# A unified approach for hierarchical time series forecasting

Marco A. Villegas, Diego J. Pedregal*

*ETSI Industriales. University of Castilla-La Mancha. 13071 Ciudad Real, Spain.*

**Abstract**

In this paper an approach for hierarchical time series forecasting based on State Space modelling is proposed. Previous developments provide solutions to the hierarchical forecasting problem by algebra manipulations based on forecasts produced by independent models for each time series involved in the hierarchy. The solutions produce optimal reconciled forecasts for each individual forecast horizon, but the link along time that is implied by the dynamics of the models is completely ignored. Therefore, the novel approach in this paper improves upon past research at least in two key points. Firstly, the algebra is already encoded in the State Space system and the Kalman Filter algorithm, giving an elegant and clean solution to the problem. Secondly, the State Space approach is optimal both across the hierarchy, as expected, but also along time, something missing in past developments. In addition, the present approach provides an unified treatment of *top-down*, *bottom-up*, *middle-out* and *reconciled* approaches reported in the literature; it generalizes the optimization of hierarchies by proposing *combined* hierarchies which integrate the previous categories at different segments of the hierarchy; and it allows for multiple hierarchies to be simultaneously adjusted. The approach is assessed by comparing its forecasting performance to the existing methods, through simulations and using real data of a Spanish grocery retailer.

*Keywords:* Forecasting, Hierarchical forecasting, Reconciliation, State Space, Forecast combination.

## 1. Introduction

Business and economic applications relying on time series are usually aggregated across geographic or logical dimensions to form hierarchical structures. The data is then forecast at different levels of aggregation which have been reported as beneficial in terms of robustness and accuracy. However, independent forecasts of all time series at all levels of aggregation have the undesired property of being inconsistent across the hierarchy. Therefore, a strategy for reconciling forecasts across aggregation levels to improve consistency is required.

Research on hierarchical time series have been reported in the literature for several decades. Examples of early studies include Grunfeld and Griliches (1960); Orcutt et al. (1968); Shlifer and Wolff (1979);

*Corresponding author. Tel.:+34 (9)26 295 430.

*Email addresses:* marco.villegas@uclm.es (Marco A. Villegas), diego.pedregal@uclm.es (Diego J. Pedregal)

Barnea and Lakonishok (1980); Gross and Sohl (1990); Dangerfield and Morris (1992); Fliedner (1999); Weatherford et al. (2001); etc. In supply chain scenarios, for example, disaggregated demand data is usually available for every shop or distribution center. At this level, natural hierarchies can be formed aggregating data corresponding to the criteria of interest. Logical hierarchies can be formed when data are grouped based on relevant criteria to the business. Families (or groups) of products in grocery industries, for example, have an important role as a business variable for supply chain orchestration (Fliedner and Lawrence, 1995; Muir, 1979). *Synthetic* hierarchies can be defined simply by attending to the nature of time series themselves. In these cases, time series are first classified into groups according to their structural components so as to optimize modeling, and then are aggregated to form the next level of the hierarchy (Fliedner, 2001).

A number of different approaches for optimizing hierarchical forecasts have been proposed in recent years. The *bottom-up* approach entails forecasting the lowest level in the hierarchy and propagating the forecasts upwards according to the hierarchy structure (Orcutt et al., 1968; Dangerfield and Morris, 1992). The *top-down* strategy forecasts the uppermost level time series and then disaggregate down the forecasts to lower levels using different approaches, normally *proration* or by ratios of every single time series with respect to the aggregate, as described by Gross and Sohl (1990); Strijbosch et al. (2008); Boylan (2010). *Middle-out* approach combines both previous approaches by forecasting a middle level and then aggregating *bottom-up* as well as disaggregating *top-down*. *Optimal reconciliation* is proposed as a way of modifying individual forecasts and making them compatible, without the need to fix any of them on a priori grounds, the procedure has been also extended to temporal hierarchies (Hyndman et al., 2011, 2016; Athanasopoulos et al., 2017). Recently, an allegedly new procedure appeared, known as *integrated hierarchies*, based on multivariate State Space systems (Pennings and van Dalen, 2017).

Debate as to which approach is best in forecasting terms has been going on for some time. Some studies are in favor of *top-down* approaches (Grunfeld and Griliches, 1960; Gross and Sohl, 1990; Fliedner, 1999), others prefer the *bottom-up* strategy (Orcutt et al., 1968; Dangerfield and Morris, 1992; Zellner and Tobias, 2000; Weatherford et al., 2001), while others advocate a *reconciled* approach (Hyndman et al., 2011, 2016). Some more theoretically oriented studies have focused on analyzing the conditions under which approach produces more accurate forecasts than others (Shlifer and Wolff, 1979; Lütkepohl, 1984; Widiarta et al., 2009; Sbrana and Silvestrini, 2013; Rostami-Tabar et al., 2016).

At present, discussions remain inconclusive and will probably persist over the following decades, for several reasons. Firstly, empirical comparisons suggest that there is no single approach that outperforms the rest, as the results depend on the complexity of the hierarchy, the quality of the data at each level, the accuracy of the models, etc. Secondly, setting out general conditions on which any hierarchy is better optimized on theoretical grounds presupposes ideal conditions that may or may not be met in experiments with real data, thus weakening importance for day to day practice. Finally, discussions are becoming more complicated because of the emergence of novel ways to optimize a hierarchy (for example, the *reconciled* approach). This paper seeks to take the debate a step further by proposing a new strategy to carrying out all the aggregation approaches, and for generalizing a hierarchy by defining each node type.

In this paper, a State Space (SS) approach for the general treatment of hierarchies is proposed, providing a framework that gives an unified solution for *top-down*, *bottom-up*, *middle-out* and *reconciled* hierarchical forecasting approaches. This framework permits a new category, *combined*, defined as any feasible logical combination of the aforementioned categories. In this framework, all the previous cases are seen as particular examples of a *combined* hierarchy, in which each node is defined in any of such categories. In other words, our approach to the problem would consist on defining the hierarchy structure and the type of each single node.

This approach improves upon recent research (mainly Hyndman et al., 2011, 2016) at least in two fundamental ways. Firstly, the appropriate SS form and the associated optimal recursive algorithms provide an elegant optimal solution to the hierarchical forecasting problem without the need to turn up to any additional algebra. Secondly, given the recursive nature of the solution to the state estimation problem given by the Kalman Filter, the optimality is propagated along time, i.e., the solution preserves the time consistency implied by the individual models for each time series. This fact is completely disregarded in previous studies.

Time consistency is, on the contrary, taken into account in Pennings and van Dalen (2017). But, though developed also in a State Space framework, the approach is rather different and more restricted than the one in this paper. Certainly, it consists of a multivariate Unobserved Components or Structural model *a la Harvey* (Harvey, 1989) with components arbitrarily constrained, in which the hierarchical restrictions are taken into account solely in a *bottom-up* manner. The univariate modelling does not play any rule and no dynamical models are proposed for aggregated time series. In this paper, other models than Unobserved Components are possible, all *bottom-up*, *top-down* and *reconciled* structures are possible, dynamical models are specified for aggregated series and time consistency is preserved.

The layout of the paper is as follows. Section 2 states the problem of forecasting a hierarchy. Section 3 reviews SS fundamentals. Section 4 shows how a general hierarchy may be set up in the SS framework. Several reflections emerge from the results of a numerical simulation (Section 5) and the experimentation on real data of a Spanish grocery retailer (Section 6). The paper concludes with a final discussion and afterthoughts in Section 7.

## 2. Hierarchical forecasting

The problem of hierarchical forecasting consists in finding optimal forecasts for all the time series involved in a hierarchy in such a way that the forecasts add up according to the constraints imposed by the hierarchy. There are many types of hierarchies and even more ways to organize the same data into different hierarchies. The total sales of a company could be disaggregated first into regions and then in product families, or vice-versa, leading to different hierarchies in each case. Figure 1 shows an example of a simple hierarchy, composed of three levels and eight nodes or time series, five of which are at the bottom level, as in Hyndman et al. (2016).

A hierarchy may be represented as a multivariate model with a number of constraints. However, relevant hierarchies are assumed to be large, making their multivariate representation infeasible. Thus, a more

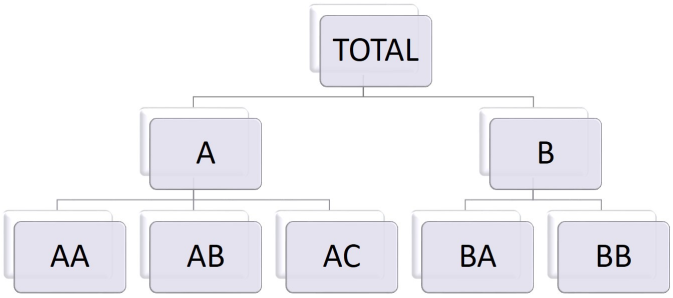


Figure 1: An example of a simple hierarchy

appropriate approach consists in identifying univariate models for each time series at every level in the hierarchy, and achieving forecast reconciliation in a second stage.

Any hierarchy or grouping data structure may be expressed as a linear transformation of the bottom level variables. The ensuing matrix transforms the data or the forecasts at the bottom layer into the whole data structure in a straightforward manner. Taking $y_{node}$ the data at any *node* in the hierarchy of Figure 1, such transformation is

$$\begin{pmatrix} y_{TOTAL} \\ y_A \\ y_B \\ y_{AA} \\ y_{AB} \\ y_{AC} \\ y_{BA} \\ y_{BB} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} y_{AA} \\ y_{AB} \\ y_{AC} \\ y_{BA} \\ y_{BB} \end{pmatrix} \tag{1}$$

or simply $y_t = Sb_t$, where $y_t$ is the vector of data or forecasts of all the time series, $S$ is the summing matrix (Hyndman et al., 2011) defining the hierarchy and $b_t$ is the data or forecasts at the bottom level.

## 3. State Space modelling

The general linear Gaussian SS model implemented in this paper is shown in Equation (2).

$$\begin{aligned} \text{Transition equation:} \quad & \alpha_{t+1} = T\alpha_t + R\eta_t, \quad \eta_t \sim N(0, Q) \\ \text{Observation equation:} \quad & y_t = Z\alpha_t + C\epsilon_t, \quad \epsilon_t \sim N(0, H) \end{aligned} \tag{2}$$

In these equations $\alpha_t$ is a non-observable state vector of length $n$; $\eta_t$ and $\epsilon_t$ are the state and observational vectors of zero mean Gaussian noises, with dimensions $r \times 1$ and $h \times 1$, respectively; both are considered independent or each other along this paper, but in general applications may be allowed to be correlated with covariance $S = Cov(\eta_t, \epsilon_t)$ of dimension $r \times h$; $y_t$ is the $k \times 1$ vector of output data. The initial state vector is assumed to be stochastic with Gaussian distribution, i.e., $\alpha_1 \sim N(a_1, P_1)$, and independent of all data and noises involved in the system.

The remaining elements in (2) are the so called system matrices with appropriate dimensions, i.e.:

$$\begin{array}{llllll} T: & n \times n; & R: & n \times r; & Q: & r \times r; \\ Z: & k \times n; & C: & k \times h; & H: & h \times h. \end{array}$$

All the system matrices may be time varying, but they are not explicitly considered here simplify the notation.

Further sophistication may be introduced in both equations by means of additional matrices taking into account either linear or non-linear input-output relationships. These particular configurations with a number of extensions not quoted here (such as non-gaussian and non-linear models) are implemented in the **SSpace** toolbox written by the authors for the MATLAB environment (The MathWorks, Inc, 2018; Villegas and Pedregal, 2018). This toolbox is based on several references, like Harvey (1989), Young et al. (1999), Durbin and Koopman (2012), Casals et al. (2016). Other pieces of software related to this are, among many others, Commandeur et al. (2011), Koopman et al. (2009), Taylor et al. (2007), Gómez (2015).

Many standard univariate techniques considered for forecasting actually fit into the present framework. Typical examples are ExponenTial Smoothing (ETS) or ARIMA models, but others not so common in this area are also implementable, like Unobserved Components (UC) models (Harvey, 1989; Durbin and Koopman, 2012; Young et al., 1999). In addition, SS systems offer an exceptional framework in which all the hierarchical approaches, *bottom-up*, *top-down*, *middle-out* and *reconciled*, fit in quite naturally. Even any logical mixture of these approaches, combined with any mixture of models, may be used simultaneously across any hierarchy, as described in the next section. This allows to use the most convenient model and approach for each part of the hierarchy. To sum up, SS methodology provides the possibility of a unified treatment of all possible cases.

### *3.1. The Kalman Filter*

It is well-known that the Kalman Filter (KF) is a recursive algorithm that provides the optimal estimation of the states and their covariances of a SS system at any point in time by minimizing the mean squared error, conditional on all information available up to that point in time (Kalman, 1960). There have been many formulations of the recursive equations, of which the decomposition in prediction and correction steps is especially illustrative in the present context. The algorithm is shown in Equation (3), where $a_t = E(\alpha_t|Y_{t-1})$, $a_{t|t} = E(\alpha_t|Y_t)$, $P_t = Var(\alpha_t|Y_{t-1})$, $P_{t|t} = Var(\alpha_t|Y_t)$, $Y_{t-1}$ and $Y_t$ stand for all information available up to $t-1$ or $t$, respectively. The variable $v_t$ is known as the innovation (with time covariance matrix $F_t$), the information that cannot be predicted from all the information available at the previous step.

$$\begin{array}{rlrl} & \textit{Correction step:} & & \textit{Prediction step:} \\ v_t &= y_t - Za_t & & \\ F_t &= ZP_tZ' + CHC' & & \\ K_t &= P_tZ'F_t^{-1} & & \\ a_{t|t} &= a_t + K_t v_t & a_{t+1} &= Ta_{t|t} \\ P_{t|t} &= P_t - K_tF_tK_t' & P_{t+1} &= TP_{t|t}T' + RQR' \end{array} \tag{3}$$

Additionally, the forecast for the output data is produced by $y_{t+1} = Za_{t+1}$. Given an appropriate initialization of the KF, i.e. appropriate values for $a_1$ and $P_1$, the algorithm proceeds recursively applying the correction equation after the prediction equation for $t = 1, 2, \cdots, N$.

Correction equations are skipped if missing data is present. Forecasts are the natural outcome of the KF if the output data is set to missing information, because the correction equations do not apply and the prediction equations are applied recursively. The issue of initialization, especially for non-stationary models, is a complex one for which the solution applied in this paper and implemented in **SSpace** is the exact initial KF as in Durbin and Koopman (2012).

The KF assumes all system matrices are known. However, in real situations this is very rarely the case. Therefore, prior to the application of the KF, estimation of the unknown parameters should be carried out. There are many ways to do this task, of which Maximum Likelihood in time domain via error decomposition is very often preferred, because of its good statistical properties in very general situations. See details in Durbin and Koopman (2012), also Young et al. (1999) for frequency domain approaches.

### *3.2. An example*

A simple example is proposed here in order to illustrate the SS approach. Let's consider an ARMA(2, 1) model in Equation (4).

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t + \theta \epsilon_{t-1}, \quad Var(\epsilon_t) = \sigma^2 \tag{4}$$

The SS representation of this model is not unique, but one that is often used in the literature is

$$\begin{aligned} \begin{pmatrix} \alpha_{1,t+1} \\ \alpha_{2,t+1} \end{pmatrix} &= \begin{pmatrix} \phi_1 & 1 \\ \phi_2 & 0 \end{pmatrix} \begin{pmatrix} \alpha_{1,t} \\ \alpha_{2,t} \end{pmatrix} + \begin{pmatrix} 1 \\ \theta \end{pmatrix} \epsilon_t \\ y_t &= \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} \alpha_{1,t} \\ \alpha_{2,t} \end{pmatrix} \end{aligned} \tag{5}$$

Matching system (5) with the general SS system (2) identifies the system matrices for this particular case, i.e.,

$$T = \begin{pmatrix} \phi_1 & 1 \\ \phi_2 & 0 \end{pmatrix}; \quad R = \begin{pmatrix} 1 \\ \theta \end{pmatrix} \epsilon_t \quad Q = \sigma^2$$
$$Z = \begin{pmatrix} 1 & 0 \end{pmatrix} \qquad C = 0 \qquad H = 0$$

Matrices $T$, $Q$ and $R$ depend on the parameters of the ARMA(2,1) model and given the data may be estimated by Maximum Likelihood. With the estimated values placed in these matrices, the KF in Equation (3) gives the optimal estimation of the states, forecasts and their covariances. Generalizations of this example to ARMA($p, q$) models is straightforward.

## 4. Hierarchies in SS form

The problem stated in Section 2 may be solved by building an overall SS system that comprises all the models in a hierarchy with the constraints imposed by the hierarchy structure and all the information involved. It is assumed that the base models have been identified previously in some way, either automatically or manually. In order to build the overall SS structure, such models ought to be cast in SS form, though not necessarily identified within a SS framework.

The overall SS system representing the hierarchy is built in two steps:

- Build up the overall system by block concatenation.
- Add hierarchical constraints depending on the type of hierarchy.

### *4.1. Build overall system*

Assuming that the univariate models for all variables are set up in SS form, the model for the $i$-th variable may be written as in Equation (2) with a superscript $(^{(i)})$ added to every single element in the SS formulation. Such superscript runs from 1 to the number of variables $k$.

The first step, then, consists of building an overall SS system based on $k$ systems of type (2) ($k = 8$ in example of Figure 1). This is possible by block concatenation of the individual systems, see Equation (6).

$$\begin{pmatrix} \alpha_{t+1}^{(1)} \\ \vdots \\ \alpha_{t+1}^{(k)} \end{pmatrix} = \begin{pmatrix} T^{(1)} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & T^{(k)} \end{pmatrix} \begin{pmatrix} \alpha_{t}^{(1)} \\ \vdots \\ \alpha_{t}^{(k)} \end{pmatrix} + \begin{pmatrix} R^{(1)} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & R^{(k)} \end{pmatrix} \begin{pmatrix} \eta_{t}^{(1)} \\ \vdots \\ \eta_{t}^{(k)} \end{pmatrix}$$
$$\begin{pmatrix} y_{t}^{(1)} \\ \vdots \\ y_{t}^{(k)} \end{pmatrix} = \begin{pmatrix} Z^{(1)} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & Z^{(k)} \end{pmatrix} \begin{pmatrix} \alpha_{t}^{(1)} \\ \vdots \\ \alpha_{t}^{(k)} \end{pmatrix} + \begin{pmatrix} C^{(1)} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & C^{(k)} \end{pmatrix} \begin{pmatrix} \epsilon_{t}^{(1)} \\ \vdots \\ \epsilon_{t}^{(k)} \end{pmatrix} \tag{6}$$
$$Cov \begin{pmatrix} \eta_{t}^{(1)} \\ \vdots \\ \eta_{t}^{(k)} \end{pmatrix} = \begin{pmatrix} Q^{(1)} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & Q^{(k)} \end{pmatrix}; \quad Cov \begin{pmatrix} \epsilon_{t}^{(1)} \\ \vdots \\ \epsilon_{t}^{(k)} \end{pmatrix} = \begin{pmatrix} H^{(1)} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & H^{(k)} \end{pmatrix}$$

Once more, this system is in form (2). Given its block diagonal structure, using system (6) on the whole dataset is exactly equivalent to using the $k$ individual SS systems independently.

### 4.2. *Add hierarchical constraints*

Let's rename all the system matrices in (6) in compact form as their counterparts of system (2) with superscript ($^{(K)}$) added to them. Then, the second step consists in finding out a modified observation equation that replicates all the univariate models and the constraints imposed by the hierarchy at the same time. This is achieved by pre-multiplying the overall observation equation by a matrix $S^{(*)}$, given in Equation (7).

$$S^{(*)} = \begin{pmatrix} S^{(K)} & S \end{pmatrix} \tag{7}$$

Matrix $S^{(*)}$ is actually a combination of matrix $S$ that defines the hierarchy structure and a new matrix $S^{(K)}$ that has to be found depending of the type of system (*bottom-up*, etc.). The dimension of $S^{(K)}$ is $k \times m$, where $m$ is the number of aggregated variables.

- *Bottom-up*: $S^{(K)} = 0$. Pre-multiplying the observation equation by $S^{(*)}$ in this case leaves the system as it is. In this case, forecasts may be produced simply as the aggregation of individual forecasts without any need of the overall system. A more efficient version of this model, with less states, is found by declaring $S^{(K)}$ as an empty matrix and using an overall system that only comprises the models at the bottom level. Indeed, this is the preferred form when the system is purely *bottom-up*. However, this less efficient version is preferred for the sake of the *combined* systems defined below.

- *Top-down*: as in the case of *bottom-up*, but inserting the forecasts obtained by the individual model of $y_{TOTAL}$ as actual data. As in the previous case, the *top-down* system may be defined more efficiently with less states, but this form is retained.

- *Reconciled*: $S^{(K)} = \begin{pmatrix} -I & 0 \end{pmatrix}'$, where $I$ is an identity matrix of a dimension equal to the number of aggregated variables ($m$) and the rest are a block of zeros of appropriate dimension ($(k-m) \times m$). In this case, $y_t^{(K)}$ is a composition of zeros for the aggregated data and the disaggregated series $b_t$.

- *Combined*: the way each case above is set up allows for their combination in a single hierarchy by selecting appropriate values for the diagonal elements of matrix $S^{(K)}$ and the data affecting it. For example, by setting $S^{(K)}(i,i) = -1$ and zeros for the $i-th$ aggregate variable that node is chosen as *reconciled*. Otherwise, set $S^{(K)}(i,i) = 0$ and missing data or the forecasts of the univariate model for $i-th$ aggregated series, for *bottom-up* or *top-down* options, respectively. One must bear in mind that the *middle-out* case quoted in the literature is just a particular case of a *combined* hierarchy. Therefore, rather than speaking about hierarchy types (i.e., *bottom-up*, *reconciled*, etc.) a thorougher nomenclature would be to talk of node types.

### *4.3. Discussion*

The path followed to build the SS hierarchical system illustrates how the transition equation embodies the dynamical systems of all the independent univariate models, while the observation equation just defines both the hierarchy structure and node types. Thus, the two-step KF in Equation (3) renders the optimal one-step-ahead forecasts of the states: in the prediction step, which only uses the transition equation system matrices, the independent forecasts are computed regardless of the hierarchy, while the correction step updates them by imposing the appropriate constraints according to the hierarchy structure. Such a correction is done by minimizing the Mean Squared Error, conditional on all information available at any time, on the univariate models for all series and on the constraints imposed by the hierarchy.

An additional benefit of this approach is that it neatly takes advantage of the KF optimality without any further assumptions or any further algebra. In the particular case of the *top-down* systems, there is no need to elaborate on how the top forecasts have to be disaggregated (Gross and Sohl, 1990). The advantages also apply to *reconciled* systems, in which additional objective functions ought to be assumed, opening the door to different options (Hyndman et al., 2011, 2016).

Still more important is the fact that any hierarchical forecasts are produced based on past optimal forecasts obtained with the hierarchy and the forecasts are consistent both hierarchical-wise and time-wise, i.e., with the constraints imposed by the hierarchy and the time dynamics of the univariate models optimally identified. Clearly this is not the case of the Hyndman et al. (2011) approach, where the reconciled forecasts are produced based solely on the initial univariate/independent forecasts without any reference to the previous reconciled forecasts, losing the consistency across the temporal dimension. The *top-down* approach has similar consistency problems along time.

The following simple example illustrates the point of time consistency of the SS reconciled forecasts. Let's assume two independent AR(1) processes,

$$\begin{aligned} x_{1,t} &= 0.8x_{1,t-1} + a_{1,t}, \qquad var(a_{1,t}) = 1 \\ x_{2,t} &= -0.8x_{2,t-1} + a_{2,t}, \qquad var(a_{2,t}) = 1 \end{aligned} \tag{8}$$

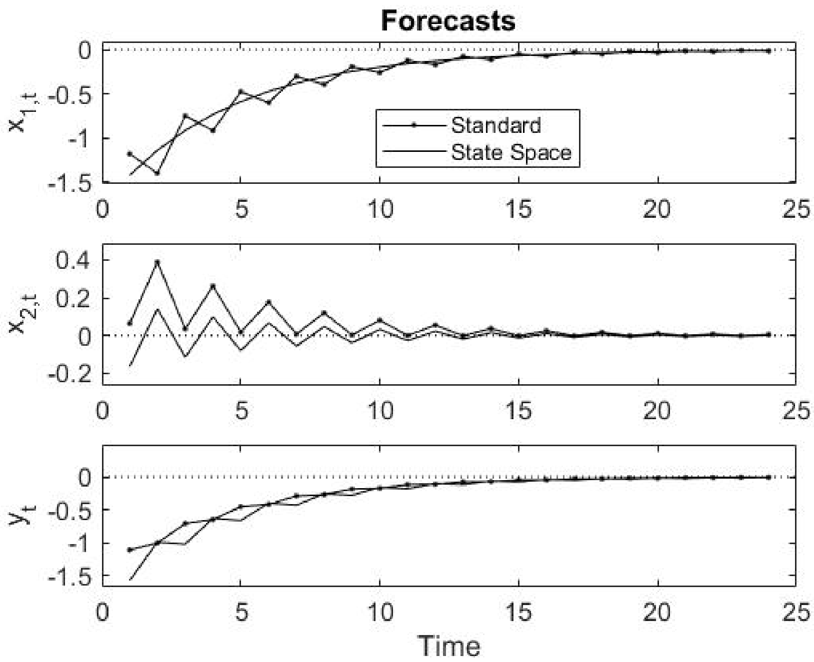


Figure 2: Forecasts of a simulation of aggregation of two AR(1) systems

It is well known that the eventual forecast function of both processes decay exponentially to zero, the first one from either the positive or negative side, while the second alternates sign around zero. The aggregation of the two processes ($y_t = x_{1,t} + x_{2,t}$) is an AR(2) because the coefficients of the two AR(1) processes are equal and opposite in sign and the variance of both noises is the same (Granger and Morris, 1976). A simulation of the above models produces forecasts like the ones shown in Figure 2. The standard procedure is done in the R environment with the **hts** package (Hyndman et al., 2011), while the SS forecasts are done in MATLAB with the **SSpace** toolbox (Villegas and Pedregal, 2018). The example is done in a way such that the base AR models are exactly the same in both cases, making sure that the observed differences are strictly and solely due to the aggregation procedures. The SS reconciled forecasts resemble what is expected in theory, while the standard procedure produces forecasts that are combinations of the expected. In this sense, SS approach is time-consistent, while the standard method computed with **hts** package is not.

The possibilities of *combined* systems open up the door to further considerations. Taking the hierarchy in Figure 1, the adjusted forecasts will be different depending on the type assigned to each node. Figure 3 shows just a few possibilities that naturally differ from those in which all the nodes are of the same type. Then, for the same hierarchy there are many different ways to produce forecasts, all optimal from the point of view of minimizing the mean squared error, but with different sets of information/constraints. The selection of each node type may be carried out by a lengthy empirical process, resembling data mining procedures, or by imposing constraints based on previous experience. Not every possible combination is logical, but the number of logical combinations is certainly very high, increasing with the number of nodes and the depth and width of the hierarchy. This is an area that requires more research and is beyond the scope of this paper.

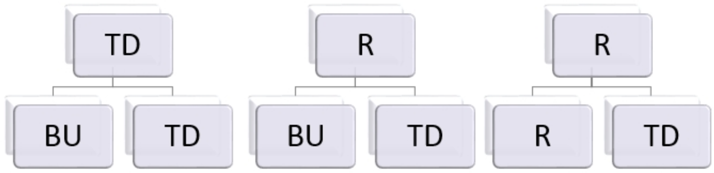


Figure 3: Some of the possible options for the top structure of the hierarchy in Figure 1. TD, BU and R stand for *top-down*, *bottom-up* and *reconciled*, respectively.

Kahn (1998) mentioned the idea of a hybrid approach combining the advantages of *top-down* and *bottom-up*. As previously discussed, our approach provides a framework that takes this idea of flexibility to a new extent, allowing to assign a combination strategy to each node in the hierarchy. Moreover, further considerations are:

- The proposed approach makes it easy to integrate multivariate models as part of the hierarchy. Instead of a block concatenation of univariate models, the corresponding block would be replaced by the appropriate multivariate module.
- Judgmental forecasts may be incorporated into the hierarchy at any node by selecting them as a *top-down* node.
- Node types may be also time dependent, keeping the hierarchy and time consistency and building forecasts on past optimal forecasts based on the hierarchy structure. A system might be *bottom-up* at some points in time and *top-down* at others, meaning that by means of the latter the forecasts are forced to reach a certain level at certain time stamps.
- Simulations are also possible, for example, it is possible to set scenarios to answer questions such us what future paths are compatible with certain future values at certain nodes.
- Optimal confidence bands or prediction intervals may be estimated automatically from the standard KF output in the usual way.

## 5. Simulations

In this section we develop a simulation study to compare the proposed methodology with existing alternatives. The simulation process is basically a replication of the one described in Hyndman et al. (2011) with minor modifications. All bottom level series are simulated as ARIMA$(p, d, q)$ processes with $d = 1$, $p \in (0, 2)$ and $q \in (0, 2)$ with parameter values chosen randomly with equal probability within the stationary and invertible regions. The hierarchy structure remains the same, i.e., 3 levels, with eight series at bottom-level that are aggregated by pairs to form the four series at level 2, two series at level 1, and the most aggregated series at the top level.

The bottom-level series are first generated and then aggregated to form the hierarchy. The same covariance matrix is used to allow for correlation between bottom level series from the same *level 1* group. Such matrix is

$$\begin{bmatrix} 7 & 3 & 2 & 1 & 0 & 0 & 0 & 0 \\ 3 & 7 & 2 & 1 & 0 & 0 & 0 & 0 \\ 2 & 2 & 6 & 3 & 0 & 0 & 0 & 0 \\ 1 & 1 & 3 & 6 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 7 & 3 & 2 & 1 \\ 0 & 0 & 0 & 0 & 3 & 7 & 2 & 1 \\ 0 & 0 & 0 & 0 & 2 & 2 & 6 & 3 \\ 0 & 0 & 0 & 0 & 1 & 1 & 3 & 6 \end{bmatrix}$$

Datasets of 100 samples were generated, taking the first 90 samples as the training partition for identifying and fitting the models, and the last 10 observations as the out-of-sample partition. Automatic identification of ARIMA models was done using **SSpace** (Villegas and Pedregal, 2018) following the algorithm implemented in the **forecast** package in **R** (Hyndman and Khandakar, 2008). The fitted models were used to produce 1-to-7 steps-ahead predictions for each of the 15 series.

Once the individual models and predictions were calculated, reconciled forecasts were produced according to the different approaches discussed in Section 4, i.e., *bottom-up*, *top-down* and *reconciled*, using both **hts** (Hyndman et al., 2011) and **SSpace** (Villegas and Pedregal, 2018) implementations. The simulation was repeated 1000 times. Table 1 shows the Root Mean Squared Error (RMSE) in the out-of-sample for each of the approaches and for each of the 15 series. For convenience, an entry for the independent forecast has been also included, though not consistent with the hierarchy. The best alternative for each horizon and level is shown in bold. Since the *bottom-up* approach simply aggregates the bottom level forecasts all the way to the top level, results are equivalent for both **hts** and **SSpace** and they are shown in a single entry. On the other hand, different results are drawn for *top-down* and *reconciled* strategies, and RMSE values are shown separately for each implementation.

One important point is that all models automatically identified are exactly the same in both implementations (**hts** and **SSpace**). Thence, the accuracy differences observed in Table 1 are only due to the differences in reconciling approaches.

Since random ARIMA processes were used both to generate the series and (automatically) identify the models, the *bottom-up* strategy results favored, explaining its good performance. However, the value of this particular setup relays on providing a common ground for comparing existing approaches (**hts**) with the proposed approach (**SSpace**), leaving apart the aforementioned discussion on *top-down* vs *bottom-up*. The mean RMSE value for *bottom-up* (9.77) is indeed the best of all, even better than the *independent* forecasts (10.52 mean RMSE).

Some observations can be highlighted from Table 1. Regarding global averages (last row), results seem to be scattered around the *independent* forecasts (10.52), except for *top-down*[hts] value sitting apart

Table 1: RMSE for out-of-sample forecasting of the simulated data.

| | Independent | Bottom-up | Top-down | | reconciled | |
|---|---|---|---|---|---|---|
| | | | hts | sspace | hts | sspace |
| Total | 28.59 | **24.30** | 28.59 | 28.59 | 26.66 | 24.63 |
| A | 18.50 | **16.64** | 35.28 | 18.50 | 17.97 | 16.89 |
| B | 18.93 | **16.88** | 34.95 | 18.74 | 18.42 | 17.24 |
| AA | 11.19 | **10.58** | 24.89 | 11.43 | 11.05 | 10.62 |
| AB | 10.78 | **9.86** | 25.05 | 10.83 | 10.70 | 10.09 |
| BA | 11.18 | **10.59** | 26.60 | 11.64 | 11.35 | 10.77 |
| BB | 10.58 | **9.71** | 28.95 | 10.37 | 10.54 | 9.88 |
| AAA | **6.01** | **6.01** | 15.92 | 6.42 | 6.32 | 6.11 |
| AAB | 6.54 | 6.54 | 21.39 | 6.99 | 6.77 | **6.48** |
| ABA | **5.88** | **5.88** | 16.28 | 6.39 | 6.30 | 5.93 |
| ABB | **5.76** | **5.76** | 17.52 | 6.18 | 6.23 | 5.95 |
| BAA | **6.25** | **6.25** | 19.85 | 6.76 | 6.61 | 6.33 |
| BAB | **6.19** | **6.19** | 22.43 | 6.72 | 6.59 | 6.28 |
| BBA | **5.69** | **5.69** | 29.05 | 6.00 | 6.12 | 5.77 |
| BBB | **5.72** | **5.72** | 24.85 | 6.02 | 6.12 | 5.77 |
| *Average* | *10.52* | *9.77* | *24.77* | *10.77* | *10.52* | *9.91* |

(24.77), a finding in common with Hyndman et al. (2011). *Reconciled*[sspace] method (9.91) outperforms all alternatives except *bottom-up* (9.77), including the *independent* baseline (10.52).

Regarding *top-down* and *reconciled* approaches, **SSpace** systematically provides better results than the counterpart in **hts** for all series. This result comes to support the claim about the convenience of keeping consistency across time while reconciling across hierarchy. As previously discussed in Section 4, the proposed approach for reconciling forecasts also maintains consistency with models across time, which is expected to increase coherence and performance of the final system.

Pairwise *t*-tests for the 6 approaches were computed. As expected from results in Table 1, *top-down*[hts] vs the rest returned a p-value smaller than $10^{-8}$ for all possible comparisons. The best of all approaches is the *bottom-up* (being consistent with the experiment design as previously discussed) which compared to *independent* is said to be different with 99% of confidence (p-value $5.2 \times 10^{-4}$), but not significantly different from reconciled[sspace] (p-value 0.51). Finally, both *reconciled* approaches are different with 99% of confidence (p-value $7 \times 10^{-3}$).

## 6. Hierarchical demand forecasting

In this section we apply the proposed approach to predict the demand of a Spanish grocery retailer. The dataset contains daily observations on the sold units of 39 products covering the period 2013:Q1-2014:Q2. The hierarchical structure used is described in Table 2. The top level contains aggregated demand for the whole company. At level 1, demand is disaggregated by type of dish: first courses, pastas and meats. At level 2 products are grouped in seven different families: salads, creams, omelets, spaghetti, macaroni, beef and chicken. And finally, at the bottom level we have 39 disaggregated series for the individual products, and the hierarchy therefore involves 50 time series (39 disaggregated and 11 aggregated).

Table 2: Hierarchy for grocery demand data

| Level | N$^{o}$ of series per level | Aggregation vector |
|---|---|---|
| Company | 1 | [3] |
| Type of food | 3 | [3 2 2] |
| Family of products | 7 | [10 2 4 3 3 12 5] |
| Products | 39 | |

An optimal ARIMA model was selected for each series using the automatic identification algorithm in Hyndman and Khandakar (2008), as done in the simulation study discussed above. A rolling window procedure with 8 forecasting origins was implemented to produce 7-step-ahead forecasts, and the models were re-identified at every forecasting origin. Once more, *top-down*, *bottom-up* and *reconciled* approaches were computed using **hts** package and **SSpace**.

Table 3 contains the mean absolute percentage error (MAPE) for each forecast horizon and hierarchical forecasting approach. The table is organized in several blocks, one for each hierarchical level and the mean MAPE across all levels. MAPE for independent forecasts is shown in the first row of each section. The approach with smallest MAPE is shown in bold for each horizon. The last column contains the average MAPE across all forecast horizons. Since the *bottom-up* approach produces identical results in both implementations, they are shown in a single entry, while results for *top-down* and *reconciled* approaches are shown separately.

Some ideas emerge from this empirical study: (1) the *bottom-up* strategy is systematically out-performed in all horizons and levels by either *top-down* or *reconciled* approaches. (2) As we move down in the hierarchy, the *reconciled* approach excels at longer horizons (5, 6 and 7 steps ahead), while the *top-down* approach does better at shorter horizons (1, 2 and 3 steps ahead). (3) The **SSpace** approach seems to systematically out-perform **hts** implementation for longer horizons (4, 5, 6 and 7 steps ahead) both in *top-down* and *reconciled* approaches, while **hts** work better at shorter horizons in both approaches as well (with some exceptions, see for example Level 1 of Table 3).

## 7. Conclusions

In this paper, a new approach for hierarchical time series forecasting based on SS modelling is proposed. The principal merit is that it provides an unified treatment of *top-down*, *bottom-up*, *middle-out* and *reconciled* approaches reported in literature. In this way, the hierarchical forecasting problem is converted into a standard SS problem for which the Kalman Filter provides an optimal solution. This fact simplifies the assumptions on which the approach relies on. In particular, there is no need to add any way to disaggregate down in *top-down* approaches (Gross and Sohl, 1990), or to propose any additional objective function in *reconciled* approaches (Hyndman et al., 2011).

Table 3: MAPE for out-of-sample forecasting of the alternative hierarchical approaches applied to a Spanish retailer demand data.

| | | Forecast horizon | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | *AVG* |
| Top level | | | | | | | | | |
| Independent | | **6.86** | **7.94** | 7.99 | 15.29 | 16.23 | 15.71 | 15.54 | *12.22* |
| Bottom-up | | 7.48 | 9.90 | 9.78 | 16.41 | 17.68 | 17.07 | 17.00 | *13.62* |
| Top-down | hts | **6.86** | **7.94** | 7.99 | 15.29 | 16.23 | 15.71 | 15.54 | *12.22* |
| | sspace | **6.86** | **7.94** | 7.99 | 15.29 | 16.23 | 15.71 | 15.54 | *12.22* |
| reconciled | hts | 6.87 | 7.97 | **7.82** | **15.17** | **16.20** | 15.76 | 15.48 | ***12.18*** |
| | sspace | 7.07 | 8.92 | 10.48 | 17.86 | 17.22 | **15.59** | **14.94** | *13.15* |
| Level 1 | | | | | | | | | |
| Independent | | **9.44** | 10.70 | 10.34 | 17.29 | 18.80 | 18.98 | 17.42 | *14.71* |
| Bottom-up | | 10.66 | 12.53 | 12.86 | 19.29 | 20.19 | 19.16 | 18.06 | *16.11* |
| Top-down | hts | 9.66 | 10.63 | 10.77 | 17.51 | 18.88 | 18.89 | 17.29 | *14.80* |
| | sspace | 9.69 | **9.61** | **10.08** | **17.02** | 17.59 | 16.80 | 15.67 | *13.78* |
| reconciled | hts | 9.68 | 10.83 | 10.87 | 17.52 | 18.87 | 18.84 | 17.25 | *14.84* |
| | sspace | 10.44 | 11.28 | 12.51 | 17.83 | **16.71** | **15.16** | **14.38** | ***14.04*** |
| Level 2 | | | | | | | | | |
| Independent | | **13.28** | 14.47 | 15.39 | **22.08** | 22.56 | 23.55 | 21.42 | *18.96* |
| Bottom-up | | 15.10 | 17.20 | 18.24 | 24.44 | 25.44 | 25.36 | 23.69 | *21.35* |
| Top-down | hts | 13.33 | **13.99** | 14.95 | 22.32 | 22.83 | 23.64 | 21.37 | ***18.92*** |
| | sspace | 16.23 | 17.48 | 17.36 | 22.70 | 23.93 | 23.42 | 21.54 | *20.38* |
| reconciled | hts | 13.48 | 14.68 | **14.91** | 22.37 | 23.07 | 23.78 | 21.45 | *19.11* |
| | sspace | 15.23 | 17.19 | 17.44 | 22.27 | **21.64** | **20.60** | **19.14** | *19.07* |
| Bottom level | | | | | | | | | |
| Independent | | 28.09 | 30.19 | 31.99 | 36.80 | 37.23 | 37.72 | 35.95 | *33.99* |
| Bottom-up | | 28.09 | 30.19 | 31.99 | 36.80 | 37.23 | 37.72 | 35.95 | *33.99* |
| Top-down | hts | **27.68** | **29.42** | **30.67** | 36.31 | 36.48 | 37.12 | 35.12 | *33.26* |
| | sspace | 31.41 | 31.72 | 32.55 | 36.43 | 36.69 | 36.15 | 34.58 | *34.22* |
| reconciled | hts | 27.99 | 30.44 | 31.52 | 37.46 | 37.98 | 38.67 | 36.43 | *34.36* |
| | sspace | 29.82 | 30.55 | 31.30 | **35.43** | **34.96** | **34.29** | **33.31** | ***32.81*** |
| Average | | | | | | | | | |
| Independent | | 24.48 | 26.37 | 27.89 | 33.14 | 33.65 | 34.17 | 32.39 | *30.30* |
| Bottom-up | | 24.82 | 26.90 | 28.47 | 33.61 | 34.16 | 34.46 | 32.78 | *30.74* |
| Top-down | hts | **24.17** | **25.71** | **26.82** | 32.81 | 33.11 | 33.71 | 31.74 | *29.72* |
| | sspace | 27.49 | 27.93 | 28.58 | 32.92 | 33.35 | 32.80 | 31.24 | *30.62* |
| reconciled | hts | 24.44 | 26.61 | 27.48 | 33.71 | 34.31 | 34.94 | 32.77 | *30.61* |
| | sspace | 26.16 | 27.09 | 27.82 | **32.18** | **31.64** | **30.85** | **29.82** | ***29.37*** |

Most importantly, due to the use of the Kalman Filter, the solution preserves the time consistency of the forecasts, a property that is missing in many of the previous approaches. Moreover, forecasts are always produced on the basis of past optimal forecasts, instead on the raw individual forecasts, enhancing the coherency of the approach. The SS framework permits the generalization of the problem by defining the type of each single node in a hierarchy by means of *combined* hierarchies. Such heterogeneity may be even time-varying, if required. Confidence intervals or prediction intervals may be built with the aid of the standard output of the Kalman Filter.

The new approach is evaluated on simulations and on real data of a Spanish grocery retailer. Both experiments show that the proposed approach provides significantly better results than existing approaches.

## 8. Acknowledgements

This work was supported by the European Regional Development Fund and Spanish Government (MINECO/FEDER, UE) under the project with reference DPI2015-64133-R and by the Vicerrectorado de Investigación y Política Científica from UCLM by DOCM 31/07/2014 [2014/10340].

## 9. References

Athanasopoulos, T., Hyndman, R., Kourentzes, N., and Petropoulos, F. (2017). Forecasting with temporal hierarchies. *European Journal of Operational Research*, 262(1):60–74.

Barnea, A. and Lakonishok, J. (1980). An analysis of the usefulness of disaggregated accounting data for forecasts of corporate performance. *Decision Sciences*, 11(1):17–26.

Boylan, J. (2010). Choosing levels of aggregation for supply chain forecasts. *Foresight*, 18:9–13.

Casals, J., Garcia-Hiernaux, A., Jerez, M., Sotoca, S., and Trindade, A. (2016). *State-Space Methods for Time Series Analysis: Theory, Applications and Software.* Forthcoming by Chapman and Hall/CRC.

Commandeur, J., Koopman, S., and Ooms, M. (2011). Statistical software for state space methods. *Journal of Statistical Software*, 41(1):1–18.

Dangerfield, B. and Morris, J. (1992). Top-down or bottom-up: Aggregate versus disaggregate extrapolations. *International Journal of Forecasting*, 8(2):233 – 241.

Durbin, J. and Koopman, S. (2012). *Time series analysis by state space methods*. Number 38. Oxford University Press.

Fliedner, E. B. and Lawrence, B. (1995). Forecasting system parent group formation: An empirical application of cluster analysis. *Journal of Operations Management*, 12(2):119–130.

Fliedner, G. (1999). An investigation of aggregate variable time series forecast strategies with specific subaggregate time series statistical correlation. *Computers & Operations Research*, 26(10):1133 – 1149.

Fliedner, G. (2001). Hierarchical forecasting: Issues and use guidelines. *Industrial Management & Data Systems*, 101(1):5–12.

Granger, C. and Morris, M. (1976). Time series modelling and interpretation. *Journal of the Royal Statistical Society. Series A*, 139(2):246–257.

Gross, C. W. and Sohl, J. E. (1990). Disaggregation methods to expedite product line forecasting. *Journal of Forecasting*, 9(3).

Grunfeld, Y. and Griliches, Z. (1960). Is aggregation necessarily bad? *The Review of Economics and Statistics*, 42(1):1–13.

Gómez, V. (2015). Ssmmatlab: A set of matlab programs for the statistical analysis of state space models. *Journal of Statistical Software*, 66(9):1–37.

Harvey, A. (1989). *Forecasting, structural time series models and the Kalman filter*. Cambridge university press.

Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G., and Shang, H. L. (2011). Optimal combination forecasts for hierarchical time series. *Computational Statistics & Data Analysis*, 55(9):2579 – 2589.

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

Hyndman, R. J., Lee, A. J., and Wang, E. (2016). Fast computation of reconciled forecasts for hierarchical and grouped time series. *Computational Statistics & Data Analysis*, 97:16 – 32.

Kahn, K. B. (1998). Revisiting top-down versus bottom-up forecasting. *Journal of Business Forecasting Methods & Systems*, 17(2):14.

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

Koopman, S., Harvey, A., Doornik, J., and Shephard, N. (2009). *STAMP 8.2: Structural Time Series Analyser and Modeller and Predictor*. Timberlake Consultants Limited.

Lütkepohl, H. (1984). Forecasting contemporaneously aggregated vector arma processes. *Journal of Business & Economic Statistics*, 2(3):201–214.

Muir, J. W. (1979). The pyramid principle. In *Proceedings of 22nd Annual Conference, American Production and Inventory Control Society*, pages 105–7.

Orcutt, G., Watts, H., and Edwards, J. (1968). Data aggregation and information loss. *The American Economic Review*, 58(4):773–787.

Pennings, C. and van Dalen, J. (2017). Integrated hierarchical forecasting. *European Journal of Operational Research*, 263(2):412–418.

Rostami-Tabar, B., Babai, M. Z., Ducq, Y., and Syntetos, A. (2016). Non-stationary demand forecasting by cross-sectional aggregation. *International Journal of Production Economics*, 170:297–309.

Sbrana, G. and Silvestrini, A. (2013). Forecasting aggregate demand: Analytical comparison of top-down and bottom-up approaches in a multivariate exponential smoothing framework. *International Journal of Production Economics*, 146(1):185 – 198.

Shlifer, E. and Wolff, R. (1979). Aggregation and proration in forecasting. *Management Science*, 25(6):594–603.

Strijbosch, L. W. G., Heuts, R. M. J., and Moors, J. J. A. (2008). Hierarchical estimation as a basis for hierarchical forecasting. *IMA Journal of Management Mathematics*, 19(2):193–205.

Taylor, C., Pedregal, D., Young, P., and Tych, W. (2007). Environmental time series analysis and forecasting with the captain toolbox. *Environmental Modelling & Software*, 22(6):797–814.

The MathWorks, Inc (2018). *MATLAB - The Language of Technical Computing, Version R2015b*. Natick, Massachusetts. URL http://www.mathworks.com/products/matlab/.

Villegas, M. A. and Pedregal, D. J. (2018). Sspace: A toolbox for state space modelling. *Journal of Statistical Software, in press*.

Weatherford, L., Kimes, S., and Scott, D. (2001). Forecasting for hotel revenue management: Testing aggregation against disaggregation. *The Cornell Hotel and Restaurant Administration Quarterly*, 42(4):53 – 64.

Widiarta, H., Viswanathan, S., and Piplani, R. (2009). Forecasting aggregate demand: An analytical evaluation of top-down versus bottom-up forecasting in a production planning framework. *International Journal of Production Economics*, 118(1):87 – 94.

Young, P., Pedregal, D., and Tych, W. (1999). Dynamic harmonic regression. *Journal of forecasting*, 18(6):369–394.

Zellner, A. and Tobias, J. (2000). A note on aggregation, disaggregation and forecasting performance. *Journal of Forecasting*, 19(5):457–465.