Forecasting for food demand

Fotios Petropoulos[a,*], Shawn Carver[b]

[a] *School of Management, University of Bath, UK*
[b] *Fiddlehead Technology Inc., Moncton, New Brunswick, Canada*

Abstract

The sustainability of food supply chains depends on accurately predicting future demand. Forecasting will form the basis for making decisions with regards to replenishment from the distribution centers and ordering from the suppliers. This chapter explores methods and approaches for forecasting for food demand.

Keywords: statistical forecasting, demand, supply chain, method selection, hierarchies, judgement

1. Introduction

Accurate forecasting of food demand has significant economic and environmental consequences. Unreliable forecasts can result in a multitude of problems that ripple across the food supply chain ranging from frequent changes to production schedules, expedited shipments, and high inventory carrying costs to poor customer service levels, stock-outs, and significant waste.

This chapter focuses on methods and approaches for forecasting for food demands. In section 2, we describe methods that could be employed if just univariate time series data are available; such methods are able to capture and model a variety of predictable series patterns, such as trends and seasonality. Then, section 3 discusses how such methods can be expanded to include causal information, such as promotional activity. The chapter continues (section 4) with approaches for selecting the best across different forecasting methods for a particular series via validation and cross-validation approaches. Finally, in sections 5 and 6 we explore how hierarchical structures (both cross-sectional and temporal) can enhance the forecasting process and further improve the forecasting performance. The last section of this chapter concludes and provides a brief discussion on the role of judgement in the forecasting process.

2. Univariate forecasting

Time series data comprise of the interaction of several predictable and unpredictable patterns, such as level, trend, seasonality, randomness (noise) and cycle. In some cases, time series may also contain irregular (outlying) or missing values and as such data pre-processing may be required prior to extrapolation. Statistical forecasting targets on the modelling and extrapolation

*Correspondance: F Petropoulos, East Building, School of Management, University of Bath, Claverton Down, Bath, BA2 7AY, UK.

Email address: f.petropoulos@bath.ac.uk (Fotios Petropoulos)

of the predictable series patterns (level, trend and seasonality) while effectively smoothing out the unpredictable ones. Univariate forecasting methods solely focus on historical observations of the variable of interest (usually the recorded demand patterns for a particular stock keeping unit). In that sense, univariate forecasting can be seen as driving a car just by seeing through the rear mirror. Regardless, numerous empirical studies have demonstrated the effectiveness of univariate forecasting and its importance to demand planning.

Time series data can be widely grouped in four categories with regards to their predictable patterns: level-only (no trend nor seasonality is evident), trend-only, seasonal-only and trend and seasonal data. More refined categories may be considered for distinguished between different types of trend (additive or multiplicative; linear or damped) and seasonality (additive or multiplicative). In the next four subsections, we explore simple univariate methods to extrapolate time series data from each of the four main categories.

2.1. Forecasting level data

When the past data do not exhibit neither trend nor seasonality, then level-only methods are considered suitable for extrapolating such signals. Simple level-only methods include:

- Naive method, where the forecast for the next period equals to the last observed actual. Note that this method has no parameters and requires just one past data point. Naive method reacts fast when the level of the data changes, however it is not robust against outliers as it simply copies forward the noise in the data.
- Global average, where the forecast for the next period equals to the arithmetic mean of all past observations. As with the Naive method, global average has no parameters. Contrary to the Naive method, global average is robust against outliers however it is quite slow in reacting to level changes.
- Simple moving average (SMA), where the forecast for the next period equals to the arithmetic mean of the last k observations, where k can take positive integer values in $[1, n]$ and n is the number of available past observations. Note that if $k = 1$ then SMA corresponds to the Naive method whereas if $k = n$ then SMA is equivalent to the global average method. Even if unequal weighting schemes might be considered, in its simpler version SMA assumes equal weights for each of the k last observations.

A more robust method for forecasting for level-only data is provided by the Simple Exponential Smoothing (SES) method (Brown, 1956). SES is the simplest method in the Exponential Smoothing family of methods. The SES forecast for the next period is calculated as

$$\hat{y}_{t+1} = \alpha y_t + (1 - \alpha)\hat{y}_t. \tag{1}$$

where y_t represents the actual value for period t, while $\hat{y}_t$ is the forecast for the same period. Note that equation 1 suggests that the forecast for period $t + 1$ is a linear combination of the actual at period t and the forecast at period t. In fact, the weights of the combination are controlled by the α smoothing parameter which reflects to the combination weight of y_t. Note that α takes values in $[0, 1]$. Also note that when $\alpha = 1$ then SES is equivalent to the Naive method.

Let us now rewrite equation 1 by replacing t with n, which suggests that we create forecasts for the out-of-sample, and $\hat{y}_n$ with its respective linear combination using again equation 1:

$$\hat{y}_{n+1} = \alpha y_n + (1 - \alpha)[\alpha y_{n-1} + (1 - \alpha)\hat{y}_{n-1}] = \alpha y_n + (1 - \alpha)\alpha y_{n-1} + (1 - \alpha)^2 \hat{y}_{n-1}. \tag{2}$$

By repeating this process and replacing $\hat{y}_{n-1}$, then $\hat{y}_{n-2}$ and so on so forth until we reach the beginning of the data, we get

$$\hat{y}_{n+1} = \alpha y_n + (1-\alpha)\alpha y_{n-1} + (1-\alpha)^2 \alpha y_{n-2} + \cdots + (1-\alpha)^{n-2}\alpha y_1 + (1-\alpha)^{n-1}\hat{y}_1. \quad (3)$$

The last equation suggests that:

- the forecast for the next period $n+1$ is a linear combination of all past data (y_t with t in $1, 2, ..., n$) and the very first forecast ($\hat{y}_1$). Also, given that $0 \leq \alpha \leq 1$, the weights for the actuals exponentially decay. For example, if $\alpha = 0.5$, then the weights for y_n, y_{n-1} and y_{n-2} are 0.5, 0.25 and 0.125 respectively.
- SES has, apart from the α smoothing parameter, another parameter which is the initial forecast ($\hat{y}_1$); this forecast cannot be calculated through equation 1 thus it has to be estimated.

The optimal value for the α smoothing parameter is usually automatically selected by the forecasting software by minimising the one-step-ahead in-sample mean squared error (MSE). Linear or non-linear optimisation techniques may be used for this task. The value of the initial forecast could be either calculated using simple initialisations (such as $\hat{y}_1 = y_1$) or optimised along with α.

While in-sample forecasts are usually produced using equation 1 for one-step-ahead, multi-horizon forecasts can also be calculated as

$$\hat{y}_{t+h} = \alpha y_t + (1-\alpha)\hat{y}_t, \quad (4)$$

which suggests that forecasts for more than one-step ahead are simply equal to the one-step-ahead forecast; or else, the out-of-sample forecasts from SES can be represented as a straight horizontal line.

SES can also be expressed in components form:

$$l_t = \alpha y_t + (1-\alpha) l_{t-1}, \quad (5)$$

$$\hat{y}_{t+h} = l_t. \quad (6)$$

The above set of equations suggest that SES estimates just one component (the current level) and the forecast for the next period(s) equals to the estimation of the current level. Note that when expressed in components form, SES initialisation parameter is called initial level (l_0). Equations 5 and 6 are equivalent to equation 1; in the following, more complex exponential smoothing models will be directly expressed in components form.

Figure 1 presents an example of forecasting a demand series with SES. The values of smoothing parameter and initial level are optimised using the `forecast` package of the R statistical software ($\alpha = 0.565$ and $l_0 = 923.17$).

2.2. Forecasting trended data

If the data exhibit trend (change of the level over time), either this is upwards or downwards, then SES will produce biased forecasts (forecasts that will systematically either under or over the real outcomes). In such cases, methods that are able to capture the trend component should be considered as well. One such method is the Holt's exponential smoothing method (HES, Holt,

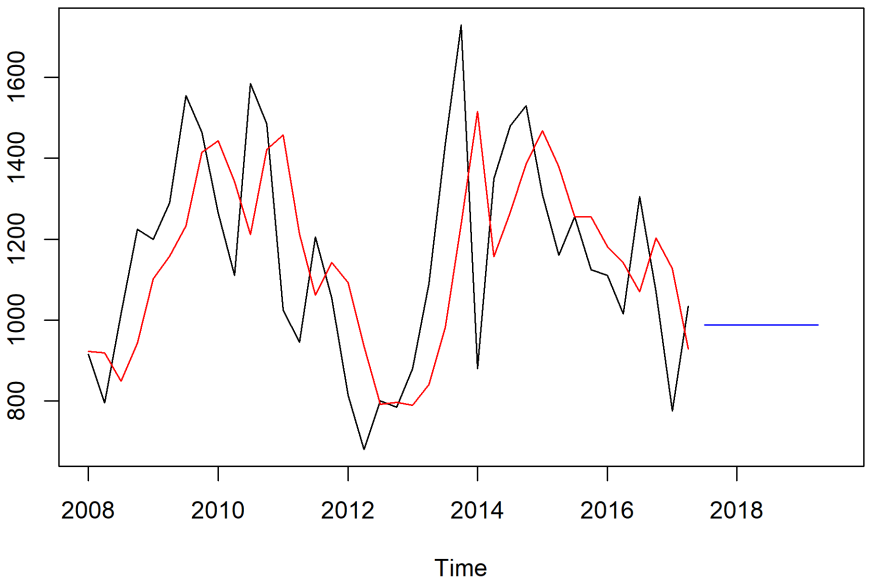

Figure 1: Forecasting with SES (black line: past actual values; red line: fitted values; blue line: forecasts).

1957) which can be expressed as:

$$l_t = \alpha y_t + (1-\alpha)(l_{t-1} + b_{t-1}), \quad (7)$$

$$b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1}, \quad (8)$$

$$\hat{y}_{t+h} = l_t + hb_t. \quad (9)$$

Compared to equation 5, equation 7 estimates the current level as a linear combination of the current actual observation, y_t, and the current forecast, $\hat{y}_t = l_{t-1} + b_{t-1}$ (note that the current forecast in SES consisted only by the level, $\hat{y}_t = l_{t-1}$). Equation 8 estimates the trend component. This works in a similar fashion with the estimation of the level component. The current trend, b_t, is the linear combination of the difference of the latest two level estimations, $l_t - l_{t-1}$ which can also be seen as the local trend, and the previously estimated trend, b_{t-1}. The weights of the linear combination for trend component are controlled by β smoothing parameter, which (as was the case with α) takes values in $[0, 1]$. Finally, the h-step-ahead forecast is the sum of the latest estimated level plus the latest estimated trend multiplied by the respective horizon, h. In other words, HES forecast is a straight line that exhibits a linear trend (either upwards or downwards based on the sign of b_t). HES has in total four parameters: α and β smoothing parameters, initial level, l_0, and initial trend, b_0. We suggest that the initial states are optimised along with the smoothing parameters. Lastly note that equation 8 suggests that the estimation and smoothing of the trend component depends on the smoothing of the level component. At the same time, the estimation of the level component depends on the trend estimation for the previous period. As such, the selection of optimal α and β values should be done concurrently rather than serially.

HES assumes that trend will always be linear and the forecast will increase/decrease with the same rate regardless the forecasting horizon. However, this is not a reasonable assumption especially if one considers the typical life-cycle of products. Thus, Gardner and McKenzie (1985)

proposed a damped-trend variation of HES, the Damped exponential smoothing (DES) method. DES is expressed as follows:

$$l_t = \alpha y_t + (1-\alpha)(l_{t-1} + \phi b_{t-1}), \quad (10)$$

$$b_t = \beta(l_t - l_{t-1}) + (1-\beta)\phi b_{t-1}, \quad (11)$$

$$\hat{y}_{t+h} = l_t + \sum_{i=1}^{h} \phi^i b_t. \quad (12)$$

The above set of equations suggests that the estimation of the trend for one-step-ahead forecasts is multiplied by a damping parameter, ϕ, which usually takes values in $[0.8, 1]$. The trend for multi-step-ahead forecasts is further dampened. Assuming $\phi = 0.9$, the trend, b_t, for the 1, 2 and 3-steps ahead forecasts of DES is multiplied by $\phi = 0.9$, $\phi + \phi^2 = 1.71$ and $\phi + \phi^2 + \phi^3 = 2.439$ respectively. In contrast, the trend for the respective forecasts of HES is multiplied by 1, 2 and 3. The DES forecasts do not exhibit a linear trend like HES; the long-term DES forecast is an almost horizontal line. DES falls to HES when $\phi = 1$. Also, note that if we allow $\phi > 1$ then an exponential trend is assumed. The damping parameter, ϕ, is optimised along with the other parameters so that the one-step-ahead MSE is minimised.

Figure 2 provides a visual example of applying HES and DES on quarterly data that exhibit trend.

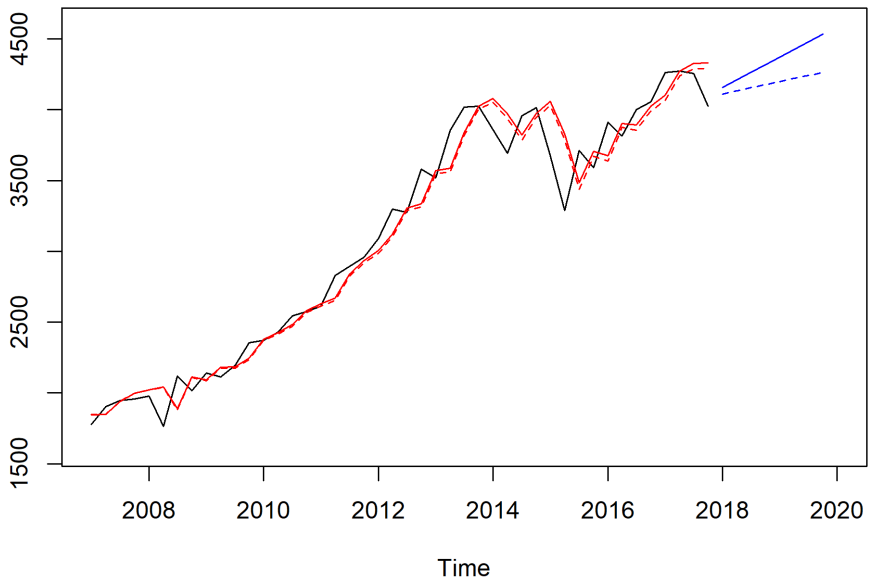

Figure 2: Forecasting with HES and DES (black line: past actual values; red lines: fitted values; blue lines: forecasts; solid lines: HES; dashed lines: DES).

2.3. Forecasting seasonal data

Food demand patterns often have single and/or multiple seasonal cycles. For example, monthly demand of ice cream will increase during the summer months. Other products, such as alcoholic

beverages, exhibit weekly seasonality when observed on a daily frequency, with the demand being higher on Fridays and Saturdays.

A simple seasonal method arises from an extension of the naive method, where the forecast for the next period (for example, Saturday if the frequency was daily) is the realised demand of the previous respective period (the demand for last Saturday). However, we suggest the more robust exponential smoothing models, which can be suitably expanded to include such seasonal patters. For example, a component that smooths the seasonal component could be added to the SES (equations 5 and 6) so that:

$$\begin{aligned} l_t &= \alpha(y_t - s_{t-m}) + (1-\alpha)l_{t-1}, && (13)\\ s_t &= \gamma(y_t - l_t) + (1-\gamma)s_{t-m}, && (14)\\ \hat{y}_{t+h} &= l_t + s_{t+h-m}. && (15) \end{aligned}$$

Equation 14 estimates the seasonal component with m representing the number of periods within a full seasonal cycle (12 for monthly data, 4 for quarterly, 7 for daily). A separate smoothing parameter, γ, is used for smoothing the seasonal component; γ, as with α and β, takes values in $[0, 1]$, however low values should be avoided unless sufficient cycles of data are available. Equation 13 suggests that the level is smoothed on the seasonally adjusted data, $y_t - s_{t-m}$ whereas equation 14 suggests that the seasonal index is updated based on the value of the local seasonal index, $y_t - l_t$.

There exist m separate seasonal indices (for example, for monthly data, one for January one for February and so on so forth). Each seasonal index is updated every m periods (whereas the level component is updated every single period). Estimation of a seasonal method requires at least two full cycles of data, however we suggest that 3 or even 4 cycles should be used. Note that the method expressed in equations 13-15 has in total $m + 3$ parameters, the smoothing parameters (α and γ), initial level (l_0) and initial seasonal indices for each period ($s_{1-m}, s_{2-m}, ..., s_0$). The forecast is the sum of the estimation of the level and the seasonal index for the respective period. An example of forecasting using the seasonal SES method described above is given in figure 3.

The seasonality of the above method is assumed to be additive in form (i.e. the effect of the seasonal cycles does not interact with the level and the seasonal indices are expressed absolute values); thus, this method is called Additive Seasonal SES. In any case, a Multiplicative Seasonal SES can also be expressed as follows:

$$\begin{aligned} l_t &= \alpha(y_t / s_{t-m}) + (1-\alpha)l_{t-1}, && (16)\\ s_t &= \gamma(y_t / l_t) + (1-\gamma)s_{t-m}, && (17)\\ \hat{y}_{t+h} &= l_t s_{t+h-m}. && (18) \end{aligned}$$

Contrary to the Additive Seasonal SES, the current seasonally adjusted data and seasonal index are derived by division: y_t / s_{t-m} and y_t / l_t. Similarly, the forecast is the product of the level by the respective seasonal index. Note that in multiplicative seasonal methods the seasonal indices are expressed in values around unity (representing percentages) and it is assumed that the seasonal effect interacts with the level. Also note that multiplicative seasonal methods are not applicable when the data contain negative or zero values.

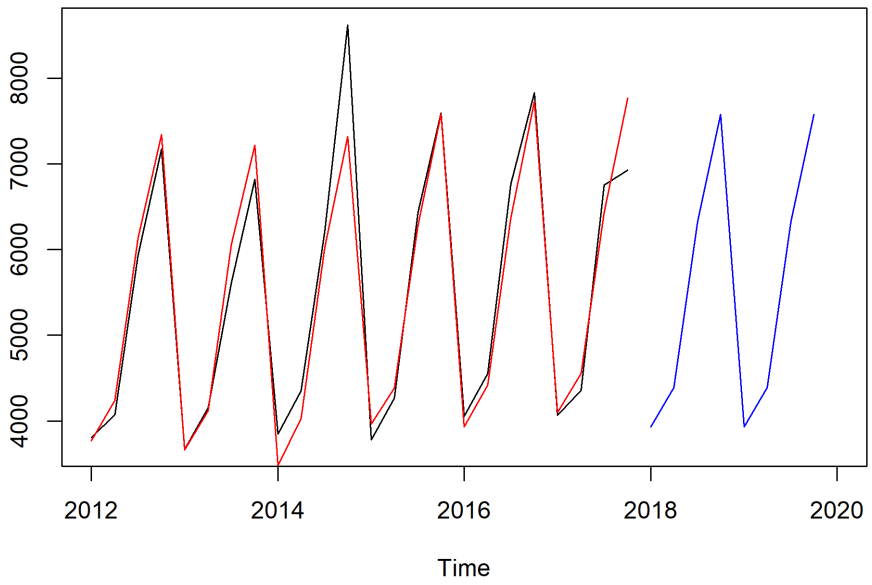

Figure 3: Forecasting with Additive Seasonal SES (black line: past actual values; red line: fitted values; blue line: forecasts).

2.4. Forecasting trended and seasonal data

In the most complex form, exponential smoothing methods have three separate components, level, trend and seasonality. Such methods are widely known as Holt-Winters exponential smoothing methods. Similarly to section 2.3, seasonality may be expressed in either additive or multiplicative forms.

Holt-Winters method with additive seasonality is expressed as:

$$\begin{aligned} l_t &= \alpha(y_t - s_{t-m}) + (1-\alpha)(l_{t-1} + b_{t-1}), && (19)\\ b_t &= \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1}, && (20)\\ s_t &= \gamma(y_t - l_t) + (1-\gamma)s_{t-m}, && (21)\\ \hat{y}_{t+h} &= l_t + hb_t + s_{t+h-m}. && (22) \end{aligned}$$

Holt-Winters method with multiplicative seasonality is expressed as follows.

$$\begin{aligned} l_t &= \alpha(y_t/s_{t-m}) + (1-\alpha)(l_{t-1} + b_{t-1}), && (23)\\ b_t &= \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1}, && (24)\\ s_t &= \gamma(y_t/l_t) + (1-\gamma)s_{t-m}, && (25)\\ \hat{y}_{t+h} &= (l_t + hb_t)s_{t+h-m}. && (26) \end{aligned}$$

Each component is smoothed separately using a different smoothing parameter. Equations 20 and 24 are exactly the same as the respective equations in section 2.2; similarly, equations 21 and 25 reflect to the equations 14 and 17 of section 2.3. The estimation of the level is a linear combination of the seasonally adjusted actual, $y_t - s_{t-m}$ or y_t/s_{t-m}, and the seasonally adjusted forecast, $l_{t-1} + b_{t-1}$ (equations 19 and 23). Finally, the Holt-Winters h-step-ahead forecast consists

of the sum of the level and trend estimations, $l_t + hb_t$, which is subsequently added to (multiplied by) the respective seasonal index, s_{t+h-m}, when the seasonality is additive (multiplicative). Holt-Winters method has $m + 5$ parameters to be estimated/optimised, including the three smoothing parameters for the level the trend and the seasonality. Lastly, note that Holt-Winters methods could be suitably adjusted to include a damped trend component rather than a linear one. Figure 4 provides an illustrative example of the HW with additive seasonality.

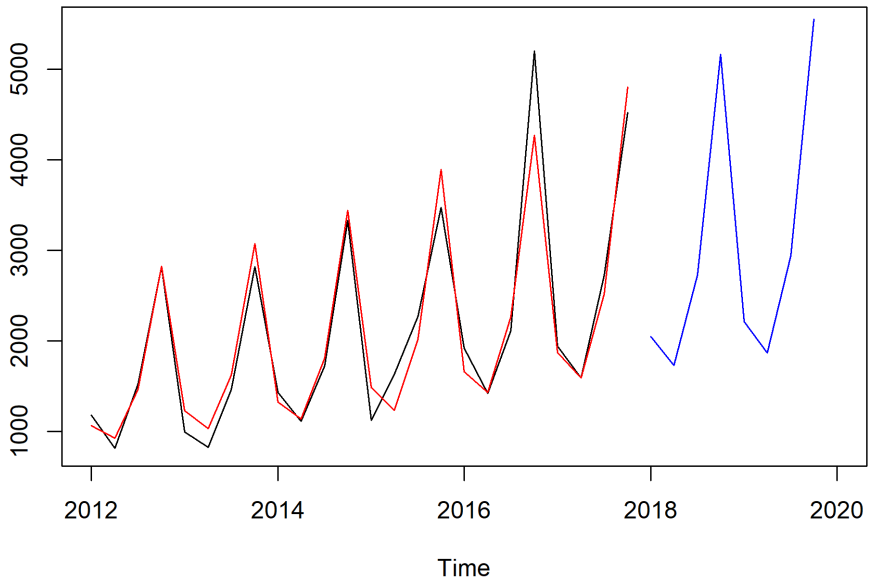

Figure 4: Forecasting with HW with additive seasonality (black line: past actual values; red line: fitted values; blue line: forecasts).

3. Including exogenous variables

Quite often, looking just at past demand values to make predictions is not enough. Several drivers may influence the demand, such as weather, own or competitors' promotions, sport events and other special events or actions. The effect of such drivers should be taken into account when building forecasting models. Traditionally, regression-type models have been used to incorporate the effect of such causal variables. However, we can also expand the formulation of exponential smoothing methods as to include information from external variables (Kourentzes and Petropoulos, 2016). For example, SES can be extended to include additional components for estimating the effect of exogenous variables, X_i with $i = 1, 2, ..., N$:

$$l_t = l_{t-1} + \alpha e_t, \tag{27}$$

$$d_{i,t+1} = b_i x_{i,t+1} \tag{28}$$

$$\hat{y}_{t+1} = l_t + \sum_{i=1}^{N} d_{i,t+1}. \tag{29}$$

where $d_{i,t}$ refers to the effect at time t of a variable X_i with observations $x_{i,t}$ and b_i is the respective coefficient. Similar to regression models, the set of coefficients show the additive effect of the

respective variables. Estimation of the b coefficients is performed together with the estimation of the rest parameters of the exponential smoothing methods (for the example of SES, together with α and initial level) by minimising the in-sample MSE. The extension shown above for SES can be applied to any of the exponential smoothing methods. Note that the (natural) assumption here is that exogenous information is available for the future periods (as is the case for future planned promotions) or it can be predicted.

In some cases, the correlation between the exogenous variables may be high enough to lead to multicollinearity issues and its negative effects on the estimation of the coefficients. According to Kourentzes and Petropoulos (2016), such issues are especially relevant when temporal aggregation is applied on the data (we will further discuss temporal aggregation on section 6). As such, Kourentzes and Petropoulos (2016) suggest that the variables are transformed to become orthogonal. Principal components analysis returns a set of principal components orthogonal to each other, that have no redundant information and are ordered in terms the variance extracted. To maintain the model as simple and at the same time as effective as possible, we suggest the inclusion of the first k principal components, so that the sum of the extracted variance is between 60 and 80%. As the estimated coefficients for the principal components cannot be directly interpreted with regards to the original variables, it is suggested that a back-transformation is applied so that the coefficients of the original variables are derived.

Kourentzes and Petropoulos (2016) showed that exponential smoothing with exogenous variables outperforms significantly univariate methods as well as regression models, both in terms of bias and accuracy.

4. Selecting between methods

Given the plethora of available methods to select from coupled with the fact that food and beverage related stock keeping units usually counts tens of thousands of items within a retailer, suitable automatic selection strategies seem to be necessitate. In short, there exist three classes of automatic method selection strategies, namely selection based on past performance, selection based on information criteria and selection based on rules. We suggest that, if enough data are available, selection is performed via the first option, which we also describe in detail in this section. In the last paragraph of this section we also briefly discuss the other two options.

Selection on past performance can be achieved by suitably splitting the available historical data (in-sample) into two sets, the training set and the test set. Subsequently, using the training set as input forecasts that correspond to the periods of the test set are produced from all available methods. Finally, the forecasts of each method are compared against the with-held actuals on the training set observations and the methods are ranked in terms of performance given a cost function, such as MSE, mean absolute error (MAE) and mean absolute percentage error (MAPE). More formally, assume that the available historical data are presented in a vector y with values $y_1, y_2, ..., y_n$. Also assume that the in-sample is divided so that the training set consists of the $n-p$ first observations of y where as the test set consists of the last p observations. The value of p can be decided so as to match the required forecast horizon, h, however other options are available (such as $p = m$). In any case, the training set should include an adequate number of observations that would allow fitting of the applied forecasting methods. The p-steps-ahead forecasts of method j, $\hat{y}^j_{n-p+1}, \hat{y}^j_{n-p+2}, ..., \hat{y}^j_n$, are produced using the first $n-p$ observations and with-holding the last p. These are then compared against the actuals; for instance, if mean absolute error is used as a cost function, then $MAE_j = p^{-1} \sum_{i=n-p+1}^{n} |y_i - \hat{y}^j_i|$. This process is repeated for every of the J

forecasting methods available, or $j = 1, 2, ..., J$. Finally, the method, j, with the minimum value of MAE_j is selected. Once a method is selected, forecasts beyond end of the in-sample data (n) are produced using all available observations, $y_1, y_2, ..., y_n$. The approach described above is known as fixed origin evaluation and some times is also referred to as validation for time series data. Forecast origin is the period from which the forecasts originate; in the case of validation, there is only one forecast origin referring to the selection process, which is the last period of the training set, $n - p$. The above process is illustrated in figure 5.

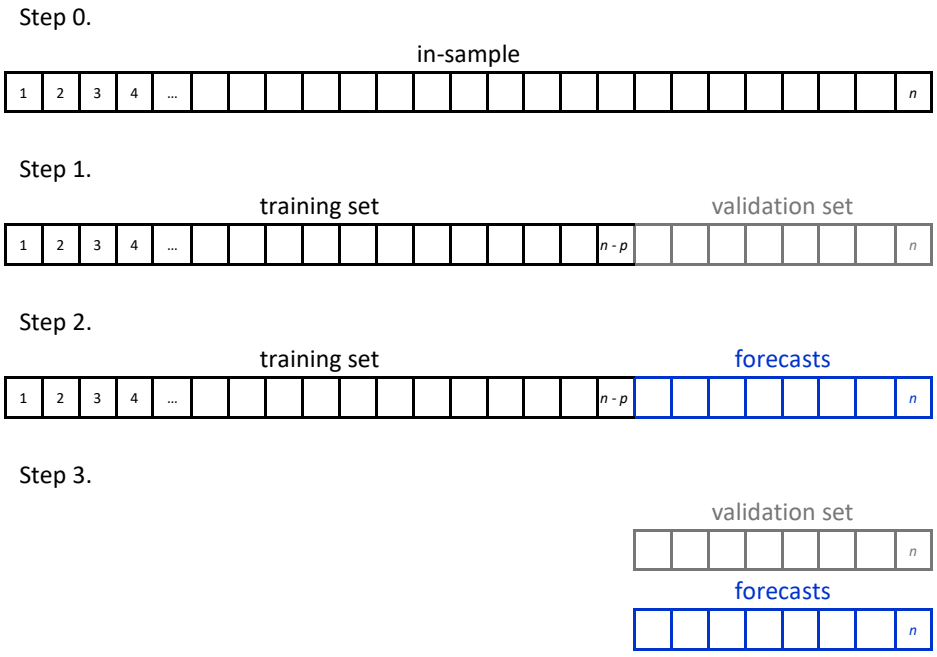

Figure 5: Method selection through validation. Step 0: the in-sample consists of n observations. Step 1: the in-sample is divided to the training set and the validation set. Step 2: the training set is fitted to produce forecasts for the validation periods. Step 3: the forecasts are evaluated against the actuals of the validation set. Steps 2 and 3 are repeated for every method available.

However, a fixed origin evaluation has the disadvantage of overly focusing on a single validation window, on which irregularities might occur. A better way to select between methods based on their past performance would be through a multiple validation approach, also called cross-validation or rolling origin evaluation. Forecasts for each method are not produced just once from one origin (reflecting to a single validation period) but multiple times (rolling origin). Most usually, overlapping validation windows are considered and the rolling through the origins takes place by one period at a time. However, non-overlapping windows can also be defined. The case of overlapping cross-validation can be formally described as follows. The in-sample data may be divided in the initial training set, consisting of $n - p$ observations, and the cross-validation set, consisting of p observations. Forecasts for the next h' periods (where $h' < p$) are produced by each method using the initial training set. Then, the training set increases by one period, so that it now contains $n - p + 1$ observations; h'-periods-ahead forecasts are again produced from the new origin (y_{n-p+1}). This process is repeated until the origin y_{n-p+q} where $p - q = h'$ so that all

h'-step-ahead forecasts can be evaluated. Consequently, $q+1$ sets of h' forecasts each are produced for each of the J available methods. These forecasts are compared against the respective actuals of the cross-validation period that were withhold. Finally, the method with the best cross-validated performance is selected. The process of selecting via cross-validation is graphically depicted in figure 6.

For the curious readers that would like to see more details on fixed versus rolling origin evaluation schemes, we suggest the paper by Tashman (2000).

5. Refocusing the scope: cross-sectional aggregation

Data within an organisation are often organised in cross-sectional hierarchical structures. Each level of the hierarchy refers to a different degree of data aggregation. For example, possible hierarchical levels include sales by stock keeping units (SKUs), category, location, channels, clients or combinations of the above. A simple hierarchy is depicted in figure 7. This hierarchy consists from three levels, company, department and SKU. The first department consists of three SKU, while the second department consists of two SKUs. Data management processes should generally focus on the systematic recording and storing of data on the most granular level of the hierarchy within a company. In the case of figure 7, that would be the SKU-level, however in other cases the most granular level could be a combination of SKU per location sales.

Historical data within a hierarchical structure are consistent. This means that for every period t the sum of the values of all the bottom level nodes is equal to the value of the top-level node. Data summation also holds for every intermediate aggregation level. However, this is not true for forecasts. Following the structure of figure 7, the sum of the forecasts for SKUs 4 and 5 (S_4 and S_5) is not generally equal to the forecast of the respective department, D_2. This leads to two important issues:

- Given that different functions of the organisation (operations, sales, marketing, finance, etc) focus on different levels of aggregations, separate sets of forecasts are independently produced within the company. Such forecasts are the basis for decisions within the different functions. However, forecasts inconsistency will inevitably lead to decision inconsistency. In other words, there is a need for approaches that lead to forecast consistency.
- While naturally forecasts needed at the SKU-level would be produced at the same level, this process does not guarantee maximisation of the forecasting performance. Time series at higher aggregation levels will generally exhibit less noise and lower levels of intermittence; at the same time, individual series characteristics are better captured and modelled at lower aggregation levels. So, a process to identify optimal levels of aggregation is required.

Four hierarchical approaches can lead to forecast consistency:

- The *bottom-up* approach is possibly the most widely applied hierarchical approach. Forecasts are produced on the most granular hierarchical level (SKU level for the hierarchy depicted in figure 7). Consequently, forecasts for all other levels are calculated as the respective sums of the lowest level forecasts following the hierarchical structure.
- In the *top-down* approach, forecasts are produced only for a single hierarchical node that represents the highest hierarchical level (Company level for the hierarchy depicted in figure 7). Forecasts for lower aggregation levels are calculated as proportions of the forecasts of the highest level. Historical or forecasted proportions may be considered (Athanasopoulos et al., 2009). In the former case, the top-down approach requires model fitting, parametrisation

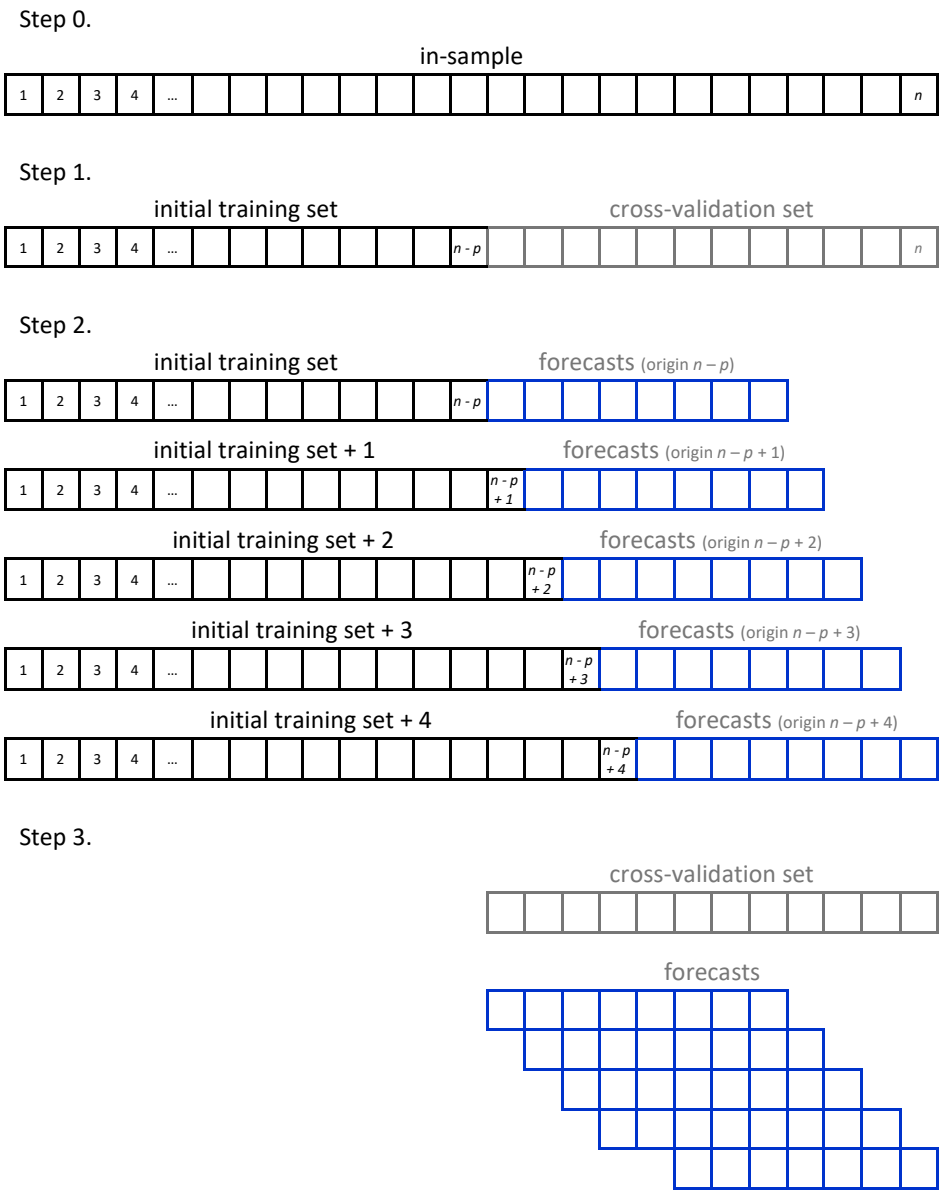

Figure 6: Method selection through cross-validation. Step 0: the in-sample consists of n observations. Step 1: the in-sample is divided to the initial training set and the cross-validation set. Step 2: the initial training set is fitted to produce forecasts; one more observation is added at the end of the initial training set; forecasts are produced again from the next origin; this is repeated until the forecasts cover all the cross-validation set. Step 3: the multiple sets of forecasts are evaluated against the actuals of the cross-validation set. Steps 2 and 3 are repeated for every method available.

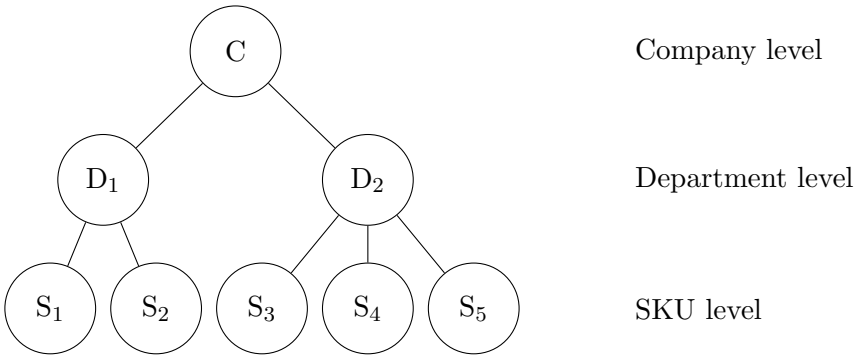

Figure 7: A standard hierarchy.

and forecasting of a single series (time series data on the very top level), rendering it the least expensive hierarchical approach.

- The *middle-out* approach is conceptually a combination of bottom-up and top-down approaches. Initially, forecasts are produced at a middle hierarchical level (Department level for the hierarchy depicted in figure 7). Then, forecasts for higher aggregation levels are calculated using structurally-imposed summations (similarly to the bottom-up approach). On the other hand, forecasts for lower levels are derived by appropriate forecast decomposition (similarly to the top-down approach). Fiddlehead's fieldwork suggests that this is the most commonly used hierarchical approach among the food and beverage manufacturers.
- The most recently proposed hierarchical approach is called *optimal combination* (Hyndman et al., 2011). In this approach, forecasts at all levels are generated. These forecasts are subsequently reconciled in a statistically optimal way. Optimal combination is the single hierarchical approach that directly takes into account forecasts produced at all levels. However, this also renders it the most computationally intensive approach.

Regardless the fact that all four aforementioned approaches lead to forecast consistency that is essential to consistency in terms of decisions, still a selection between these fours approaches is required so that forecasting performance is maximised. Such a selection can follow the guidelines for forecasting method selection presented in section 4. The performance of the various hierarchical approaches may be measured through a validation or cross-validation fashion. The approach with the lowest error through a validation set is taken forward. Hierarchical approach selection can be completed by calculating the performance across the hierarchy or by focusing at specific hierarchical levels or even by just focusing on a single hierarchical level. If the latter is the case, it should be noted that the optimal aggregation level might not necessarily match the level where the cost-function is to be minimised. This suggests that forecasting through cross-sectional hierarchical structures allows for refocusing the scope. Decision makers are invited not to blindly extrapolate the historical data that directly refer to the variable of interest but to also investigate the predictability of alternative cross-sectional aggregation levels. This process is usually referred to as hierarchical level optimisation. However, it is worth noting that the role of portfolio segmentation, the trade-off between accuracy and the number of forecasting units to be maintained at each level, as well as the level at which judgmental adjustments are most frequently applied are key elements to take into account when selecting the most appropriate hierarchical level.

6. Extracting information from the data: multiple temporal aggregation

On top of the cross-sectional aggregation of the time series data, as explored in the previous section, temporal aggregation can also be considered. Temporal aggregation refers to within series data transformation that focuses on changing the observed frequency. For instance, a monthly time series can be transformed to quarterly one via non-overlapping temporal aggregation that considers time buckets of size 3 (the aggregation level equal in this case is 3). It suggested that instead of focusing on a single frequency of the observed data (which usually matches the frequency on which the data are collected), to transform the series in many different frequencies (Kourentzes et al., 2014). For instance, a monthly time series may be transformed to bi-monthly, quarterly, four-monthly, semesterly, ...up to the yearly or even bi-yearly frequency.

Temporal aggregation of the originally observed series will help us view the same data through different lenses. Specific series characteristics are amplified or attenuated in different frequencies. When dealing with fast-moving series, within-year periodicity and seasonality is more apparent and better modelled in high-frequency time series whereas the long-term trend is easier to be identified in the lower-frequency (temporally aggregated) data (Petropoulos and Kourentzes, 2014). If the data in hand are intermittent, then temporal aggregation will reduce the degree intermittence, allowing the application of the fast-moving methods described in section 2 (Nikolopoulos et al., 2011). Moreover, the variability of the data will decrease as we move towards higher temporal aggregation levels (Spithourakis et al., 2014; Petropoulos et al., 2016b). Essentially, temporal aggregation is a cheap and efficient way that allows us to extract more information from the data in hand.

Selection of the optimal aggregation level might be difficult in practice. However, it has been empirically demonstrated that producing a forecast via combining the series components estimates from multiple aggregation levels can lead to increased forecasting performance especially for the longer horizons (Kourentzes et al., 2014). Improvements in performance through multiple temporal aggregation have also been observed in the cases where exogenous information is available (Kourentzes and Petropoulos, 2016). Such improvements are linked with tackling the model and parameters uncertainty, which refers to the holy grail of forecasting: optimally selecting model and set of parameters for the data in hand. However, an even more important benefit that arises from the application of the multiple temporal aggregation approach (on top of the improved accuracy) is the alignment of decisions on different levels: operational (usually a few weeks to a few months ahead), tactical (usually one to two quarters ahead) and strategic (usually one to two years ahead).

In a recent study, Athanasopoulos et al. (2017) proposed that multiple temporal aggregation can be expressed as hierarchical structured, the so-called temporal hierarchies. An example of a temporal hierarchy is depicted in figure 8, where the data are observed in the quarterly frequency (lowest aggregation level) and are temporally aggregated to semesters and years. Structuring multiple temporal levels as hierarchies allows the application of the hierarchical approaches discussed in section 5. More importantly, one can also consider super-hierarchies that incorporate both cross-sectional and temporal information, or cross-temporal hierarchies. Such hierarchies will allow the calibration of decisions with regards to different functions of the company (sales, operations, logistics, manufacturing, marketing, finance, upper-level management), different divisions and business operating units as well as different horizons (short- versus long-term decisions). In other words, cross-temporal hierarchies will allow for the "one number forecast" that will lead to harmonized decisions within the company.

Regardless the advantages of multiple temporal aggregation approaches, empirical evidence

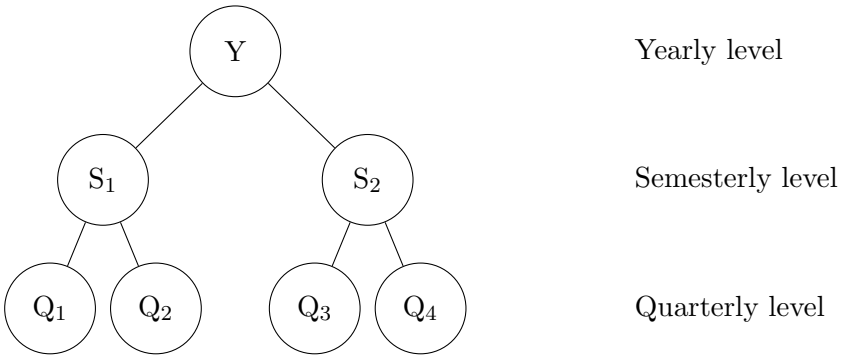

Figure 8: A temporal hierarchy.

from Fiddlehead's work has identified a few cases where the original series should be used instead. These included sudden level changes, extreme seasonality, and localised trends. In such cases, pre-processing of the original data might be considered, such as seasonal adjustments. At the same time, the performance of multiple temporal aggregation increases together with the length of the series, while de-promotionalising the data allows for better capturing the underlying characteristics.

7. Concluding remarks and the role of judgement

In this chapter, we explored methods and approaches to statistically produce forecasts for food demand. We presented univariate forecasting techniques suitable to model a variety of predictable series patterns but we also discussed how exogenous information may be incorporated into such methods. We also described approaches of automatically selecting the most suitable method for each time series individually, through validation and cross-validation processes. Finally, we explored the role of hierarchical structures, how cross-sectional and temporal aggregation can lead to increased performance and discussed issues with forecasts reconciliation.

We would also like to very briefly touch on the role of managerial judgement in the forecasting process. It is usually observed that the final forecast within any company it is not solely based on statistical models. In fact, managerial judgement may appear in different stages of the forecasting function. Sometimes, a purely judgemental forecast is produced and taken forward for decision making. We advise against such practice, as numerous studies have demonstrated that such judgemental forecasts are inferior to statistical ones. Other times, the statistical output is adjusted/modified so that the impact of anticipated special circumstances and conditions (such as forthcoming promotions or other special events) are taken into account. Several recent studies have empirically examined the conditions under which such adjustments might be beneficial in practice (for example, see: Fildes et al., 2009; Petropoulos et al., 2016a). In any case, we suggest that a forecast-value-added analysis (Gilliland, 2013) is applied so that the benefit of performing such judgemental adjustments is analysed and monitored. Finally, a recent has showed that judgement can complement to the method selection phase of the forecasting process (Petropoulos et al., 2017). In fact, wisdom of crowds (weighted combinations from the judgemental method selections of multiple experts) or a 50-50% combination of statistics + expert can bring significant benefits in terms of forecast accuracy.

References

Athanasopoulos, G., Ahmed, R. A., Hyndman, R. J., 2009. Hierarchical forecasts for Australian domestic tourism. International Journal of Forecasting 25 (1), 146–166.

Athanasopoulos, G., Hyndman, R. J., Kourentzes, N., Petropoulos, F., 10 2017. Forecasting with temporal hierarchies. European Journal of Operational Research 262 (1), 60–74.

Brown, R. G., 1956. Exponential smoothing for predicting demand. Arthur D. Little Inc, Cambridge, Massachusetts.

Fildes, R., Goodwin, P., Lawrence, M., Nikolopoulos, K., 2009. Effective forecasting and judgmental adjustments: An empirical evaluation and strategies for improvement in supply-chain planning. International Journal of Forecasting 25 (1), 3–23.

Gardner, Everette S., J., McKenzie, E., 1985. Forecasting trends in time series. Management Science 31 (10), 1237–1246.

Gilliland, M., 2013. FVA: A reality check on forecasting practices. Foresight: The International Journal of Applied Forecasting 29, 14–18.

Holt, C. C., 1957. Forecasting seasonals and trends by exponentially weighted moving averages. O.N.R. Memorandum 52/1957.

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

Kourentzes, N., Petropoulos, F., 2016. Forecasting with multivariate temporal aggregation: The case of promotional modelling. International Journal of Production Economics 181, Part A, 145–153.

Kourentzes, N., Petropoulos, F., Arenas, J. R. T., 2014. Improving forecasting by estimating time series structural components across multiple frequencies. International Journal of Forecasting 30 (2), 291–302.

Nikolopoulos, K., Syntetos, A. A., Boylan, J. E., Petropoulos, F., Assimakopoulos, V., 2011. An aggregate - disaggregate intermittent demand approach (ADIDA) to forecasting: An empirical proposition and analysis. Journal of the Operational Research Society 62 (3), 544–554.

Petropoulos, F., Fildes, R., Goodwin, P., 2016a. Do 'big losses' in judgmental adjustments to statistical forecasts affect experts' behaviour? European Journal of Operational Research 249 (3), 842–852.

Petropoulos, F., Kourentzes, N., 2014. Improving forecasting via multiple temporal aggregation. Foresight 34 (Summer 2014), 12–17.

Petropoulos, F., Kourentzes, N., Nikolopoulos, K., 2016b. Another look at estimators for intermittent demand. International Journal of Production Economics 181, Part A, 154–161.

Petropoulos, F., Kourentzes, N., Nikolopoulos, K., Siemsen, E., 2017. Judgmental selection of forecasting models. Working paper.

Spithourakis, G., Petropoulos, F., Nikolopoulos, K., Assimakopoulos, V., 2014. A systemic view of ADIDA framework. IMA Management Mathematics 25, 125–137.

Tashman, L. J., 2000. Out-of-sample tests of forecasting accuracy: an analysis and review. International Journal of Forecasting 16 (4), 437–450.