A New Taxonomy for Vector Exponential Smoothing and Its Application to Seasonal Time Series

Ivan Svetunkov[a,*], Huijing Chen[b], John E. Boylan[a]

[a] *Centre for Marketing Analytics and Forecasting*
Lancaster University Management School, Lancaster, LA1 4YX, UK
[b] *Operations and Systems Management, Faculty of Business and Law, University of Portsmouth, Portsmouth, PO1 3DE, UK*

Abstract

In short-term demand forecasting, it is often difficult to estimate seasonality accurately, owing to short data histories. However, companies usually have multiple products with similar seasonal demand patterns. A possible solution in this case is to use the components of several time series from a homogeneous family, thus estimating seasonal coefficients based on cross-sectional information. Motivated by this practical problem, we propose a new taxonomy of Parameters, Initial States and Components (PIC), which exploits homogeneous features of time series. We then apply this framework to vector exponential smoothing. We develop a model selection mechanism based on information criteria to select the appropriate PIC restrictions. We then conduct a simulation experiment and empirical analysis on retail data in order to assess the performance of point forecasts and prediction intervals of the models within this framework.
Keywords: Forecasting, Multivariate statistics, Seasonal data, Vector exponential smoothing, Retailing

1. Introduction

Seasonal time series are ubiquitous. In settings such as the prediction of weather or other physical events (e.g. earthquakes), long time series are often available, containing records of many seasonal cycles. Such data are amenable to sophisticated analysis at the level of individual series. In a retail business setting, long time series are the exception rather than the norm. Products are often subject to short life cycles, owing to fast changing markets, opening of new retail outlets or introduction of new modes of ordering. Even when products, outlets and ordering modes are well established, records are not always retained for periods long enough to capture the seasonality accurately. Therefore, it is common for retailers to be required to forecast seasonal time series which have very few complete seasonal cycles of data. For example, accurate forecasts and prediction intervals are needed to determine cycle stock and

*Correspondence: Ivan Svetunkov, Centre for Marketing Analytics and Forecasting
Lancaster University Management School, Lancaster, LA1 4YX, UK
Email address: i.svetunkov@lancaster.ac.uk (Ivan Svetunkov)

safety stock requirements (a classic OR problem). Accuracy is not easy to attain for seasonal products when there are few observations corresponding to each period in the seasonal cycle.

Although data histories are often short, they may be available for a large cross-section of homogeneous series, especially in a retail setting. For example, there may be many variations of an item of clothing, such as size and colour. If the item is seasonal, it is often reasonable to assume that the variations share similar seasonal demand patterns. Similarly, seasonality in several geographical regions may often be assumed to be common. Therefore, in many cases there are readily available homogeneous groups to explore. This paper analyses such situations in depth.

In practice, exponential smoothing is a well-established approach to short-term forecasting. It is included in many commercial software packages, and there are comprehensive suites of open source software available to implement the approach (see, for example, Hyndman and Khandakar, 2008; Svetunkov, 2021b). It is more intuitive than other approaches such as Seasonal ARIMA, making it more attractive to businesses which lack personnel with well-developed skills in statistics and time series analysis. Moreover, exponential smoothing has a well-developed underpinning model framework both for a collection of univariate models (ETS – "Error-Trend-Seasonal" model proposed and then developed in Hyndman et al., 2002, 2008; Svetunkov, 2021c) or to a multivariate framework (de Silva et al., 2010).

Univariate seasonal ETS models (Koehler et al., 2001) are straightforward to apply but may be heavily parameterised. Suppose that 24 months of data are available and the model includes a trend component and monthly multiplicative seasonal indices. This requires the estimation of 11 initial seasonal indices (assuming normalisation of indices), an initial level, an initial trend and three smoothing parameters (for level, trend and seasonality). This amounts to 16 parameters to be estimated on the basis of just 24 data points. This situation is not unusual in practice, and calls for consideration of alternatives.

In this paper, we propose a taxonomy of Parameters, Initial values and Components (PIC) of models based on their commonality between time series. It is important to note that the PIC taxonomy can be applied to a multivariate model or a set of univariate models. We apply the PIC framework to Vector ETS (VETS) models to show how this leads naturally to utilising cross-sectional data for estimation, taking into account trend and seasonality in additive and multiplicative forms. Although vector models, including ETS, have previously been proposed in the literature (more details in Sections 2 and 3), the PIC taxonomy is an original contribution of this research. The usefulness of vector models in a retail context is often limited, because of the large number of parameters to estimate, often with short data histories. We argue and demonstrate that the PIC taxonomy affords flexibility and makes the VETS models more feasible, especially for seasonal models. Simulation experiments are conducted on synthetic

data to quantify the benefits of the seasonal VETS-PIC approach and to identify the effect of such factors as group size and length of data histories. The new approach is assessed empirically using retail data from the M5 competition (Makridakis et al., 2021), and the paper concludes with suggestions for future research directions.

2. Literature review

Armstrong (1985) pointed out that the estimation of seasonal factors is much more challenging when data history is short. Duncan et al. (1993) developed a multivariate multiple source of error approach for hierarchical seasonal time series forecasting, demonstrating that using cross-sectional information can lead to improvements in forecasting accuracy. This conclusion was supported by Armstrong (2004), who argued for making effective use of additional information from analogous time series.

Early research explored the potential benefits of applying group seasonal indices (GSI), rather than the conventional individual seasonal indices (ISI). Dalhart (1974) proposed a method based on simple averages of ISIs (Dalhart's Group Seasonal Index, DGSI). Withycombe (1989) suggested calculating seasonal indices from aggregated data, which are equivalent to weighted averages of ISIs (Withycombe's Group Seasonal Index, WGSI). Both methods assume that seasonality is not changing over time and use classical decomposition for calculation of seasonal indices.

Bunn and Vassilopoulos (1993) empirically compared the performance of ISI, DGSI and WGSI. They applied all three methods on 54 weekly series from five product groups, with 42 observations in each series. Assessing forecasting accuracy by Mean Squared Error (MSE) and Mean Absolute Error (MAE), the authors concluded that both DGSI and WGSI outperformed ISI, and WGSI was better than DGSI overall.

Gorr et al. (2003) applied forecasting approaches on deseasonalised data, comparing ISI and GSI (as reseasonalisation techniques) for crime forecasting, focussing on the city of Pittsburg and its six police precincts. Forecasting results, measured by MSE, MAE and Mean Absolute Percentage Error (MAPE) (which gave similar ranking of methods), showed that there was clear evidence that forecasts using GSI (pooled seasonality) were more accurate than forecasts using ISI. The authors recommended applying GSI to obtain seasonal estimates at the city-wide level, for use in forecasts at the spatially-disaggregated precinct level.

Although the empirical comparisons showed potential benefits in applying the GSI methods, there was no theoretical understanding of conditions under which these methods perform better than ISI. Chen and Boylan (2007) was the first paper to address this gap. Assuming two simple non-trended models, for both

additive and multiplicative seasonality, appropriately underpinning the ISI and GSI methods, the authors developed easily implementable rules for selection between these methods. The key insights from this research were that noisy series can "borrow strength" from less noisy ones in a seasonally homogeneous group, and that the benefits of the GSI methods are greater when data histories are shorter and/or the noise components are negatively correlated. Further empirical evidence was established in Chen and Boylan (2008), in which the ISI and GSI methods, and the previously established selection rules, were tested on a dataset of 218 series from seven product groups. The authors found that applying the selection rules resulted in the greatest benefits in terms of forecasting accuracy, while universal application of DGSI outperformed both WGSI and ISI.

For evolving seasonality, Dekker et al. (2004) highlighted the poor performance of the classical Holt-Winters exponential smoothing method in an empirical study. The authors found that it failed to estimate seasonality accurately in the presence of high demand uncertainty and stochastic seasons. The Holt-Winters method showed the highest forecasting errors of all the compared methods. It was even worse than single exponential smoothing. By applying product aggregation, and estimating seasonal indices from the product family aggregate, the authors found a reduction of 12-59% in MSE.

Ouwehand et al. (2007) defined a state space model that also utilised the idea of estimating seasonality from a group. This model extends the ETS(M,A,M) model (from the taxonomy of Hyndman et al., 2008) to multivariate forecasting. The seasonal component is defined as common, with common smoothing parameter and common initial seasonal indices. Level and trend are estimated individually. This model can be summarised as:

$$\begin{aligned} y_{i,t} &= (l_{i,t-1} + b_{i,t-1})\, s_{t-m}(1 + \epsilon_{i,t}) \\ l_{i,t} &= (l_{i,t-1} + b_{i,t-1})(1 + \alpha_i \epsilon_{i,t}) \\ b_{i,t} &= b_{i,t-1} + \alpha_i \beta_i \,(l_{i,t-1} + b_{i,t-1})\, \epsilon_{i,t} \\ s_t &= s_{t-m} \left(1 + \gamma \sum_{i=1}^{n} w_i \epsilon_{i,t}\right) \end{aligned} \tag{1}$$

where $y_{i,t}$ is the actual value of the series i at time t, $l_{i,t}$ is the level component, $b_{i,t}$ is the trend component, s_t is the common seasonal component, $\epsilon_{i,t}$ is the i.i.d. error term, α_i, β_i and γ are the smoothing parameters and n is the number of time series under consideration. Note that in this mixed model the trend can become negative (typically the error term $\epsilon_{i,t}$ is assumed to follow a normal distribution). Therefore the level and trend components can become negative as well, leading to unreasonable predictions. The mixture of multiplicative seasonality and additive trend in the model may lead to the reverse of the seasonal pattern, because in this case the negative values are multiplied by the seasonal indices.

The assumption of normality of the error term is reasonable in many contexts and is used often in research. However, Akram et al. (2009) argued that as the normal distribution extends over the whole real line, it may cause problems if the observations need to be positive, as is the case with typical demand/sales forecasting. They found that additive and mixed ETS models (with both additive and multiplicative components) can lead to prediction intervals with negative values and, as the forecasting horizon extends, even the point forecasts may become negative. However, pure multiplicative models can guarantee a positive sample space. The authors also discussed several alternative pure multiplicative models and concluded that the most reasonable one is the model constructed on the logarithms of the original data, implying a log normal distribution of the error term in the original scale. Consistent with Akram et al. (2009)'s definition, this research demonstrates that using pure multiplicative models, i.e. with all ETS elements being multiplicative, might be a more desirable approach than using the mixed ones for forecasting in context of ETS. Pure additive and multiplicative models are formally defined in the following section.

Given that GSI approaches are applied to groups of time series, it is only logical to consider the multivariate models in this context. Interestingly, the earlier work of Duncan et al. (1993) used a multivariate state space model in order to forecast hierarchical seasonal data. This idea was adopted much later by Pennings and van Dalen (2017) who compared the model with existing hierarchical forecast reconciliation approaches and found that the proposed approach performs well in terms of MAPE. This example demonstrates that there is a connection between the hierarchical approaches and vector models. The interest in the former was stimulated by the original paper of Hyndman et al. (2011) and has been promulgated in many areas, including hierarchical forecasting of products (Pennings and van Dalen, 2017), intermittent demand (Li and Lim, 2018; Kourentzes and Athanasopoulos, 2021), and energy forecasting (see, for example, Taieb et al., 2020). Related research has explored temporal aggregation (examples include Kourentzes et al., 2014; Kourentzes and Petropoulos, 2016; Athanasopoulos et al., 2017), and the combination of cross-temporal hierarchies, with aggregations over different products and time buckets (see, for example, Kourentzes and Athanasopoulos, 2019; Spiliotis et al., 2020). All these papers assume that univariate models are applied to the data and their forecasts are then reconciled, but there is an apparent connection between hierarchical approaches and multivariate models, which has not been explored in the literature before. So, developing further the theory of vector models is an important topic in forecasting.

The main limitation of conventional vector models (such as Vector Autoregression Lütkepohl, 2005) is that they typically require more parameters to be estimated than simpler models. A typical solution for this problem in the context of VAR models is to shrink estimates of parameters and elements of the

covariance matrix (see, for example, Schäfer and Strimmer, 2005; Lee et al., 2016; Chan et al., 2020). However, Vector Exponential Smoothing (de Silva et al., 2010; Snyder et al., 2017), having similarities with univariate exponential smoothing, allows the use of similar parameters for different time series. For example, Chen et al. (2019) used such a multivariate forecasting model and showed that it performs better than simple models, when time series share some components.

In this paper, we extend the existing vector exponential smoothing framework, and propose a new taxonomy to take account of components being set in common (across a range of series), or individually (for each time series).

3. Vector ETS

When we deal with several products, we can consider the problem as multivariate, with demands on those products being represented in the form of a vector:

$$\mathbf{y}_t = \begin{pmatrix} y_{1,t} \\ y_{2,t} \\ \vdots \\ y_{n,t} \end{pmatrix}, \tag{2}$$

where n is the number of time series included in the group. In order to model the demand on these products, we will use Vector ETS (VETS). This is similar to the vector innovations structual time series (VISTS) framework developed by de Silva et al. (2010), and later extended by Athanasopoulos and de Silva (2012) to include seasonality. The additive version of VETS can be formulated as:

$$\begin{aligned} \mathbf{y}_t &= \mathbf{W}\mathbf{v}_{t-\boldsymbol{l}} + \boldsymbol{\epsilon}_t \\ \mathbf{v}_t &= \mathbf{F}\mathbf{v}_{t-\boldsymbol{l}} + \mathbf{G}\boldsymbol{\epsilon}_t \end{aligned}, \tag{3}$$

where $\mathbf{W}$ is the measurement matrix, $\mathbf{F}$ is the transition matrix, $\mathbf{G}$ is the persistence matrix, containing smoothing parameters, $\mathbf{v}_t$ is the vector of states for time series, and $\boldsymbol{\epsilon}_t' = \begin{pmatrix} \epsilon_{1,t} & \dots & \epsilon_{n,t} \end{pmatrix}$ is the error term vector which, in the case of an additive model, is assumed to have a multivariate normal distribution with a mean vector of zeroes and $\boldsymbol{\Sigma}$ covariance matrix. The states for each time series may contain up to three conventional components of ETS: level, trend and seasonality. Finally, $\boldsymbol{l}$ is the vector of lags of components, which denotes that the components of the vector $\mathbf{v}_t$ can have different lags (similar to ETS models in Svetunkov, 2021c) and the operation $t - \boldsymbol{l}$ returns a vector of values. For example, the level component can have a lag of 1, while the seasonal component can have a lag equal to the seasonal

frequency m.

While the additive model can be considered as appropriate in many contexts, a multiplicative form may be more useful when seasonality needs to be evaluated across different products. This is because a standard percentage increase (say 41%) in sales of a product can be considered as more natural for several products than a fixed increase (say 410 units) for all products, especially if there are widely varying scales of sales volumes and units of measurement (e.g. kilograms, units and litres). The pure multiplicative model can be achieved by taking logarithms of the data, leading to:

$$\begin{aligned} \log \mathbf{y}_t &= \mathbf{W} \log \mathbf{v}_{t-\boldsymbol{l}} + \log \boldsymbol{\epsilon}_t \\ \log \mathbf{v}_t &= \mathbf{F} \log \mathbf{v}_{t-\boldsymbol{l}} + \mathbf{G} \log \boldsymbol{\epsilon}_t \end{aligned}. \tag{4}$$

Note that the log function is applied element-wise. An example of a pure multiplicative VETS model is discussed in Subsection refsec:VETS-PIC-Example. The assumption for the error term in this model corresponds to the conventional in this case: $\log \boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}_n, \boldsymbol{\Sigma})$, where $\mathbf{0}_n$ is the vector of n zeroes, and that the exponent of this variable follows the multivariate log normal distribution. Given the properties of the log normal distribution, the point forecasts from the model (4) correspond to geometric means (equal to medians in this case) rather than arithmetic means. If arithmetic means are needed, then they can be obtained via the well known formula $\mu_{y_i,t} = \exp\left(\mu_{\log y_i,t} + \frac{\sigma^2_{\log y_i}}{2}\right)$, where $\mu_{\log y_i,t}$ is the geometric mean and $\sigma^2_{\log y_i}$ is the geometric variance for the series i (i-th element on the diagonal of $\boldsymbol{\Sigma}$). We note at this stage that the arithmetic means will result in increasing trajectories even for the local level model due to this connection. Recalling that the point forecasts for the pure multiplicative models in the original ETS do not correspond to conditional expectations (Hyndman et al., 2008, p.23), we argue that it is more natural, for pure multiplicative models, to produce geometric means instead.

In this paper, we will call the models based on (3) as pure additive and refer to them as VETS(A,N,N) / VETS(A,N,A) / VETS(A,A,N) / etc, where 'V' stands for 'Vector', while the pure multiplicative models based on (4) are referred to as VETS(M,N,N;LN) / VETS(M,N,M;LN) / VETS(M,M,N;LN) / etc. to reflect that the models are applied on log-transformed data. Note that the previous work in the vector exponential smoothing, the VISTS framework (de Silva et al., 2010) and its extension (Athanasopoulos and de Silva, 2012), considered pure additive models only. The letters in the brackets are taken from the taxonomy proposed by Hyndman et al. (2008) for ETS models. Our definition of pure multiplicative models differs from the original ETS framework and is in line with Akram et al. (2009). We do not discuss mixed models (with both additive and multiplicative components) as they would need to be formulated differently, introducing complex interactions between the components and, as discussed earlier, may, under

some circumstances, lead to negative values in both prediction intervals and point forecasts. Therefore, in this research it is not our intention to fully extend the univariate ETS models to a general vector framework, but rather to examine the effect of the PIC taxonomy through a subset of pure additive and multiplicative models.

Given that we deal with the state space model in (3) and (4), we can define different components for the vector $\mathbf{v}_t$ and specify different structures of matrices $\mathbf{W}$, $\mathbf{F}$ and $\mathbf{G}$. This flexibility leads to a variety of VETS models, imposing different types of restrictions on the parameters and components of the model.

3.1. PIC taxonomy

In this subsection we introduce a simple PIC taxonomy for VETS models. Although our main interest is in seasonal models, this taxonomy applies to all of the pure additive and multiplicative VETS models.

3.1.1. The main elements of the taxonomy

The taxonomy we propose allows different forms of the VETS framework, taking into account the following aspects:

- Parameters

  Traditionally, smoothing parameters are set individually for each time series. However, benefits of using common parameters have been reported in the literature. In a non-seasonal setting, Fildes et al. (1998) examined common parameters for level and trend. They applied single exponential smoothing, Holt's method and damped Holt's on 261 telecommunications series previously analysed by Fildes (1992). Their results showed common smoothing parameters to perform better than individually optimised ones, either once or every time period. This is an example of the common parameters restriction applied to a collection of univariate models. In that study, the common smoothing parameters were set arbitrarily, rather than optimised. Ouwehand et al. (2007) applied a common seasonal smoothing parameter for the common seasonal component, but the parameters for level and trend were still individual.

  In the proposed taxonomy, we allow all of the smoothing parameters for specific components across time series to be either individual or common, and all of them are optimised in one model. This further extends the findings of Fildes et al. (1998) and provides insights on using common smoothing parameters for similar time series.

  For models with damped trends, an additional damping parameter is applied. Damped trend models show good forecasting performance in a univariate setting (McKenzie and Gardner, 2010; Gardner

and McKenzie, 2011). But to the best of our knowledge, there is no evidence in the literature to explore the benefit of common damping parameters. In our taxonomy, damping parameters are allowed to be either individual or common, and modelled separately from smoothing parameters, to provide more flexibility.

- Initial values

  All of the components need to have starting values but no previous study has considered the effect of common initial values (rather than individual starting points). In this taxonomy, initial values can also be estimated either individually or commonly. It is intuitive, for example, to allow seasonal indices to have common starting values if the group of items is indeed seasonally homogeneous. Also, estimating these initial values commonly can significantly reduce the number of parameters to estimate; this will be discussed in the next section.

- Components

  In univariate models, level, trend and seasonality are all modelled at the level of the individual series. There is evidence in the literature that assuming common seasonality can improve forecasting accuracy if seasonal behaviour in a group of series is similar. Within the ETS framework, Ouwehand et al. (2007) proposed a common seasonal component for ETS(M,A,M). In our taxonomy, seasonal components can be either individual or common. This can also be extended to the level and trend components, each of which can be either individual or common. Note, however, that not all combinations of common components are reasonable. For example, it is not possible for several time series to have common levels but individual trends, because different trends will lead to the levels becoming uncommon.

3.1.2. Proposed taxonomy for VETS

The ETS(*,*,*) notation can apply to a model for an individual series or for a set of series sharing a common model. However, it cannot cater for a set of series which may have some elements that are individual and others that are shared. The notation, outlined below, is designed to address this need.

The first step is to identify the essential elements that need to be considered with regard to their commonality. We propose three categories:

- **P**arameters (Smoothing constants for Level, Trend, Seasonality, and Damping parameter)
- **I**nitial values (Level, Trend, Seasonality)
- **C**omponents (Level, Trend, Seasonality)

Taken together, this gives a framework of VETS(*,*,*)PIC(*,*,*) models, supporting only the pure additive and pure multiplicative models. This framework can be readily extended to cater for more general models (e.g. double seasonal models), although such models are not discussed further in this paper.

For the **P** (parameters) element, * may take the value of 'N' if no parameters are in common, 'L' for the common level, 'T' for the common trend and 'S' for the common seasonal smoothing parameters. These three are supplemented by 'D', for the damping parameter. Combinations of letters indicate that more than one element is in common, e.g. 'TS' indicates that trend and seasonal smoothing parameters are common (but not the level), and 'LTS' indicates that all three smoothing parameters are in common.

For the **I** (initial values) element, the options are similar to P, but without 'D': 'N' if there is no commonality in initials, 'L' for the common initial level, 'T' for the common initial trend and 'S' for the common initial seasonal indices between time series.

For the **C** (components) element, the options are the same as for the I element.

Using this taxonomy, for example, VETS(M,Md,M;LN)PIC(N,N,N) means all components, smoothing parameters and initial values are set individually. The main difference of this model from ETS(M,Md,M;LN) (from Akram et al., 2009) applied to all time series individually is the estimation procedure, discussed in Section 4.3. VETS(M,Md,M;LN)PIC(LTSD,LTS,LTS), on the other hand, has the most restrictive form of everything being common. There are many variations in between, such as VETS(M,Md,M;LN)PIC(TS,TS,TS). This shows common components, parameters and initial values for trend and seasonality (level is individual).

The work of Ouwehand et al. (2007) may be considered as a collection of univariate ETS(M,A,M) models with PIC(N,S,S) restrictions according to our taxonomy. Although Ouwehand et al. (2007) use the same smoothing parameter for seasonal components, they estimate individual weights for each time series, thus making the effect of smoothing individual. In their original paper they have the following for the seasonal component (from the equation (1)):

$$s_t = s_{t-m}\left(1 + \gamma \sum_{i=1}^{n} w_i \epsilon_{i,t}\right),$$

which can be reformulated into:

$$s_t = s_{t-m}\left(1 + \sum_{i=1}^{n} \gamma w_i \epsilon_{i,t}\right).$$

This is equivalent to the model with PIC(N,S,S) restrictions, where $\gamma_i = \gamma w_i$:

$$s_t = s_{t-m}\left(1 + \sum_{i=1}^{n} \gamma_i \epsilon_{i,t}\right).$$

As noted earlier, not all variations in the taxonomy are feasible. For example, if any of the components (level, trend or seasonality) are common, then the corresponding initial values have to be common as well. Similarly individual trends cannot be combined with a common level component. So while VETS(M,N,M;LN)PIC(*,*,LS) is feasible, VETS(M,M,M;LN)PIC(*,*,LS) is not. Individual parameters can be mixed with common components, implying that the impact of different errors on the common component can be different.

The proposed VETS taxonomy provides a conceptual contribution to the literature. Previous studies showed some evidence of the potential benefits of using common / group seasonality, but there is no work that brings together the smoothing parameters, damping parameter, initial seed values and time series components, to enable a comparison of common and individual components in a systematic way.

The main difference between the models, in terms of their construction, is in the number of parameters requiring estimation. The following formula (5) quantifies the number (k) of parameters and initial values that need to be estimated, assuming likelihood estimation and n time series with a seasonal frequency of m. The formula is based on the choices of: i) not estimated because not included ($\eta_a^b = 0$), ii) estimated in common ($\eta_a^b = 1$) or iii) estimated individually ($\eta_a^b = n$), where b indicates parameters (P) or initial values (I), and a indicates level (L), or trend (T), or seasonality (S), or damping (D, for parameters only). Note that the commonality of components (C) does not change the number of estimated parameters, but rather impacts the structure of the model.

$$k = \frac{n(n+1)}{2} + \sum_{a \in \{L,T,S,D\}} \eta_a^P + \sum_{a \in \{L,T\}} \eta_a^I + (m-1)\eta_S^I \tag{5}$$

The first term in (5) accounts for the n variance terms and $\frac{n(n-1)}{2}$ covariance terms to be estimated in likelihood estimation. If it is assumed that the error terms are independent (thus making the covariance matrix diagonal), then it should be changed to n. The equation allows the calculation of the overall number of parameters to be estimated. Models are estimable if this number, per series, $\frac{k}{n}$ is lower than the number of observations in each series T (similar to VAR models studied in Lütkepohl, 2005).

In some cases, the formula may need to be modified to take into account other considerations. For example, in theory, the model VETS(M,N,M;LN)PIC(N,N,N) can be applied to the data when each series has at least $T > 4 + (m-1)$, assuming a diagonal covariance matrix. However, in the case when $T < 2m$, there may be difficulties with the estimation of the seasonal indices in such a model, and so it is advisable to have at least $2m$ observations in the sample for each series, for any model. Otherwise some restrictions on components may be necessary in order to undertake cross-sectional estimation of seasonal components.

Next we provide two contrasting examples, demonstrating the application of the matrices in (3) and (4) for a case of two time series.

4. Constructing VETS-PIC models

4.1. VETS-PIC Example

To demonstrate how specific VETS-PIC model may look like, we consider an example of a restricted model in the framework, VETS(M,Md,M;LN)PIC(LTSD,LTS,S). In this case, we assume that all the seasonal components in (4) are shared between the series, $s_{1,t} = s_{2,t} = s_t$ for all t together with the initial level, trend and seasonal components. Furthermore, the smoothing and the dampening parameters are common for all the series, $\alpha_1 = \alpha_2 = \alpha$, $\beta_1 = \beta_2 = \beta$, $\gamma_1 = \gamma_2 = \gamma$, $\phi_1 = \phi_2 = \phi$. We rewrite the original equation (4) in a multiplicative form to show what such model implies:

$$\begin{aligned}
y_{1,t} &= l_{1,t-1} b_{1,t-1}^{\phi} s_{t-m} \epsilon_{1,t} \\
y_{2,t} &= l_{2,t-1} b_{2,t-1}^{\phi} s_{t-m} \epsilon_{2,t} \\
l_{1,t} &= l_{1,t-1} b_{1,t-1}^{\phi} \epsilon_{1,t}^{\alpha} \\
l_{2,t} &= l_{2,t-1} b_{2,t-1}^{\phi} \epsilon_{2,t}^{\alpha} \qquad . \\
b_{1,t} &= b_{1,t-1}^{\phi} \epsilon_{1,t}^{\beta} \\
b_{2,t} &= b_{2,t-1}^{\phi} \epsilon_{2,t}^{\beta} \\
s_t &= s_{t-m} \left(\epsilon_{1,t} \epsilon_{2,t}\right)^{\gamma}
\end{aligned} \tag{6}$$

This model implies the following values in the original formulation (4):

$$\mathbf{W} = \begin{pmatrix} 1 & 0 & \phi & 0 & 1 \\ 0 & 1 & 0 & \phi & 1 \end{pmatrix}, \mathbf{v}_t = \begin{pmatrix} l_{1,t} \\ l_{2,t} \\ b_{1,t} \\ b_{2,t} \\ s_t \end{pmatrix}, \mathbf{v}_{t-\boldsymbol{l}} = \begin{pmatrix} l_{1,t-1} \\ l_{2,t-1} \\ b_{1,t-1} \\ b_{2,t-1} \\ s_{t-m} \end{pmatrix},$$

$$\mathbf{F} = \begin{pmatrix} 1 & 0 & \phi & 0 & 0 \\ 0 & 1 & 0 & \phi & 0 \\ 0 & 0 & \phi & 0 & 0 \\ 0 & 0 & 0 & \phi & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \mathbf{G} = \begin{pmatrix} \alpha & 0 \\ 0 & \alpha \\ \beta & 0 \\ 0 & \beta \\ \gamma & \gamma \end{pmatrix} \text{ and } \boldsymbol{\epsilon}_t = \begin{pmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \end{pmatrix}.$$

The same matrices can be rewritten in a more compact form:

$$\mathbf{W} = \begin{pmatrix} \mathbf{I}_2 & \phi\mathbf{I}_2 & \mathbf{1}_2 \end{pmatrix}, \mathbf{F} = \begin{pmatrix} \mathbf{I}_2 & \phi\mathbf{I}_2 & \mathbf{0}_2 \\ \mathbf{O}_2 & \phi\mathbf{I}_2 & \mathbf{0}_2 \\ \mathbf{0}_2' & \mathbf{0}_2' & 1 \end{pmatrix}, \mathbf{G} = \begin{pmatrix} \alpha\mathbf{I}_2 \\ \beta\mathbf{I}_2 \\ \gamma\mathbf{1}_2' \end{pmatrix},$$

where $\mathbf{1}_2$ is the unity 2-vector (containing ones), $\mathbf{0}_2$ is the 2-vector of zeroes and $'$ is the symbol of transposition.

This example shows how a specific model looks and how it connects with a bigger framework. In the next subsection, we show how VETS-PIC models can be constructed using block values of matrices.

4.2. Constructing VETS-PIC in general

In order to show how any of the VETS-PIC models can be constructed we introduce the following block structure for measurement, transition and persitence matrices:

$$\mathbf{W} = \begin{pmatrix} \mathbf{W}_1 & \mathbf{W}_2 & \mathbf{W}_3 \end{pmatrix}, \mathbf{F} = \begin{pmatrix} \mathbf{F}_{1,1} & \mathbf{F}_{1,2} & \mathbf{F}_{1,3} \\ \mathbf{F}_{2,1} & \mathbf{F}_{2,2} & \mathbf{F}_{2,3} \\ \mathbf{F}_{3,1} & \mathbf{F}_{3,2} & \mathbf{F}_{3,3} \end{pmatrix}, \mathbf{G} = \begin{pmatrix} \mathbf{G}_1 \\ \mathbf{G}_2 \\ \mathbf{G}_3 \end{pmatrix}, \tag{7}$$

where each block can be a matrix, a vector or a scalar depending on the restrictions imposed on the model. The possible values of blocks are summarised in Table 1. This table includes all feasible combinations. Note that the model with common level and individual trend components has been omitted because it does not make sense, as discussed earlier in the paper.

Any VETS-PIC model with some restrictions on parameters and components of the original VETS-PIC(N,N,N) implies the use of specific block matrices values from Table 1. In order to construct desired matrices for a selected VETS-PIC model, one would need to identify the specific imposed restrictions in the table and insert them in (7) in the style of a construction kit. For example, if we consider the combination for VETS(M,M,M;LN)PIC(LT,LT,L), then we would need to have a structure for the model with common level, individual trend and individual seasonal components. We then construct transition, measurement and persistence matrices by searching the fields level='c', trend='i' and seasonal='i' in Table 1 and inserting them in (7). Furthermore, given the commonality of level and trend smoothing parameters, we would also need to use the rows α='c' and β='c'.

We do not provide details for state vector restrictions, noting that the following will hold for the components of the vector for all t:

1. If level is common, then $l_{1,t} = \cdots = l_{n,t} = l_t$;

Type of restriction	**Block Values**		
Restrictions on components			
level='i'	$\mathbf{F}_{1,1} = \mathbf{I}_n$	$\mathbf{W}_1 = \mathbf{I}_n$	$\mathbf{G}_1 = \mathbf{A}$
trend='i'	$\mathbf{F}_{2,2} = \mathbf{\Phi}$	$\mathbf{W}_2 = \mathbf{\Phi}$	$\mathbf{G}_2 = \mathbf{B}$
seasonal='i'	$\mathbf{F}_{3,3} = \mathbf{I}_n$	$\mathbf{W}_3 = \mathbf{\Phi}$	$\mathbf{G}_3 = \mathbf{\Gamma}$
level='c'	$\mathbf{F}_{1,1} = 1$	$\mathbf{W}_1 = \mathbf{1}_n$	$\mathbf{G}_1 = \boldsymbol{\alpha}'_n$
trend='c'	$\mathbf{F}_{2,2} = \phi$	$\mathbf{W}_2 = \phi\mathbf{1}_n$	$\mathbf{G}_2 = \boldsymbol{\beta}'_n$
seasonal='c'	$\mathbf{F}_{3,3} = 1$	$\mathbf{W}_3 = \mathbf{1}_n$	$\mathbf{G}_3 = \boldsymbol{\gamma}'_n$
level='i' & trend='i'	$\mathbf{F}_{1,2} = \mathbf{\Phi}$	$\mathbf{F}_{2,1} = \mathbf{O}_n$	
level='i' & trend='c'	$\mathbf{F}_{1,2} = \phi\mathbf{1}_n$	$\mathbf{F}_{2,1} = \mathbf{0}'_n$	
level='c' & trend='c'	$\mathbf{F}_{1,2} = \phi$	$\mathbf{F}_{2,1} = 0$	
level='i' & seasonal='i'	$\mathbf{F}_{1,3} = \mathbf{O}_n$	$\mathbf{F}_{3,1} = \mathbf{O}_n$	
level='c' & seasonal='i'	$\mathbf{F}_{1,3} = \mathbf{0}'_n$	$\mathbf{F}_{3,1} = \mathbf{0}_n$	
level='i' & seasonal='c'	$\mathbf{F}_{1,3} = \mathbf{0}_n$	$\mathbf{F}_{3,1} = \mathbf{0}'_n$	
level='c' & seasonal='c'	$\mathbf{F}_{1,3} = 0$	$\mathbf{F}_{3,1} = 0$	
trend='i' & seasonal='i'	$\mathbf{F}_{2,3} = \mathbf{O}_n$	$\mathbf{F}_{3,2} = \mathbf{O}_n$	
trend='c' & seasonal='i'	$\mathbf{F}_{2,3} = \mathbf{0}'_n$	$\mathbf{F}_{3,2} = \mathbf{0}_n$	
trend='i' & seasonal='c'	$\mathbf{F}_{2,3} = \mathbf{0}_n$	$\mathbf{F}_{3,2} = \mathbf{0}'_n$	
trend='c' & seasonal='c'	$\mathbf{F}_{2,3} = 0$	$\mathbf{F}_{3,2} = 0$	
Restrictions on parameters			
α='c'	$\mathbf{A} = \alpha\mathbf{1}'_n$	$\boldsymbol{\alpha}_n = \alpha\mathbf{1}_n$	
β='c'	$\mathbf{B} = \beta\mathbf{1}'_n$	$\boldsymbol{\beta}_n = \beta\mathbf{1}_n$	
γ='c'	$\mathbf{\Gamma} = \gamma\mathbf{1}'_n$	$\boldsymbol{\gamma}_n = \gamma\mathbf{1}_n$	
ϕ='c'	$\mathbf{\Phi} = \phi\mathbf{I}_n$	$\boldsymbol{\phi}_n = \phi\mathbf{1}_n$	

Table 1: Values of block matrices for (7), depending on imposed restrictions. 'i' stands for the individual, while 'c' stands for the common element. Note that the combination level='c' & trend='i' is not feasible and is omitted from the table.

2. If trend is common, then $b_{1,t} = \cdots = b_{n,t} = b_t$;

3. If seasonal is common, then $s_{1,t} = \cdots = s_{n,t} = s_t$.

These restrictions will reduce the number of components in the state vector.

Finally, the commonality of initial states does not impact the architecture of VETS-PIC model; it will change only the number of estimated parameters. We do not aim to cover all the theoretically possible 1,428 models in this paper, but the values provided in this section should allow any of them to be constructed.

4.3. Estimation of VETS models

There are different ways of estimating VETS; the most straightforward is via maximisation of the likelihood function. In case of the pure additive model, assuming that $\mathbf{y}_t \sim \mathcal{N}(\mathbf{0}_n, \mathbf{\Sigma})$ the log-likelihood can be written as:

$$\ell(\boldsymbol{\theta}, \mathbf{\Sigma}|\mathbf{Y}) = -\frac{T}{2}\left(n\log(2\pi) + \log|\mathbf{\Sigma}|\right) - \frac{1}{2}\sum_{t=1}^{T}\left(\boldsymbol{e}'_t\mathbf{\Sigma}^{-1}\boldsymbol{e}_t\right), \tag{8}$$

where $\mathbf{Y}$ is the matrix of all actual observations, $\boldsymbol{\theta}$ is the vector of all estimated parameters, $\boldsymbol{e}_t = \mathbf{y}_t - \boldsymbol{\mu}_{y,t|t-1}$ is the estimate of the error term, $\boldsymbol{\mu}_{y,t|t-1}$ is the one step ahead conditional expectation of

the model, and $\mathbf{\Sigma}$ is the covariance matrix, which can be estimated via maximisation of likelihood as:

$$\hat{\mathbf{\Sigma}} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{e}_t \boldsymbol{e}_t'. \tag{9}$$

Inserting (9) in (8), similar to Snyder et al. (2017), it can be shown that the concentrated likelihood is then:

$$\ell(\boldsymbol{\theta}, \hat{\mathbf{\Sigma}}|\mathbf{Y}) = -\frac{T}{2}\left(n\log(2\pi e) + \log|\hat{\mathbf{\Sigma}}|\right). \tag{10}$$

The maximisation of the likelihood (10) implies the minimisation of the determinant of the covariance matrix $\hat{\mathbf{\Sigma}}$, which is sometimes called in the literature "Generalised Variance". This typically implies that the variances of each series are minimised, but at the same time the covariances between them are increased.

As for the pure multiplicative model, the log-likelihood will be similar to the pure additive one, but with the addition of the actual values to reflect that the original data follows a log normal distribution as in the model (4):

$$\begin{aligned}\ell(\boldsymbol{\theta}, \mathbf{\Sigma}|\mathbf{Y}) = &-\frac{T}{2}\left(n\log(2\pi) + \log|\mathbf{\Sigma}|\right) \\ &-\frac{1}{2}\sum_{t=1}^{T}\left((\log \boldsymbol{e}_t)'\mathbf{\Sigma}^{-1}\log \boldsymbol{e}_t\right) - \sum_{t=1}^{T}\sum_{j=1}^{n}\log y_{j,t},\end{aligned} \tag{11}$$

where $\log \boldsymbol{e}_t = \log \mathbf{y}_t - \log \boldsymbol{\mu}_{y,t|t-1}$ and $\boldsymbol{\mu}_{y,t|t-1}$ is the vector of geometric one step ahead expectations from the model. The formula for the covariance matrix $\mathbf{\Sigma}$ will be the same as in (9), only with $\log \boldsymbol{e}_t$ instead of $\boldsymbol{e}_t$, and the concentrated log-likelihood will be similar to (10), but with the addition of the sum of logarithms of actual values as in (11).

In order to start the estimation of parameters of the model, we need to calculate the initial values of the level, trend and the seasonal indices. Linear regression of actual observations on the deterministic trend and seasonal dummy variables for all the observations and all the series in the available group of data is used in the initialisation. After calculating the parameters of this form of regression model for each series, we either average the seasonal indices (if the seasonal indices are shared between the series) or use them for each series (if the seasonality is individual). Similarly, if the initials of level or trend are shared, the respective parameters are averaged across series. Then, the estimation procedure starts, maximising the likelihood function, which leads to a set of optimal initial values of parameters.

We also need to make sure that the constructed models are stable as defined in Hyndman et al. (2008). In order to do so, we need to make sure that all eigenvalues of the following discount matrix lie in the

unit circle:

$$\mathbf{D} = \mathbf{F} - \mathbf{GW}. \tag{12}$$

This means that some of the smoothing parameters might become negative or greater than one, which is common in ETS with admissible bounds. Note that in the case of VETS, restricting parameters to the (0, 1) interval bound is not required. The model contains many smoothing parameters and there is no reason for them to be restricted to this interval, given potential complex interactions between time series; see also, for example, Snyder et al. (2017).

The other important thing to note is how the model is constructed. Given the state space form, we initialise the components of the model before the start of the sample, i.e. before the observation $t = 1$. So, for example, we define the initial seasonal indices on the observations $-m + 1, -m + 2, \ldots, 0$, not sacrificing the first m observations, as used to be the case for the conventional seasonal exponential smoothing method (Winters, 1960).

While it is possible to estimate VETS using other loss functions, the MLE is preferable from a statistical point of view, as it gives consistent and efficient estimates of parameters, also permitting model selection via information criteria. The criteria themselves are calculated differently than in the case of univariate models. For example, Bedrick and Tsai (1994) developed the corrected version of AIC for multivariate models on small samples, which we would recommend for model selection.

4.4. Model selection in VETS

Given the number of possible variations in the VETS-PIC model framework, model selection becomes challenging. Maximum likelihood estimation, described in Section 4.3, allows information criteria to be used for model selection.

While there are different ways to select models in the VETS-PIC framework, we do not aim to discuss how to select the ETS part of the model. This is because in order to select the ETS components, an analyst needs to first group time series and decide, which components could potentially be common in the series. For example, some series might exhibit seasonality, and thus would form a natural group. So, there is no need to do ETS components selection in this case - it can be done in the preliminary analysis (see Section 6 for an example). Instead, we propose to focus a model selection process on the PIC part via the following procedure:

1. Select initial states restrictions,
2. Try different parameters restrictions,
3. Try restricted components.

The main rationale for starting from the initial states is to reduce the number of estimated parameters. This becomes extremely important in case of seasonal models, where reducing the number of estimated seasonal initials from $n \times (m-1)$ to $(m-1)$ might be critical for the next steps of the algorithm. Testing parameters is a safe next step, keeping some flexibility in the model and potentially further reducing the number of estimated parameters. As a final step, we test whether the restriction on components is needed or not. As discussed earlier in this paper, this step does not change the number of estimated parameters but changes the architecture of the model, making it as restrictive as possible. We expect that the last step would only be needed for time series that exhibit very similar patterns (e.g. seasonal shapes or trends). Much of the rationale for the sequencing in the second stage has been drawn from simulations (see Section 5).

5. Simulation experiment

5.1. Purposes of simulation

From this point forward, we define (true) models as the underlying processes that describe how data is generated (DGP, Data Generating Processes). We use the term 'methods' to denote estimation procedures to derive forecasts (i.e. models applied to the data).

A simulation experiment is conducted, with the following aims:

1. Understand the relative performance of different PIC restrictions for a given data generating process within the proposed taxonomy;
2. Evaluate the robustness of the proposed methods to mis-specification of the underlying data generating processes;
3. Quantify the benefits of seasonal VETS methods for point forecasts and prediction intervals, within controlled parameter settings;
4. Understand the effects of group size and length of data histories on the accuracy of methods.

Given our interest in seasonal behaviour and the large number of possible variations in the taxonomy, the simulation experiment focuses on the VETS(M,N,M;LN) model and compares performances of its various PIC variations. This model is chosen following earlier discussions in Section 2 favouring pure multiplicative models. Focusing on one model in a controlled environment aids in understanding of the impact of the PIC variations. To achieve this, we focus on the application of the taxonomy to seasonal time series in the simulation experiments. This restriction is relaxed in the next section, when the full taxonomy and model selection is examined in empirical analysis.

Even with this restriction, the PIC taxonomy can have 64 variations, because each of the PIC elements can take 4 values: N,L,S and LS. Note that some of the variations are not feasible, either because their models cannot be constructed (e.g. VETS(M,N,M;LN)PIC(N,N,S)), or are not reasonable (can be constructed but hard to find domains of application, e.g. it is difficult to motivate the commonality of the level component in VETS(M,N,M;LN)PIC(LS,LS,LS)). Since we are interested in common seasonality, we further reduce the variations by focusing on a subset of:

- **P**: N, S, LS
- **I**: N, S
- **C**: N, S

This reduces the possible variations to 9 (after removing 3 infeasible variations of PIC(N,N,S), PIC(S,N,S) and PIC(LS,N,S)), allowing for meaningful comparisons and useful insights to be gained.

5.2. Data generation

In this experiment, we use the `sim.ves()` function from the `legion` package v0.1.0 (Svetunkov and Pritularga, 2021) for R (Team, 2020) to generate the data. It was originally generated according to the VETS(A,N,A), and then exponents were taken in order to create VETS(M,N,M;LN). All the parameters of the simulation are given in Table 2.

Parameter	Values
Seasonal cycle	$m = 4$
Initial level	Selected randomly from region [4, 10]
Seasonal indices	Randomly chosen from (-1, 1), then normalised
Number of observations	$T = \{24, 64\}$ quarters (6 and 16 years)
Holdout size	8 quarters (2 years)
Forecasting horizon	$h = 1, \dots, 8$ quarters (up to 2 years ahead)
Group sizes	$n = \{2, 10, 20, 30, 40, 50\}$
Noise	$\epsilon_{i,t} \sim$ i.i.d. $\mathcal{N}(0, \sigma_i^2)$ for all i
Smoothing parameters for PIC:	
(LS,*,*)	$\alpha = 0.3$, $\gamma = 0.1$
(S,*,*)	$\alpha \in [0, 0.3]$, $\gamma = 0.1$
(N,*,*)	$\alpha \in [0, 0.3]$, $\gamma \in [0, 1 - \alpha]$
Number of iterations	1000

Table 2: Simulation setup

In terms of seasonality, we focused on quarterly seasonality in the experiment, although monthly seasonality was checked separately as well, giving similar findings. The region [4, 10] for the initial level corresponds to the [54, 22000] region after taking the exponent. This covers many real life situations for

demand. A similar approach was used for the initial seasonal indices: (-1, 1) corresponds to (0.37, 2.72) for multiplicative indices in the data scale.

We have focused on two cases for sample sizes: 6 years and 16 years of quarterly data, where the last two years (8 observations) were used for testing the performance of the models. While some of the models (e.g. PIC(LS,S,S)) can be estimated on short data histories of 2 years or even less, not all of them work in this case. As for the group sizes, in order to cover several possible real life situations, we have used groups of 2, 10, 20, 30, 40 and 50 series. This allows analysis of how the models work in case of very small groups and larger ones, when the number of products is larger than the number of observations.

In this experiment, we have set all standard deviations for the error term equal to 0.3 and generated the errors using a normal distribution (which becomes log normal after exponentiation). We have tried lower standard deviations as well, but the results were similar to those reported in this experiment. The results we report in the following sections do not consider the case when the error terms are correlated between the series, because the parameters to be estimated easily outnumber the observations when the group sizes are large and/or histories are short. Recall from equation (5) that the number of parameters in a full variance-covariance matrix is $\frac{n(n+1)}{2}$, which gives 1275 parameters for a group of 50 series, and 55 even for 10 series. However, we have conducted a separate small-scale experiment with cross correlations, focusing on groups of 2 and 10 series. The results were consistent with the main findings.

Finally, we have restricted the range for the smoothing parameters α and γ as shown in Table 2, which does not cover the whole parameter space, but is sufficient to cover common cases in practice. It is also worth pointing out that usually the smoothing parameters tend to be smaller in pure multiplicative models than in the additive ones, due to linearisation. This is because taking logarithms reduces the variability of the data, typically making it better behaved.

5.3. Error measures

After generating data from each DGP, we have applied all 9 VETS methods via the `vets()` function from the `legion` package, measuring their performance in terms of point forecasts and prediction intervals. For point forecast accuracy, the relative Mean Absolute Error (rMAE) and the relative Root Mean Squared Error (rRMSE) have been calculated. The former is optimised on the median and is more appropriate for the multiplicative VETS models, whilst the latter is optimised on the mean and is provided just for information. Forecast bias is assessed using the relative Absolute Mean Error (rAME). Prediction interval performance is evaluated using 95% coverage (as commonly used in the statistics literature) and the relative pinball values for both the lower and upper bounds.

Table 3 summarises the error measures used in our experiments. e_{t+j} is the j-steps ahead forecast

Error measure	**Formula**
Root Mean Squared Error (RMSE)	$\sqrt{\frac{1}{h}\sum_{j=1}^{h} e_{t+j}^2}$
Relative RMSE (rRMSE)	$\frac{\text{RMSE}_a}{\text{RMSE}_b}$
Mean Absolute Error (MAE)	$\frac{1}{h}\sum_{j=1}^{h} \lvert e_{t+j}\rvert$
Relative MAE (rMAE)	$\frac{\text{MAE}_a}{\text{MAE}_b}$
Absolute Mean Error (AME)	$\left\lvert\frac{1}{h}\sum_{j=1}^{h} e_{t+j}\right\rvert$
Relative AME (rAME)	$\frac{\text{AME}_a}{\text{AME}_b}$
Pinball value (pinball)	$(1-\tau)\sum_{j,y_{t+j}<b_{t+j}} \lvert y_{t+j}-b_{t+j}\rvert + \tau\sum_{j,y_{t+j}\geq b_{t+j}} \lvert y_{t+j}-b_{t+j}\rvert$
Relative pinball (rPinball)	$\frac{\text{pinball}_a}{\text{pinball}_b}$
Coverage	$\frac{1}{h}\sum_{j=1}^{h} \mathbb{1}(y_{t+j} < u_{t+j} \vee y_{t+j} >= l_{t+j})$

Table 3: Error measures used in the experiments. h is the forecast horizon.

error, b_{t+j} is the α-quantile, u_{t+j} is the upper bound and l_{t+j} is the lower bound of the prediction interval. In order to distinguish the pinball value for the lower and the upper bounds, we denote them as 'rPinballL' and 'rPinballU' respectively. The subscript a refers to the method being evaluated and b refers to the benchmark method.

In all cases, we used either PIC(N,N,N) as a benchmark or the correct method for the DGP, the latter being explicitly specified in the text. After obtaining relative error measures for each time series, we aggregate them using geometric means over all the iterations, as proposed by Davydenko and Fildes (2013), except for the 95% coverage, which is summarised by arithmetic means. Result tables are presented in a heat map format: the darker the grey shade is, the worse the results are.

5.3.1. Forecasting accuracy results

We group the results based on the flexibility of the models. The first group has the most restrictive models, with shared seasonal components between the series: PIC(LS,S,S), PIC(S,S,S) and PIC(N,S,S). The second group has restrictions on the initial seasonal indices: PIC(LS,S,N), PIC(S,S,N) and PIC(N,S,N). Finally, the last group has the most flexible models with independent seasonal components and initial values: PIC(LS,N,N), PIC(S,N,N) and PIC(N,N,N).

Table 4 summarises the performance of the methods on two types of samples: 56 quarters and 16 quarters. These are overall accuracy results summarised across all DGPs. It is apparent that the most restrictive methods perform worse than the benchmark of PIC(N,N,N), while all other models show very similar accuracy results. This is consistent between all three measures: rMAE, rRMSE and rAME.

To gain greater insights into the methods' behaviour, we now proceed to show the results in more

Methods	56 + 8 quarters			16 + 8 quarters		
	rMAE	rRMSE	rAME	rMAE	rRMSE	rAME
LS,S,S	1.23	1.21	1.11	1.17	1.16	1.08
S,S,S	1.23	1.22	1.10	1.17	1.15	1.07
N,S,S	1.24	1.22	1.11	1.17	1.15	1.07
LS,S,N	0.99	0.99	0.98	1.00	1.00	1.01
S,S,N	0.99	0.99	0.99	0.99	0.99	1.01
N,S,N	1.00	1.00	0.99	0.99	0.99	1.01
LS,N,N	1.00	1.00	0.99	1.02	1.02	0.99
S,N,N	1.00	1.00	0.99	1.00	1.00	1.00
N,N,N	1.00	1.00	1.00	1.00	1.00	1.00

Table 4: Overall Forecasting Accuracy (Benchmark PIC(N,N,N))

detail, examining how the methods perform on different DGPs, on different group sizes and lengths of data history.

We start by examining the results for a small sample size, with time series aggregated over all group sizes. Tables 5 and 6 show the performance of the methods for each DGP, measured by rMAE and rRMSE respectively. In order to highlight whether the correct method is indeed performing better than the others, the method from the respective DGP is used as the benchmark (the diagonal line).

Methods	Data Generating Processes								
	LS,S,S	S,S,S	N,S,S	LS,S,N	S,S,N	N,S,N	LS,N,N	S,N,N	N,N,N
LS,S,S	1.00	0.98	1.01	1.00	1.00	1.16	1.58	1.66	1.70
S,S,S	1.01	1.00	1.02	1.01	1.01	1.15	1.55	1.64	1.67
N,S,S	1.01	0.98	1.00	1.01	1.01	1.16	1.55	1.64	1.68
LS,S,N	1.02	1.02	1.03	1.00	1.02	1.01	1.02	1.06	1.02
S,S,N	1.02	1.02	1.03	1.00	1.00	1.00	1.02	1.07	1.01
N,S,N	1.03	1.02	1.02	1.01	1.01	1.00	1.01	1.06	1.00
LS,N,N	1.11	1.06	1.02	1.06	1.07	1.02	1.00	1.03	1.01
S,N,N	1.08	1.06	1.04	1.04	1.04	1.02	0.98	1.00	1.01
N,N,N	1.07	1.06	1.03	1.03	1.04	1.02	0.97	1.00	1.00

Table 5: rMAE results by models (16+8 periods) (Benchmark correct methods)

Broadly, the correct methods perform better than the incorrect ones for each DGP. In many cases there is hardly any difference between different methods. It is interesting to notice that the top right corners of Tables 5 and 6 show much worse results than the remainder of the tables. This highlights the weakness of the most restrictive methods when the DGP is very flexible (meaning that parts of the model for each time series are different). Applying these most restrictive methods can cause significant damage in terms of accuracy. This comes mainly from the common seasonal component element. When this assumption is relaxed to individual seasonal components, even with the restrictions of smoothing parameters and initial seasonal indices (the middle block of the tables), the methods tend to perform

Methods	Data Generating Processes								
	LS,S,S	S,S,S	N,S,S	LS,S,N	S,S,N	N,S,N	LS,N,N	S,N,N	N,N,N
LS,S,S	1.00	0.98	1.01	1.00	1.00	1.14	1.52	1.60	1.63
S,S,S	1.01	1.00	1.02	1.00	1.01	1.14	1.49	1.58	1.61
N,S,S	1.01	0.98	1.00	1.01	1.01	1.15	1.50	1.58	1.61
LS,S,N	1.02	1.02	1.02	1.00	1.02	1.01	1.02	1.06	1.02
S,S,N	1.02	1.01	1.03	1.00	1.00	1.00	1.02	1.06	1.01
N,S,N	1.03	1.02	1.02	1.01	1.01	1.00	1.01	1.06	1.01
LS,N,N	1.11	1.06	1.02	1.06	1.07	1.02	1.00	1.03	1.01
S,N,N	1.08	1.06	1.03	1.04	1.04	1.02	0.98	1.00	1.01
N,N,N	1.07	1.05	1.03	1.03	1.04	1.01	0.97	1.00	1.00

Table 6: rRMSE results by models (16+8 periods) (Benchmark correct methods)

better.

The bottom left corners of the tables shows that applying the most flexible methods to the most restrictive DGPs also leads to reduced accuracy in comparison with the benchmarks. However, the loss in accuracy is much less serious in this case than in the top right corners of the tables.

Since the two tables convey a consistent message and results are indeed very similar, we have decided to report the rMAE hereafter, dropping the other two from the discussion.

We then examine the effect of group sizes, for two contrasting models: PIC(LS,S,S) (the most restrictive) and PIC(N,N,N) (the most flexible). For consistency and ease of comparison, the method PIC(N,N,N) is used as the benchmark.

Methods	2 series	10 series	20 series	30 series	40 series	50 series
LS,S,S	0.97	0.95	0.95	0.95	0.95	0.95
S,S,S	0.98	0.95	0.95	0.97	0.95	0.95
N,S,S	0.98	0.96	0.96	0.96	0.96	0.95
LS,S,N	1.00	0.96	0.96	0.96	0.95	0.95
S,S,N	0.99	0.96	0.96	0.96	0.96	0.96
N,S,N	0.99	0.98	0.98	0.98	0.98	0.97
LS,N,N	1.01	1.02	1.02	1.02	1.02	1.02
S,N,N	1.00	1.00	1.01	1.01	1.01	1.01
N,N,N	1.00	1.00	1.00	1.00	1.00	1.00

Table 7: Effect of group sizes on rMAE for PIC(LS,S,S) DGP (Benchmark PIC(N,N,N))

When PIC(LS,S,S) is used as the DGP, the PIC(*,S,S) methods perform well, similarly to each other, as seen in the top area in Table 7. Increasing group size from 2 to 10 helps these methods, but thereafter larger groups do not seem to have much effect. The gains over the benchmark of PIC(N,N,N) are moderate. The more flexible PIC(*,S,N) methods perform similarly to the most restrictive ones, especially on the data with large groups. Furthermore, there is a slight improvement in terms of accuracy when the

smoothing parameters become more restrictive: the errors for PIC(N,S,*) are typically higher than the errors of PIC(S,S,*), which are slightly higher than PIC(LS,S,*).

Methods	2 series	10 series	20 series	30 series	40 series	50 series
LS,S,S	1.53	1.82	1.86	1.87	1.88	1.88
S,S,S	1.52	1.80	1.83	1.84	1.85	1.85
N,S,S	1.53	1.80	1.83	1.84	1.85	1.85
LS,S,N	1.00	1.01	1.02	1.02	1.02	1.02
S,S,N	1.00	1.01	1.02	1.02	1.01	1.01
N,S,N	0.99	1.00	1.00	1.00	1.00	1.00
LS,N,N	1.01	1.01	1.01	1.01	1.01	1.01
S,N,N	1.01	1.01	1.01	1.01	1.01	1.01
N,N,N	1.00	1.00	1.00	1.00	1.00	1.00

Table 8: Effect of group sizes on rMAE for PIC(N,N,N) DGP (Benchmark PIC(N,N,N))

When PIC(N,N,N) is used as DGP (Table 8), the most flexible methods (bottom block) perform the best, as expected. However, the PIC(*,S,N) methods are also performing well, showing little difference from the true ones. The PIC(*,S,S) methods do not perform as well. The increase in group sizes shows a similar effect to the one observed in Table 7: more pronounced change in accuracy from 2 to 10, and stabilising after that. Both tables show that increasing group size from 10 to 50 does not have any noticeable effect; therefore, groups larger than 50 were not included in our experiments.

In addition, as expected, when the smoothing parameters are restricted in the models applied to PIC(N,N,N) data, there is a decrease in the accuracy in all cases, although it is not substantial.

In summary, although, as expected, the best performing methods tend to be those that align with the DGPs, the middle group, PIC(*,S,N) performs well in all the cases, being slightly less accurate than the 'correct methods'. Based on this finding, this group of methods can be considered as the most robust for VETS(M,N,M;LN) to model mis-specification.

5.3.2. 95% prediction interval results

Having assessed relative accuracy of point forecasts of the methods in the proposed taxonomy, we now examine the overall performance in terms of 95% prediction intervals.

In Table 9, the further away from the 95% target, the darker the shading is. The coverage results are reported to three decimal places, in order to distinguish between the methods. The PIC(*,S,S) and PIC(*,S,N) methods get closer to the nominal 95% coverage, when the sample size is larger. PIC(*,N,N) methods on the other hand produce narrower intervals and provide a little less coverage on larger samples. This could be due to the overfitting effect, which may appear in the more flexible methods.

As for the pinball values, the group of the most restrictive methods (top block) is worse than the

Methods	56 + 8 quarters			16 + 8 quarters		
	rPinballU	rPinballL	Coverage	rPinballU	rPinballL	Coverage
LS,S,S	1.23	1.12	0.946	0.99	1.04	0.944
S,S,S	1.31	1.12	0.950	1.05	1.05	0.953
N,S,S	1.29	1.12	0.949	1.09	1.06	0.960
LS,S,N	0.97	0.99	0.953	0.86	0.98	0.937
S,S,N	1.01	1.00	0.954	0.96	1.00	0.953
N,S,N	1.04	1.01	0.954	1.02	1.01	0.961
LS,N,N	1.00	0.99	0.943	0.86	0.97	0.938
S,N,N	1.01	1.00	0.944	0.91	0.98	0.954
N,N,N	1.00	1.00	0.946	1.00	1.00	0.970

Table 9: Overall forecasting interval results (Pinball Benchmark PIC(N,N,N))

benchmark of PIC(N,N,N) in capturing both upper and lower quantiles for longer histories. Improvements are observed in the other two groups, especially when data histories are short.

In order to better understand why this happens, we provide a more detailed analysis on relative pinballs for the upper bound (the results for the lower bound are similar).

Methods	Data Generating Processes								
	LS,S,S	S,S,S	N,S,S	LS,S,N	S,S,N	N,S,N	LS,N,N	S,N,N	N,N,N
LS,S,S	1.00	0.95	0.96	1.03	0.98	1.01	1.71	1.73	1.70
S,S,S	1.06	1.00	1.00	1.08	1.03	1.05	1.87	1.93	1.87
N,S,S	1.09	1.01	1.00	1.12	1.06	1.08	2.00	2.05	1.99
LS,S,N	0.98	0.95	0.98	1.00	0.92	0.89	1.26	1.23	1.19
S,S,N	1.03	0.98	0.98	1.05	1.00	0.96	1.52	1.60	1.47
N,S,N	1.10	1.03	0.99	1.11	1.06	1.00	1.70	1.78	1.65
LS,N,N	1.17	1.09	1.11	1.19	1.07	0.92	1.00	0.93	0.85
S,N,N	1.24	1.17	1.17	1.22	1.15	0.98	1.01	1.00	0.91
N,N,N	1.39	1.28	1.26	1.36	1.28	1.07	1.13	1.11	1.00

Table 10: rPinballU results by models (16+8 periods) (Benchmark correct methods)

Table 10 shows the relative Pinball values for the upper bound for all of the methods, with the true models as benchmarks. The PIC(*,S,S) and PIC(*,S,N) methods perform well overall, when the DGPs have similar restrictions. But again we observe poor results in the top right corner; the very restrictive common seasonal components do not do well when the underlying models are flexible. The middle block of methods also does poorly in this case, but to a lesser degree than the PIC(*,S,S) methods. So, simply changing the seasonal component element of the taxonomy from common to individual can bring improvements in terms of the pinball values.

Furthermore, the bottom left corner demonstrates similar tendencies as in Table 5 with the rMAE: when the most flexible methods are applied to the most restrictive DGPs, they tend to perform more poorly than the PIC(*,S,N) and PIC(*,S,S) methods.

Once again the PIC(*,S,N) methods seem to be robust in comparison with the other methods.

It is also worth pointing out that, within each block of methods, there is an improvement in terms of pinball, when restrictions are imposed on the smoothing parameters: it is better to have PIC(LS,*,*) than PIC(S,*,*), which is subsequently better than PIC(N,*,*) methods. This finding holds for all the methods under consideration, no matter what the DGP is. An explanation for this performance is that the more restrictive methods tend not to overfit the data, because the smoothing parameters are averaged out across all the time series.

Next we focus on two contrasting models, PIC(LS,S,S) and PIC(N,N,N), and examine the effect of group sizes on the pinball values. The corresponding tables can be found in Appendix A.

When the DGP is PIC(LS,S,S) (Table A.15), the correct method does indeed show considerable improvement over PIC(N,N,N). Notably, PIC(LS,S,N) is performing equally well and, in several cases, slightly better than PIC(LS,S,S). In fact, we can see this pattern for almost all the methods in the first two blocks. Even though the differences are very small, this highlights the benefits of applying individual seasonal components. It also appears that the benefits are greater when group sizes increase: when there are 50 series in a group, the best performing method, PIC(LS,S,N), has a 35% lower pinball value than PIC(N,N,N).

When the underlying model is PIC(N,N,N) (Table A.16), we observe the opposite picture: the group of the most flexible methods performs better than PIC(*,S,N), which performs better than PIC(*,S,S). This aligns with the previous findings, that applying the most restrictive methods to the most flexible data leads to losses in accuracy.

Finally, similarly to the previous parts of the experiment, we observe that restrictions on smoothing parameters tend to improve the performance of methods: PIC(S,*,*) outperforms PIC(N,*,*), with PIC(LS,*,*) being the best. This can be explained in the same way as in the case of PIC(LS,S,S): the methods with the restricted smoothing parameters tend not to overfit the data as much as the most flexible ones.

In summary, the interval results suggest that PIC(*,S,S) methods can be too restrictive, and switching to PIC(*,S,N) leads to more robust results. Another interesting finding is that, keeping everything else the same, PIC(LS,*,*) performs better than PIC(S,*,*), which in turn is better than PIC(N,*,*). This may be due to the parsimonious nature of the PIC(LS,*,*) methods, given that the more flexible methods need more parameters to estimate, especially on larger groups of time series.

5.3.3. Summary of the simulation results

Overall, the findings from the simulation experiment provide useful insights in understanding how to apply common smoothing parameters, initials and seasonal components. It appears that common smoothing parameters for all components bring the most benefits in terms of accuracy due to the mitigation of the overfitting effect.

The methods with common initial seasonal indices (PIC(*,S,N)), overlooked in the literature, perform robustly in all the settings, being outperformed by the flexible methods (PIC(*,N,N)) only, when the DGP favours those methods. This means that these methods can be used in many cases, as long as the time series exhibit similar seasonal patterns. If they are substantially different, then the methods with the individual components (PIC(*,N,N)) should be preferred.

Simulation results also suggest that the common seasonal components (PIC(*,S,S) models) can become too restrictive in some situations, and modelling seasonal components individually brings more flexibility. It is recommended to apply PIC(*,S,S) methods only to homogeneous time series, which exhibit very similar seasonal patterns; in the other cases these methods may perform poorly. The insights gained from the simulation experiments are helpful in informing the model selection mechanism, especially in the second stage of the algorithm from Section 4.4.

Finally, the simulation results show that PIC(LS,S,N) and PIC(LS,N,N) are the most robust methods from the proposed taxonomy. They perform well under a range of different conditions, in terms of both forecasting accuracy and prediction intervals. These methods achieve a good balance between accuracy and flexibility. All of these findings are valid for data with both diagonal and full covariance matrices (Appendix A).

6. Empirical analysis

In order to see how VETS works on real data, we conducted an experiment on data from the M5 competition (Makridakis et al., 2021). We use this data because our initial motivation came from a retail context and M5 contains data from the retail company, Walmart, together with a natural hierarchy, allowing us to apply multivariate models. We aggregated the original time series to monthly frequency and to department-store level in order to avoid intermittence in the original data, which gave us 70 products, each with 63 observations. We used the `adam()` function from the `smooth` package (Svetunkov, 2021b) for R with `bounds="admissible"` (aligning the parameter space with the one of VETS) in order to classify the data into trended and non-trended, seasonal and non-seasonal. This resulted in the classification shown in Figure 1. As can be seen, the majority of time series (1+27+1+28=57) are seasonal and mainly

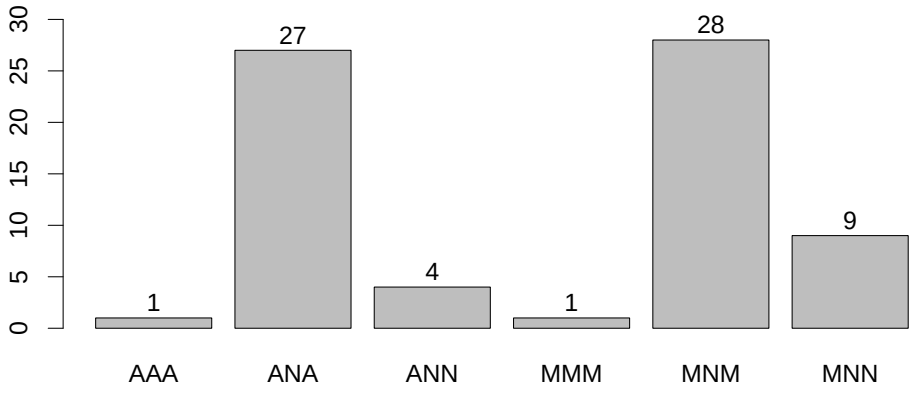

Figure 1: ETS models selected for M5 data by the `adam` function.

do not exhibit trends. The trends in the two time series are not very strong, and so models with trend components have been excluded from consideration. We have classified the series as seasonal and non-seasonal, keeping the original department-store groups. This resulted in the grouping of time series in Appendix B. The last group contained only one time series, and so we removed it from the experiment, as a vector model would not make sense in this situation.

We then decided to conduct two experiments:

1. On the full data, using 51 observations for model estimation and withholding the last 12 observations in order to measure the accuracy of models;
2. On the short data, using the last 36 observations with 24 for model fitting and the next 12 for the accuracy measure.

This ensures that all the results are comparable, as the same holdout sample is used in both cases.

We have applied the VETS(M,N,M;LN) model to each group of seasonal data and VETS(M,N,N;LN) for each non-seasonal group, automatically selecting the PIC restrictions for both using the `auto.vets()` function from the `legion` package, implementing the algorithm described in Subsection 4.4. We argue that the pure multiplicative models are more appropriate for such data, because the scale of sales for each product would differ and assuming that they would have common components (e.g. common seasonal index of 400 units each January) is unreasonable. This model is called "VETS F" further in the text, where "F" stands for "Full" covariance matrix. Given that the likelihood estimation involves many parameters to estimate, we also estimated VETS assuming that the covariance matrix for the error term is diagonal (`loss="diagonal"`), implying that the errors are independent. This model is denoted "VETS D". The grouping we have done in this section yields 11 VETS F and 11 VETS D models for our experiment. Finally, we used the ETS model (Svetunkov, 2021c), implemented in `adam()` from the `smooth` package,

applying ETS(A,N,N) and ETS(A,N,A) to the data in logarithms with `distribution="dnorm"` and then exponentiating the results, to make them consistent with a VETS approach. We denote this model as "ETS LN". Furthermore, we have measured the performance of ETS(M,N,N) and ETS(M,N,M) models implemented in `ets()` function from `forecast` (denoted "ETS", Hyndman and Khandakar, 2008) and `adam()` function from `smooth` with `distribution="dnorm"` (denoted "ETS (ADAM)") packages for R. We include ETS (ADAM) model in order to see whether the potential differences between the models are due to their implementation or due to the differences between the conventional multiplicative ETS and the one applied on data after logarithmic transformations.

In order to measure the accuracy of models, we calculated geometric means and medians of the rMAE, taking Naïve forecasts as the benchmark, aligning the measure with the one used in Section 5. Having both (geometric) means and medians provides some additional information about the distribution of errors: if the two are close to each other, then the distribution of rMAE is close to being symmetric. In addition, medians are more robust to potential outliers and will not be impacted by rare cases of poor performance. Furthermore, we have calculated separately rRMSE for the same setting to see whether the ranking of models would differ, but the results were similar and so we do not present them here.

Sample size	Model	Overall		Seasonal		Non-Seasonal	
		Geom Mean	Median	Geom Mean	Median	Geom Mean	Median
Long	ETS LN	0.917	0.959	0.906	**0.878**	0.972	1.031
	ETS (ADAM)	0.955	1.000	0.946	0.964	0.996	**1.015**
	ETS	0.933	0.957	0.920	0.924	0.996	**1.015**
	VETS D	**0.894**	**0.934**	**0.878**	0.888	0.975	1.029
	VETS F	0.906	0.960	0.894	0.934	**0.965**	1.028
Short	ETS LN Short	0.997	0.988	0.993	**0.974**	1.018	0.993
	ETS (ADAM)	0.991	1.000	0.979	0.999	1.047	1.027
	ETS	1.000	1.000	0.994	1.003	1.029	0.997
	VETS D	**0.936**	**0.974**	**0.923**	**0.974**	0.999	**0.974**
	VETS F	1.011	1.000	1.020	1.006	**0.969**	0.976

Table 11: Overall empirical performance on M5 data in terms of rMAE: overall, on seasonal and on non-seasonal data.

The results of accuracy performance are reported in Table 11. For the full data, the VETS models are almost always better than Naïve and almost always better than ETS LN, ETS (ADAM) and ETS, although the ETS LN is slightly more accurate than the other two models. For the short data, VETS typically performs better than Naive, while ETS can sometimes be slightly worse than Naïve. These results indicate that using cross-sectional data for forecasting is beneficial for accuracy in this case. Focusing on the seasonal series, VETS D tends to perform better than VETS F, although the improvement is small. Finally, having longer history leads to improvements for all models in almost all categories, as may be expected. Overall, this indicates that having longer history and using VETS D is a robust option for

forecasting.

In order to see if the performance of ETS and VETS is statistically different, we conduct an MCB test (Koning et al., 2005) on rMAE using the `rmcb()` function from the `greybox` package (Svetunkov, 2021a).

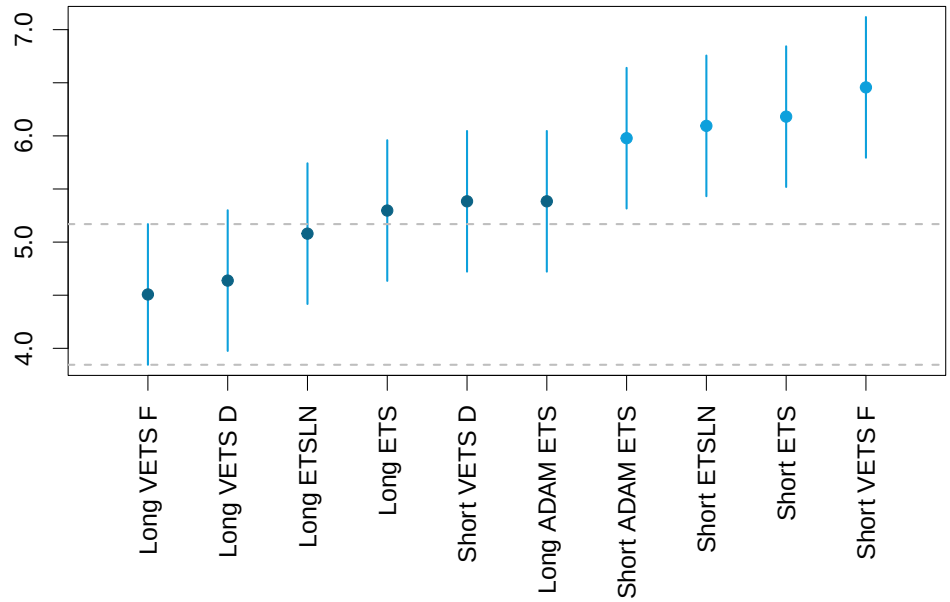

Figure 2: MCB test applied to rMAE of VETS and ETS on M5 data.

Figure 2 demonstrates results of the test. It is apparent that VETS F on long history of data has the lowest mean rank, but its performance is not statistically different from VETS D, ETS LN, ETS (ADAM) and ETS on the long sample and VETS D on the short history. When it comes to short histories, ETS LN, ETS (ADAM) and ETS are doing worse than the other approaches in terms of mean ranks, only outperforming VETS F on short history of data, although the difference between these approaches is not statistically significant.

We have also measured the performance of models in terms of prediction intervals, calculating coverage and relative pinball values, similar to how it was done in Section 5. We used prediction intervals from the ETS model on full data as a benchmark, when calculating relative pinball values. The results of this part of experiment are summarised in Table 12.

From Table 12, it is immediately apparent that ETS provides lower coverage than the required 95% in all cases. The situation is worse, when it is applied to short history data. This can have serious consequences in many forecasting applications. For example, in an inventory replenishment setting, it would result in inadequate stocks and availability falling well behind targets.

The VETS F approach performs much better than ETS in terms of coverage, producing wider pre-

Model	Long			Short		
	Coverage	pinballU	pinballL	Coverage	pinballU	pinballL
ETS LN	0.934	1.000	**1.000**	0.854	1.061	**0.895**
ETS (ADAM)	0.925	0.943	1.080	0.815	1.155	0.908
ETS	0.937	**0.912**	1.300	0.886	**0.938**	1.194
VETS D	**0.947**	0.990	1.006	**0.970**	2.515	1.567
VETS F	0.982	2.383	1.494	0.972	4.414	1.804

Table 12: Performance of models in terms of interval measures on M5 data.

diction intervals which are, in fact, a little too wide for both long and short data. This is reflected by the high relative pinball values for VETS F. This could be due to the estimation difficulties in the model, with many more parameters to estimate than ETS.

The VETS D Model on the other hand, performs the best in terms of coverage, getting very close to the nominal 95% on long data, and with a slightly smaller surplus coverage than VETS F on short data. On long data, VETS D clearly outperforms ETS models, with much better coverage with the cost of the slightly worse performance in terms of pinball values. On short data, VETS D has better relative pinball values than VETS F, reflecting the saving in the number of parameters to be estimated. However, for short data, VETS D does not perform as well as ETS models in terms of relative pinball values. The trade-off between coverage and relative pinball values is not straightforward and will depend on the domain of application. An interesting avenue for further research would be to explore the implications of this trade off using different financial loss functions.

In order to understand where the main benefits of VETS come from, we extracted the PIC parts of all 11 VETS D models (via `modelType(object, pic=TRUE)`) and calculated the frequencies of different restrictions. These are summarised in Table 13.

PIC	N,N,N	L,N,N	L,L,N	LS,S,N	LS,S,S
Frequency	1	1	2	6	1

Table 13: Frequency of different PIC restrictions for VETS on M5 data.

As can be seen from the table, the most frequently imposed restriction is for the seasonal initials and common smoothing parameters for both level and seasonal, PIC(LS,S,N). This validates our simulation findings that PIC(LS,S,N) is one of the most robust models that performs consistently well across different settings. One of the seasonal groups also had more strict restrictions with common seasonal components, PIC(LS,S,S). We argue that the main benefits observed in terms of forecasting accuracy arise from these restrictions. As for the non-seasonal groups, the most common restriction is for initial level and common smoothing parameter, although this group is small, and so it is difficult to draw reliable conclusions.

Finally, we present the results of an additional empirical evaluation, which aligns with capacity plan-

ning in retail stores. Walmart wishes to optimise allocation of retail space to their departments and product categories and, naturally, the number of units sold is an indicator they use to decide how much space to allocate (Souza, 2017). The planning horizon will vary, depending on the extent of space reallocation. For illustrative purposes, we have examined departmental volume forecasts for 3, 6 and 12 months ahead. Effective plans depend not only on the accuracy of forecasts but also on the reliability of prediction intervals. This is important to ensure that enough space is allocated, especially in peak seasons, and to ensure appropriate contraction of those departments with anticipated reductions in volume sales.

We measure performance of the same set of models on the same data (both long and short samples) using the Geometric Mean Relative Absolute Error (GMRAE, Fildes, 1992) with Naïve forecasts for the benchmark. We also show the coverage of the forecasts but not the pinball measures, as they do not align naturally with space planning decisions. The GMRAE is used instead of rMAE measures, because the point forecasts in this case are not aggregated over the forecast horizon. The results of this experiment are shown in Table 14.

Length	Model	GMRAE			Coverage		
		h3	h6	h12	h3	h6	h12
Long	ETS LN	0.748	0.962	1.092	0.928	0.928	**0.942**
	ETS (ADAM)	**0.716**	0.942	0.988	0.855	0.899	0.913
	ETS	0.888	**0.827**	**0.939**	0.899	0.899	0.928
	VETS D	0.897	0.938	1.007	0.928	**0.957**	0.971
	VETS F	0.923	0.894	0.980	**0.942**	0.971	0.971
Short	ETS LN	1.162	1.167	0.905	0.841	0.826	0.899
	ETS (ADAM)	1.170	1.198	0.926	0.826	0.797	0.841
	ETS	1.058	**0.946**	0.936	0.870	0.855	0.913
	VETS D	**1.037**	1.112	0.938	0.870	0.841	0.899
	VETS F	1.065	1.031	**0.872**	**0.957**	**0.942**	**0.942**

Table 14: GMRAE and Coverage for specific horizons for M5 data.

As can be seen from Table 14, VETS models do not perform best in terms of Geometric Mean Relative Absolute Error for the specified horizons on longer time series, but on shorter ones they do much better. In these cases, the cross-sectional information helps in the estimation. Furthermore, VETS F performs consistently better than the other models, for short data history, in terms of 95% coverage. This is a strong point in its favour, given the importance of prediction intervals for space planning decisions.

Summarising the empirical evaluation, we would recommend using VETS, estimated via likelihood with the diagonal covariance matrix (assuming that the errors of different time series are not correlated) on real data, as it is easier to estimate than VETS with the full covariance matrix or even ETS, especially on small samples. The only exception when we would not recommend using VETS would be when pinball

measure is of the main importance and the available history is short. Using cross sectional information for estimation brings benefits in terms of forecasting accuracy, which comes mainly from the restrictions on the seasonal component, but assumes that the data is correctly grouped.

7. Conclusions and further research

In this paper we studied the effect of cross-sectional information on the performance of forecasting models on seasonal data. Using the Vector ETS (VETS) framework for pure additive and pure multiplicative models, we proposed a PIC taxonomy to examine the forecasting accuracy effect of having common or individual **P**arameters, **I**nitial values, and **C**omponents. This taxonomy is general and can be applied to models with any combination of level, trend (including damped trend) and seasonality. The number of variations proposed in the taxonomy can become large; therefore, we also propose a model selection mechanism and it is tested empirically using data from the M5 competition. The advantage of the full taxonomy is that it opens up the widest pool of methods to the practitioner and invites consideration of the potential benefits of estimating, in common across series, the relevant parameters, initial values and components. However, there are some situations when some of these elements do not need any common treatment. In our empirical experiments, it was mainly the parameters (particularly for level and seasonality) and initial values (for seasonality) that benefited from common estimation. Only for one group were common (seasonal) components required. For all other groups, the 'C' element could have been dropped from the PIC taxonomy and consideration restricted to 'PI' variations. (This facility is also available in the `vets()` function from the `legion` package for R, which was used in the experiments in this paper). Our practical recommendation, to operational researchers, therefore, is to consider the full PIC framework in the first instance, but to dispense with any elements that are not necessary and to use a simpler framework when appropriate. In this way, the model selection process can be simplified, thus lending itself to practical implementation, with fewer models to be evaluated.

Given our interest in applications to seasonal time series, through the simulation experiment we focused on VETS(M,N,M;LN) and examined the relative performances of nine PIC variations in point forecasts and prediction intervals. Methods with individual seasonal components and common smoothing parameters (PIC(LS,*,N) and PIC(S,*,N)) show good performance and are robust across various scenarios. This becomes especially apparent with interval results, and overall it is observed that PIC(LS,*,N) methods perform better than PIC(S,*,N) ones. These methods offer a good balance in terms of accuracy and flexibility.

We also found that restrictions on the smoothing parameters tend to lead to improvements in terms

of accuracy and prediction intervals performance for all the methods in the taxonomy. This is because the more restrictive methods lead to averaged-out smoothing parameters, which protects them from the effect of overfitting.

The empirical analysis of the M5 data shows strong performance of the VETS models against the benchmark of univariate ETS models. It was found that VETS D (diagonal covariance matrix) performs better than VETS F (the full matrix) on both long and short samples of data. ETS performs better when there are more observations available, but, even then, VETS D outperforms it.

The evidence from this paper suggests that using common smoothing parameters for all components has the biggest positive impact in terms of accuracy and prediction intervals performance out of the taxonomy. Common seasonal initial values are also beneficial because of increased efficiency in estimating the number of parameters. As a result, PIC(LS,S,N) performs the best in both simulations and empirical analysis (selected in 6 out of the 11 groups). In many practical settings, where data histories are short and there are many series to forecast, this approach can be particularly useful in offering parsimony and robustness, while the traditional individual approach can become too cumbersome or even infeasible.

Our current paper makes a contribution to the forecasting literature in the area of dynamic models; it also opens up many potentially fruitful opportunities for further investigations. The more restrictive methods can be applied to shorter data histories (in theory, at least a year of data), but it is not yet clear whether this strategy would be effective. One of potential strategies for the estimation of high dimensional multivariate models on small samples is shrinkage of parameters, which could be investigated in context of VETS in future papers. In practice, many businesses are collecting data on a much more granular level, so efforts in examining seasonality in higher frequencies such as weekly or even daily patterns are needed. Additional seasonal indices and the trend element may change the dynamics of how these methods behave and reveal fresh insights. Future work can also look at pre-processing, for example, by grouping seasonally homogeneous series. In our view, this current research and its many possible extensions promise to deepen understanding in seasonal demand forecasting and aid better planning. The benefits of improved planning are worthy of investigation, not only at retailers, but also in many other domains of application.

Appendix A. Pinball values for the upper bound of prediction interval from the simulation experiment

	2 series	10 series	20 series	30 series	40 series	50 series
LS,S,S	0.87	0.74	0.71	0.69	0.68	0.66
S,S,S	0.89	0.81	0.77	0.72	0.72	0.70
N,S,S	0.90	0.82	0.79	0.76	0.74	0.73
LS,S,N	0.87	0.73	0.70	0.68	0.66	0.65
S,S,N	0.88	0.76	0.74	0.72	0.70	0.68
N,S,N	0.89	0.82	0.79	0.76	0.75	0.73
LS,N,N	0.91	0.83	0.81	0.80	0.78	0.78
S,N,N	0.96	0.89	0.89	0.88	0.87	0.85
N,N,N	1.00	1.00	1.00	1.00	1.00	1.00

Table A.15: rPinballU effect on group sizes PIC(LS,S,S) DGP (Benchmark PIC(N,N,N))

	2 series	10 series	20 series	30 series	40 series	50 series
LS,S,S	1.60	1.76	1.73	1.73	1.70	1.68
S,S,S	1.66	1.92	1.90	1.90	1.96	1.83
N,S,S	1.72	2.07	2.05	2.04	1.99	1.97
LS,S,N	1.23	1.19	1.18	1.18	1.16	1.15
S,S,N	1.30	1.50	1.56	1.50	1.43	1.41
N,S,N	1.33	1.48	1.58	1.62	1.67	1.74
LS,N,N	0.93	0.85	0.83	0.82	0.81	0.79
S,N,N	0.96	0.91	0.91	0.91	0.89	0.88
N,N,N	1.00	1.00	1.00	1.00	1.00	1.00

Table A.16: rPinballU effect of group sizes on PIC(N,N,N) DGP (Benchmark PIC(N,N,N))

Appendix B. Grouping of time series from M5 competition data

Using the encoding from the M5 competition, we can distinguish the following groups of time series, based on the algorithm described in Section 6:

- Seasonal:
  - Group 1, Hobbies 1: CA_1_HOBBIES_1, CA_3_HOBBIES_1, TX_1_HOBBIES_1, TX_3_HOBBIES_1, WI_1_HOBBIES_1, WI_3_HOBBIES_1;
  - Group 2, Hobbies 2: CA_2_HOBBIES_2, CA_4_HOBBIES_2, TX_1_HOBBIES_2, TX_2_HOBBIES_2, TX_3_HOBBIES_2, WI_2_HOBBIES_2, WI_3_HOBBIES_2;
  - Group 3, Household 1: CA_1_HOUSEHOLD_1, CA_2_HOUSEHOLD_1, CA_3_HOUSEHOLD_1, CA_4_HOUSEHOLD_1, TX_1_HOUSEHOLD_1, TX_2_HOUSEHOLD_1, TX_3_HOUSEHOLD_1, WI_1_HOUSEHOLD_1, WI_2_HOUSEHOLD_1, WI_3_HOUSEHOLD_1;
  - Group 4, Household 2: CA_1_HOUSEHOLD_2, CA_2_HOUSEHOLD_2, CA_3_HOUSEHOLD_2, CA_4_HOUSEHOLD_2, TX_1_HOUSEHOLD_2, TX_2_HOUSEHOLD_2, TX_3_HOUSEHOLD_2, WI_1_HOUSEHOLD_2, WI_2_HOUSEHOLD_2, WI_3_HOUSEHOLD_2;
  - Group 5, Foods 1: CA_1_FOODS_1, CA_2_FOODS_1, CA_3_FOODS_1, CA_4_FOODS_1, TX_1_FOODS_1, TX_2_FOODS_1, TX_3_FOODS_1, WI_1_FOODS_1;
  - Group 6, Foods 2: CA_1_FOODS_2, CA_2_FOODS_2, CA_3_FOODS_2, TX_1_FOODS_2, TX_2_FOODS_2, TX_3_FOODS_2, WI_3_FOODS_2;
  - Group 7, Foods 3: CA_1_FOODS_3, CA_2_FOODS_3, CA_3_FOODS_3, CA_4_FOODS_3, TX_1_FOODS_3, TX_2_FOODS_3, TX_3_FOODS_3, WI_1_FOODS_3, WI_2_FOODS_3;
- Non-seasonal:

- Group 8, Hobbies 1: CA_2_HOBBIES_1, CA_4_HOBBIES_1, TX_2_HOBBIES_1, WI_2_HOBBIES_1;
- Group 9, Hobbies 2: CA_1_HOBBIES_2, CA_3_HOBBIES_2, WI_1_HOBBIES_2;
- Group 10, Foods 1: WI_2_FOODS_1, WI_3_FOODS_1;
- Group 11, Foods 2: CA_4_FOODS_2, WI_1_FOODS_2, WI_2_FOODS_2;
- Group 12, Foods 3: WI_3_FOODS_3;

References

Akram, M., Hyndman, R.J., Ord, J.K., 2009. Exponential smoothing and non-negative data. Australian & New Zealand Journal of Statistics 51, 415–432. doi:10.1111/j.1467-842X.2009.00555.x.

Armstrong, J., 2004. Damped seasonality factors: Introduction. International Journal of Forecasting 20, 525–527.

Armstrong, J.S., 1985. Long-Range Forecasting (2nd Edition). Wiley.

Athanasopoulos, G., Hyndman, R.J., Kourentzes, N., Petropoulos, F., 2017. Forecasting with temporal hierarchies. European Journal of Operational Research 262, 60–74. doi:10.1016/j.ejor.2017.02.046.

Athanasopoulos, G., de Silva, A., 2012. Multivariate Exponential Smoothing for Forecasting Tourist Arrivals. Journal of Travel Research 51, 640–652. doi:10.1177/0047287511434115.

Bedrick, E.J., Tsai, C.L., 1994. Model Selection for Multivariate Regression in Small Samples. Biometrics 50, 226. doi:10.2307/2533213.

Bunn, D.W., Vassilopoulos, A., 1993. Using group seasonal indices in multi-item short-term forecasting. International Journal of Forecasting 9, 517–526.

Chan, J.C., Eisenstat, E., Strachan, R.W., 2020. Reducing the state space dimension in a large TVP-VAR. Journal of Econometrics 218, 105–118. doi:10.1016/j.jeconom.2019.11.006.

Chen, H., Boylan, J.E., 2007. Use of individual and group seasonal indices in subaggregate demand forecasting. Journal of the Operational Research Society 58, 1660–1671.

Chen, H., Boylan, J.E., 2008. Empirical evidence on individual, group and shrinkage seasonal indices. International Journal of Forecasting 24, 525 – 534.

Chen, J.L., Li, G., Wu, D.C., Shen, S., 2019. Forecasting seasonal tourism demand using a multiseries structural time series method. Journal of Travel Research 58, 92–103.

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

Davydenko, A., Fildes, R., 2013. Measuring forecasting accuracy: The case of judgmental adjustments to sku-level demand forecasts. International Journal of Forecasting 29, 510–522.

Dekker, M., Van Donselaar, K., Ouwehand, P., 2004. How to use aggregation and combined forecasting to improve seasonal demand forecasts. International Journal of Production Economics 90, 151–167.

Duncan, G., Gorr, W., Szczypula, J., 1993. Bayesian forecasting for seemingly unrelated time series: Application to local government revenue forecasting. Management Science 39, 275–293.

Fildes, R., 1992. The evaluation of extrapolative forecasting methods. International Journal of Forecasting 8, 81–98.

Fildes, R., Hibon, M., Makridakis, S., Meade, N., 1998. Generalising about univariate forecasting methods: further empirical evidence. International Journal of Forecasting 14, 339–358.

Gardner, E.S., McKenzie, E., 2011. Why the damped trend works. Journal of the Operational Research Society 62, 1177–1180. doi:10.1057/jors.2010.37.

Gorr, W., Olligschlaeger, A., Thompson, Y., 2003. Short-term forecasting of crime. International Journal of Forecasting 19, 579–594.

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

Hyndman, R.J., Khandakar, Y., 2008. Automatic time series forecasting: the forecast package for R. Journal of Statistical Software 26, 1–22.

Hyndman, R.J., Koehler, A.B., Ord, J.K., Snyder, R.D., 2008. Forecasting with Exponential Smoothing. Springer Berlin Heidelberg.

Hyndman, R.J., Koehler, A.B., Snyder, R.D., Grose, S., 2002. A state space framework for automatic forecasting using exponential smoothing methods. International Journal of Forecasting 18, 439–454.

Koehler, A.B., Snyder, R.D., Ord, J.K., 2001. Forecasting models and prediction intervals for the multiplicative Holt-Winters method. International Journal of Forecasting 17, 269–286.

Koning, A.J., Franses, P.H., Hibon, M., Stekler, H.O., 2005. The M3 competition: Statistical tests of the results. International Journal of Forecasting 21, 397–409. doi:10.1016/j.ijforecast.2004.10.003.

Kourentzes, N., Athanasopoulos, G., 2019. Cross-temporal coherent forecasts for Australian tourism. Annals of Tourism Research 75, 393–409. doi:10.1016/j.annals.2019.02.001.

Kourentzes, N., Athanasopoulos, G., 2021. Elucidate structure in intermittent demand series. European Journal of Operational Research 288, 141–152. doi:10.1016/j.ejor.2020.05.046.

Kourentzes, N., Petropoulos, F., 2016. Forecasting with multivariate temporal aggregation: The case of promotional modelling. International Journal of Production Economics 181, 145–153. doi:10.1016/j.ijpe.2015.09.011.

Kourentzes, N., Petropoulos, F., Trapero, J.R., 2014. Improving forecasting by estimating time series structural components across multiple frequencies. International Journal of Forecasting 30, 291–302. doi:10.1016/j.ijforecast.2013.09.006.

Lee, N., Choi, H., Kim, S.H., 2016. Bayes shrinkage estimation for high-dimensional VAR models with scale mixture of normal distributions for noise. Computational Statistics and Data Analysis 101, 250–276. doi:10.1016/j.csda.2016.03.007.

Li, C., Lim, A., 2018. A greedy aggregation–decomposition method for intermittent demand forecasting in fashion retailing. European Journal of Operational Research 269, 860–869.

Lütkepohl, H., 2005. New Introduction to Multiple Time Series Analysis. Springer Berlin Heidelberg.

Makridakis, S., Spiliotis, E., Assimakopoulos, V., 2021. The M5 competition: Background, organization, and implementation. International Journal of Forecasting doi:10.1016/j.ijforecast.2021.07.007.

McKenzie, E., Gardner, E.S., 2010. Damped trend exponential smoothing: A modelling viewpoint. International Journal of Forecasting 26, 661–665. doi:10.1016/j.ijforecast.2009.07.001.

Ouwehand, P., Hyndman, R.J., de Kok, T.G., van Donselaar, K.H., 2007. A state space model for exponential smoothing with group seasonality. Monash Econometrics and Business Statistics Working Papers 7/07. Monash University, Department of Econometrics and Business Statistics. URL: https://ideas.repec.org/p/msh/ebswps/2007-7.html.

Pennings, C.L., van Dalen, J., 2017. Integrated hierarchical forecasting. European Journal of Operational Research 263, 412–418.

Schäfer, J., Strimmer, K., 2005. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statistical Applications in Genetics and Molecular Biology 4, 1–30. doi:10.2202/1544-6115.1175.

de Silva, A., Hyndman, R.J., Snyder, R., 2010. The vector innovations structural time series framework. Statistical Modelling: An International Journal 10, 353–374. doi:10.1177/1471082X0901000401.

Snyder, R.D., Ord, J.K., Koehler, A.B., McLaren, K.R., Beaumont, A.N., 2017. Forecasting compositional time series: A state space approach. International Journal of Forecasting 33, 502–512.

Souza, K., 2017. Walmart exec discusses store layout challenges as online retail grows. Talk Business and Polytics. URL: https://talkbusiness.net/2017/10/wal-mart-exec-discusses-store-layout-challenges-as-online-retail-grows/. (version: 2022-02-01).

Spiliotis, E., Petropoulos, F., Kourentzes, N., Assimakopoulos, V., 2020. Cross-temporal aggregation: Improving the forecast accuracy of hierarchical electricity consumption. Applied Energy 261, 114339. doi:10.1016/j.apenergy.2019.114339.

Svetunkov, I., 2021a. greybox: Toolbox for Model Building and Forecasting. URL: https://github.com/config-i1/greybox. R package version 0.6.9.

Svetunkov, I., 2021b. smooth: Forecasting Using State Space Models. URL: https://github.com/config-i1/smooth. R package version 3.1.1.

Svetunkov, I., 2021c. Time series analysis and forecasting with adam. OpenForecast. URL: https://openforecast.org/adam/. (version: 2021-04-13).

Svetunkov, I., Pritularga, K.F., 2021. legion: Forecasting Using Multivariate Models. URL: https://github.com/config-i1/legion. R package version 0.1.0.

Taieb, S.B., Taylor, J.W., Hyndman, R.J., 2020. Hierarchical probabilistic forecasting of electricity demand with smart meter data. Journal of the American Statistical Association 0, 1–17.

Team, R.C., 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Winters, P.R., 1960. Forecasting sales by exponentially weighted moving averages. Management Science 6, 324–342.

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