KU LEUVEN

RESEARCH CENTRE FOR
INFORMATION SYSTEMS
ENGINEERING (LIRIS)

Sales forecasting with temporal hierarchies at Christeyns

Group report

Hai Yen Nguyen R0866689

Huyen Linh Dinh R0821236

Nhat Linh Hoang R0967749

Thesis submitted to obtain the degree of
Master of Information Management

Promoter: Jente Van Belle

Daily Supervisor: Daan Caljon

Academic year: 2023-2024

Abstract.

While traditional approaches of time series forecasting often apply to one single level, temporal hierarchical forecasting is a technique where forecasts are generated at multiple levels of time periods and reconciled to produce coherent predictions. Research has shown that this method could detect time series patterns, leading to better accuracy compared to single-level forecasts. This paper examines whether temporal hierarchies yield better forecasts for Christeyns' sales, using their top 18 SKUs as the dataset. The study compared various time series forecasting approaches, including temporal hierarchical methods, and evaluated their accuracy using Mean Absolute Scaled Error (MASE). It was found that forecasts at lower granularity, using the thief (ETS) variance scaling and structural scaling approach offer superior accuracy in comparison to other methods discussed for times series forecasting.

Keywords: Sales forecasting, Temporal Hierarchies, Univariate Time Series.

Acknowledgement.

Despite our efforts, we couldn't have accomplished our thesis without the help of others.

We are grateful to our thesis supervisor, Daan Caljon, for his insightful remarks and assistance. Each thesis meeting resulted in fresh discoveries that helped further the investigation.

We would like to extend our sincere appreciation to our thesis promotor, Jente Van Belle for his persistent help during the research. The completion of this thesis would not have been possible without his exceptional competence.

In addition, we would like to thank Christeyns, Simon Lambert, Data Management Director and Delphine Verstraete, Data Scientist, for providing us with the data insights and giving us the chance to experiment with real-world observations.

Finally, we would like to express our heartfelt gratitude to our family and friends for their unwavering faith and assistance.

Table of Contents

1 Introduction

Every successful retail business in the competitive market of today has to be able to forecast its sales with accuracy since sales are crucial for businesses. Retail companies that want to effectively distribute resources and satisfy customer demand need precise order volume forecasts. Excessive allocation of resources may lead to higher expenses, whilst inadequate allocation may cause delays and dissatisfaction among consumers.

Despite time-series analytic methods and data complexity, finding the optimum prediction model is tough. Time series forecasting uses past data and other information to anticipate future values (Kotu & Deshpande, 2019). Time-series forecasting is effective for assessing data patterns and trends and developing forecasts (Kotu & Deshpande, 2019). Business, economics, and finance utilize this strategy to make choices and offer inputs.

The purpose of this study is to examine if the use of temporal hierarchical forecasting techniques may enhance the accuracy of Christeyns' product sales estimates. Our objective is to conduct a comparative analysis be-tween traditional univariate time series and advanced temporal hierar-chical forecasting techniques. Temporal hierarchical forecasting methods generate forecasts at different aggregation levels (e.g., weekly, biweekly, monthly, and quarterly) using non-overlapping temporal aggregation. Modeling data at various levels concurrently captures aggregate and granular trends, resulting in more precise forecast.

Christeyns is a global business-to-business manufacturer of hygiene chemicals. Christeyns offers installation services for cleaning and energy recovery systems to end clients, including industrial laundries and food processors (Christeyns, 2024). Improving the precision of their sales forecasting will greatly benefit Christeyns. These predictions are essential for the supply chain planning process, which include activities like pro-duction planning and raw material procurement.

The paper is organized in the following manner: Section 2 introduces the ideas of univariate time series forecasting, temporal hierarchies, and advanced forecasting algorithms. In Section 3, we provide more details and explanations of the suggested technique. Section 4 provides a detailed description of the findings and experimental setting.

2 Literature Review

2.1 Univariate times series

Introduction to Univariate Time Series Forecasting

Economics, finance, and business frequently employ time-series forecast-ing as a research method. Researchers may draw better conclusions by studying a variable's temporal trends. Univariate time series forecasting uses past observations of a variable to make forecasts (Petropoulos & Spiliotis, 2021). Univariate time series are fundamentally composed of three components: trend, seasonality, and remainder: cycle.

Trend: A trend is defined as a pattern of data that exhibits a persistent upward or downward movement over a specified period of time (Hyndman & Athanasopoulos, 2021).

Seasonality: Seasonal patterns form when a time series is affected by seasonal factors like the season or day of the week. Seasonality has a fixed duration (Hyndman & Athanasopoulos, 2021).

Cycle: cycles occur when there are irregular fluctuations. According to Hyndman and Athana-sopoulos (2021), these changes are usually set off by the "business cycle" and general economic circumstances.

Forecasting Methods

There is no single technique that can effectively address all time-series forecasting issues (Zhang & Kline, 2007). Different time-series forecasting challenges call for different methodologies.

Mean forecasting

One of the most basic forecasting techniques is the mean forecast. Using the historical average of observed data, the mean forecast, also referred to as the average forecast, predict future values (Hyndman & Athanasopoulos, 2021). Taking the average of previous data points within a certain time range defines the mean prediction. The mean value is used to forecast future occurrences. This equation represents the mean forecast:

$$\hat{y}_{t+h|t} = \bar{y} = (y_1 + \ldots + y_t)/t$$

Where:

$y_1 \ldots y_t$ stands for every single data point in the historical dataset.

t is the total number of data points in the historical dataset.

The notation $\hat{y}_{t+h|t}$ represents the estimate of $\hat{y}_{t+h}$ is obtained from the data $y_1 \ldots y_t$ (Hyndman & Athanasopoulos, 2021).

Autoregressive Integrated Moving Average (ARIMA) Models

This statistical method was first created by Box and Jenkins (1970) and is used to make predictions about time series. It is still thought to be the best way to predict time series. This model combines the moving average and autoregressive components to produce a comprehensive view of complex temporal patterns (Hyndman & Athanasopoulos, 2021).

The ARIMA model is commonly written ARIMA (p, d, q). The variables p, d, and q represent, respectively, the order of the autoregressive component, the degree of first differencing, and the order of the moving average component.

Autoregressive (AR): The autoregressive (AR) component displays the impact of past values on the current. The delayed observations of the model are represented by parameter p.

Integrated (I): Differencing raw data in this component makes the time series stationary. The number d shows how many differencing changes are needed to reach stationarity. Differencing demonstrates a method for transforming a time series that is not stationary into a stationary one by calculating the differences between subsequent observations. By using differencing, we achieve stabilization of the mean of the time series and removal of fluctuations in the data's level in ARIMA models (Hyndman & Athanasopoulos, 2021).

Moving Average (MA): This component shows the association between the current observation and a residual error from a moving average model on lagged data. The parameter q represents the moving average window size.

The whole ARIMA model is presented:

$$y'_t = c + \phi_1 y'_{t-1} + \ldots + \phi_p y'_{t-p} + \theta_1 e_{t-1} + \ldots + \theta_q e_{t-q} + e_t$$

The differenced series are y'_t and the random error e_t is given at time t. On the right side of the equation are lagged values of the differenced series and errors (Hyndman & Athanasopoulos, 2021). Once the numbers of p and q are known, the model parameters $c, \phi_i (i = 1,2, \ldots p)$ and θ_j(j $= 1,2, \ldots$ q) estimated (Hyndman & Athanasopoulos, 2021).

The ARIMA model is versatile enough to handle different time series data such as seasonal patterns and trends. Its parameters can also provide insight into the dynamics of the data. ARIMA models could handle many kinds of data, including stationary and non-stationary series, without preprocessing or alteration. Stationary time series have statistical features independent of time. Thus, time series containing trends or seasonality are not stationary since they change value over time (Hyndman & Athanasopoulos, 2021).

Exponential Smoothing (ETS)

Exponential smoothing techniques aim to capture the trend and seasonality patterns in the data, whereas ARIMA models are specifically designed to capture the autocorrelations that are inherent in the data (Hyndman & Athanasopoulos, 2021). Exponential smoothing is a technique that is frequently used for forecasting time series data. In this approach, previous observations are assigned decreasing weights based on an exponential function. This makes the technique adaptable, since the weights are automatically adjusted depending on the data, capturing any underlying patterns or trends (Malkari, 2023). It is especially beneficial for producing short-term predictions and reducing unpredictable variations in the data.

The ETS models include error, trend, and seasonality components, enabling the capture of diverse patterns with flexibility. ETS models are dynamic, which makes them well-suited for datasets that exhibit changing features over time (Yen, 2023).

The simple exponential smoothing model is presented:

$$\hat{y}_{t+1|t} = \alpha y_t + \alpha(1-\alpha)y_{t-1} + \alpha(1-\alpha)^2 y_{t-2} + \ldots$$

The smoothing parameter, α, ranges between 0 and 1. The weighted average of all observations in the series $y_1 \ldots y_t$ is used to create the one-step-ahead forecast for time t + 1 (Hyndman & Athanasopoulos, 2021).

Hyndman and Athanasopoulos (2021) propose the Holt-Winters method for trending and seasonal time series which is also known as double exponential smoothing or exponential smoothing with seasonal component and additive trend. According to Hyndman et al. (2005), the taxonomy of exponential smoothing methods includes (E) Error, (T) Trend and (S) Seasonal. Error can be either A (additive) or M (multiplicative), Trend can either N (none), A (additive) or A_d (damped) and Seasonal can either be N (none), A (additive) or M (multiplicative) (Hyndman et al., 2005). For example, AAA is the exponential smoothing method with additive trend and additive seasonal component and additive underlying error model.

ETS models may have difficulties in accurately representing long-term trends and may be influenced by the selection of smoothing parameters. Typically, they are better suited for predicting shorter-term outcomes, and their accuracy may decrease when used with data that has complicated or nonlinear patterns (Yen, 2023).

2.2 Temporal aggregation and temporal hierarchical forecasting.

Temporal aggregation

The idea behind temporal aggregation, as outlined by Kaltsounis et al. (2023), is to address model uncertainty by aggregating forecasts from multiple models, which is also known as ensemble method. As referred to from Rostami-Tabar et al.'s study in 2023, temporal aggregation involves transforming a time series from a higher granularity to a lower granularity. The motivation is to detect patterns that may be challenging to discover when analyzing the original series. By altering the frequency of the data, it results in a change in the series characteristics, enabling different patterns to be detected, and therefore enhancing forecast accuracy (Kaltsounis et al., 2023).

Temporal aggregation was proved to be applicable for predicting both slow-moving and fast-moving series. For slow-moving objects - where data exhibit intermittent or erratic behaviour, aggregation eliminates randomness, exposing the underlying signal of the series. For fast-moving series, the transformation reveals trend patterns while accommodating seasonality and level modelling at various levels. Apart from improving accuracy, forecasting with temporal aggregation, when combined with hierarchical approach, offers the advantage of producing reconciled predictions that facilitate synchronized decisions across different planning horizons (Kaltsounis et al., 2023).

The granularity of a time series can be altered by employing overlapping or non-overlapping methods and following either bottom-up, top-down, or middle-out approaches hierarchically. Recent research suggests that the effectiveness of these approaches varies based on factors such as autocorrelation presence, aggregation level, forecast horizon, and the chosen forecasting method (Rostami-Tabar et al., 2023).

In overlapping temporal aggregation, time buckets with sizes equal to the aggregation level are created. At each aggregation, the time window is shifted one step backward or one period ahead, so that the newest observation is dropped and an older one is included or vice versa. Therefore, the main advantage of the overlapping temporal aggregation approach is that more blocks are available compared to when the time buckets do not overlap. On the other hand, this induces correlations between time blocks (Babaï et al., 2021).

In non-overlapping aggregation, time series are created by adding up the values inside consecutive non-overlapping buckets, where the length of the time bucket equals the aggregation level. One benefit of non-overlapping temporal aggregation is that it preserves demand's autocorrelation structures. Nonetheless, a primary drawback is that if the aggregation level is long or the demand history is short, only a small number of blocks are obtained, and part of the oldest data is removed if the history length is not a multiple of the aggregation level (Babaï et al., 2021).

Temporal hierarchical forecasting

When several temporal aggregations combine, Kaltsounis et al. (2023) refer to this as "multiple temporal aggregations". Athanasopoulos et al. (2017) introduced a new approach to multiple temporal aggregation which resembles a hierarchical forecasting framework, called temporal hierarchies. Abolghasemi et al. (2019) define hierarchical time series as several time series grouped hierarchically across distinct levels, allowing for aggregation and disaggregation depending on criteria such as location, size, and product type.

Given a time series $\{y_t,\ t = 1, \ldots, T\}$ with observations recorded at frequency *m*, Athanasopoulos et al. (2017) demonstrated the construction of the temporal hierarchies when aggregating *k* observations (*k* is a factor of *m*) using the following equation:

$$y_j^{[k]} = \sum_{t=t^* + (j-1)k}^{t^* + jk - 1} y_t, \tag{1}$$

for $j = 1, \ldots, \lfloor T/k \rfloor$ and $M_k = m/k$ is the seasonal period of the aggregated series. Because of the non-overlapping aggregation requirement, the total number of observations must be a multiple of *m*. This is ensured by starting the aggregation from $t^* = T - \lfloor T/m \rfloor m + 1$.

The factors *k* of *m* is denoted in descending order as $\{k_p, \ldots, k_3, k_2, k_1\}$ where $k_p = m,\ k_1 = 1$ and *p* represents the distinct count of aggregation levels.

Acknowledge that for each different aggregation level *k*, the observation index *j* changes, the observation index *j* is set equal to *i* for the most aggregated series to have a consistent index across all levels, i.e. $y_i^{[m]}$.

Applying the equation (1) and index $i = 1, \ldots, \lfloor T/m \rfloor$, each observation at each level of aggregation is demonstrated as $y^{[k]}_{M_k(i-1)+z}$, for $z = 1, \ldots, M_k$. As such, when index *i* increases by 1 unit, the timeseries advances M_k periods.

For each aggregation level below the annual level, observations are organized using column vectors defined as

$$y^{[k]}_i = \left(y^{[k]}_{M_k(i-1)+1}, y^{[k]}_{M_k(i-1)+2}, \ldots, y^{[k]}_{M_k i}\right)' . \tag{2}$$

By collecting these vectors into a single column vector:

$$y_i = \left(y^{[m]}_i, \ldots, y^{[k_3]'}_i, y^{[k_2]'}_i, y^{[k_1]'}_i\right)'$$

we can have:

$$y_i = S y^{[1]}_i \tag{3}$$

The “summing” matrix, denoted as *S*, is derived from research by Hyndman et al. (2011).

For example, with *m = 4*, we have the temporal hierarchy as follow:

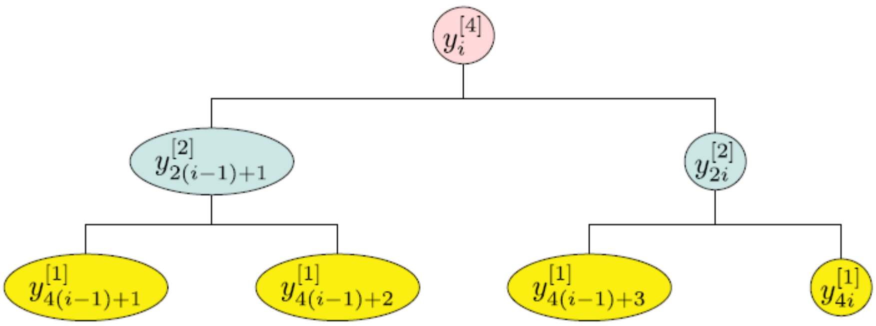


Fig. 1. Temporal hierarchy using the common index i for all levels of aggregation (m = 4) (Athanasopoulos et al. (2017))

And the summing matrix is written as:

$$S = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

It was also mentioned by Athanasopoulos et al. (2017) that representing aggregated time series data in a single tree structure is not always feasible. In some cases, aggregated levels can be organized into more than one separate hierarchy (i.e. for monthly data with *k* in *{12, 6, 4, 3, 2, 1}*). A unique hierarchy can only be constructed when there are no coprime pairs among the values of *k*.

A forecasting approach related to the hierarchical structure is known as hierarchical forecasting, which involves generating forecasts at multiple levels by aggregating or disaggregating processes. Studies conducted by Kourentzes et al. (2017) suggest that employing forecasts with temporal hierarchies leads to greater accuracy compared to conventional forecasting methods, especially under increased modelling uncertainty. The temporal hierarchical model can be constructed by combining base forecasts generated for each time series, such as ETS or ARIMA.

According to Hyndman & Athanasopoulos (2021), for temporal hierarchical forecasting, there is a requirement for forecasts to stay coherent across all temporal levels in the hierarchy. In order to compile with this requirement, the traditional way was to use single level approaches including top-down, bottom-up, and middle-out approaches. The efficiency of these models is determined by the time series characteristics, forecasting level, forecasting horizon, and hierarchy structure (Abolghasemi et al., 2019).

The bottom-up method involves forecasting each series at the bottom level, and then aggregate them to obtain forecasts at the higher levels of the hierarchy. While this approach ensures that no information is lost due to aggregation, it may perform poorly on highly aggregated data. Additionally, in large hierarchies, bottom-up forecasting can be labour-intensive, computationally expensive, and susceptible to noise (Oliveira & Ramos, 2019).

The top-down method involves forecasting the most aggregated series at the top level, and then disaggregating these, using either historical or forecasted proportions, to obtain bottom-level forecasts. However, top-down approaches relying on historical proportions often yield less accurate forecasts at lower hierarchy levels, because of significant information loss and challenges in disaggregating forecasts downwards (Oliveira & Ramos, 2019).

The middle-out approach is a combination of both bottom-up and top-down methods. Forecasts for series at an intermediate level of the hierarchy are initially obtained, with forecasts for series above the intermediate level derived using the bottom-up approach, and those below using the top-down approach (Oliveira & Ramos, 2019).

The traditional approaches outlined earlier only produced at one specific level of the aggregation structure before either aggregated for higher levels, or disaggregated for lower levels. These approaches disregard potential correlations across hierarchical levels (Athanasopoulos et al., 2019). To overcome this drawback, Wickramasuriya et al. (2019) proposed an advanced forecast reconciliation approach called Minimum Trace (MinT). MinT approach's objective is to find an optimal P matrix that minimises the total forecast variance of the set of coherent forecasts. The purpose of P is to map the base forecasts into disaggregated forecasts at the lowest level in the hierarchy, which are subsequently summed by the summing matrix S. Considering that certain levels of aggregation or specified groupings can uncover important data characteristics, one key advantage of this method over traditional ones is its ability to utilize all the accessible information contained in a hierarchical or grouped framework (Hyndman & Athanasopoulos, 2021).

For h-step ahead forecast horizon, the base forecasts are denoted by $\hat{y}_h$, then all reconciled forecasts for the hierarchy is calculated by:

$$\tilde{y}_h = SP\hat{y}_h$$

The variance – covariance matrix of the h-step-ahead coherent forecasts errors is:

$$V_h = \mathrm{Var}[y_{T+h} - \tilde{y}_h] = SPW_hP'S'$$

with $W_h = \mathrm{Var}[y_{T+h} - \hat{y}_h]$ as the variance – covariance matrix of the corresponding base forecasts errors.

According to Wickramasuriya et al. (2019), the equation of matrix P to minimize the trace of V_h is:

$$P = (S'W_h^{-1}S)^{-1}S'W_h^{-1}$$

Given the equation for all reconciled forecast for the hierarchical structures above, the optimal reconciled forecasts are achieved by:

$$\tilde{y}_h = S(S'W_h^{-1}S)^{-1}S'W_h^{-1}\hat{y}_h$$

As it is challenging in practice to estimate the variance of the forecast error W_h, Hyndman & Athanasopoulos (2021) presented some estimators:

1. The ordinary least squares (OLS) estimator assumes W_h = I, where I is an identity matrix, this simplifies P to $P = (S'S)^{-1}S'$. However, it was indicated by the authors that this OLS approach does not account for scale differences between hierarchy levels due to aggregation nor does it take into account the correlations across series.

2. The weighted least squares (WLS) estimator using variance scaling considers W_h= diag($\widehat{W}_1$) where diag($\widehat{W}_1$) is a diagonal matrix consisting of the sample variances of base forecast error:

$$\widehat{W}_1 = \frac{1}{T}\sum_{t=1}^{T} e_t e_t'$$

with e_t as a vector of n-dimensions of residuals from the models that produced the base forecasts, arranged in the same sequence as the data.

3. The structure scaling estimator with $\widehat{W}_1 = \Lambda$, where $\Lambda = diag(S1)$ and 1 is a unit column vector with the same dimension as the forecasts $\hat{y}_h^{[1]}$ from the bottom-level series. This approach assumes equal variance and uncorrelated errors at the bottom level of the hierarchy and not across all levels. It was found to be useful especially when forecast errors are unavailable, such as in judgmental forecasting scenarios (Athanasopoulos et al., 2019 and Yang et al., 2017).
4. The shrinkage estimator that shrinks the sample covariance to a diagonal matrix.

Thief package

The R package thief was developed by Rob J Hyndman and Nikolaos Kourentzes. The implements the methods described in the article Forecasting with temporal hierarchies (Athanasopoulos et al., 2017) to generate forecasts at different temporal frequencies using a hierarchical time series approach. The study's forecasting framework makes use of three scaling methods for WLS that are hierarchy variance scaling, series variance scaling, and structural scaling. For the hierarchy variance scaling, the diagonal of covariance matrix is used as the estimator, resulting in fewer variables to be estimated. On the other hand, the series variance scaling assumes homogeneous error variance within level but not across levels, leading to increase in sample size.

Thief package allows for the use of different combination methods of temporal hierarchies that are Structural scaling - weights from temporal hierarchy, Variance scaling - weights from in-sample MSE, Unscaled OLS combination weights, Bottom-up combination, GLS using a shrinkage (to block diagonal) estimate of residuals, and GLS using sample covariance matrix of residuals. Among the available combination methods, 4 combination methods that allowed for the use of estimator: variance scaling (mse), ordinal least squares (ols), structural scaling (struc), and shrinkage estimate of residuals (shr) respectively were selected into our experimental design.

3 Methodology

3.1 Overview

This study aimed to assess the use of temporal hierarchical forecasting methods in sales forecasting at our partner company – Christeyns. In order to achieve the goal, the study followed 2 key steps: (1) developing 5 forecasting models which are Mean forecast model, ARIMA model, ETS model, and 2 temporal hierarchical models with ARIMA and ETS as corresponding models used for forecasting at each aggregation level; (2) evaluate the accuracy of each model using scale independent error metrics (MASE). R programming language was chosen for this study because of its rich collection of forecasts and analysis package. Specifically, the “forecast” package was used for developing univariate models such as Mean, ARIMA and ETS, and the "thief" package was used for its suitability in generating advanced temporal hierarchical forecasts, aligning with the research question. Moreover, R offers robust visualization tools for detecting patterns and trends through graphical representation.

3.2 Dataset

The dataset used in this study is Christeyns’ historical sales data of 100 SKUs from 4th January 2016 to 4th December 2023. As agreed with Christeyns regarding the scope of this study, the study objects were narrowed down to top 20 SKUs with the focus on when (order date), how much (weight) and which product (SKU) was ordered per customer.

Prior to developing forecast models, the data preprocessing process was implemented by aggregating sales data to produce weekly data and transforming the leap year into non-leap year (52 weeks) by distributing an equal amount from the total weight of 53rd week to each of the remaining 52 weeks. This avoided problems with uneven seasonal periods. Additionally, it facilitated the aggregation consistency required for the implementation of thief to develop temporal hierarchical forecasting models. As explained above in the literature review section, in temporal hierarchical forecasting, data is often aggregated at various levels. Leap years add an extra day to the aggregation, which needs to be accounted for to ensure that all aggregations are based on equivalent time periods.

3.3 Experimental design

Models

Currently, Christeyns uses Lanham Software which relies on some basic statistical models to produce the forecast. Due to technical issues, Lanham's forecasting results were not fully available. Alternative baseline model was decided as mean forecast with our partner's acknowledgment that the mean forecasting model resulted in similar forecast accuracy in comparison to Lanham's.

Mean forecast: forecast model were employed by the 'meanf()' function.

ARIMA: forecast model were derived through the utilization of the 'auto.arima()' function, as embedded within the 'forecast' package of the R statistical software. This function is meticulously designed to autonomously determine the optimal ARIMA model, guided by the algorithm articulated by Hyndman and Khandakar (2008).

ETS: forecast model were facilitated by the 'ets' function—also a constituent of the 'forecast' package in R. The ets() function will automatically select an appropriate model based on the Akaike Information Criterion (AIC), with error, trend, and seasonality are automatically selected by the function models derived from by specifying the arguments model = "ZZZ".

Temporal hierarchical forecasting with ETS as base forecast model: using thief() function in Thief package, with 4 different estimators (mse, ols, struc, shr) for forecast error's variance for optimal reconciled forecasts.

Temporal hierarchical forecasting with ARIMA as base forecast model: using thief() function in Thief package, with 4 different estimators (mse, ols, struc, shr) for forecast error's variance for optimal reconciled forecasts.

Rolling origin evaluation scheme setup

The use of rolling origin evaluation scheme is indicated to help facilitate the forecasting accuracy for time series (Tashman, 2000). This methodological framework optimizes the segmentation of data by dynamically sliding the training set towards and progressively updating the test set following each division, thereby generating forecasts for each distinct origin.

Acknowledge the benefit of using rolling origin evaluation scheme in out-of-sample forecasting, that methodology was applied in our study by fitting each model multiple times per SKU. The rolling origin evaluation protocol was set up with a fixed length of 104 weeks (equivalent to 2 years) observations. Since temporal hierarchical forecasting involves disaggregating forecasts across different temporal granularities, having more data can improve the model's ability to capture and reconcile the patterns at each of aggregation levels. The start point of the training set is moving by a sliding step of 26 weeks. Forecasts were extended over a forecasting horizon of 26 weeks ahead, approximately equivalent to six months.

3.4 Performance measures

In this study, MASE (Hyndman & Koehler, 2006) was the selected metrics to evaluate the forecast performance. MASE is the mean absolute scaled error. The metric allows for calculating forecasting accuracy across time series of different magnitudes. MASE is calculated as:

$$MASE = \frac{MAE^a}{Q}$$

MAE^a is the mean absolute error for forecasts of method *a* and can be calculated as:

$$MAE^a = \frac{1}{h}\sum_{j=1}^{h} |y_j - \hat{y}_j|$$

in which y_j is the actual value at period *j* and $\hat{y}_j$ is the forecasted value at period *j*.

Q is the scaling factor and can be defined as follow:

$$Q = \frac{1}{T-m}\sum_{t=m+1}^{T} |y_t - y_{t-m}|$$

with the historical data is denoted by $y_1,\dots,y_T$ and the seasonal period is denoted by m.

4 Results

To determine whether temporal hierarchical forecasting outperforms traditional methods, we evaluated five forecasting techniques: Mean forecasts, ETS, ARIMA, thief (ARIMA), and thief (ETS). Our evaluation consists of two stages which are comparing the accuracy of the techniques mentioned above for weekly forecasts (1 -to -26 step ahead) and aggregated six months ahead forecasts (sum over the 1- to 26-step ahead forecasts), using average MASE (Mean Absolute Scaled Error) values as the metric. The analysis was conducted on two levels of granularity to compare their accuracy and assess their consistency. While the initial dataset comprises 20 SKUs, only 18 SKUs have sufficient time lengths to be investigated.

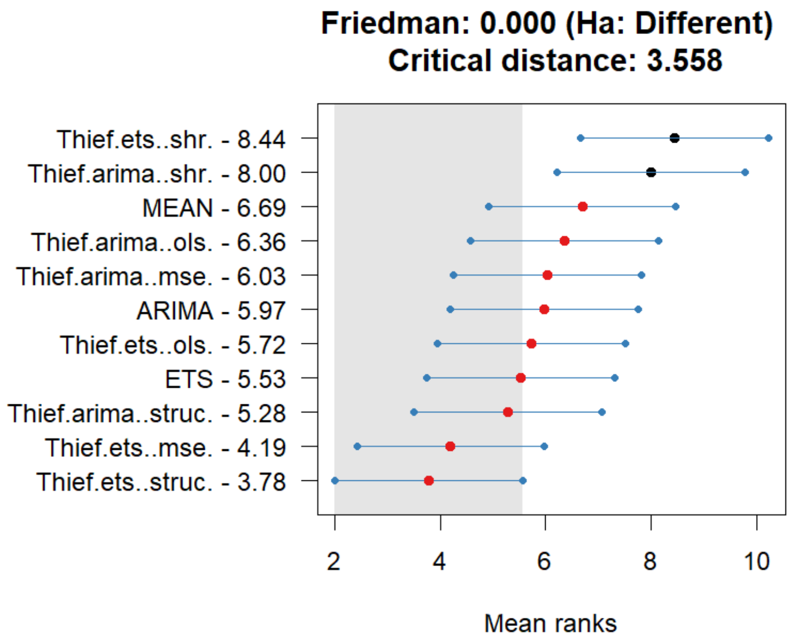
Table 1 summarizes the average MASE values for 1- to 26-week ahead forecasts across 18 different SKUs. As can be seen from the table, all five methods produced MASE scores below 1, indicating that they all perform better naïve forecasts. Thief (ETS) method with variance scaling yields the lowest error, suggesting that it is the most accurate among the methods tested, followed by the structural scaling approach.

Table 1. Average MASE for 1-26 step ahead forecasts

Mean	ARIMA	ETS	Thief (ARIMA)				Thief (ETS)			
			MSE	SHR	OLS	STRUC	MSE	SHR	OLS	STRUC
0.8160	0.7820	0.7889	0.7774	0.7876	0.7959	0.7779	0.7687	0.7979	0.7910	0.7693

Figure 2 illustrates the plot from the Friedman test, which compares the mean ranks of these forecasting methods based on their MASE scores. The dots represent the average MASE values, with horizontal lines indicating critical value ranges. This chart shows that thief (ETS) following structural scaling approach obtains the lowest mean rank, followed by the thief (ETS) with variance scaling approach. Although variance scaling's mean rank is slightly lower, it is not significantly different from the structural scaling approach. In contrast, the shrinkage approach performed on thief (ARIMA) and thief (ETS) both showed the poorest accuracy among these forecasting techniques.

Fig. 2. Friedman test on mean ranks MASE for 1-26 step ahead forecasts



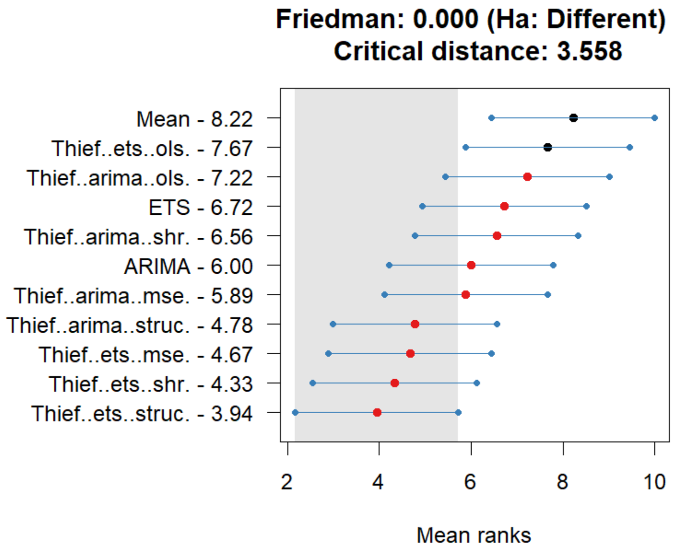
In the second stage of our evaluation, we examined the forecasting methods' performance on aggregated sales over the next six months. As shown in Table 2, these forecasts generally yield higher MASE values than those for weekly forecasts. In addition, all measures are larger than 1, indicating poorer performance than naïve forecasts. Nevertheless, the thief (ETS) following structural scaling method achieves the lowest average error score, suggesting that it might be the optimal option among the methods discussed for forecasting aggregated data.

Table 2. Average MASE for aggregated forecasts

Mean	ARIMA	ETS	Thief (ARIMA)				Thief (ETS)			
			MSE	SHR	OLS	STRUC	MSE	SHR	OLS	STRUC
4.0729	3.2759	3.1268	3.2148	3.3306	3.4187	2.9729	2.9148	2.6996	3.6697	2.6923

The Friedman plot in Figure 3 reveals that thief (ETS) with structural scaling, variance scaling, and shrinkage approaches produce the lowest mean errors, with little significant differences among their ranges.

Fig. 3. Friedman test on mean ranks MASE for sum over 1- to 26-step ahead forecasts



In summary, for weekly forecasts, the thief (ETS) using variance scaling and structural scaling delivered the most accurate predictions across 18 SKUs, while for aggregated forecasts, the best performing method is thief(ETS) with structural scaling.

5 Discuss and Critical Reflection

5.1 Insights from Literature Review and Performance Discussion

The results above align with our literature review to some extent. Firstly, the mean forecasts generally provide relatively good baseline forecasts, showing slightly higher errors than more advanced methods. Specifically, mean forecasts outperformed ARIMA in four cases out of eighteen when comparing their MASE score for weekly forecasts.

Secondly, we observe that aggregated forecasts generate higher error values than weekly forecasts, suggesting that forecasting models perform better at low granularity. The higher error for aggregated forecasts might be due to the larger scale of these sums compared to individual weekly forecasts. Despite the large differences in MASE scores, the accuracy ranking of the methods remained fairly consistent for both granularity levels.

Finally, as the studies suggest, thief(ETS) performed better overall compared to the traditional forecasting techniques, although that was not the case for thief(ARIMA).

Given this outcome, we would recommend performing forecasts at lower granularity using either thief(ETS) variance scaling or structural scaling approaches to achieve optimal accuracy.

5.2 Limitations

There are constraints to our analysis that should be taken into account. First, our analysis was limited to the top 18 Stock Keeping Units (SKUs) from Christeyns. It accounts for 40.8% of their total sales, which could be representative across Christeyns' entire product inventories but might not be representative of the chemical production industry. Therefore, the use of these results beyond this specific context should be done with caution.

Additionally, existing research indicates that univariate methods such as ARIMA and ETS are more useful for datasets that exhibit strong seasonal trends and patterns or those that prioritize recent observations while adjusting based on historical data. However, our analysis did not investigate the difference between seasonal and non- seasonal times series to determine how patterns, trends and cycles impact the forecast accuracy of different techniques. As a result, our findings do not provide the outcomes based on these seasonal variations.

Moreover, MASE was used as an evaluation metric for this study to assess the forecasts' performance. While this metric was recommended by Hyndman and Koehler in their research in 2006, it is scale- dependent and cannot be used to compare with different time series (Tim, 2020**)**. Incorporating more metrics such as Mean Absolute Percentage Error (MAPE) or Root Mean Squared Error (RMSSE) could offer more insights into the accuracy and robustness of these forecasting approaches. Nevertheless, although MAPE is scale-independent and can be used to compare different time series, it is not suitable when there are zeros in the data (Tim, 2020). As for RMSSE, for non-stationary series containing a unit root, the denominator grows with the sample size leading to higher error values (Hyndman and Koehler, 2006).

5.3 Future research

Practical implications could be retrieve from this findings and they could offer insights for Chrysteyns' supply chain managers and time series forecasting researchers. This study highlighted that advanced forecasting methods like thief (ETS) do offer better accuracy in predicting Christeyns' sales inventories. However, the same results could not be drawn for thief(ARIMA).

To better investigate the effectiveness of hierarchical forecasting approaches, further study could extend the dataset to the entire sales SKUs of the company or experience on new sample of data using data within and outside of the industry. For time series that exhibit similar trends and/ or seasonal characteristics, they can be studied together to detect possible factors that could affect the accuracy of temporal hierarchical forecasting. To enhance forecasting accuracy, advanced forecasting methods such as machine learning and neural network could be explored and adopted. Finally, adding more evaluation metrics can help better determine the most accurate forecasting techniques.

6 Conclusion

We compared temporal hierarchical forecasting to mean predictions, ETS, ARIMA, thief (ARIMA), and thief (ETS) to see whether it outperforms traditional approaches. Our review compares the accuracy of the various methodologies for weekly predictions (1 to 26 weeks ahead) and aggregated forecasts (total of 1- to 26-week forecasts) by analyzing their average MASE values. Two degrees of granularity were analyzed to assess consistency and accuracy.

Since the two SKUs' time series data was less than 52 weeks, we now only have 18 SKUs instead of the original 20. The average MASE values for 18 SKUs' 1-26 step ahead forecasts showed that all five approaches yielded MASE scores below 1, indicating improved naïve forecasts. Thief (ARIMA) and thief (ETS) had the poorest predicting accuracy using shrinkage approaches. The thief (ETS) technique with variance scaling produces the most accurate results.

We then assessed the forecasting methods' aggregated forecast performance. These forecasts had greater MASE values than weekly forecasts. In addition, all indicators exceed 1, suggesting worse performance than naïve forecasts. However, the thief (ETS) using structural scaling approach had the lowest error, suggesting it may be the best way for anticipating aggregated data. The structural scaling, variance scaling, and shrinkage approaches generated the lowest mean errors for detection (ETS), while mean forecasts produced the most.

According to the average MASE value, thief (ETS) with variance scaling produces the best weekly forecasts, whereas structural scaling performed best for the aggregated forecasts.

The findings somewhat correspond to the existing body of literature. Mean forecasts are solid baseline forecasts with somewhat larger errors than advanced approaches. Aggregated forecasts had larger error values than weekly forecasts, indicating low granularity forecasting models perform better. Thief (ETS) outperformed standard forecasting methods, but not ARIMA. Given the result, we advocate lower-granularity forecasts employing either (ETS) variance scaling or structural scaling for best accuracy.

References

Abolghasemi, M., Hyndman, R., Tarr, G., & Bergmeir, C. (2019). Machine learning applications in time series hierarchical forecasting. arXiv (Cornell University). https://arxiv.org/pdf/1912.00370.pdf

Athanasopoulos, G., Ahmed, R. A., & Hyndman, R. J. (2009). Hierarchical forecasts for Australian domestic tourism. *International Journal Of Forecasting*, 25(1), 146–166. https://doi.org/10.1016/j.ijforecast.2008.07.004

Athanasopoulos, G., Kourentzes, N., & Petropoulos, F. (2017). Forecasting with temporal hierarchies. *European Journal Of Operational Research*, 262(1), 60–74. https://doi.org/10.1016/j.ejor.2017.02.046

Athanasopoulos, G., Gamakumara, P., Panagiotelis, A., Hyndman, R., & Affan, M. (2019). *Hierarchical forecasting.* In Advanced studies in theoretical and applied econometrics (Print) (pp. 689–719). https://doi.org/10.1007/978-3-030-31150-6_21

Babaï, M. Z., Boylan, J. E., & Rostami-Tabar, B. (2021). Demand forecasting in supply chains: a review of aggregation and hierarchical approaches. *International Journal Of Production Research*, 60(1), 324–348. https://doi.org/10.1080/00207543.2021.2005268

Christeyns (2024). About Us. https://www.christeyns.com/be-nl/over-ons/

Box, G. E. P., & Jenkins, G. M. (1976). Time series analysis: Forecasting and control (Revised ed.). San Francisco, CA: Holden-Day

Hyndman, R. J., Koehler, A. B., Ord, J. K., Snyder, R. D., 2005. Prediction intervals for exponential smoothing using two new classes of state space models. *Journal of Forecasting* 24, 17–37.

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.

Hyndman, R.J., & Athanasopoulos, G. (2021) *Forecasting: principles and practice*, 3rd edition, OTexts: Melbourne, Australia. OTexts.com/fpp3. Accessed on April 30th, 2024.

Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The fore- cast package for R. Journal of Statistical Software, 27 (3), 1–22.

Hyndman, R., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001

Hyndman RJ, Kourentzes N (2023). *thief: Temporal HIErarchical Forecasting*. R package version 0.3, http://pkg.robjhyndman.com/thief.

Kaltsounis, A., Spiliotis, E., & Assimakopoulos, V. (2023). Conditional Temporal Aggregation for Time Series Forecasting Using Feature-Based

Meta-Learning. *Algorithms*, 16(4), 206. https://doi.org/10.3390/a16040206

Kotu, V., & Deshpande, B. (2019). Time series forecasting. *Data science*, 395-445.

Kourentzes, N., Rostami-Tabar, B., & Barrow, D. K. (2017). Demand forecasting by temporal aggregation: Using optimal or multiple aggregation levels? *Journal Of Business Research*, 78, 1–9. https://doi.org/10.1016/j.jbusres.2017.04.016

Malkari, N. (2023). Exponential Smoothing: A Method for Time Series Forecasting. Medium. https://medium.com/@nikhilmalkari18/exponential-smoothing-a-method-for-time-series-forecasting-7ea35ca2c781

Mancuso, P., Piccialli, V., & Sudoso, A. M. (2021). A machine learning approach for forecasting hierarchical time series. *Expert Systems With Applications*, 182, 115102. https://doi.org/10.1016/j.eswa.2021.115102

Moon, S., Hicks, C., & Simpson, A. (2012). The development of a hierarchical forecasting method for predicting spare parts demand in the South Korean Navy—A case study. *International Journal Of Production Economics*, 140(2), 794–802. https://doi.org/10.1016/j.ijpe.2012.02.012

Oliveira, J. M., & Ramos, P. (2019). Assessing the Performance of Hierarchical Forecasting Methods on the Retail Sector. *Entropy*, 21(4), 436. https://doi.org/10.3390/e21040436

OpenAI. (2023). *ChatGPT* (Mar 14 version) [Large language model]. https://chat.openai.com/chat

Petropoulos, F., & Spiliotis, E. (2021). The Wisdom of the Data: Getting the Most Out of Univariate Time Series Forecasting. *Forecasting*, 3(3), 478–497. https://doi.org/10.3390/forecast3030029

Rostami-Tabar, B., Goltsos, T. E., & Wang, S. (2023). Forecasting for lead-time period by temporal aggregation: Whether to combine and how. *Computers in Industry*, 145, 103803. https://doi.org/10.1016/j.compind.2022.103803

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

Tim. (2020, June 8). *Mean absolute scaled error: definition, example - statistics how to*. Statistics How To. https://www.statisticshowto.com/mean-absolute-scaled-error/

Theodosiou, F., & Kourentzes, N. (2021). Forecasting with Deep Temporal Hierarchies. *Social Science Research Network*. https://doi.org/10.2139/ssrn.3918315

Wickramasuriya, S. L., Athanasopoulos, G., & Hyndman, R. J. (2019). Optimal Forecast Reconciliation for Hierarchical and Grouped Time Series Through Trace Minimization. *Journal of the American Statistical*

Association, *114*(526), 804–819. https://doi.org/10.1080/01621459.2018.1448825

Yang, D., Quan, H., Disfani, V. R., & Rodríguez-Gallegos, C. D. (2017). Reconciling solar forecasts: Temporal hierarchy. Reconciling Solar Forecasts: Temporal Hierarchy, 158, 332–346. https://doi.org/10.1016/j.solener.2017.09.055

Yen, L. (2023). What is Time Series Analysis? Definition, Types, and Examples. Datamation. https://www.datamation.com/big-data/what-is-time-series-analysis/

Zhang, G., & Kline, D. (2007). Quarterly Time-Series Forecasting With Neural Networks. *IEEE Transactions on Neural Networks*, 18(6), 1800–1814. https://doi.org/10.1109/tnn.2007.896859

Appendix

Table 3. MASE for 1-26 step ahead forecasts

SKU	Mean	ARIMA	ETS	Thief(ARIMA)				Thief(ETS)			
				MSE	SHR	OLS	STRUC	MSE	SHR	OLS	STRUC
SA0086720	0.7729	0.7796	0.8158	0.7757	0.7805	0.7715	0.7724	0.7883	0.8096	0.7738	0.7831
SA0062880	0.6940	0.6980	0.6924	0.7148	0.6996	0.7072	0.7104	0.6954	0.7159	0.6970	0.6925
SA0016160	0.8630	0.8166	0.8131	0.8310	0.8527	0.8242	0.7918	0.7932	0.9008	0.8229	0.7758
SA0063980	0.8547	0.7910	0.7813	0.7854	0.7964	0.8297	0.7971	0.7840	0.7890	0.8392	0.7942
SA0024160	0.7896	0.7947	0.7845	0.7830	0.7901	0.7866	0.7831	0.7776	0.8046	0.7830	0.7757
SA0040350	0.7984	0.8182	0.8795	0.8111	0.8152	0.8017	0.8055	0.8440	0.8447	0.8084	0.8291
SA0062872	0.7693	0.7719	0.7721	0.7719	0.7845	0.7690	0.7672	0.7681	0.7873	0.7651	0.7633
SA0039300	0.7825	0.7537	0.7525	0.7604	0.7805	0.7796	0.7647	0.7525	0.7575	0.7775	0.7585
SA0018250	0.7910	0.7158	0.6979	0.7018	0.7028	0.7469	0.7100	0.6819	0.6927	0.7566	0.6958
SA0062852	0.7599	0.7626	0.7587	0.7698	0.7683	0.7663	0.7675	0.7621	0.7838	0.7614	0.7589
SA0031182	0.8384	0.5501	0.4825	0.5063	0.5361	0.6717	0.5175	0.4671	0.4791	0.7166	0.5050
SA0086752	1.0421	1.0094	0.9776	1.0006	1.0053	1.1333	1.0517	0.9776	0.9994	1.0317	0.9851

SA0038902	0.6314	0.5361	0.6875	0.5379	0.5464	0.5379	0.5485	0.4997	0.6280	0.4997	0.5236
SA0062892	0.7624	0.8144	0.8655	0.8175	0.8212	0.7747	0.7983	0.8026	0.8296	0.7628	0.7809
SA0031592	0.7712	0.6624	0.6629	0.6643	0.6787	0.6643	0.6594	0.6689	0.7038	0.6689	0.6557
SF00583BE	0.8571	0.8574	0.8623	0.8580	0.8626	0.8580	0.8551	0.8691	0.8763	0.8691	0.8694
SF00599BE	1.2386	1.2385	1.2413	1.2438	1.2524	1.2438	1.2426	1.2395	1.2565	1.2395	1.2410
SA0072360	0.6724	0.7059	0.6724	0.6603	0.7039	0.6603	0.6596	0.6646	0.7042	0.6646	0.6590
Average	0.8160	0.7820	0.7889	0.7774	0.7876	0.7959	0.7779	0.7687	0.7979	0.7910	0.7693

Table 4. MASE for aggregated forecasts

SKU	Mean	ARIMA	ETS	Thief(ARIMA)				Thief(ETS)			
				MSE	SHR	OLS	STRUC	MSE	SHR	OLS	STRUC
SA0086720	3.8037	3.8898	2.6510	3.8394	3.7914	3.7334	3.7872	3.1407	2.9150	3.8744	3.2498
SA0062880	2.9587	1.5577	2.1149	2.0861	2.1207	2.5825	2.1422	2.0643	2.0398	2.5823	1.9873
SA0016160	2.6817	2.8451	3.5610	2.6630	2.6845	2.0420	1.7883	2.6836	2.8119	2.0213	1.8735
SA0063980	12.5473	11.0613	11.5682	11.2033	11.4418	12.1648	11.3700	10.4936	7.9144	11.6277	10.0507

SA0024160	9.7560	7.0750	4.3903	5.1955	6.5048	4.2962	4.0071	4.4813	4.6793	6.7343	2.2845
SA0040350	0.6812	0.6847	0.8735	0.6984	0.6993	0.6712	0.6818	0.8584	0.8012	0.7343	0.8107
SA0062872	2.3051	2.3216	1.2339	2.9377	3.2105	2.9339	2.4145	1.7108	1.5906	2.2050	1.4915
SA0039300	1.7010	1.2764	1.5019	1.0458	1.0486	1.4974	1.0916	1.2027	1.1526	1.3932	1.2435
SA0018250	3.4291	1.6375	1.4680	1.5829	1.6843	2.3953	1.4475	1.3097	1.5956	2.5696	1.2853
SA0062852	1.1343	1.2696	1.2825	1.1824	1.1891	1.1464	1.1695	0.9937	1.0231	1.1464	0.9629
SA0031182	4.8830	2.5570	0.6601	2.0104	1.8725	3.2069	0.9041	0.6747	0.7099	3.5961	1.1170
SA0086752	12.3336	12.1007	12.0077	12.1648	12.2185	12.5894	12.4071	12.3695	11.7644	13.6741	12.6124
SA0038902	3.2805	2.8442	3.2888	2.7900	2.7879	3.1608	2.8482	1.4708	1.3946	2.8631	1.8104
SA0062892	3.8065	3.6896	4.4399	3.7136	3.6914	3.8032	3.6655	3.4955	3.1505	4.0246	3.6129
SA0031592	6.0069	2.1954	3.2278	2.0811	2.0221	2.8143	1.5032	2.6642	1.9801	4.8403	1.6791
SF00583BE	1.3872	1.2495	1.3695	1.2540	1.2696	1.3869	1.2655	1.1357	1.1589	1.2234	1.2085
SF00599BE	0.6040	0.5747	0.6307	0.5931	0.5900	0.5997	0.5909	0.6217	0.7276	0.5997	0.6216
SA0072360	0.0133	0.1369	0.0128	0.8252	1.1230	0.5127	0.4281	1.0960	1.1828	0.3449	0.5594
Average	4.0729	3.2759	3.1268	3.2148	3.3306	3.4187	2.9729	2.9148	2.6996	3.6697	2.6923

List of Figures

List of Tables