array} \right),$$

where I_n is an n-dimensional identity matrix for the level $\mathbf{x}_t^{(l)}$ and A_s is the transition matrix for the shared seasonal part $\mathbf{x}_t^{(s)}$, which is (4) for HA or (6) for FA (and TA with omission of some G_r). Using the specifications above, the state space equation of HA obeys

$$\mathbf{x}_t = \begin{bmatrix} l_{1,t} \\ \vdots \\ l_{n,t} \\ \hline s_t \\ s_{t-1} \\ \vdots \\ s_{t-p+2} \end{bmatrix} = \left[\begin{array}{c|cccc} I_n & 0 & \cdots & 0 & 0 \\ \hline 0 & -1 & \cdots & -1 & -1 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{array}\right] \begin{bmatrix} l_{1,t-1} \\ \vdots \\ l_{n,t-1} \\ \hline s_{t-1} \\ s_{t-2} \\ \vdots \\ s_{t-p+1} \end{bmatrix} + \begin{bmatrix} \epsilon_{1,t}^{(l)} \\ \vdots \\ \epsilon_{n,t}^{(l)} \\ \hline \epsilon_t^{(s)} \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

where the noise term for the level states is $\epsilon_{i,t}^{(l)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}[0, Q^{(l)}], \forall i$ and the noise term for the seasonal states is $\epsilon_t^{(s)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}[0, Q^{(s)}]$.

The state space equation for FA obeys

$$\mathbf{x}_t = \begin{bmatrix} l_{1,t} \\ \vdots \\ l_{n,t} \\ \hline \mathbf{f}_{1,t} \\ \vdots \\ \mathbf{f}_{p/2,t} \end{bmatrix} = \left[\begin{array}{c|ccc} I_n & 0 & \cdots & 0 \\ \hline 0 & G_1 & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & G_{p/2} \end{array}\right] \begin{bmatrix} l_{1,t-1} \\ \vdots \\ l_{n,t-1} \\ \hline \mathbf{f}_{1,t-1} \\ \vdots \\ \mathbf{f}_{p/2,t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{1,t}^{(l)} \\ \vdots \\ \epsilon_{n,t}^{(l)} \\ \hline \boldsymbol{\epsilon}_{1,t}^{(f)} \\ \vdots \\ \boldsymbol{\epsilon}_{p/2,t}^{(f)} \end{bmatrix}, \tag{10}$$

where $\epsilon_{i,t}^{(l)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}[0, Q^{(l)}]$ is the same as that described above for HA and $\boldsymbol{\epsilon}_{r,t}^{(f)}$ is the same as (8) that described earlier in subsection 2.1.2. The state vector $\mathbf{x}_t$ and the state space equation for TA are similar to FA (10) but with omission of some harmonics $\mathbf{f}_{r,t}$ (and their corresponding G_r).

The observation equation for seasonal aggregation models (HA,FA,TA) is:

$$\mathbf{y}_t = \begin{bmatrix} y_{1,t} \\ \vdots \\ y_{n,t} \end{bmatrix} = C\mathbf{x}_t + \begin{bmatrix} \epsilon_{1,t}^{(y)} \\ \vdots \\ \epsilon_{n,t}^{(y)} \end{bmatrix}, \text{ with } \epsilon_{i,t}^{(y)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}[0, R], \tag{11}$$

The observation matrix is given by $C = [I_n, \mathbf{1}_n \times C_s]$, where $\mathbf{1}_n$ is an $n \times 1$ summing vector. Here, I_n deals with the level part and $\mathbf{1}_n \times C_s$ deals with the shared seasonal part, where C_s is the same as described in subsection 2.1.1 for HA or subsection 2.1.2 for FA (and TA).

To simplify the model, we employ the same noise parameters R for $\epsilon_{i,t}^{(y)}$ and $Q^{(l)}$ for $\epsilon_{i,t}^{(l)}$ across all time series $\mathbf{y}_i$. As a preprocessing step, we standardize all time series to have the same variance, which might help to scale all the $\epsilon_{i,t}^{(y)}$ to have the same magnitudes. The same goes for $\epsilon_{i,t}^{(l)}$.

2.3 Initialization

The initial hidden state $\mathbf{x}_1$ consists of the level part $\mathbf{x}_1^{(l)}$ and the seasonal part $\mathbf{x}_1^{(s)}$. We assume that $\mathbf{x}_1$ is Gaussian, so that in this case we only need to compute its mean and covariance. In this subsection, we

describe how we obtain the mean of $\mathbf{x}_1$ (by computing the means of $\mathbf{x}_1^{(s)}$ and $\mathbf{x}_1^{(l)}$) and then how to compute the covariance of $\mathbf{x}_1$.

To find the initial seasonal mean $\hat{\mathbf{x}}_1^{(s)} = E(\mathbf{x}_1^{(s)})$, first we make a seasonal decomposition to obtain a seasonal factor $z_{i,b}$, by subtracting the p-order centered moving average $[y_{i,b-(p/2)}, \ldots, y_{i,b+(p/2)-1}]$ from $y_{i,b}$, as suggested by Makridakis et al. (1998). In our implementation, b runs from $p/2+1$ to $3p/2$ in the initialization region, thus after time-index reordering we have a $n \times p$ seasonal factor matrix $Z_{=}[\mathbf{z}_1, \ldots, \mathbf{z}_n]^T$ with $\mathbf{z}_i = [z_{i,t}, \ldots, z_{i,p}]$. Because we assume a joint seasonality, we take the average of Z across all time series i to get a $1 \times p$ *joint* seasonal factor $\mathbf{j} = [j_1, \ldots, j_p]$. To transform the joint seasonal factor $\mathbf{j}$ to the seasonal effect mean $\hat{\mathbf{x}}_t^{(s)}$, we make use of the observability matrix of the seasonal components $O_s = [C_s^T, (C_s A_s)^T, \ldots, (C_s A_s^{p-1})^T]^T$. It provides a mapping between the seasonal state mean and the observed seasonal factor $\mathbf{j}$: $O_s\, \hat{\mathbf{x}}_t^{(s)} = \mathbf{j}^T$. The initial state mean then follows from $\hat{\mathbf{x}}_1^{(s)} = O_s^{-1}\, \mathbf{j}^T$ where O_s^{-1} is the inverse of the observability matrix O_s.

To find the initial seasonal mean $\hat{\mathbf{x}}_1^{(l)} = E(\mathbf{x}_1^{(l)})$, we compute $\hat{l}_{i,1} = E(l_{i,1})$ as $\hat{l}_{i,1} = y_{i,1} - C_s \hat{\mathbf{x}}_t^{(s)}$, where C_s is the observation matrix for the seasonal part.

For the covariance of $\mathbf{x}_1$, we use the diffuse prior method as described by Durbin and Koopman (2001).

We found that the seasonal states in our model are sensitive to the initialization procedure. Although the Kalman filter adjusts the seasonal states every time slice (every week in our sales data context), effectively the adaptation of seasonal states is not as fast as the adaptation of the level states. The level states adapt quite fast, hence the level states are less sensitive to the initialization procedure.

For inference and prediction, we make use of standard Kalman filter from Kalman (1960). Given the mean and variance of the initial hidden state $\mathbf{x}_1$, the observed data $y_{i,t}$ (for $i = 1...n, t = 1...a$) and the noise parameters (R, $Q^{(l)}$, $Q^{(s)}$ for HA/HS or R, $Q^{(l)}$, $Q^{(f)}$ for FA/FS/TA/TS), it computes the mean and variance of the unobserved $\mathbf{x}_a$ and then predicts the mean and variance of $y_{i,a+1}$ and beyond.

3 Forecasting with the dummy seasonal aggregation model

3.1 Sales forecasting using seasonal aggregation

We use weekly log sales data of soft-drinks and beers from the distribution centers of Schuitema, one of the biggest supermarket chains in the Netherlands. We choose 52 weeks as the period. The data consist of 260 weeks, in which the first 104 weeks were used for initialization ($2p$ weeks are needed for the moving average procedure explained in subsection 2.3), the weeks 105-207 were used to estimate the model parameters (training session) and the weeks 208-260 for out-of-sample forecast accuracy measurement.

Our aggregation models work with a group of time series that share a similar seasonal pattern. We chose this group using a hierarchical clustering method based on the Euclidean distance and the group average linkage measure as described in e.g. Hand et al. (2001). Figure 1 shows the clustering result on the seasonal components obtained by p-order moving-average additive seasonal decomposition as described by Makridakis et al. (1998). Notice that the peaks around the summer (week 27, 79, 131, ...) and new year (week 53, 105, ...) indicate strong seasonality patterns in these sales data. Based on this clustering of seasonal components, we chose a large aggregation group consisting of 9 soft drinks as marked by the ellipse in Figure 1. One can also choose other clusters based on Figure 1. Since in this paper we only aim to demonstrate the benefit of aggregation *given* a group of time series, we leave a more detailed discussion and analysis of how to cluster for further research. We also chose a small aggregation group of beers using the same clustering method.

Optimization algorithms (e.g. simplex search or sequential quadratic programming) are used to find the best estimate of model parameters (R, $Q^{(l)}$, $Q^{(s)}$ for HA/HS or R, $Q^{(l)}$, $Q^{(f)}$ for FA/FS/TA/TS) using different cost functions (MAD, MSE, MAPE and likelihood) measured in the parameter estimation region. We use MAPE (which is averaged over all time series) to measure the forecast accuracy in the out-of-sample (test) region since MAPE is a scale independent error measure and it is widely used in the forecasting literature. To measure the benefit of aggregation (compared to the non aggregation prediction), we define MAPE percent benefit of aggregation as

$$\text{MAPE percent benefit} = 100\% \, \frac{(\text{MAPE}_{\text{non aggregation}} - \text{MAPE}_{\text{aggregation}})}{\text{MAPE}_{\text{non aggregation}}} \, . \tag{12}$$

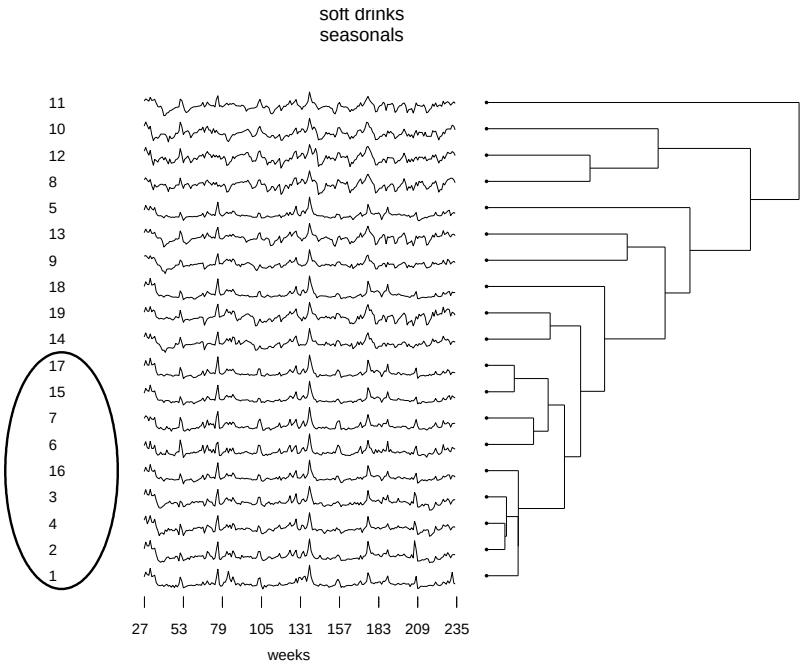

Figure 1: A hierarchical clustering on the seasonal components of log soft drink sales. The data consists of 19 soft drink products (as indicated by the numbers 1..19 in the left). The length of the raw data is 260 weeks (5 years), but the seasonal components run from week 27 to week 234 due to moving-average seasonal decomposition procedure we use. The clustering is based on the seasonal components from week 27 to week 183.

Both $\text{MAPE}_{\text{aggregation}}$ and $\text{MAPE}_{\text{non aggregation}}$ are averaged over all the time indices and all the time series within a group. We estimate the standard deviation of MAPE percent benefit using block bootstraps as proposed by Kunsch (1979), a variant of bootstrapping procedure by Efron (1979) for non i.i.d. sequences.

Table 1 shows the MAPE percent benefits of HA along with their standard deviations for soft drink and beer sales. These results confirm the possibility of forecast improvement using this aggregation model. As a comparison, in Table 1 we also show the MAPE percent benefit of another aggregation method using multiplicative Holt-Winters (HWA henceforth) developed by the third author [Ouwehand et al. (2004)]. There are some similarities between our models and the HWA approach. According to Chatfield et al. (2001), the additive Holt-Winters method is fairly similar to the BSM. Furthermore, the multiplicative approach (as used in HWA) on non-log data is similar to the log-additive approach (additive models on log data) considered in this paper (see e.g. Chatfield (2004, p. 14)). Therefore, the multiplicative Holt-Winters on non-log data in HWA is similar to the BSM on log data which we use here. Comparing the performance of our models with HWA, which model yields better accuracy strongly depends on the data and the error criterion used.

3.2 Seasonal aggregation on US industrial production data

Following a suggestion by Keogh and Kasetty (2002), we compare our method using data which are publicly available for benchmarking. For this purpose, we use US industrial production data, which is publicly available from the Federal Reserve Board's web site: `http://www.federalreserve.gov/`. These monthly data have a period of 12 months and span from 1986 to 2003: the years 1986-1987 are used for initialization, 1987-1998 for parameter estimation and 1999-2003 for out-of-sample forecast evaluation. Although the data consist of a huge amount of product families of time series, most of the product families (e.g. aluminum, steel) are not as close as the product families in the sales data (e.g. coca cola, pepsi) to construct an aggregation group. Using the clustering procedure described in subsection 3.1 above, we chose an aggregation group consisting of "pig iron" and "raw steel". These results in Table 1 again confirm the possibility of forecast improvement

using HA on the data.

	Dummy seasonal aggregation (HA)				Holt-Winters aggregation (HWA)		
Data	MAD	MSE	likelihood	MAPE	MAD	MSE	MAPE
Soft drink sales	5.6±2.3	3.7±2.6	5.9±3.3	5.1±2.6	8.7±5.0	10.9±4.6	7.4±4.7
Beer sales	17.2±4.1	9.0±4.0	18.5±4.7	7.0±3.7	6.6±6.4	10.1±6.6	5.8±6.4
US industrial production	4.6±3.4	4.2±3.2	5.5±2.9	4.2±1.9	-0.3±3.7	4.6±3.0	3.9±2.8

Table 1: MAPE percent benefits of aggregation using dummy seasonal model (HA) and Holt-Winters's aggregation (HWA) using prediction horizon 1 and different cost functions for model parameter estimation.

4 Forecasting with the trigonometric seasonal aggregation model

Figure 2 shows a comparison of the forecast accuracy and the aggregation benefits of TA and TS, as we vary the number of harmonics for a fixed cluster group of soft drinks and beers mentioned earlier. Notice that in this figure, a TA/TS with $p/2 = 26$ harmonics corresponds to a full-harmonic model (FA/FS) and a TA/TS with 0 harmonics corresponds to a model without seasonality (level model). We tried different cost functions for model parameter estimation to investigate the robustness of our study.

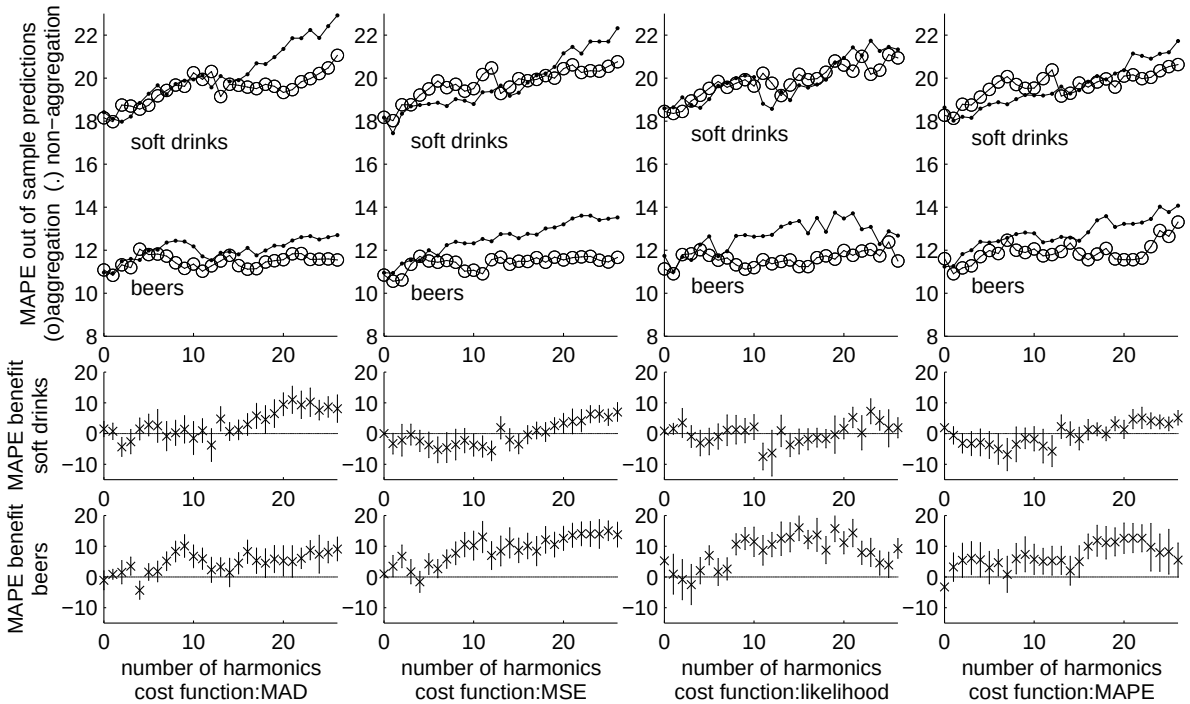

Figure 2: Comparison of forecast accuracy (MAPE) (without standard deviations for the clearness of presentation) and MAPE percent benefit (with its standard deviation) for truncated trigonometric seasonal models with aggregation (TA) and without aggregation (TS), as a function of the number of Fourier harmonics for soft drinks and beers data. The results are shown for different cost functions for model parameters estimation (MAD, MSE, likelihood and MAPE) with prediction horizon 1

The number of harmonics was varied as follows: first, we ranked the harmonics based on their averaged

amplitudes obtained using FA (for TA) or FS (for TS) in the parameter-estimation region. In an n-number of harmonics TA/TS, we then only retained the n-highest rank harmonics and discarded the other harmonics.

We found that the predictions with horizon=1 in Figure 2 are strongly influenced by the *level* components rather than the seasonal components. Consequently, it is difficult to study the effect of varying the number of *seasonal* harmonics with this figure since the influence of the levels overshadows the seasonal effects. Therefore we did other experiments with TA and TS but now with longer prediction horizons. Figure 3 shows the forecast accuracy and the aggregation benefit as a function of the number of harmonics for different prediction horizons, using soft drinks data with MAD as the cost function for parameter estimation using one-step-ahead prediction. As we increase the prediction horizon, the level influence becomes less significant and the seasonal influences become stronger. Hence, the seasonal pattern and the aggregation benefit are more noticeable as the horizon increases (as long as the prediction horizon is not too long, which makes the noise dominant over the seasonality pattern). However as the horizon increases, the overall prediction errors also increases (especially for low numbers of harmonics) and the figures become more noisy due to the increase of uncertainty.

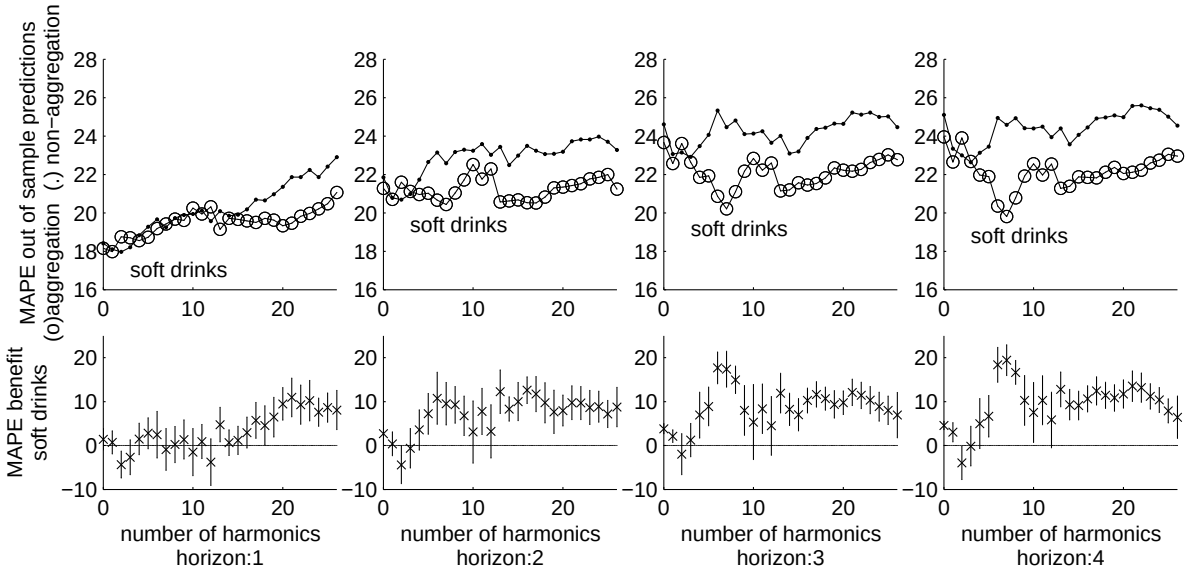

Figure 3: Comparison of forecast accuracy (MAPE) and MAPE percent benefit using different prediction horizons (1, 2, 3, 4) for truncated trigonometric seasonal models with aggregation (TA) and no aggregation (TS), as a function of the number of Fourier harmonics for soft drinks sales.

Figures 2 and 3 indicate that as we leave out a harmonic from TA/TS, the prediction error can either decrease or increase. The error may increase if the truncated model omits an important harmonic contained in the data, thus decreasing its explaining capability. Conversely, the error might decrease since a truncated harmonic model leads to a simpler model which has less risk of overfitting. Another way to see this mechanism is that a truncated harmonic model performs a kind of time averaging which smoothes irregularities. These irregularities might not only be due to a careless data-entering process, but also to exogenous factors (e.g. a special event: beer sales peaks as many people drink together while watching an important football match) or time warping/shifting of influential factors (e.g. the weather). Figure 4 illustrates this smoothing mechanism: the seasonal reconstruction from a truncated harmonics model is smoother than the seasonality of a full harmonic model. Hence, a truncated model will less overshoot and will be less sensitive to irregularities, and thus might yield better forecast accuracy. The resulted error pattern in Figures 2 and 3 is a combination of these increasing and decreasing factors, which strongly depends on the data at hand.

Figures 2 and 3 show that we hardly observe aggregation benefit for low numbers of harmonics, possibly since in this region a non aggregation model is already simple enough to avoid overfitting, thus making such a model simpler by aggregation will hardly give any added value. The benefit becomes close to zero as the

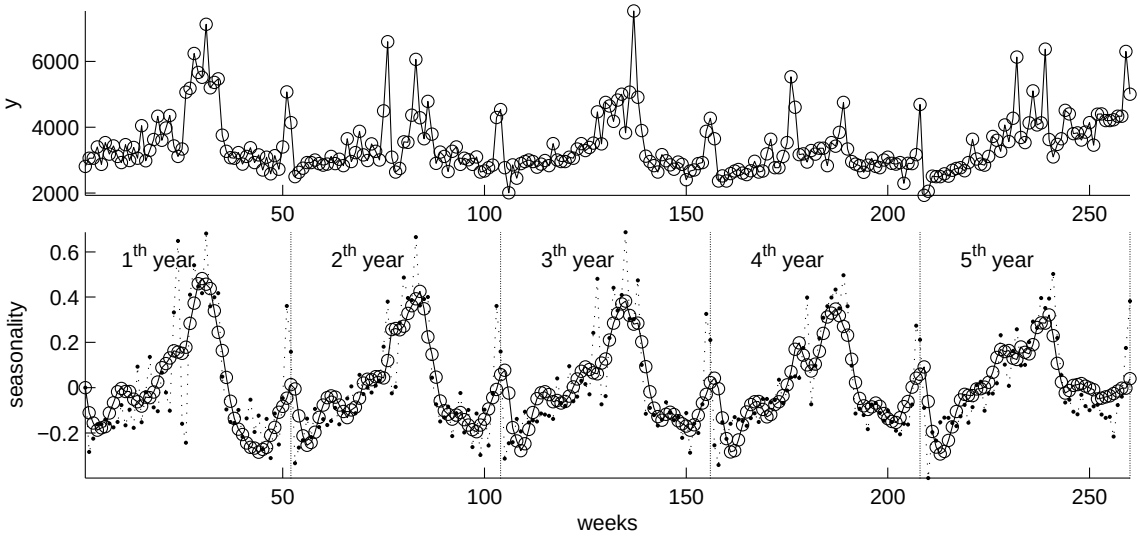

Figure 4: A comparison of (.) the seasonality from a full-harmonics seasonal model and (o) a truncated model with number of harmonics is 4 (out of 26). This figure shows that the truncated model smooths the irregularities, however the number of harmonics used in this example (TA using 4 harmonics) might be too small to capture the actual underlying seasonality in the data.

number of harmonics approaches zero (i.e. the model without seasonality).

In summary, Figures 2 and 3 show that for both aggregation and non-aggregation cases, the truncated harmonics trigonometric models (TA,TS) often yield better *forecast accuracy* than the full harmonics trigonometric models (FA,FS). Comparing all of our aggregation models, Table 2 shows that TA often yield better *aggregation benefit* than HA (horizon 1) and FA (especially for horizon 3 and 4).

Data	Horizon	HA	FA	The best TA
Soft drink sales	1	5.6±2.3	8.1±4.5	10.9±4.4 (using 21 harmonics)
	2		8.7±4.5	12.3±4.9 (using 13 harmonics)
	3		6.9±5.2	17.4±4.1 (using 7 harmonics)
	4		6.4±4.8	19.4±3.5 (using 7 harmonics)

Table 2: Comparison of MAPE percent benefits of all our aggregation models: HA (extracted from Table 1, only for horizon 1), FA (extracted from Figure 3 at 26 harmonics), and the best TA (extracted from Figure 3).

5 Comparing our probabilistic modeling approach with other aggregation methods

By analyzing the mean predictions and mean updates of the Kalman filter of our aggregation model, we can write the updating equations of our aggregation model and compare them with those of the other seasonal aggregation methods based on seasonal indices averaging, e.g. Dalhart (1974) or Holt-Winters aggregation by Dekker et al. (2004). The joint seasonality approach in our aggregation model can be seen as an averaging over the seasonality of grouped time series. Thus this is somewhat related to the seasonal aggregation method proposed by Dalhart (1974), where he averaged the seasonal indices from individual time series to construct a joint "group seasonal index." It appears that the updating equation of the level variables in our aggregation model for each time series has correction terms (i.e. Kalman gains) associated with the level variables from

the other time series. In other words, level variables of these time series become coupled to each other as we join their seasonal variables. This subtle phenomenon becomes visible naturally in probabilistic modeling but is absent in the *ad hoc* aggregation methods which take the average of the seasonal indices.

Another advantage of the probabilistic approach is that the probabilistic modeling approach offers a principled way to handle missing data, instead of using *ad hoc* ways (e.g. filling missing sales data with zeros). However, the other aggregation methods may have less computational demands than probabilistic models (i.e. it is computationally cheaper to run the Holt-Winters's algorithm and then take the average of their seasonal indices than running a Kalman filter in an aggregation model).

6 Conclusions and possible extensions

This paper deals with an effort to improve forecasts by simultaneously forecasting a group of products that exhibit a similar seasonal pattern. We develop state space models in which aggregation is naturally incorporated. Our primary contribution is the seasonal aggregation extension of the BSM: the Harvey's dummy seasonal model and the trigonometric seasonal model.

We show the possibility of improvement of forecasting accuracy using these aggregation models, compared with forecasting individual time series. The truncated harmonic aggregation models are computationally less demanding and often yield better aggregation benefits compared with the dummy-seasonal aggregation models and the full-harmonic trigonometric aggregation models. These suggest that the truncated harmonic aggregation model (with an appropriate number of harmonics) is the most promising approach to perform seasonal aggregation forecasting in terms of prediction accuracy and computational cost compared with the other aggregation models.

In this paper, we use *off line* hierarchical clustering, which can be time consuming and may not be practical for huge datasets encountered in industrial practices. Besides, the seasonality might change with time, thus the forecaster might need to redo this inspection every once in a while. The clustering result itself depends on many factors such as the clustering method, the linkage method, distance measure and how to extract seasonal effects used for clustering inputs. It takes time to optimize these factors by trial and error. Therefore, we are trying to include the clustering mechanism in the model instead of doing clustering off line.

In this paper we use a *hard clustering* approach, i.e. each time series has to be assigned to a single cluster. This assumption might be too strict; as a result our models do not tolerate any mistakes in the clustering decision. A possible improvement is a *soft clustering* approach that assigns each time series to all of the clusters along with degrees of cluster assignments. With this approach, the aggregation models will be more tolerant to the misclustering due to the changing of seasonality patterns or due to the "fuzzy" nature of clustering itself. Currently, we are developing an aggregation model that incorporates the "soft" clustering using a probabilistic framework.

We use a *hard aggregation* approach in this paper, which assumes that each time series within an aggregation group has *exactly the same* the seasonal effect. This assumption also might be too strict, it could be better if we use a *soft aggregation* approach by allowing each of the individual time series to have its own *individual* seasonal states besides the *shared* seasonal states.

In section 4, we mentioned about the problem with exogenous factors. One possible solution is to build a model that incorporates exogenous variables. Gaffney (2004) proposed a probabilistic clustering model with a time alignment to deal with the irregularities due to the shifting and warping of the influential factors (e.g. weather) discussed in section 4.

7 Acknowledgments

The authors would like to thank Onno Zoeter for valuable discussions and Prof. Eamonn Keogh who provided the basic source code to plot the hierarchical clustering. We appreciate the help from Schuitema and OPG for providing sales data. This research is supported by Dutch Science and Technology Foundation (STW).

References

Bunn, D.W. & Vassilopoulos, A. I. (1999). Comparison of seasonal estimation methods in multi-item short-term forecasting. *International Journal of Forecasting 15*, 431–443.

Chatfield, C. (2004). *The Analysis of Time Series*. London: Chapman & Hall.

Chatfield, C., A. Koehler, J. Ord, and R. Snyder (2001). A new look at models for exponential smoothing. *The Statistician 50*, 147–159.

Dalhart, G. (1974). Class seasonality a new approach. *American Production and Inventory Control Society 1974 Conference Proceedings*, 11–16.

Dekker, M., M. van Donselaar, and P. Ouwehand (2004). How to use aggregation and combined forecasting to improve seasonal demand forecasts. *International Journal of Production Economics 90*(2).

Durbin, J. and S. Koopman (2001). *Time series analysis by state space methods*. Oxford: Oxford University Press.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics 7*(1), 126.

Gaffney, S. (2004). *Probabilistic curve-aligned clustering and prediction with regression mixture models*. Ph. D. thesis, Department of Computer Science, University of California, Irvine.

Hand, D., H. Mannila, and P. Smyth (2001). *Principles of Data Mining*. The MIT Press.

Harvey, A. (1989). *Forecasting, structural time series models and the Kalman filter*. Cambridge: Cambridge University Press.

Harvey, A. and P. Todd (1983). Forecasting economic time series with structural and box-jenkins models: a case study. *Journal of Business and Economic Statistics 1*(4), 313–315.

Kalman, R. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME Journal of Basic Engineering 82*(Series D), 35–45.

Keogh, E. and S. Kasetty (2002). On the need for time series data mining benchmarks: a survey and empirical demonstration. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 102–111. ACM Press.

Kunsch, H. R. (1979). The jackknife and the bootstrap for general stationary observations. *Annals of Statistics 17*(4), 12171241.

Makridakis, S., S. Wheelright, and R. Hyndman (1998). *Forecasting: Methods and Applications*. New York: John Wiley & Sons.

Ouwehand, P., K. van Donselaar, and A. de Kok (2004). The impact of the forecasting horizon when forecasting with group seasonal indices. *under review*.

Proietti, T. (2000). Comparing seasoal components for structural time series models. *International Journal of Forecasting 16*, 247–260.

West, M. and P. Harrison (1997). *Bayesian Forecasting and Dynamic Models*. New York: Springer.

Withycombe, R. (1989). Forecasting with combined seasonal indices. *International Journal of Forecasting 5*, 547–552.