https://doi.org/10.26219/heal.aueb.6516

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

SCHOOL OF INFORMATION SCIENCES & TECHNOLOGY

DEPARTMENT OF STATISTICS

POSTGRADUATE PROGRAM

Applying Time Series Models to Hierarchical Data

By

Petroutsos. Nikolaos

A THESIS
Submitted to the Department of Statistics
of the Athens University of Economics and Business
in partial fulfillment of the requirements for
the degree of Master of Science in Statistics

Athens,Greece
MARCH 2023

https://doi.org/10.26219/heal.aueb.6516

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ & ΤΕΧΝΟΛΟΓΙΑΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ

Εφαρμογή Μοντέλων Χρονοσειρών σε Ιεραρχικά Δεδομένα

Πετρούτσος Νικόλαος

ΔΙΑΤΡΙΒΗ
Που υποβλήθηκε στο Τμήμα Στατιστικής
του Οικονομικού Πανεπιστημίου Αθηνών
ως μέρος των απαιτήσεων για την απόκτηση
Μεταπτυχιακού Διπλώματος Ειδίκευσης στη Στατιστική

ΑΘΗΝΑ
Μάρτιος 2023

https://doi.org/10.26219/heal.aueb.6516

Abstract

The present Master Thesis focuses on the description and forecasting of hierarchical and grouped time series. As "Time-series" are referred to a strictly time-ordered sequence of observations taken at equal time intervals. Time series can often be naturally disaggregated by various attributes of interest. For example, the total number of food products sold by a manufacturer can be disaggregated by product types such as cookies, chocolates, and snacks. Each of these can be disaggregated into finer categories. For example, the cookies can be divided into cookies with or without sugar and so on. These categories are nested within the larger group categories, and so the collection of time series follows an hierarchical aggregation structure. Therefore we refer to these as "hierarchical time series".

Data of this nature can be found across a broad spectrum of scientific fields, including medicine, education, psychology and finance. The present master's thesis will analyze marketing data, specifically products from four distinct categories. Each category contains five manufacturers and each manufacturer produces between one and three products.

The aim of this thesis is to identify the best model with the smallest forecast error, based on specific accuracy measures at the hierarchical level of manufacturers. Subsequently, the thesis will employ forecast reconciliation methods to compare the predictions of the best model with those obtained using reconciliation techniques within the best model.

Keywords

Time-series, Hierarchical data, Arima models, Dynamic Regressions Models, Forecast reconciliation, Measures accuracy

https://doi.org/10.26219/heal.aueb.6516

https://doi.org/10.26219/heal.aueb.6516

Περίληψη

Η παρούσα Μεταπτυχιακή Διατριβή επικεντρώνεται στην περιγραφή και πρόβλεψη ιεραρχικών και ομαδοποιημένων χρονοσειρών. Ως ¨χρονοσειρές' αναφέρονται αυστηρά χρονικά διατεταγμένες ακολουθίες παρατηρήσεων που λαμβάνονται σε ίσα χρονικά διαστήματα. Οι χρονοσειρές μπορούν συχνά να διαμεριστούν φυσικά, σε επιμέρους χρονοσειρές, βάσει διάφορων χαρακτηριστικών ενδιαφέροντος. Για παράδειγμα, ο συνολικός αριθμός των προϊόντων που πωλούνται από έναν κατασκευαστή μπορεί να αναλυθεί ανά τύπο προϊόντος, όπως μπισκότα, σοκολάτες και σνακ. Κάθε ένας από αυτούς τους τύπους μπορεί να αναλυθεί σε επιμέρους κατηγορίες. Για παράδειγμα, τα μπισκότα μπορούν να χωριστούν σε μπισκότα με ή χωρίς ζάχαρη και ούτω καθεξής. Αυτές οι κατηγορίες είναι ένθετες μέσα στις μεγαλύτερες ομάδες κατηγοριών, και έτσι η συλλογή των χρονοσειρών ακολουθεί μια ιεραρχική δομή συνάθροισης. Επομένως, αυτές αναφέρονται ως "ιεραρχικές χρονοσειρές".

Η διαθεσιμότητα τέτοιων δεδομένων είναι ευρεία σε πολλούς τομείς της επιστήμης, όπως η ιατρική, η εκπαίδευση, η ψυχολογία και οι οικονομικές επιστήμες. Στην παρούσα μεταπτυχιακή διατριβή, θα αναλυθούν δεδομένα που προέρχονται από τον τομέα του μάρκετινγκ. Συγκεκριμένα, θα εξεταστούν προϊόντα από τέσσερις διαφορετικές κατηγορίες. Κάθε κατηγορία περιλαμβάνει πέντε κατασκευαστές, και κάθε κατασκευαστής παράγει από ένα έως τρία προϊόντα

Η παρούσα διπλωματική εργασία έχει ως στόχο να αναπτυχθεί ένα μοντέλο με το ελάχιστο δυνατό σφάλμα πρόβλεψης, χρησιμοποιώντας συγκεκριμένα μέτρα ακρίβειας σε ιεραρχικό επίπεδο κατασκευαστών. Στη συνέχεια, θα χρησιμοποιηθούν μέθοδοι συμβιβασμού (reconciliation) για να συγκριθούν οι προβλέψεις του καλύτερου μοντέλου με τις προβλέψεις των μεθόδων συμβιβασμού (reconciliation).

Λέξεις Κλειδιά

Χρονοσειρές, Ιεραρχικά δεδομένα, Μοντέλα Arima, Μοντέλα δυναμικής παλινδρομήσης, Συμβιβαστική πρόβλέψη, Μέτρα αποτίμησης της πρόβλεψης.

https://doi.org/10.26219/heal.aueb.6516

to my mother

https://doi.org/10.26219/heal.aueb.6516

https://doi.org/10.26219/heal.aueb.6516

Acknowledgements

This thesis would not have been possible without the valuable guidance and support of my supervisors, Ioannis Ntzoufras and Dimitris Karlis. Their expertise, dedication, and encouragement helped me to navigate the challenges of this research and to produce work that I am proud of. I am grateful for the time they spent reading my drafts, providing feedback, and challenging me to think more deeply about my topic.

I also want to express my gratitude to Stefanos Kechagias for his thoughtful suggestions and feedback throughout the process, especially when he was making use of his personal time. His knowledge of data analysis was really helpful in developing this thesis.

Lastly, I would like to express my heartfelt gratitude to my family and friends for their unwavering support throughout my academic journey. Their love and encouragement have been a constant source of strength for me, and I could not have done this without them.

https://doi.org/10.26219/heal.aueb.6516

 https://doi.org/10.26219/heal.aueb.6516

Contents

https://doi.org/10.26219/heal.aueb.6516

https://doi.org/10.26219/heal.aueb.6516

List of Figures

https://doi.org/10.26219/heal.aueb.6516

 https://doi.org/10.26219/heal.aueb.6516

List of Tables

https://doi.org/10.26219/heal.aueb.6516

https://doi.org/10.26219/heal.aueb.6516

Chapter 1

Introduction

1.1 Forecast

Every aspect of daily living requires predictions. What kind of weather can we expect the next day? Will there be significant traffic on the way to work? These are common queries that we have all at some point considered. The answers to these questions can be found in weather forecasts and the analysis of historical traffic data. We can also extract answers to both questions from our personal experience.

This issue is not unique to the business world. The management of the company is carefully planning production to satisfy demand, fully monetize it, and prevent accumulating an excessive amount of inventory, which may be associated with high expenses. Accurately forecasting the demand for a company's products is key to its success. Forecasting involves estimating future values using all available information, including historical data and projections of upcoming events that can have an impact on the estimated reality.

Recognizing when something can be properly predicted and when forecasts are no better than throwing a coin is a crucial first step. Good forecasts capture the genuine patterns and relations existing in the historical data but do not replicate past events. This thesis will explore the multiple methods to distinguish the difference between a random fluctuation in the past data that should be ignored, and a genuine pattern that should be modeled and extrapolated.

It is a common belief that projections are difficult in an environment that is changing, however, we should point out that all ecosystems change constantly. A good prediction should be able to predict when and how these changes will occur. The most typical forecasting presumption is that the current environmental change

https://doi.org/10.26219/heal.aueb.6516

will persist over time. In other words, a highly volatile environment will remain highly volatile; a company with erratic sales will still have erratic sales; etc.

A few applications of daily life in such forecasts could be useful are; accounting and finance (forecast of general trends & their tendency), social research (study of demographic data, forecast of viruses spread among communities, etc.), the entrepreneurial world (prediction of demand of particular products/materials) and so on.

1.2 Forecast Value

We can think of the object we are trying to predict as a random variable because it is unknown (otherwise, we wouldn't be forecasting it). For instance, the total sales for the upcoming month could take on a variety of conceivable values, and we won't know what they will be until we tally up the actual sales at the end of the month. It is therefore a random quantity until we know the sales for the following month.

We have a reasonable notion of what the likely sales values could be because next month is so close. On the other hand, if we are predicting sales for the same month the following year, the range of probable values is substantially wider. In the majority of forecasting scenarios, the variance related to the object we are expecting will decrease as the event nears.

We typically refer to the "forecast" as the average value of the forecast distribution and indicate this by placing a "hat" over y. As a result, we represent the forecast of y_t as $\hat{y}_t$, which stands for the average of every possible value that y_t could take given everything we know.

1.3 Qualitative & Quantitative Forecasting

In case of no available data or if the data are not representative enough of the population for trustworthy forecasts, qualitative forecasting techniques should be employed. These methods are not solely based on speculation; there are well-developed structured approaches to obtaining good forecasts without using historical data. For example, a business that has just begun operating in the market and lacks extensive historical data must make use of qualitative projections.

https://doi.org/10.26219/heal.aueb.6516

when the following requirements are satisfied —namely, numerical data for the past is available and it is plausible to infer that some parts will show emerge—quantitative forecasting can be used. Regression models, exponential smoothing models, Box-Jenkins ARIMA models, mixed-effects models, and other mathematical techniques are only a few of the tools available to produce these predictions. Most quantitative forecasting employs cross-sectional data or time-series data.

1.4 Forecast Horizon

Another way of assessing prediction problems is to take into consideration the projection's timeline depth, in other words, how far in the future are we trying to predict? There can be short-term, middle-term, and long-term predictions. However, the very same definition of these aforementioned categories varies depending on the subject of research. For example, the predictions of energy demand, in the context of planning the construction of electricity generating stations, that span from five to ten years are considered short-term, and 50 years long-term; while for the prediction of consumer demand, six months are considered to fall in the short-term spectrum, and two years on the long-term.

Short-term predictions are more accurate than middle-term or long-term predictions. Many things could change in three months, one year, three years, or more. The factors potentially affecting these predictions are changing every day. For example, if we wanted to predict the total sales of a product for the upcoming month, the projected prices showcase probably high accuracy, while repeating the prediction for the same month but of the upcoming year, then the projected prices may decline more from the real ones. In most prediction problems, the divergence of predicted, expected value from real value is decreasing as we are getting closer to the event we seek to predict. In other words, the further in time our predictions reach, the more uncertainty increases. This dissertation studies short-term predictions, within a horizon of four weeks (h=28).

1.5 Different Case Studies

On the one hand, studies that make use of data collected at a single point in time for each subject, are known as cross-sectional studies. More specifically, cross-sectional studies' data are comprised of a variety of variables, which depending on the case, can be related to people, companies, and cities, and are collected at some particular point in time or describing some short period..

On the other hand, studies that analyze data containing values of a single variable, spread in time (normally in regular periods), and aim at the prediction of future values based on past ones are called time-series studies. Time series studies' data are comprised of values recorded between regular periods. Furthermore, longitudinal data are derived from a combination of cross-sectional and time series data.

Every subject in cross-sectional studies contributes exactly one value for each response variable, in contrast to longitudinal, where the different subjects used may contribute a different number of variable measurements. Usually, in time series, just one subject contributes all the measurements.

Response variables in models with cross-sectional data are independent of each other. On the contrary, in time series models the response variables are correlated with each other. In longitudinal models the situation is more complex; while the repeated measures (response variables) of the same subject are correlated with each other, the repeated measures of two distinct subjects are uncorrelated.

The analytical methods used for the data of a study differ and depend on the study's design. This becomes clear when applying three different study plans on 100 blood pressure measurements. If there are 100 blood pressure measurements available for 100 people, respectively, then linear regression models are used. If there are 100 measurements just for one person, then time series is used as an analytical method. Finally, if there are five repeated blood pressure measurements in different, regular, periods in time for each subject, then longitudinal models are used.

Most of the statistical techniques have been developed, in such a way as to assume that the data used are independent, namely that the measurements of the sample derive from independent experiments. This is considered to be the case with cross-section data where the order of the sample is indifferent and does not influence the results of statistical analysis.

On the contrary, when the sample is time series, then the independence assumption has no practical application and the measurements are assumed to be dependent to some extent, constituting regular statistical techniques untrustworthy. This is the case when the order of the measurements is changing as of the point of reference of time, but also when there are missing values for one or more time periods.

https://doi.org/10.26219/heal.aueb.6516

1.6 Thesis Objective

The purpose of this thesis is to compare the forecasts generated by models applied to individual time series with methods applied to hierarchical time series structures that use these models. The forecasts generated directly from the models are referred to as "basic forecasts", while the forecasts resulting from applying the models to hierarchical time series are called "coherent forecasts". To examine which of the forecasts, whether basic or coherent, exhibit the smallest error, specific accuracy measures are used.

1.7 Thesis Structure

The structure of the thesis is the following, the second chapter of this thesis will provide an overview of the fundamental concepts related to time series analysis, including the definition of time series, the different categories of time series, and their various characteristics.

In the third chapter, we will discuss ARIMA models and their variations, as well as hierarchical and grouped time series. The fourth chapter will outline the initial data used in our case study, and the results obtained from various time series models, including ARIMA, ETS, and Dynamic Regression. We will then compare and contrast the results among these models, as well as with other benchmark models with measures of forecast accuracy.

For each level, the optimal model with the least amount of error will be determined, either by utilizing reconciliation techniques or by utilizing one of the previously discussed models from the outset. Ultimately, by utilizing the cross-validation technique for each time series model at each level, we can determine the optimal model for each individual case.

https://doi.org/10.26219/heal.aueb.6516

https://doi.org/10.26219/heal.aueb.6516

Chapter 2

Time series

The time series is primarily defined in this chapter. We also go over numerous qualitative traits as well as different time series classifications. Finally, we will examine several metrics for measuring the precision of the point forecasts we will employ in this study and time-series statistical functions.

While there are variations between the different time series prediction techniques, they all have the following traits. We will highlight some of the most prevalent characteristics we have identified below.

1. Most forecasting techniques make the assumption that causal relationships that have existed in the past will remain to exist in the future.

2. Because most forecasts are based on previous data, they are rarely accurate. So, a forecasting process' objective is to reduce mistakes as much as possible in order to get closer to the object's actual value.

3. Without a doubt, the accuracy of forecasts tends to decrease as the forecast range becomes longer. It has been found that short-term projections are more accurate than long-term ones.

2.1 Time-series

A time series is a collection of observations made sequentially through time. Examples occur in a variety of fields, ranging from economics to engineering, and methods of analyzing time series constitute an important area of statistics (Chapman & Hall (2019). Examples of time series include the monthly sales figures of a retail store and the daily number of website visitors for an online platform.

The data points of a time series are typically taken at equally spaced intervals and can be represented by a set of ordered pairs $(t, Y(t))$, where "t" represents the

https://doi.org/10.26219/heal.aueb.6516

time index or timestamp, and "$Y(t)$" represents the corresponding value or measurement at the time "t". Time series data often exhibit temporal dependencies and can display trends, seasonality, and other patterns that can be analyzed and modeled.

Also, each time series can be considered to be the result of some stochastic process and the values of the time series are the result of a realization of this stochastic process. It is recalled that the stochastic process is a collection of random variables, representing the evolution of a system with random values over time. An equivalent is a set of random variables $\{Y_t, t \in T\}$ in the same probability space. $(\Omega, \mathcal{F}, P)$. The expanded definition of a probability space is the following:

- the sample space Ω — an arbitrary non-empty set,
- the σ-algebra $\mathcal{F} \subseteq 2^{\Omega}$ (also called σ-field) — a set of subsets of Ω, called events, such that:
  1. $\mathcal{F}$ contains the sample space: $\Omega \in \mathcal{F}$,
  2. $\mathcal{F}$ is closed under complements: if $A \in \mathcal{F}$, then also $(\Omega \setminus A) \in \mathcal{F}$,
  3. $\mathcal{F}$ is closed under countable unions: if $A_i \in \mathcal{F}$ for $i = 1, 2, \ldots$ then also $(\bigcup_{i=1}^{\infty} A_i) \in \mathcal{F}$
     - The corollary from the previous two properties and De Morgan's law is that $\mathcal{F}$ is also closed under countable intersections: if $A_i \in \mathcal{F}$ for $i = 1, 2, \ldots$then also $(\bigcap_{i=1}^{\infty} A_i) \in \mathcal{F}$
- the probability measure $P : \mathcal{F} \to [0, 1]$ — a function on $\mathcal{F}$ such that:
  1. P is countably additive (also called σ-additive): if $\{A_i\}_{i=1}^{\infty} \subseteq \mathcal{F}$ is a countable collection of pairwise disjoint sets, then $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$,
  2. the measure of entire sample space is equal to one: $P(\Omega) = 1$.

2.2 Terminology symbols

Then the main mathematics symbols in the time series used in the following chapters of the work are mentioned. The introduction of symbols at the beginning of the work helps to convey as many mathematical concepts as possible with as few words as possible and also to have a better understanding of mathematical concepts. It will be noted if any new mathematical symbols are employed as the process progresses. So we will use

https://doi.org/10.26219/heal.aueb.6516

- Y: The observed subject, the values of the observed object change randomly.
- y_t: The value of the random variable Y
- T: The number of observations.
- $\{y_1, y_2, \cdots, y_T\}$: The set of values of the variable Y, for a certain period.
- $\hat{y}_{T+h|T}$: The forecast of y_{T+h} taking account of $y_1, \ldots, y_T$.
- h: The forecast horizon.

2.3 Categories of time-series

Time series can be distinguished according to the range of values taken by the index t ($t \in T$) in continuous-time $T = (-\infty, +\infty)$ and discrete-time $T = (0, 1, 2, \ldots)$ time series.

- Continuous time series are those whose values can theoretically be measured at any point in time. A corresponding example of continuous time series is the recording of temperature which can be done at any time of the day.
- Discrete-time series are those whose observations are taken at predetermined, discrete equal time intervals such as year, quarter, month, and day. An example of a discrete-time series is the stock price log by day.

2.4 Time series patterns

The **trend** in time series occurs when there is a long-term increase or decrease in data values. It may, in some cases, require data over a long period of time in order to be able to identify and describe it. The trend of the values of a variable can be upward, downward or constant over a certain period of time. These long-term increases or decreases may be described by linear, logarithmic, exponential, and polynomial functions of time. However, the trend in a time series may not be possible is a known function of time, hence its name stochastic trend.

Seasonality characterizes time series when the value variation is influenced by seasonal factors with a fixed periodicity and duration that concerns intervals of less than or equal to one year. This variation is immediately understandable and predictable because the data of these time series is usually repeated in the same way. An example of the appearance of seasonality in the time series is the increase in

https://doi.org/10.26219/heal.aueb.6516

sales of ice cream during the summer.

Cyclicality occurs when the data exhibit rises and falls that are not of a fixed frequency. A cycle represents a change that occurs due to exogenous factors over long periods. The periods are longer than a year and usually five, ten, or even twenty years.

Although they are frequently mistaken, seasonal and cyclical behavior are extremely different. The fluctuations are cyclical if there is no fixed frequency; the pattern is seasonal if the frequency is fixed and related to a certain aspect of the calendar. The average length of a cycle is often longer than the average length of a seasonal pattern, and the magnitudes of cycles are typically more variable than those of seasonal patterns.

A stationary time series is one whose statistical characteristics are independent of the observation time. Time series with trends or seasonality are therefore not stationary because the trend and seasonality will change the time series' value at different points in time. A white noise series, on the other hand, is stationary; regardless of the time you examine it, it should appear about the same. A time series with cyclic behavior (but no trend or seasonality) is stationary, for example. This is due to the cycles' variable length; thus, we cannot predict where the peaks and troughs of the cycles will be before we examine the series.

2.5 Stationarity

A stationary time series is one whose statistical characteristics are independent of the observation time. Time series with trends or seasonality are therefore not stationary because the trend and seasonality will change the time series' value at different points in time. A white noise series, on the other hand, is stationary; regardless of when you examine it, it should appear substantially the same.

As a stochastic process, a time series can be characterized by its probability distribution function. However, since its probability distribution function is practically impossible to determine, we define the process using its first and second moments. For each time point, we specifically employ the means $\mu_t = E(Y_t)$, variances $\sigma_t^2 = Var(Y_t)$, and autocovariances $\gamma_{t,s} = Cov(Y_t, Y_s)$ for $t \neq s$.

The problem that arises is that the moments are not constant but change at

https://doi.org/10.26219/heal.aueb.6516

each time moment. Due to the stochastic process having only one realization or one observation for each time period, we must estimate the T means, T variances, and $\binom{T}{2} = \frac{T*(T-1)}{2}$ autocovariances. It is necessary to find a way to lessen the amount of unknown parameters because of this. The assumptions of stationarity contribute to this.

Weak stationary, also known as second-order stationary, and strict stationary are the two main categories of stationary occurrences. The following conditions must be met for a stochastic process to be considered weakly stationary. In particular, we assume that

$$E(Y_t) = \mu, \;\; \forall t \tag{2.1}$$

$$Var(Y_t) = \sigma_Y^2, \; \forall t. \tag{2.2}$$

And also the auto-covariance for two different moments depend only on the distance between them.

$$\gamma_k = Cov(Y_t, Y_{t+k}) = Cov(Y_t, Y_{t-k}) = \gamma_{-k} \tag{2.3}$$

Because Y_t, Y_s are observations of the same variable that are spaced apart time lag s, the covariance $Cov(Y_t, Y_s)$ also referred to as auto-covariance.

A stochastic process is called strictly stationary when its joint probability distribution is not affected by a change in time. This means that the joint distribution for time instants $t_1, t_2, \ldots t_T$ is the same as the joint distribution for time instants $t_{1+k}, t_{2+k}, \ldots t_{T+k}$, where k is a positive integer (2.4).

$$f(y_1, y_2, \ldots, y_T) = f(y_{1+k}, y_{2+k}, \ldots, y_{T+k}), \;\; k \in \mathbb{Z} \tag{2.4}$$

Since the joint probability function does not change with time, therefore also the probability function margins at time point t, $f(y_t)$, will be the same as the probability function margins $f(y_{t+k})$ in time point $t + k$ This is:

$$f(y_t) = f(y_{t+k}), \;\; \forall k. \tag{2.5}$$

A strictly stationary stochastic process is necessarily also weakly stationary. However, the reverse is not always true. However, the reverse is not always true. The normal distribution is an exception to this rule where strict stationarity is equivalent to weak stationarity. Verifying strict stationarity in practice is very difficult, which is why weak stationarity is often used.

https://doi.org/10.26219/heal.aueb.6516

2.6 Autocorrelation

Statistics uses the terms correlation or dependency to describe any statistical relationship between two random variables, regardless of whether it is causative or not. In statistics, "correlation" often refers to the strength of a pair of variables' linear correlations, even if it can apply to any form of association in the broadest meaning. Relationships between parent and child height and the price of a good and the amount of it that buyers are willing to purchase are two examples of dependent phenomena.

Because they can demonstrate an anticipated relationship that can be used in real-world circumstances, correlations are helpful. For instance, the correlation between the weather and the demand for electricity may cause an electrical company to offer less power on a sunny day. In this case, there is a causal relationship since people use more energy to heat or cool their homes during harsh weather. But typically, a correlation does not show causality (i.e., correlation does not prove the presence of a causal relationship).

Correlation is essentially a measurement of the relationship between two or more variables. The correlation strength is measured by several correlation coefficients, which are frequently abbreviated as ρ or r. The Pearson correlation coefficient, which solely considers a linear relationship between two variables, is the most popular of these (which may be present even when one variable is a nonlinear function of the other). It has been developed for other correlation coefficients, such as Spearman's rank correlation, to be more sensitive to nonlinear interactions than Pearson's.

It's interesting to look into how observation in a time series relates to its predecessors. The Pearson correlation coefficient is defined as

$$\rho = \frac{Cov(Y,X)}{\sqrt{Var(Y)} \cdot \sqrt{Var(X)}}$$

The autocovariance (γ_k) and autocorrelation (ρ_k) of the random variables Y_k and Y_{t-k} are defined in the same way as the covariance and correlation coefficient between two random variables Y and X, namely.

$$\gamma_k = Cov(Y_t, Y_{t-k}) = E\,(Y_t - \mu_k)(Y_{t-k} - \mu_{t-k}) \tag{2.6}$$

$$\rho_k = Corr(Y_t, Y_{t-k}) = \frac{Cov(Y_t, Y_{t-k})}{\sqrt{Var(Y_t)} \cdot \sqrt{Var(Y_{t-k})}} \tag{2.7}$$

https://doi.org/10.26219/heal.aueb.6516

If the time series is weakly stationary, for $k = 0$ and relation (2.6) then:

$$\gamma_0 = Cov(Y_t, Y_t) = E\,(Y_t - \mu)(Y_t - \mu) = E\left[(Y_t - \mu)^2\right] = Var(Y_t) = \sigma^2$$

From relations (2.6),(2.7) and if the time series is weakly stationary it is concluded that the autocorrelation coefficient ρ_k like the autocovariance γ_k do not depend on time t but only on lag k.

$$\begin{aligned}
\rho_k &= \frac{Cov(Y_k, Y_{t-k})}{\sqrt{Var(Y_t)} \cdot \sqrt{Var(Y_{t-k})}} \Rightarrow \\
\rho_k &= \frac{Cov(Y_t, Y_{t-k})}{\sqrt{Var(Y_t)} \cdot \sqrt{Var(Y_t)}} \Rightarrow \\
\rho_k &= \frac{Cov(Y_k, Y_{t-k})}{Var(Y_t)} \\
\rho_k &= \frac{\gamma_k}{\gamma_0}. \qquad (2.8)
\end{aligned}$$

From (2.8), it is clear that:

$$\rho_0 = \frac{\gamma_0}{\gamma_0} = 1$$

From the relations (2.3) and (2.8), it is obvious that:

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \frac{\gamma_{-k}}{\gamma_0} = \rho_{-k}$$

The autocovariance and autocorrelation will be estimated because we do not know the actual values. The following mathematical formulas compute the sample autocovariance at lag k and the sample autocorrelation at kth lag with $\bar{Y}$ is the sample mean in the event that we have a time series of T observations.

$$c_k = \sum_{t=k+1}^{T} (y_t - \bar{Y})(y_{t-k} - \bar{Y}) \qquad (2.9)$$

$$r_k = \frac{c_k}{c_0} = \frac{\sum\limits_{t=k+1}^{T} (y_t - \bar{Y})(y_{t-k} - \bar{Y})}{\sum\limits_{t=1}^{T} (y_t - \bar{Y})^2}, \qquad (2.10)$$

The autocorrelation function (ACF) is defined by the autocorrelations of the time series as a function of time lag k. When data exhibit a trend, the autocorrelations for small lags typically tend to be considerable due to the fact that observations that are close together in time and value are also close together. As a result, the ACF of these time series tends to have values that decrease in absolute magnitude

https://doi.org/10.26219/heal.aueb.6516

as the lags increase..

When data are seasonal, the seasonal lags' autocorrelations will be larger than those of other lags (at multiples of the seasonal period). When data are trended and seasonal, these effects are combined.

2.7 Independent & Identically Distributed

If the random variables Y_t of a stochastic process are independent and follow the same distribution with the same parameters (identically) then it is called a time series of independent and identically random variables (iid). An iid time series is entirely random, and neither a linear nor a non-linear dependency between its random variables can be seen. Independence in a time series states that there is no information to derive from studying it. Its implementation consists of random values.

The mathematical formula that defines the independence between the random variables Y_t of $\{y_t\}_1^T$ is:

$$P(Y_1 \leq y_1, Y_2 \leq y_2, \ldots, Y_T \leq y_T) = P(Y_1 \leq y_1) \cdots P(Y_2 <= y_2) \ldots P(Y_{Tt})$$

with $P(Y_1 <= y_1, Y_2 <= y_2, \ldots, Y_T <= y_T)$ the joint cumulative probability function and $P(Y_t <= y_t)$ the corresponding marginal functions.

2.8 White Noise Process

Time series that show no autocorrelation are called white noise. A time series of white noise symbols are presented as $\epsilon_t \sim WN(0, \sigma^2)$. White noise stochastic process is characterized by second-order stationarity, if the random variables follow the normal distribution ($\epsilon_t \sim N(0, \sigma^2)$) then it is also a strictly stationary process, and the random variables (ϵ_t) are independent and identically distributed $\epsilon_t \sim iid(0, \sigma^2)$. For a white noise process, we have that

$$E(\epsilon_t) = 0, \ \forall t$$
$$Var(\epsilon_t) = \sigma_Y^2, \ \forall t$$
$$Cov(\epsilon_t, \epsilon_s) = 0, \ \forall t \neq s$$

https://doi.org/10.26219/heal.aueb.6516

Figure (2.1) depicts a white noise time series with 50 observations. It is observed that all the observations are close to zero and the greater majority of the values of the observations belong to the interval (-3,3), an indication that the variation has remained constant over time.

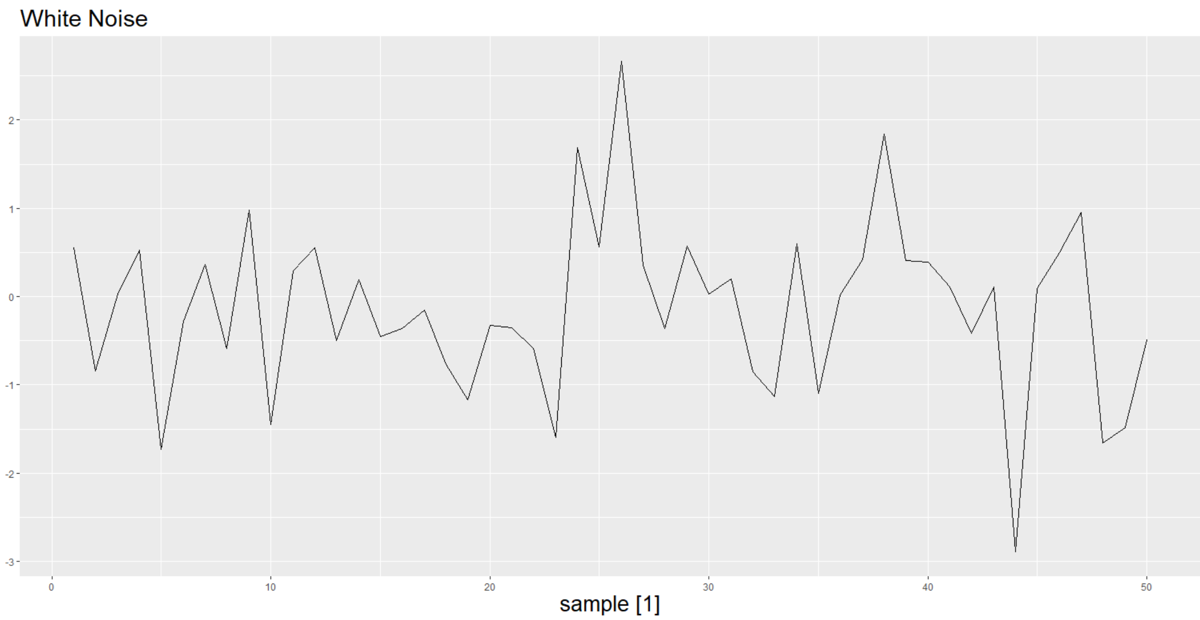

Figure 2.1: A white noise time series.

For a white noise series, we anticipate that 95% of the spikes in the ACF figure (2.2) will be contained inside $1.96/\sqrt{T}$, where T is the length of the time series, as stated in section (2.2). It is common to plot these bounds on a graph of the ACF (the blue dashed lines above). The series is probably not white noise if one or more large spikes are outside these bounds or if substantially more than 5% of spikes are outside these bounds. These boundaries are met by all of the autocorrelation coefficients, indicating that the data are white noise .

https://doi.org/10.26219/heal.aueb.6516

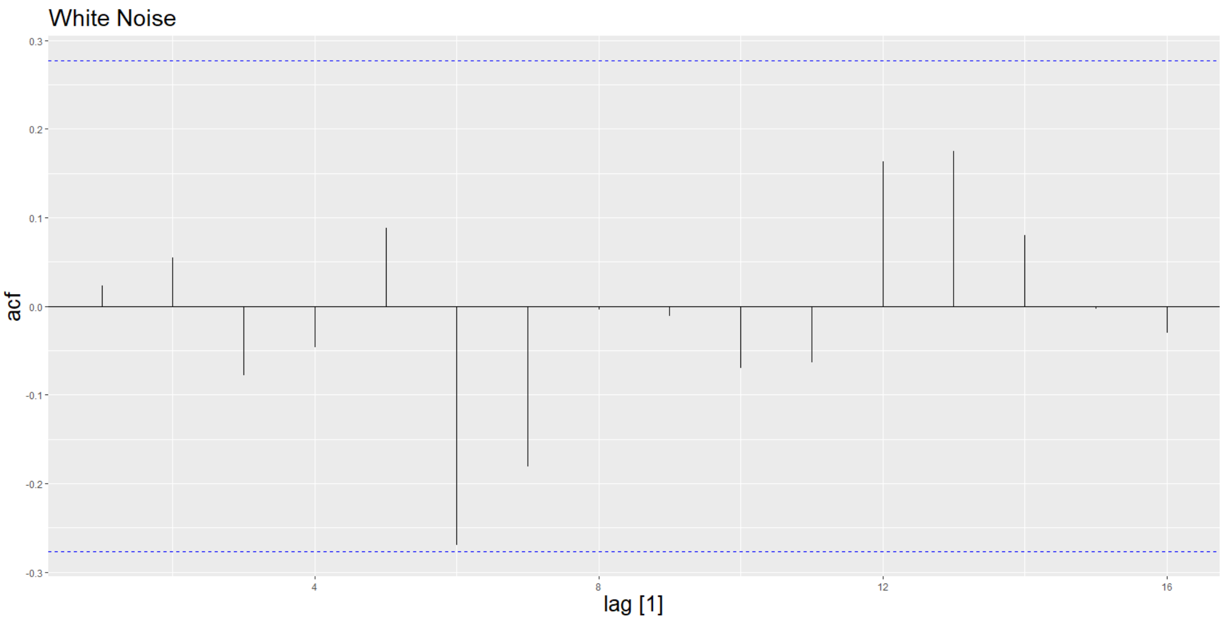

Figure 2.2: Autocorrelation function for the white noise series

2.9 Random Walk

A random walk is a non-stationary time series in which each element is produced by adding a random value to the one before it. A time series is a random walk if:

$$Y_t = Y_{t-1} + \epsilon_t, \quad \epsilon_t \sim WN(0, \sigma^2) \tag{2.11}$$

The term "random walk" describes how the time series is created by randomness motion in a straight line (in $\mathbb{R}$), which, at each instant of time, randomly moves either forward or backward ϵ_t from the position Y_{t-1} to the Y_t.

The definition of a random walk can be expressed as, starting from some value μ_0 (i.e., for $t = 0, Y_0 = \mu_0$) and iteratively substituting into the definition of a random walk (2.11).

$$\begin{aligned}
Y_0 &= \mu, \\
Y_1 &= Y_0 + \epsilon_1 = \mu + \epsilon_1, \\
Y_2 &= Y_1 + \epsilon_2 = Y_0 + \epsilon_1 + \epsilon_2 = \mu + \epsilon_1 + \epsilon_2, \\
&\vdots \\
Y_t &= Y_{t-1} + \epsilon_t = \mu + \sum_{i=1}^{t} \epsilon_i. \qquad (2.12)
\end{aligned}$$

https://doi.org/10.26219/heal.aueb.6516

applying linear properties for the mean value and variance of Y_t in relation (2.12) gives the mean value of Y_t and its variance.

$$Y_t = E(\mu + \sum_{i=1}^{t} \epsilon_i) = \mu, \tag{2.13}$$

$$Y_t = Var(\mu + \sum_{i=1}^{t} \epsilon_i) = t\sigma^2. \tag{2.14}$$

The random walk's time series is non-stationary because the variance of the walk is proportional to time t (2.14).

2.10 Statistical Inference for Sample Autocorrelations

As mentioned in section (2.6), the true values of the autocorrelation ρ are not known, so the corresponding estimates from the sample r are used.

If the random variables of the time series $Y_t \sim WN(0, \sigma_Y^2)$ then for a large sample T the Barlett (1946) had shown that $r_k \sim N(0, \frac{1}{T})$.

If $r_k \sim N(0, \frac{1}{T})$, then $\frac{r_k}{\sqrt{\frac{1}{T}}} \sim N(0, 1)$. For the $k-th$ autocorrelation, the following definition describes the statistical significance test:

$$H_0 : \rho_k = 0 \quad \text{vs} \quad H_1 : \rho_k \neq 0 \tag{2.15}$$

When $-\frac{1}{\sqrt{T}} \cdot z_{a/2} < r_k < \frac{1}{\sqrt{T}} \cdot z_{a/2}$ is true, where a is the significance level and $z_{a/2}$ is the critical value from the standard normal distribution ($P(Z >= z_a) = a$), we do not reject the null hypothesis.

The upper and lower values of the mathematical inequality are included in the autocorrelation plots so that statistically significant autocorrelations ($\rho_k \neq 0$) can be quickly identified.

In the case that we are performing k-independent statistical tests for the autocorrelation at significance level $a = 5\%$ then the probability of not rejecting the null hypothesis given that the null hypothesis is true in a statistical test is $p = 1 - a = 95\%$. The probability at least one error will occur in all tests is:

$$1 - \binom{k}{k} \cdot p^k \cdot (1-p)^{k-k} = 1 - p^k = 1 - (1-a)^k. \tag{2.16}$$

https://doi.org/10.26219/heal.aueb.6516

The Portmanteau test is designed to determine simultaneous statistical significance for k autocorrelation:

$$H_0 : \rho_1 = \rho_2 = \cdots = \rho_k = 0 \quad \text{vs} \quad H_1 : \text{at least one } \rho_i \neq 0, i = 1, 2, \ldots, k \tag{2.17}$$

If $r_i \sim N(0, \frac{1}{T})$, then $\frac{r_i}{\sqrt{\frac{1}{T}}} \sim N(0,1) \Rightarrow T \cdot r_i^2 \sim X_1^2 \Rightarrow T \cdot \sum_{i=1}^{k} r_i^2 \sim X_k^2$, where $r_1, r_2, \ldots r_k$ are iid.

For the statistical control of the relationship, (2.17) was proposed by Box & Pierce (1970), as a control function Q_k and is defined as:

$$Q_k = T \cdot \sum_{i=1}^{k} r_i^2 \tag{2.18}$$

where T is the sample size and k the number of autocoroleations. he null hypothesis will be rejected, if $Q_k > x_{k,a}^2$. where $x_{k,a}^2$ is the critical value from the chi-square distribution with k degrees of freedom $\left(P(X >= x_{k,a}^2) = a\right)$

Ljung and Box (1978) found that the Q_k statistic function is not satisfactory in small samples and suggested the following modification:

$$Q'_k = T \cdot (T+2) \sum_{i=1}^{k} \frac{r_i^2}{T-i} \tag{2.19}$$

Assuming the null hypothesis is true, then $Q'_k \sim X_k^2$. Furthermore, it is possible to check for residual autocorrelation using the Portmanteau test and Ljung-Box test. The X^2 statistics will have $k - m$ degrees of freedom, where m is the total number of model parameters.

2.11 Measures of Forecast Accuracy

It is usual practice to divide the data set into train data $\{y_1, y_2, \ldots, y_T\}$ and test data $\{y_{T+1}, y_{T+2}, \ldots\}$ and in order to select the best forecasting model. The test data is used to assess the method's accuracy, while the training data is used to estimate the method's parameters. The selection of the method with the best prediction is obtained using some evaluation measures (statistical functions), which are a function of the deviation of the forecast value from the corresponding actual value. These evaluation measures are based on so-called forecast errors.

The forecast error is defined as the deviation of the predicted value from the actual value (2.20). In the specific case of interpretation given to the word "error" does not imply error but rather expresses an aspect of the observation that cannot be predicted. We define the error at point $T + h$ as

$$e_{T+h} = y_{T+h} - \hat{y}_{T+h|T} \tag{2.20}$$

The "residuals" in the model is what is left over after fitting a model. The residuals are equal to the difference between the observations and the corresponding fitted values:

$$e_t = y_t - \hat{y}_t,\ t \in \{1, 2, \ldots, T\} \tag{2.21}$$

The forecast errors and the residuals differ from each other in two key ways that should be acknowledged. The errors are first computed on the test dataset, while the residuals are first computed on the training dataset. Second, while forecast mistakes may also include multi-step forecasts, residuals are based on one-step forecasts.

As was said in the opening sentence of this section, some statistical functions that are based on forecast errors are used to evaluate the predictions of various approaches in a dataset. The categories of accuracy measurements used in point forecasts are as follows:

- Scale-dependent errors.
- Percentage errors.
- Scaled errors.

.

Certain forecast measures that belong to the "Scale-dependent errors "category are Mean Error (ME), Mean Absolute Error (MAE), Root Mean Squared Error ($RMSE$), and Mean Squared Error(MSE). The first three forecast measures' units of measurement are the same as the data's units of measurement, whereas MSE's units of measurement are the data's squared units of measurement. If we define that the number of predictions is H then the following formulas give us the corresponding prediction measures.

https://doi.org/10.26219/heal.aueb.6516

$$\text{Forecast error: } e_t = y_t - \hat{y}_t,\ t \in \{T+1, T+2, \dots T+H\}$$

$$ME = \frac{1}{H} \sum_{t=T+1}^{T+H} e_t, \tag{2.22}$$

$$MAE = \frac{1}{H} \sum_{t=T+1}^{T+H} |e_t|, \tag{2.23}$$

$$MSE = \frac{1}{H} \sum_{t=T+1}^{T+H} e_t^2, \tag{2.24}$$

$$RMSE = \sqrt{\frac{1}{H} \sum_{t=T+1}^{T+H} e_t^2}\,. \tag{2.25}$$

Each of the above 4 measures can be used to compare methods of the same time series or time series with the same unit of measurement. Forecast measures (2.23) and (2.24) give particular weight to high deviations between forecast and actual values. The key characteristic of the measure ME (2.22) is to show the presence of bias. If the ME is positive then the prediction model gives pessimistic predictions since the actual value is systematically lower than the predicted value. In the event that the index is negative, then the predicted values are systematically higher than the actual values and there are optimistic forecasts.

The category of percentage errors includes measures such as Mean Absolute Percentage Error ($MAPE$), and Symmetric MAPE ($sMAPE$). The percentage errors are independent of measurement units so predictions of different methods can be compared on more than one data set. Measures based on percentage errors have the drawback of having extreme values if any y_t is close to zero and being infinite or undefined if $y_t = 0$ for any t in the period of interest. The percentage error,$MAPE$, and $sMAPE$ are defined by the following formulas:

$$p_t = 100 \cdot \frac{e_t}{y_t},$$

$$MAPE = \frac{1}{H} \sum_{t=T+1}^{T+H} |p_t|,$$

$$sMAPE = \frac{1}{H} \sum_{t=T+1}^{T+H} 200 \cdot \frac{|e_t|}{y_t + \hat{y}_t}.$$

Hyndman & Koehler (2006) suggested the use of scaled errors as a substitute for percentage errors for evaluating forecast accuracy between series with various units.

https://doi.org/10.26219/heal.aueb.6516

They proposed scaling the errors based on the training MAE from a simple forecast method.

Forecast measures like Mean Absolute Scaled Error ($MASE$) and Root Mean Squared Scaled Error($RMSSE$) are within the category of Scaled errors. The key benefit of these indicators is that they are independent of each time series' scale while also avoiding the undetermined values of the percentage errors. Using naive predictions, a scaled error is defined as follows:

$$q_t = \frac{e_t}{\frac{1}{T-m}\sum_2^T |y_t - y_{t-1}|}. \tag{2.26}$$

The scale of the original data is unrelated to q_t because both the numerator and denominator contain numbers on that scale. If the absolute value of a scale error is greater than one then the forecast is worse than the average one-step Naïve forecast computed on the training data. Otherwise, if the absolute value of the scale error is less than 1, it means that the prediction is better than the average one-step Naïve forecast computed on the training data.

The parameter m of formulas (2.11) characterizes the seasonality of the time series, if there is no seasonality, m is set to unity, in the case that there is monthly seasonality, m is set to 12, etc. The following formulas give us the definitions of ($MASE$) and ($RMSSE$). It should be noted that ($RMSSE$) was used as an evaluation measure in the $M5$ competition (2022).

$$MASE = \frac{\frac{1}{H} \cdot \sum_{t=T+1}^{T+H} |e_t|}{\frac{1}{T-m}\sum_2^T |y_t - y_{t-1}|}, \tag{2.27}$$

$$RMSSE = \frac{\frac{1}{H} \cdot \sum_{t=T+1}^{T+H} e_t^2}{\frac{1}{T-m}\sum_2^T (y_t - y_{t-1})^2}. \tag{2.28}$$

2.12 Validating Time Series Models via Cross-Validation

A more advanced approach to dividing data into train and test sets for time series models is time series cross-validation. This technique involves creating a series of test sets, each consisting of a single observation, with the corresponding training

https://doi.org/10.26219/heal.aueb.6516

set comprised only of observations that occurred before the observation forming the test set. This ensures that no future observations are utilized in constructing the forecast. Since it is impossible to make a reliable forecast with a small training set, the initial observations aren't thought of as test sets.

Figure (2.3) illustrates the sequence of training and test sets. The training sets are represented by blue observations, and the orange observations form the test sets. Forecast accuracy is determined by averaging across the test sets. This procedure is sometimes called "evaluation on a rolling forecasting origin" since the forecast's origin rolls forward in time.

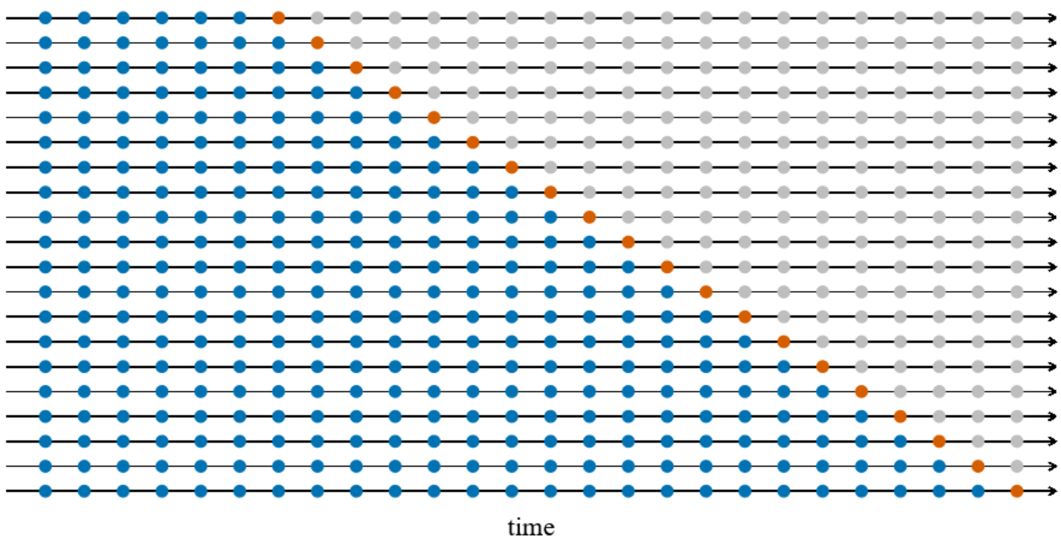

Figure 2.3: A Visual Representation of One-Step Cross-Validation in Time Series Forecasting.

Figure (2.4) illustrates the multi-step forecasts being more pertinent than one-step forecasts, the cross-validation technique based on a rolling forecasting origin can be adjusted to permit the use of multi-step errors. If we aim to generate accurate 4-step-ahead forecasts, the corresponding diagram would appear as shown below.

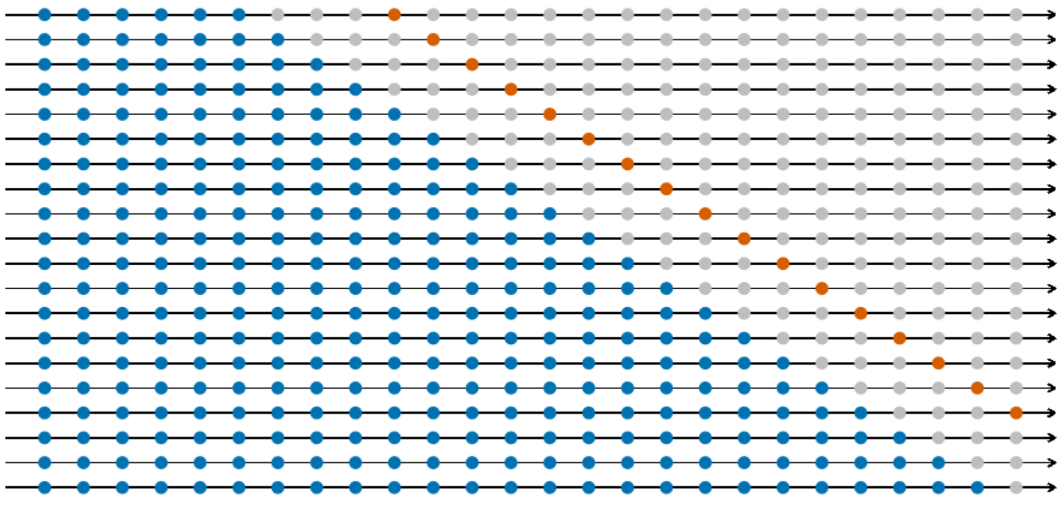

Figure 2.4: A Visual Representation of Multi-step Cross-Validation in Time Series Forecasting

In the next chapter, we will present the ARIMA models, their properties, and some of their variations. Additionally, a dynamic regression models will be described, and an automatic procedure for finding the optimal model for a time series will be discussed.

https://doi.org/10.26219/heal.aueb.6516

https://doi.org/10.26219/heal.aueb.6516

Chapter 3

ARIMA models

3.1 Introduction

ARIMA (Autoregressive Integrated Moving Average) models are a class of statistical models commonly used for analyzing and forecasting the future values of a stationary time series. They capture the linear relationships and patterns in the time series by modeling the series as a combination of three key components: autoregressive (AR), moving average (MA), and differencing (I). The AR component models the linear relationship between the current value and its past values, the MA component models the short-term patterns in the model's errors, and the differencing component transforms the non-stationary time series into a stationary time series.

ARIMA models are widely used for forecasting stock prices, economic indicators, and industrial process variables in finance, economics, and engineering fields. The model is adaptable to complex time series patterns such as trend, seasonality, and cyclic behavior.

The model parameters are estimated using methods such as maximum likelihood estimation or least squares estimation, and the estimated model is used to forecast future values of the time series.

3.2 Stationarity & Differencing

Stationarity is a key concept in time series analysis, and it refers to a time series' statistical properties remaining constant over time, as discussed in Section (2.5). The constant mean, variance, and autocorrelation structure of a stationary time series simplifies modeling and forecasting. A non-stationary time series, on the other

https://doi.org/10.26219/heal.aueb.6516

hand, has a changing mean, variance, or autocorrelation structure, making it difficult to model and predict.

Differencing is a common method for transforming a non-stationary time series into a stationary one. The process of differencing entails computing the difference between consecutive observations in a time series. The objective is to eliminate the underlying trend or seasonality that is causing non-stationary behavior.

We can use the Augmented Dickey-Fuller (ADF) test to determine the degree of differencing required to make a non-stationary time series stationary that tests the null hypothesis that the time series is non-stationary. The ADF test computes a test statistic and compares it to critical values to determine if the null hypothesis can be rejected.

If the ADF test indicates that the time series is non-stationary, we can then apply differencing to transform it into a stationary one. We can either apply first-order differencing, which subtracts each observation from its previous one, or higher-order differencing if needed. The differencing process can be repeated until the time series becomes stationary.

However, it should be used carefully, and the degree of differencing required should be determined using statistical tests, and the process should be carefully executed to avoid over-differencing (i.e. that can lead to the loss of valuable information and can make the time series too volatile).

3.3 Autoregressive Model (AR)

An Autoregressive (AR) model is a type of time series model that predicts future values based on past values. The AR model's basic premise is that the current value of a time series is modeled as a linear combination of its past values. The order of the model, denoted by p, determines the number of past values used to predict the current value.

The general form of an AR(p) model is:

$$Y_t = c + \Phi_1 Y_{t-1} + \Phi_2 Y_{t-2} + ... + \Phi_p Y_{t-p} + \epsilon_t$$

where:

Y_t is the value of the time series at time t.

$\Phi_1, \Phi_2, ..., \Phi_p$ are the autoregressive coefficients that represent the influence of past values of the time series on the current value.

ϵ_t is the white noise at time t, and c is a constant.

The order of the AR model, p, is determined by analyzing the autocorrelation plot of the time series. The autocorrelation plot shows the correlation between the current value of the series and its past values at different lags.

When time series data exhibits a trend or seasonality, and the current value of the series is influenced by its past values, an AR model is a good choice. For instance, stock prices frequently exhibit a trend over time, with periods of increasing or decreasing value. In this case, an AR model can be used to capture the impact of past values on the series' current value and forecast future values based on the trend.

3.4 Moving Average Model (MA)

A Moving Average (MA) model is also a type of statistical model that is used to analyze and forecast time series data. A MA model, as opposed to an AR model, models the current value of a time series as a linear combination of its past forecast error terms. The model's order, denoted by q, determines the number of past error terms used to predict the current value.

The general form of an MA(q) model is:

$$Y_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + ... + \theta_q \epsilon_{t-q}$$

where:

Y_t is the value of the time series at time t.

μ is the mean of the time series.

ϵ_t is the white noise term at time t.

$\theta_1, \theta_2, ..., \theta_q$ are the moving average coefficients that represent the influence of past error terms on the current value.

q is the order of the moving average model.

The order of the MA model, q, is determined by examining the time series' autocorrelation plot. The autocorrelation plot depicts the relationship between the

https://doi.org/10.26219/heal.aueb.6516

current value of the series and its previous values at various lags. The autocorrelation plot in an MA(q) model should show a significant correlation at lag q and no significant correlation at higher lags.

An MA model is a good choice when the time series data exhibits erratic or random behavior, and the current value of the series is influenced by error terms from past periods. For instance, the demand for a product may fluctuate at random over time as a result of factors such as changes in consumer preferences or economic conditions. In this case, an MA model can be used to capture the influence of past error terms on the series' current value and forecast future values based on random fluctuations.

3.5 Non-seasonal ARIMA Model

ARIMA stands for Autoregressive Integrated Moving Average. As mentioned before, a non-seasonal ARIMA model are based on three main components:

- The AutoRegressive (AR) component that models the linear dependence of the variable on its past values. It is denoted by AR(p), where p is the order of the autoregressive process.
- The Moving Average (MA) component that models the linear dependence of the variable on past errors or residuals. It is denoted by MA(q), where q is the order of the moving average process.
- The Integrated (I) component which models the differencing of the variable to make it stationary. It is denoted by I(d), where d is the order of differencing required to make the series stationary.

The notation for an ARIMA model is ARIMA(p, d, q), where p, d, and q are the parameters for the AR, I, and MA components, respectively.

The ARIMA model can be represented by the following formula:

$$Y_t' = c + \Phi_1 Y_{t-1}' + \Phi_2 Y_{t-2}' + ... + \Phi_p Y_{t-p}' + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + ... + \theta_q \epsilon_{t-q} + \epsilon_t$$

where:

Y_t' is the differenced series (it may have been differenced more than once).

https://doi.org/10.26219/heal.aueb.6516

$\Phi_1, \Phi_2, ..., \Phi_p$ are the autoregressive coefficients.
ϵ_t is the white noise at time t, and c is a constant.

The ARIMA model can be expressed as a difference equation:

$$(1 - \Phi_1 B - \Phi_2 B^2 - ... - \Phi_p B^p)(1 - B)^d Y_t = c + (1 + \theta_1 B + \theta_2 B^2 + ... + \theta_q B^q)\epsilon_t$$

where the backshift operator B is defined to perform the following operation: it causes the observation that it multiplies to be shifted backwards in time by 1 period. That is, for any time series Y and any period t:

$$BY_t = Y_{t-1}$$

and so the first difference operator $(1 - B)$, $d = 1$ equals to:

$$(1 - B)Y_t = Y_t - BY_t = Y_t - Y_{t-1}$$

Multiplication by higher powers of B correspondingly yields a backward shift by more than 1 period:

$$B^2 Y_t = B(BY_t) = B(Y_{t-1}) = Y_{t-2}$$

and in general, for any integer n:

$$B^n Y_t = Y_{t-n}$$

Thus, multiplying by B-to-the-n^{th} power has the effect of shifting an observation backwards by n periods. The differencing operator $(1 - B)^d$ is applied to the time series Y_t d times to make it stationary. The AR(p) and MA(q) components are used to model the autocorrelation and moving average of the stationary time series.

As is shown in the following table, many of the models we have already discussed are special cases of the ARIMA model.

3.6 Seasonal ARIMA model

It is important to note that ARIMA models can be extended to incorporate seasonal effects, resulting in the Seasonal ARIMA (SARIMA) model. This model is denoted by ARIMA(p,d,q)(P,D,Q)m.

https://doi.org/10.26219/heal.aueb.6516

Table 3.1: Special cases of ARIMA models.

White noise	ARIMA(0,0,0)
Random walk	ARIMA(0,1,0) with no constant
Random walk with drift	ARIMA(0,1,0) with a constant
Autoregression	ARIMA(p,0,0)
Moving average	ARIMA(0,0,q)

The mathematical representation of a SARIMA model is given by the following difference equation:

$$(1-\phi_1 B-\phi_2 B^2-...-\phi_p B^p)(1-\Phi_1 B^m-\Phi_2 B^{2m}-...-\Phi_P B^{Pm})(1-B)^d(1-B^m)^D Y_t =$$

$$c + (1+\theta_1 B+\theta_2 B^2+...+\theta_q B^q)(1+\Theta_1 B^m+\Theta_2 B^{2m}+...+\Theta_Q B^{Qm})\epsilon_t,$$

where B is the backshift operator, c is a constant term, ϵ_t is a white noise error term with mean zero and constant variance, $\phi_1, \phi_2, ..., \phi_p$, and $\theta_1, \theta_2, ..., \theta_q$ are the coefficients of the non-seasonal AR and MA components, respectively, $\Phi_1, \Phi_2, ..., \Phi_P$, and $\Theta_1, \Theta_2, ..., \Theta_Q$ are the coefficients of the seasonal AR and MA components, respectively, d is the order of non-seasonal differencing required to make the series stationary, D is the order of seasonal differencing required to make the series stationary, and m is the seasonal period.

The seasonal difference operator $(1 - B^m)$ is defined as:

$$(1 - B^m)Y_t = Y_t - Y_{t-m}$$

The seasonal part of the model consists of terms that are similar to the non-seasonal components of the model but involve backshifts of the seasonal period. For example, an ARIMA$(1, 1, 1)(1, 1, 1)_4$ model (without a constant) is for quarterly data $(m = 4)$, and can be written as

$$(1-\phi_1 B)(1-\Phi_1 B^4)(1-B)(1-B^4)Y_t = c + (1+\theta_1 B)(1+\Theta_1 B^4)\epsilon_t$$

https://doi.org/10.26219/heal.aueb.6516

Seasonal patterns in the data influence the choice between a SARIMA model and an ARIMA model. If the time series data indicate clear seasonal patterns, a SARIMA model is preferable to an ARIMA model.

A seasonal decomposition of the time series can be performed to identify the presence of seasonal components in order to determine whether a SARIMA model is required. A seasonal decomposition divides a time series into trend, seasonal, and residual components. If there is a significant seasonal component in the data, a SARIMA model should be used to capture the seasonal patterns. If the seasonal component is insignificant, an ARIMA model can be used to model the data's non-seasonal patterns.

In summary, when the time series exhibits significant seasonal patterns or cycles, a SARIMA model should be used, and when the time series does not exhibit any significant seasonal patterns or cycles, an ARIMA model should be used.

3.7 Forecast in Arima model

The following three steps can be used to calculate point forecasts.

1. Extend the ARIMA equation so that the left side of the equation contains y_t and the right side contains all other terms.
2. Substitute $T + h$ for t in the equation.
3. Replace future observations with their forecasts on the right-hand side of the equation, future errors with zero, and past errors with the corresponding residuals.

These procedures are then repeated for $h = 2, 3, \dots$ starting with $h = 1$, until the computation of all forecasts. An example makes the process the most understandable. We will use the $ARIMA(3, 1, 1)$ model, which can be expressed as follows, to show it.

$$(1 - \hat{\phi}_1 B - \hat{\phi}_2 B^2 - \hat{\phi}_3 B^3)(1 - B)y_t = (1 + \hat{\theta}_1 B)\varepsilon_t.$$

Then we expand the left hand side to obtain

$$\left[1 - (1 + \hat{\phi}_1)B + (\hat{\phi}_1 - \hat{\phi}_2)B^2 + (\hat{\phi}_2 - \hat{\phi}_3)B^3 + \hat{\phi}_3 B^4\right] y_t = (1 + \hat{\theta}_1 B)\varepsilon_t,$$

using the backshift operator results in

$$y_t - (1 + \hat{\phi}_1)y_{t-1} + (\hat{\phi}_1 - \hat{\phi}_2)y_{t-2} + (\hat{\phi}_2 - \hat{\phi}_3)y_{t-3} + \hat{\phi}_3 y_{t-4} = \varepsilon_t + \hat{\theta}_1 \varepsilon_{t-1}.$$

https://doi.org/10.26219/heal.aueb.6516

Last but not least, we shift all terms—aside from y_t to the right side:

$$y_t = (1 + \hat{\phi}_1)y_{t-1} - (\hat{\phi}_1 - \hat{\phi}_2)y_{t-2} - (\hat{\phi}_2 - \hat{\phi}_3)y_{t-3} - \hat{\phi}_3 y_{t-4} + \varepsilon_t + \hat{\theta}_1\varepsilon_{t-1}. \tag{3.1}$$

The first step is now finished. Although the equation now resembles an $ARIMA(4,0,1)$, the original $ARIMA(3,1,1)$ model is still present. The coefficients don't satisfy the stationarity requirements, hence it cannot be regarded as an $ARIMA(4,0,1)$.

In step two, we switch t in (3.1) with $T+1$:

$$y_{T+1} = (1 + \hat{\phi}_1)y_T - (\hat{\phi}_1 - \hat{\phi}_2)y_{T-1} - (\hat{\phi}_2 - \hat{\phi}_3)y_{T-2} - \hat{\phi}_3 y_{T-3} + \varepsilon_{T+1} + \hat{\theta}_1\varepsilon_T.$$

All values on the right hand side, assuming we have observations up to time T, are known, with the exception of ε_{T+1}, which we replace with zero, and ε_T, which we replace with the most recent residual e_T:

$$\hat{y}_{T+1|T} = (1 + \hat{\phi}_1)y_T - (\hat{\phi}_1 - \hat{\phi}_2)y_{T-1} - (\hat{\phi}_2 - \hat{\phi}_3)y_{T-2} - \hat{\phi}_3 y_{T-3} + \hat{\theta}_1 e_T.$$

By replacing $T+2$ for t in (3.1), a prediction of y_{T+2} is produced. Except for y_{T+1}, which we replace with $\hat{y}_{T+1|T}$, and ε_{T+2} and ε_{T+1}, both of which we replace with zero, all of the numbers on the right-hand side will be known at time T:

$$\hat{y}_{T+2|T} = (1 + \hat{\phi}_1)\hat{y}_{T+1|T} - (\hat{\phi}_1 - \hat{\phi}_2)y_T - (\hat{\phi}_2 - \hat{\phi}_3)y_{T-1} - \hat{\phi}_3 y_{T-2}.$$

This process continues for all future periods of time. Any number of point forecasts can be generated in this way.

3.8 Dynamic regression model

The previous session's time series model allows for the inclusion of information from previous observations of a series, but not for the inclusion of other information that may be relevant. Holidays, competitor activity, legal changes, and other external variables, for example, may explain some of the historical variations and lead to more accurate forecasts. They do, however, allow for the inclusion of a large amount of relevant information from predictor variables but do not allow for the subtle time series dynamics that ARIMA models can handle. In this section, we will explain how to extend ARIMA models to allow for the inclusion of additional information.

Specifically, we will allow the errors from regression to contain autocorrelation. In ordinary regression, ϵ_t is usually assumed to be an uncorrelated error term (i.e.white

https://doi.org/10.26219/heal.aueb.6516

noise). To emphasize this change in perspective, we will replace the white noise ϵ_t with η_t in the equation. The error series η_t is assumed to follow an ARIMA model.

For example, if η_t follows an ARIMA(1,1,1) model, we can write

$$Y_t = \beta_0 + \beta_1 X_{1,t} + \beta_2 X_{2,t} + ... + \beta_k X_{k,t} + \eta_t$$

,

$$(1 - \phi_1 B)(1 - B)\eta_t = (1 + \theta_1 B)\epsilon_t \Leftrightarrow \eta_t = c + \eta_{t-1} + \phi_1(\eta_{t-1} - \eta_{t-2}) + \theta_1 \epsilon_{t-1} + \epsilon_t$$

The model has two error terms here the error from the regression model, denoted by η_t, and the error from the ARIMA model, denoted by ϵ_t. Only the ARIMA model errors are assumed to be white noise.

When we estimate the model parameters, we need to minimize the $\sum \eta_t^2$. In this case, some errors may arise:

1. Estimated coefficients $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ are no longer optimal as some information is ignored.

2. Statistical tests associated with the model (e.g. t-tests on the coefficients) are incorrect.

3. P-values for coefficients are usually too small ("spurious regression").

4. AICc values of fitted models are misleading.

Minimizing the sum of squared values ($\sum \epsilon_t^2$) avoids these problems. Alternatively, maximum likelihood estimation can be used; this will give similar estimates of the coefficients with the method of minimizing the sum of squared error of Arima model.

When estimating a regression with ARMA errors, it is critical to consider that all of the variables in the model must first be stationary. As a result, we must first verify that Y_t and all of the predictors $X_{1,t}, ..., X_{k,t}$ appear to be stationary. If we estimate the model while any of these are non-stationary, the estimated coefficients will be inconsistent in most cases.

Therefore, we should first differentiate the non-stationary variables in the model. It is a good practice to keep the form of the relationship between Y_t and the predictors. So if necessary, we differentiate all of the variables if any of them require

https://doi.org/10.26219/heal.aueb.6516

differentiation. After differentiating the non-stationary variables in the model, then we only need to consider an ARMA process for the errors.

3.9 Dynamic harmonic regression model

When there are long seasonal periods, it is a good practice to implement a dynamic regression with Fourier terms. In other words, we can combine Fourier terms with ARIMA errors (η_t).

$$Y_t = \beta_0 + \beta_1 X_{1,t} + \beta_2 X_{2,t} + ... + \beta_k X_{k,t} + \eta_t,$$

$$X_{1,t} = \sin\left(\tfrac{2\pi t}{m}\right), X_{2,t} = \cos\left(\tfrac{2\pi t}{m}\right),$$
$$X_{3,t} = \sin\left(\tfrac{4\pi t}{m}\right), X_{4,t} = \cos\left(\tfrac{4\pi t}{m}\right),$$

This approach is extremely beneficial since it offers the following advantages:

1. It allows any length of seasonality.
2. For data with more than one seasonal period, you can include Fourier terms of different frequencies.
3. The smoothness of the seasonal pattern can be controlled by K, the number of Fourier sin and cos pairs.
4. The seasonal pattern is smooth for small values of K (but more wiggly seasonality can be handled by increasing K).
5. The short-term dynamics are easily handled with a simple ARMA error.

However, the disadvantage of this approach, compared to a seasonal ARIMA model, is that seasonality is assumed to be fixed since the seasonal pattern cannot change over time. Except for long-time series, seasonality is usually remarkably constant in practice, so this is not a significant disadvantage.

3.10 Automated ARIMA Modeling

The $ARIMA()$ function in the fable package utilizes a modified version of the Hyndman-Khandakar algorithm (2008) to automatically determine the appropriate ARIMA model for a given time series data. The algorithm incorporates unit root

https://doi.org/10.26219/heal.aueb.6516

tests, the minimization of Akaike Information Criterion corrected (AICc) and maximum likelihood estimation (MLE) techniques to determine the optimal ARIMA model. The arguments provided to ARIMA() can be customized to modify the algorithm's behavior.

The algorithm starts by determining the number of differences, represented by $0 \leq d \leq 2$, through repeated KPSS tests. The values of p and q are then chosen by minimizing the AICc after the data has been differenced d times. Rather than considering every possible combination of p and q, the algorithm employs a stepwise search to navigate the model space.

The algorithm fits four initial models: $ARIMA(0, d, 0)$, $ARIMA(2, d, 2)$, $ARIMA(1, d, 0)$, and $ARIMA(0, d, 1)$. If d is less than or equal to 1, an additional model is fitted: $ARIMA(0, d, 0)$ without a constant. If d is less than 2, a constant is included in the initial models.

The best model, as determined by the smallest AICc value, obtained from the 4 above models is chosen as the "current model". Variations on the current model are then considered by varying p and/or q by ± 1 and including or excluding c from the current model. The best model considered so far, either the current model or one of its variations, becomes the new current model. This process is repeated until no lower AICc value can be found.

Figure (3.1) provides a visual representation of how the Hyndman-Khandakar algorithm traverses the space of $ARMA$ orders through an example. The initial set of models considered by the algorithm is shown in orange, with the $ARMA(2, 2)$ model having the lowest $AICc$ value among these models. This model becomes the current model, and the algorithm searches over neighboring models. If a better model is found, it becomes the new current model. The algorithm continues in this fashion until no better model can be found. In this example, the model returned is an $ARMA(4, 2)$ model.

The default procedure of the Hyndman-Khandakar algorithm switches to a new current model as soon as a better model is identified, without going through all neighboring models. The full neighborhood search is performed when the greedy argument is set to FALSE. The algorithm also uses approximations to speed up the search, but these approximations can be avoided by setting the approximation argument to FALSE. However, it is possible that the minimum $AICc$ model will

https://doi.org/10.26219/heal.aueb.6516

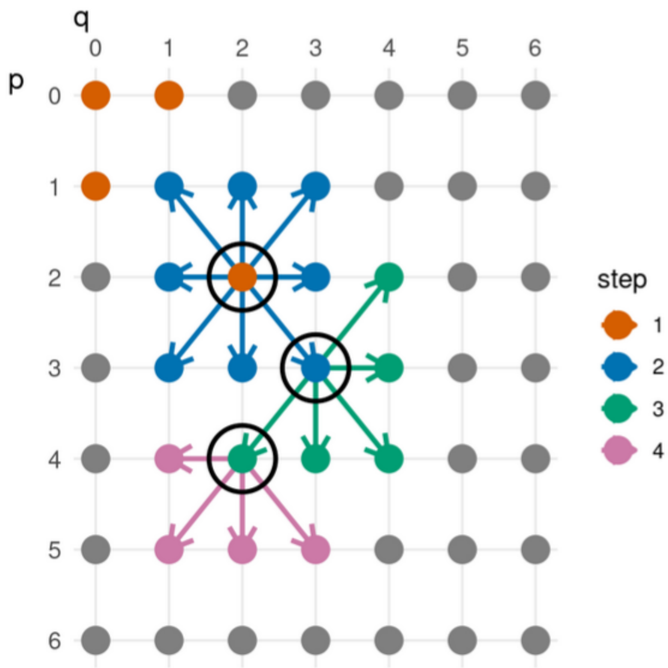

Figure 3.1: Example of the Hyndman-Khandakar Stepwise Search Process for ARIMA Modeling.

not be found due to these approximations or the use of the stepwise procedure. If a more exhaustive search is required, the stepwise argument can be set to FALSE.

https://doi.org/10.26219/heal.aueb.6516

Chapter 4

Hierarchical and grouped time series forecasting

Forecasting hierarchical and grouped time series involves making predictions for a set of related time series data that have a hierarchical or grouped structure. In both cases, the data can be divided into smaller sub-series that share a common characteristic. For example, sales data can be split by product, and region, resulting in a data hierarchy.

The challenging goal of hierarchical time series forecasting is to generate coherent forecasts across the entire aggregation structure. Forecasts that add up in a way that is consistent with the underlying aggregation structure are said to be coherent. For example, forecasts from all regions should add to country levels, and country levels should add to the higher level, and so on. Incoherent forecasts, on the other hand, can be reconciled to make them coherent.

Another point to be clarified is that hierarchical time series forecasting is not a time series forecasting methodology in and of itself (as opposed to ARIMA, ETS, or Prophet). Instead, it is a set of techniques that ensures forecasts are coherent across the given hierarchy of individual time series.

4.1 Hierarchical time series

A hierarchical time series is a collection of time series that are related and can be aggregated in a hierarchy. The hierarchy can be divided into levels, with the top level representing the overall series and the lower levels representing the sub-series. The levels of a hierarchy are frequently organized in a tree-like structure, with the

https://doi.org/10.26219/heal.aueb.6516

leaves representing individual series.

The use of hierarchical time series models can improve forecast accuracy by capturing both the individual series and the overall hierarchy (Athanasopoulos et al. (2017)). These models can be particularly useful in settings where there is a large number of related time series that need to be forecasted. For example, retailers may be interested in forecasting the sales of individual products at the store level, as well as the sales of product categories at the regional and national levels. By using hierarchical time series models, forecasters can take advantage of the hierarchical structure of the data and improve the accuracy of their forecasts.

However, one of the challenges of hierarchical time series forecasting is determining the appropriate level of aggregation. Aggregating data at a higher level can result in a loss of information, while aggregating data at a lower level can result in too much noise. Additionally, the choice of the aggregation level can impact the forecast accuracy. Therefore, it is important to carefully consider the trade-offs between the level of aggregation and forecast accuracy when building hierarchical time series models. Overall, by carefully selecting the appropriate aggregation level and using hierarchical time series models, forecasters can produce more accurate forecasts for large collections of related time series.

Forecasting hierarchical time series entails making predictions for the entire hierarchy while taking into account the relationship between individual series and aggregates. For illustration, we can consider a simple two-level hierarchical structure. At the top of the hierarchy is the "Total", the most aggregate level of the data. The tth observation of the Total series is denoted by y_t for $t = 1, \ldots, T$. The Total is disaggregated into two series, which are further divided into three and two series at the bottom of the hierarchy. Below the top level, we denote with $y_{j,t}$, the tth observation of the series corresponding to the node j. Given the above

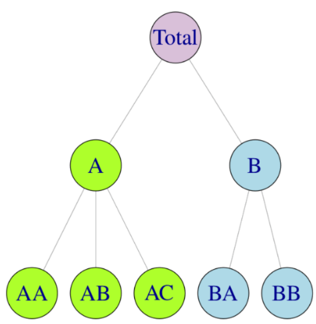

Figure 4.1: A two-level hierarchical tree diagram.

https://doi.org/10.26219/heal.aueb.6516

example, the total number of series in the hierarchy is $n = 1 + 2 + 5 = 8$, while the number of series at the bottom level is $m = 5$. Note that $m > n$ in all hierarchies.

For any time t, the observations at the bottom level of the hierarchy will sum to the observations of the series above. For this particular example, we get the following equations:

$$y_t = y_{AA,t} + y_{AB,t} + y_{AC,t} + y_{BA,t} + y_{BB,t} \tag{4.1}$$

$$y_{A,t} = y_{AA,t} + y_{AB,t} + y_{AC,t} \quad \text{and} \quad y_{B,t} = y_{BA,t} + y_{BB,t} \tag{4.2}$$

4.2 Grouped time series

A grouped time series is a set of related time series that can be grouped based on some characteristic, such as geography or product type. Unlike hierarchical time series, grouped time series do not have a clear hierarchical structure.

Again at the top of the grouped structure is the Total, the most aggregate level of the data. The Total can be disaggregated by attributes (A, B) forming series $y_{A,t}, y_{B,t}$, or by attributes (X, Y) forming series $y_{X,t}, y_{Y,t}$. At the bottom level, the data are disaggregated by both attributes.

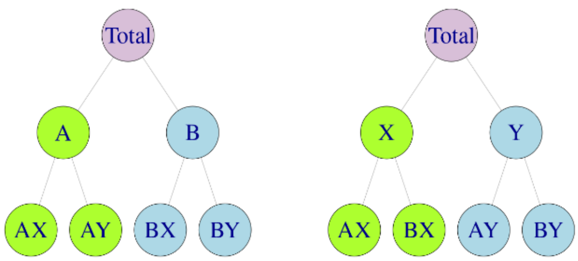

Figure 4.2: Alternative representations of a two-level grouped structure.

We can observe that there are alternative aggregation paths for grouped structures. For any time t, as with the hierarchical structure,

$$y_t = y_{AX,t} + y_{AY,t} + y_{BX,t} + y_{BY,t} \tag{4.3}$$

For the first level of the grouped structure, we have

$$y_{A,t} = y_{AX,t} + y_{AY,t} \quad \text{and} \quad y_{B,t} = y_{BX,t} + y_{BY,t} \tag{4.4}$$

$$y_{X,t} = y_{AX,t} + y_{BX,t} \quad \textit{and} \quad y_{B,t} = y_{AY,t} + y_{BY,t} \tag{4.5}$$

https://doi.org/10.26219/heal.aueb.6516

Grouped time series are similar to hierarchical time series, but they do not have a clear hierarchical structure. Instead, the data can be grouped based on some characteristic, such as geography or product type.

4.3 Mixed Hierarchical Models

Mixed hierarchical models can also be characterized as both nested and crossed at the same time. This is because the data can be organized into multiple levels of nesting, with observations nested within groups, and the groups themselves nested within higher-level units. At the same time, the groups can also be considered as crossed, as they may share common characteristics that are not dependent on their hierarchical structure. This makes the mixed hierarchy a particularly flexible and versatile approach to analyzing complex data structures.

The nested and crossed structure of the mixed hierarchy is particularly useful when analyzing data with complex dependencies between groups. For example, in product surveys, products may be nested within brands, which are in turn nested within categories. However, brands within the same category may also share common characteristics, such as location, that are not dependent on their hierarchical structure.

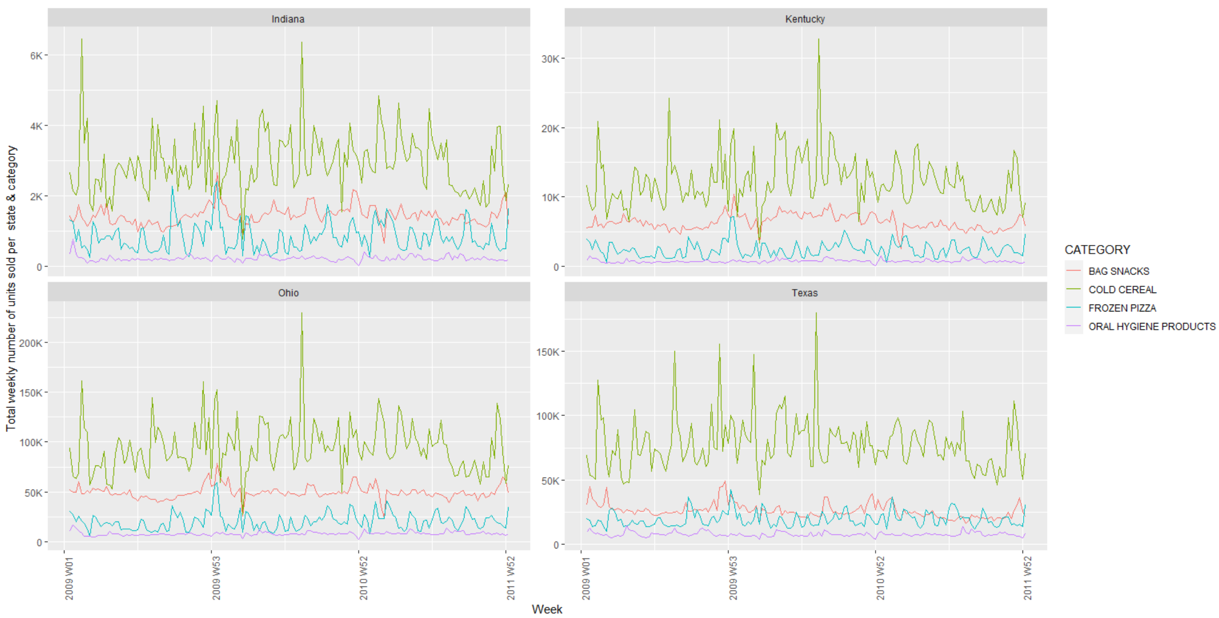

Figure 4.3: Mixed hierarchical structure.

https://doi.org/10.26219/heal.aueb.6516

In summary, the mixed hierarchy is a statistical approach that combines the advantages of nested and crossed structures, allowing for the estimation of both within-group and between-group variability, while also accounting for the commonalities between groups. This makes it a powerful and flexible approach to analyzing complex data structures with multiple levels of nesting and crossed factors.

4.4 Approaches to hierarchical or grouped time series forecasting

Forecasting hierarchical or grouped time series has traditionally involved selecting one level of aggregation and generating forecasts for that level. These are then aggregated for higher levels or disaggregated for lower levels to obtain a set of coherent forecasts for the rest of the structure.

Below, we present the main approaches to hierarchical or grouped time series forecasting.

- **The bottom-up approach**

  The bottom-up approach is a widely used method for generating forecasts in hierarchical time series analysis. This approach involves first generating forecasts for each series at the lowest level of the hierarchy, and then aggregating these forecasts to produce coherent forecasts for all the series in the hierarchy. The forecasts for bottom-level series are obtained using appropriate forecasting methods, and then summed to produce higher-level forecasts. The bottom-up approach provides flexibility in terms of forecasting models used for each series, while also ensuring consistency across different levels of the hierarchy. It is a simple and effective method for generating coherent forecasts for hierarchical time series data.

  The bottom-up approach is a method for forecasting hierarchical time series data, where the forecasts for the lowest-level series are aggregated to generate forecasts for higher-level series. This approach has several advantages and disadvantages, which are discussed below.

  For example, for the hierarchy of Figure (4.1), we first generate h-step-ahead

https://doi.org/10.26219/heal.aueb.6516

forecasts for each of the bottom-level series:

$$\hat{y}_{AAh}, \quad \hat{y}_{ABh}, \quad \hat{y}_{ACh}, \quad \hat{y}_{BAh} \text{ and } \hat{y}_{BBh}$$

Summing the h-step-ahead forecasts of the individual series at the bottom level results in coherent forecasts for the higher-level series in the hierarchy.

$$\begin{aligned} \tilde{y}_h &= \hat{y}_{AAh} + \hat{y}_{ABh} + \hat{y}_{ACh} + \hat{y}_{BAh} + \hat{y}_{BBh}, \\ \tilde{y}_{Ah} &= \hat{y}_{AAh} + \hat{y}_{ABh} + \hat{y}_{ACh}, \\ \text{and} \quad \tilde{y}_{Bh} &= \hat{y}_{BAh} + \hat{y}_{BBh}. \end{aligned}$$

we will use the “tilde”notation to indicate coherent forecasts.

Advantages:

Granularity: The bottom-up approach allows for a high level of granularity, as it generates forecasts for each individual series at the lowest level. This can be particularly useful when analyzing data with multiple levels of aggregation, such as sales data for a product across different regions, stores, and products.

Robustness: The bottom-up approach is often more robust than other methods, as it accounts for the specific characteristics and trends of each individual series. This can help to reduce errors in the forecasts, particularly when dealing with data that has high variability or seasonality.

Flexibility: The bottom-up approach is flexible and can be adapted to different levels of aggregation, allowing for forecasts at different levels of detail. This can help to meet the specific needs of different stakeholders, such as managers, analysts, or customers.

Disadvantages:

Complexity: The bottom-up approach can be complex and computationally intensive, particularly when dealing with large datasets. This can make it difficult to implement and analyze, particularly for organizations with limited computational resources.

Inefficiency: The bottom-up approach can be inefficient, as it requires forecasts to be generated for each individual series before they can be aggregated

https://doi.org/10.26219/heal.aueb.6516

to higher levels. This can result in longer processing times and delays in generating forecasts for higher-level series.

Information loss: The bottom-up approach may result in information loss, as the forecasts for the lower-level series are aggregated to generate forecasts for higher-level series. This can result in less accurate forecasts for higher-level series, particularly when dealing with data that has high variability or seasonality.

In summary, the bottom-up approach has several advantages, including its granularity, robustness, and flexibility. However, it also has several disadvantages, including its complexity, inefficiency, and potential for information loss. As such, the bottom-up approach should be carefully evaluated before being used to forecast hierarchical time series data, taking into account the specific needs and characteristics of the data and the stakeholders involved.

- **The top-down approach**
  Top-down approaches for hierarchical time series forecasting involve generating forecasts for the aggregate series at the highest level of the hierarchy, followed by disaggregating these forecasts to obtain individual forecasts for the lower-level series. The disaggregation process is determined by a set of proportions $p_1, \ldots, p_m$ (m the number of series at the bottom level), which determine the distribution of the aggregate forecasts among the lower-level series. For instance, in the hierarchy depicted in Figure (4.1), the disaggregation proportions $p_1, \ldots, p_m$ $(m = 5)$ are used to obtain the individual forecasts for the lower-level series we get:

$$\tilde{y}_{AAt} = p_1\hat{y}_t, \quad \tilde{y}_{ABt} = p_2\hat{y}_t, \quad \tilde{y}_{ACt} = p_3\hat{y}_t, \quad \tilde{y}_{BAt} = p_4\hat{y}_t \quad \text{and} \qquad \tilde{y}_{BBt} = p_5\hat{y}_t.$$

  After the generation of the individual h-step-ahead (coherent) forecasts for the series at the bottom level of the hierarchy, they are aggregated to generate coherent forecasts for the higher-level series.

  There are several top-down methods that can be used, and the two most common methods (Average historical proportions, Proportions of the historical averages) rely on the historical proportions of the data to specify the disaggregation proportions. These methods have demonstrated good performance

https://doi.org/10.26219/heal.aueb.6516

in the study conducted by Gross & Sohl (1990).

Using historical proportions to disaggregate forecasts may not be very accurate at lower levels of the hierarchy as it does not consider changes in those proportions over time. Therefore, top-down approaches based on historical proportions may not generate as accurate forecasts as bottom-up approaches.

To improve the accuracy of top-down approaches, Athanasopoulos et al. (2009) suggested using proportions based on forecasts rather than historical data.

Disadvantages:

Unbiased Forecasts: One disadvantage of all top-down approaches, is that they do not produce unbiased coherent forecasts (Hyndman et al., (2011)) even if the base forecasts are unbiased..

Lack of accuracy at lower levels: Top-down approaches may not be as accurate as bottom-up approaches at the lower levels of the hierarchy. This is because the disaggregation process relies on historical proportions, which may not hold true in the future.

- **Middle-out approach**

The mixed hierarchical approach, also known as the middle-out approach, is a hybrid of the top-down and bottom-up approaches and can only be applied to strictly hierarchical structures (no mixed hierarchical structures). Firstly, the method selects a "middle" level and generates forecasts for all the series at that level. The bottom-up approach is then used to generate coherent forecasts for the series above the middle level by aggregating the "middle-level" forecasts upwards. In contrast, the top-down approach is employed to generate coherent forecasts for the series below the middle level by disaggregating the "middle-level" forecasts downwards.

The mixed hierarchical approach offers the advantage of incorporating both bottom-up and top-down approaches, thereby improving forecast accuracy at all levels of the hierarchy. Moreover, this method is computationally efficient since it requires the generation of forecasts only for a single level rather than

https://doi.org/10.26219/heal.aueb.6516

all levels. However, the middle-out approach is only applicable to strictly hierarchical structures, and it is not always possible to identify a "middle" level that can provide the best results for all the levels in the hierarchy. Additionally, the method may be sensitive to the choice of the "middle" level, which can lead to biased results if the selected level is not optimal for the series at higher or lower levels.

4.5 Forecast reconciliation

For the hierarchical structure depicted in (4.1), Equations (4.1) and (4.2) represent the aggregation of data. Similarly, for the grouped structure depicted in (4.2), Equations (4.3) and (4.4) represent the aggregation of data. These equations serve as constraints on aggregation, expressing the requirement that the sum of lower-level series must equal the value of higher-level series.

To represent any aggregation structure, an $n \times m$ matrix $\mathbf{S}$ (known as the "**summing matrix**") can be constructed, which determines how the series at the bottom level of the hierarchy aggregate to form the higher-level series.

For the hierarchical structure in Figure (4.1), we can write

$$\begin{bmatrix} y_t \\ y_{At} \\ y_{Bt} \\ y_{AAt} \\ y_{ABt} \\ y_{ACt} \\ y_{BAt} \\ y_{BBt} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} y_{AAt} \\ y_{ABt} \\ y_{ACt} \\ y_{BAt} \\ y_{BBt} \end{bmatrix}.$$

The compact notation for expressing the hierarchical or grouped time series is given by

$$\mathbf{y}_t = \mathbf{S}\mathbf{b}_t. \tag{4.6}$$

Here, $\mathbf{y}_t$ is an n-dimensional vector that represents all the observations in the hierarchy at time t, $\mathbf{S}$ is the summing matrix, and $\mathbf{b}_t$ is an m-dimensional vector that represents all the observations in the bottom level of the hierarchy at time t. It

https://doi.org/10.26219/heal.aueb.6516

should be noted that the first row of the summing matrix S corresponds to Equation (4.1), while the second and third rows correspond to (4.2). The remaining rows in the matrix form an m-dimensional identity matrix $\mathbf{I}_m$, which ensures that each bottom-level observation on the right-hand side of the equation is equal to itself on the left-hand side.

Similarly, for the grouped structure shown in Figure (4.2), we can use the same compact notation $\mathbf{y}_t = \mathbf{Sb}_t$, where $\mathbf{y}_t$ is an n-dimensional vector that represents all the observations in the group at time t, $\mathbf{S}$ is the summing matrix, and $\mathbf{b}_t$ is a m-dimensional vector that represents all the observations in the bottom level of the group at time t.

$$\begin{bmatrix} y_t \\ y_{A,t} \\ y_{B,t} \\ y_{X,t} \\ y_{Y,t} \\ y_{AX,t} \\ y_{AY,t} \\ y_{BX,t} \\ y_{BY,t} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1` \end{bmatrix} \cdot \begin{bmatrix} y_{AX,t} \\ y_{AY,t} \\ y_{BX,t} \\ y_{BY,t} \end{bmatrix} \tag{4.7}$$

The matrix notation for expressing forecasting techniques for hierarchical or grouped time series uses the second and third rows of matrix $\mathbf{S}$ to represent Equation (4.4) and the fourth and fifth rows to represent (4.7). This notation provides a standardized approach to represent all forecasting methods.

Assuming that we generate forecasts for all series without considering any aggregation restrictions, we refer to these as the base forecasts, which are denoted by $\hat{\mathbf{y}}_\mathbf{h}$ for a forecast horizon of h. They are stacked in the same order as the data $\mathbf{y}_t$.

All coherent forecasting approaches for hierarchical or grouped structures can be expressed as

$$\tilde{\mathbf{y}}_h = \mathbf{SG}\hat{\mathbf{y}}_h, \tag{4.8}$$

in (4.8), where $\mathbf{G}$ is a matrix that maps the base forecasts into the bottom level, and the summing matrix $\mathbf{S}$ aggregates the forecasts using the hierarchical or grouped

https://doi.org/10.26219/heal.aueb.6516

structure to produce a set of coherent forecasts $\tilde{\mathbf{y}}$.

The definition of the matrix $\mathbf{G}$ depends on the chosen forecasting approach. For instance, in case of the bottom-up approach to forecast the hierarchy illustrated in (4.1), matrix $\mathbf{G}$ would be determined as follows.

$$\mathbf{G} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1‘ \end{bmatrix}$$

It should be noted that matrix $\mathbf{G}$ comprises of two partitions. The initial three columns of the matrix set the base forecasts of the series above the bottom level to zero, whereas the m-dimensional identity matrix selects only the base forecasts of the bottom level. The summing matrix $\mathbf{S}$ then aggregates these selected forecasts.

In the case of employing any of the top-down approaches to forecast the hierarchy, the matrix $\mathbf{G}$ would differ from that of the bottom-up approach. If any of the top-down approaches were used then

$$\mathbf{G} = \begin{bmatrix} \mathbf{p_1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \mathbf{p_2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \mathbf{p_3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \mathbf{p_4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \mathbf{p_5} & 0 & 0 & 0 & 0 & 0 & 0 & 0‘ \end{bmatrix}$$

In the case of implementing any of the top-down approaches to forecast the hierarchy, the matrix $\mathbf{G}$ will be distinct from that of the bottom-up approach. The first column of the matrix in top-down approaches incorporates a group of proportions that allocate the base forecasts of the top level to the bottom level. These base forecasts are subsequently aggregated by the summing matrix $\mathbf{S}$. The remaining columns of the matrix set the base forecasts beneath the highest level of aggregation to zero.

https://doi.org/10.26219/heal.aueb.6516

When adopting a middle-out approach, the matrix **G** will constitute a combination of the above two approaches. Through a set of proportions, the base forecasts of a particular pre-determined level are disaggregated to the bottom level. All other base forecasts are set to zero, and the bottom-level forecasts are subsequently aggregated up the hierarchy by means of the summing matrix.

Equation (4.8) shows that pre-multiplying any set of base forecasts with **SG** will return a set of coherent forecasts.

The conventional techniques discussed so far have a limitation as they rely on the base forecasts from a single level of aggregation that have been either aggregated or disaggregated to produce forecasts at other levels, thus utilizing limited information. However, in practice, we could incorporate other G matrices, and the SG algorithm combines and reconciles all base forecasts to produce coherent forecasts. Moreover, it is possible to identify the optimal G matrix that provides the most accurate reconciled forecasts.

The MinT optimal reconciliation approach

Wickramasuriya et al. (2019) found a **G** matrix that minimizes the total forecast variance of the set of coherent forecasts, leading to the MinT (Minimum Trace) optimal reconciliation approach.

Suppose we generate coherent forecasts using (4.8). First, we want to make sure we have unbiased forecasts. If the base forecasts $\hat{\mathbf{y}}_h$ are unbiased, then the coherent forecasts $\tilde{\mathbf{y}}_h$ will be unbiased provided that

$$\mathbf{SGS} = \mathbf{S}$$

This provides a constraint on the matrix **G**. Interestingly, no top-down method satisfies this constraint, so all top-down approaches result in biased coherent forecasts.

Next, we need to find the errors in our forecasts. Wickramasuriya et al. (2019) show that the variance-covariance matrix of the h-step-ahead coherent forecast errors is given by

$$\mathbf{V}_h = \mathrm{Var}[\mathbf{y}_{T+h} - \tilde{\mathbf{y}}_h] = \mathbf{SGW}_h\mathbf{G}'\mathbf{S}'$$

where $\mathbf{W_h} = \mathrm{Var}[(\mathbf{y}T + h - \hat{\mathbf{y}}_h)]$ is the variance-covariance matrix of the corresponding base forecast errors.

https://doi.org/10.26219/heal.aueb.6516

The objective is to find a matrix $\mathbf{G}$ that minimizes the error variances of the coherent forecasts. These error variances are on the diagonal of the matrix $\mathbf{V}_h$, and so the sum of all the error variances is given by the trace of the matrix $\mathbf{V}_h$. Wickramasuriya et al. (2019) show that the matrix $\mathbf{G}$ which minimises the trace of $\mathbf{V}_h$ such that $\mathbf{SGS} = \mathbf{S}$, is given by

$$\mathbf{G} = (\mathbf{S}'\mathbf{W}_h^{-1}\mathbf{S})^{-1}\mathbf{S}'\mathbf{W}_h^{-1}.$$

Therefore, the optimally reconciled forecasts are given by

$$\tilde{\mathbf{y}}_h = \mathbf{S}(\mathbf{S}'\mathbf{W}_h^{-1}\mathbf{S})^{-1}\mathbf{S}'\mathbf{W}_h^{-1}\hat{\mathbf{y}}_h. \tag{4.9}$$

To use this in practice, we need to estimate $\mathbf{W}_h$, the forecast error variance of the h-step-ahead base forecasts. This can be difficult, and so we provide some more simplifying approximations such as OLS reconciliation, WLS (weighted least squares) and structural scaling that have been shown to work well in both simulations and in practice.

https://doi.org/10.26219/heal.aueb.6516

https://doi.org/10.26219/heal.aueb.6516

Chapter 5

Analysis

5.1 Dataset

The dataset **Dunnhumby** includes sales and promotion information on the top three products from each of the top five brands falling under four selected categories. The data used were collected from January 2009 to January 2012. In this particular survey, data were recorded weekly, for 156 weeks, at each store and for each product. The main information reported for each product is the weekly sales recorded, the price of single units, the location where it was sold, and the use of promotional activities during that week if that was the case.

Table (5.1) refers to the number of products recorded throughout the survey's lifespan. Adding the table values gives the number of different products (55), each product is characterized by a unique code (UPC) and a product description.

Table 5.1: Number of products repetition during the weeks.

98 weeks	106 weeks	136 weeks	140 weeks	144 weeks	152 weeks	156 weeks
3	1	1	1	1	1	47

Table (5.3) introduces specific products' codes (UPCs) with remarkable contrasts regarding the amounts of UPCs sold throughout the weeks of years 2009, 2010, 2011 & 2012. More specifically, the first three UPCs (according to Table (5.2), the "frozen pizza" products manufactured by KING) had zero sales in 2009 and the first few weeks of 2010, as well. The following four products sign sales throughout the weeks of years 2010/2011, while for 2009 sales' start indicate some delay. It is worth men-

https://doi.org/10.26219/heal.aueb.6516

Table 5.2: Correspondence of products' codes (UPC) to "Description", "Manufacturer" & "Category".

UPC	DESCRIPTION	MANUFACTURER	CATEGORY
2066200530	NWMN OWN PEPPERONI PIZZA	KING	FROZEN PIZZA
2066200531	NWMN OWN 4 CHEESE PIZZA	KING	FROZEN PIZZA
2066200532	NWMN OWN SUPREME PIZZA	KING	FROZEN PIZZA
3500068914	COLG SPEARMINT WISP	COLGATE	ORAL HYG. PR.
88491212971	POST FRUITY PEBBLES	POST FOODS	COLD CEREAL
88491201427	POST FM SZ HNYBNCH OT ALM	POST FOODS	COLD CEREAL
88491201426	POST HNY BN OTS HNY RSTD	POST FOODS	COLD CEREAL
3000006560	QKER CAP N CRUNCH BERRIES	QUAKER	COLD CEREAL
1111009477	PL MINI TWIST PRETZELS	PRIVATE LABEL	BAG SNACKS

tioning that for product 3500068914 (Table (5.2) - manufactured by Colgate) sales were realized until the end of March, 2011.

The last UPC code (1111009477) shows that products have been sold throughout the duration of this research. Similar behavior to the aforementioned UPC, have 46 other products (Table 5.1). Although product 3000006560 presented a similar behavior to the above-mentioned, it did not present sales during weeks 14/12/11, 07/12/11, 23/11/11, 28/09/11.

Data has a hierarchical structure (of three levels) related to products' categories, with the lowest level of this structure being the product while the highest level is the different categories of products. Figure (5.1) describes the aforementioned hierarchical structure of data. The highest hierarchical level (the most general one) is the different product categories, which in our case are "Cold Cereal", "Bag Snacks", "Frozen Pizza"and "Oral Hygiene Products".

The orange-colored shapes in Figure (5.1) illustrate the second level of the hierarchical structure. This is the brand of the product (16 brands in total). Five brands are reported for each category. It should be noted that there is a brand ("Private Label") that has products for all 4 product categories. For category "Frozen Pizza" and brand "Home Run" there are no products. It should be noted that the "Private Label" the company develops products for all four categories. The

https://doi.org/10.26219/heal.aueb.6516

Table 5.3: Number of weeks for product sales that had no sales in all survey weeks (156).

UPC	2009	2010	2011	2012	Total_Week	First_Week	Last_Week
2066200530	0	45	52	1	98	2010-02-24	2012-01-04
2066200531	0	45	52	1	98	2010-02-24	2012-01-04
2066200532	0	45	52	1	98	2010-02-24	2012-01-04
3500068914	42	52	12	0	106	2009-03-18	2011-03-30
88491212971	31	52	52	1	136	2009-05-13	2012-01-04
88491201427	35	52	52	1	140	2009-03-18	2012-01-04
88491201426	39	52	52	1	144	2009-03-18	2012-01-04
3000006560	51	52	48	1	152	2009-01-14	2012-01-04
1111009477	51	52	52	1	156	2009-01-14	2012-01-04

last category of this particular hierarchical structure is the products (yellow-colored shapes of the figure). The name of each product is fully corresponded in Table (A.4).

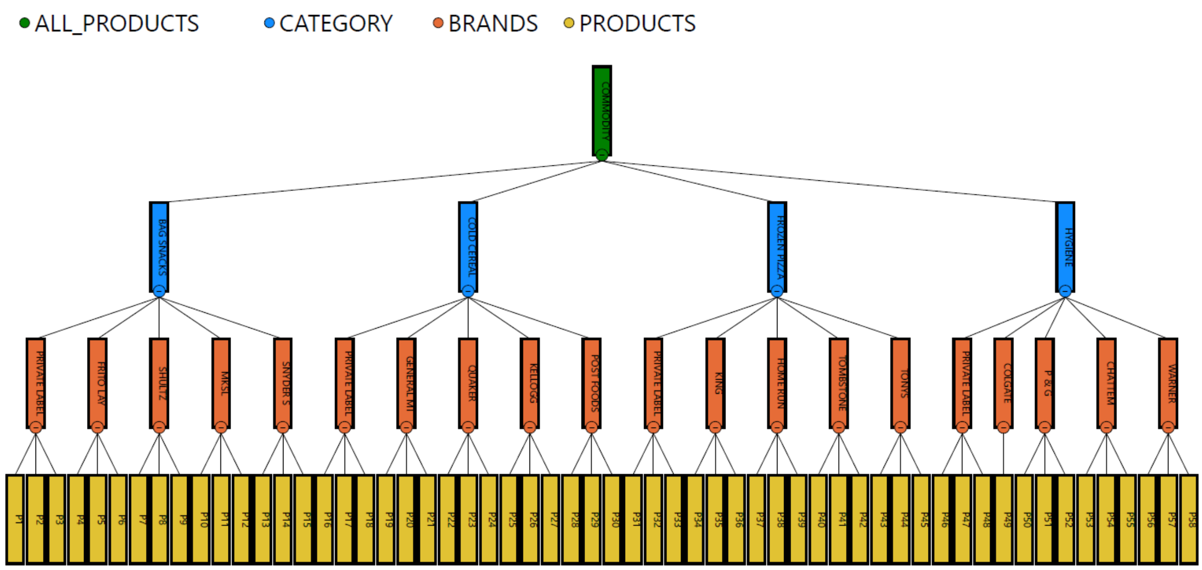

Figure 5.1: Hierarchical structure of data

Table (5.4) shows the quantities of products sold respectively for each category while Figure (5.2) shows times series for units sold per category and state. It can be inferred that most product sales fall in the “Cold Cereal”category, while the largest number of sales for each product category has been recorded in the state of “Ohio”.

https://doi.org/10.26219/heal.aueb.6516

The time series per category for the four different states follow a similar pattern.

Table 5.4: Total units sold per Category

Cold Cereal	Bag Snacks	Frozen Pizza	Oral Hygiene Products
$5.83M$	$2.60M$	$1.34M$	$0.52M$

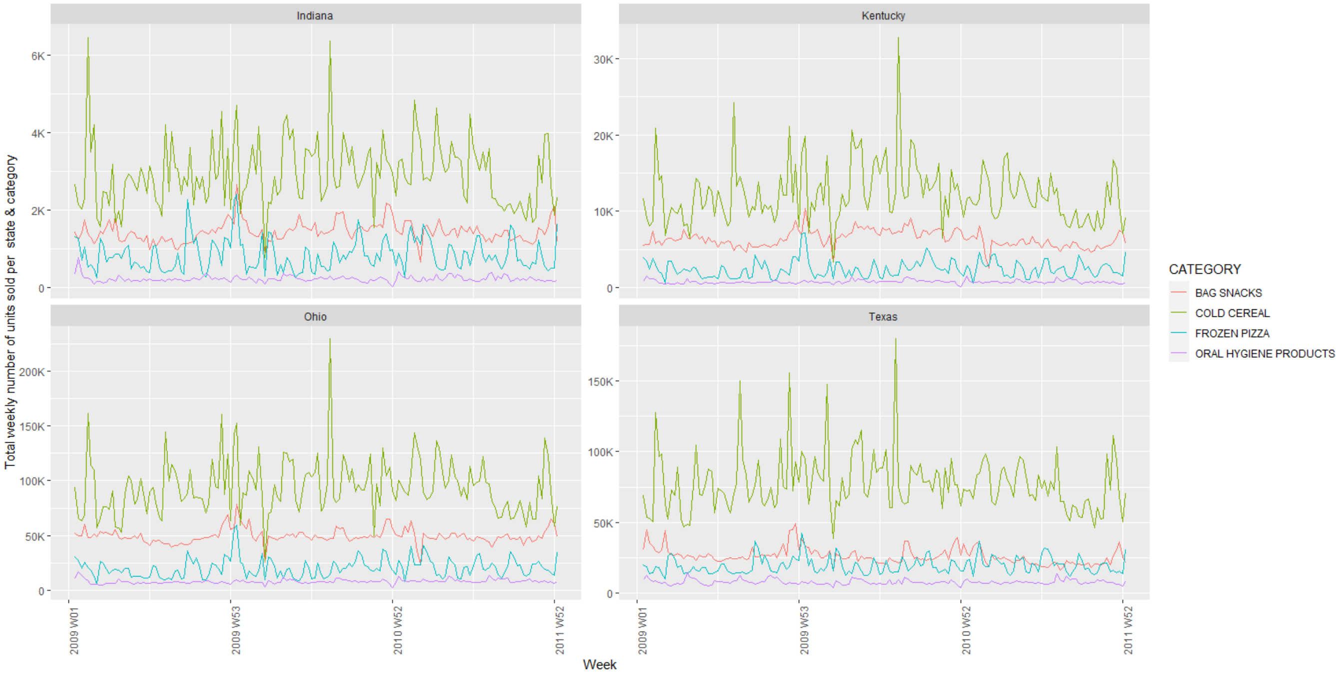

Figure 5.2: Time series for the units sold per category & state

However, from the time series of Figure (5.3), it is observed that the weekly quantities sold by category do not have the same pattern. The categories of products were compared to each other in terms of the absolute number of product units sold and not in terms of the generated revenues. That is because the price ranges for each product per category might have presented differences in terms of their prices and the comparison would not have been so fair enough.

Table (5.5) reports the total number of products sold and total revenue per brand. Table (A.1) refers to the weekly mean, median, and standard deviation of the products sold and revenue ($) per brand. From Tables (5.5) and (A.1), we can draw some direct conclusions. Initially, the brand "Private Label"occupies the 1st

https://doi.org/10.26219/heal.aueb.6516

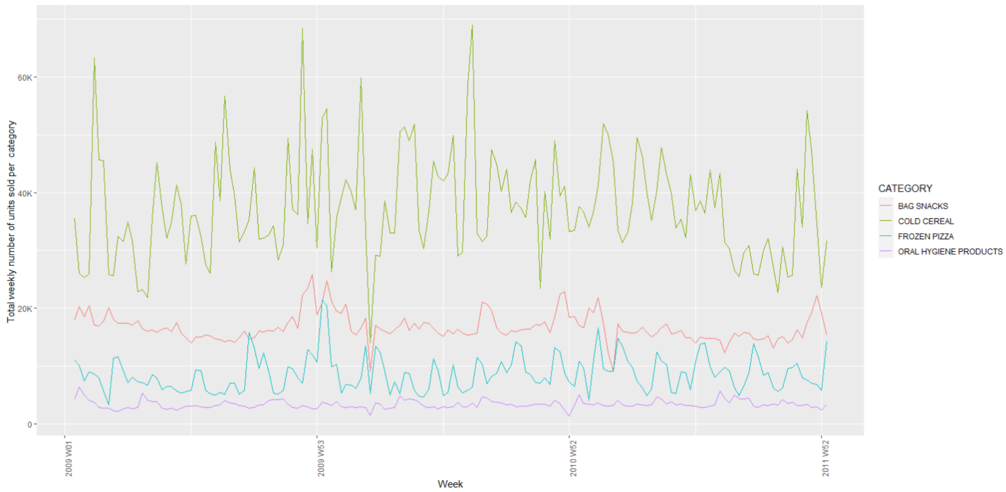

Figure 5.3: Time series for the units sold per category

or 2nd position with the highest volume of sales in each product category.

Most product sales result from a brand belonging to the “Bag Snack”category, which occupies the second place in terms of sales. Another remark is the positive correlation between the weekly products sold and the weekly standard deviation of the progenitors sold. A similar correspondence can be formulated between weekly revenues and their weekly standard deviation.

Table (5.6) represents the seven product subcategories.The table’s cells show the frequency of occurrence for each subcategory. Each brand encompasses $1-3$ subcategories while each subcategory consists of $1-3$ products. Figure (5.4) implements the comparison between the medians of weekly products sold for all subcategories. The weekly medians of the subcategories for the units sold can also be compared in the cases of the “Cold Cereal”and “Oral Hygiene Product”categories where there are more than one product subcategory.

The number of stores engaged in the survey is 77 appearing in a total of 51 cities. Table (5.7) shows the distribution of stores and cities that participated in the survey per state. It should be mentioned that the states of America are Texas, Ohio, Kentucky, and Indiana.

An interesting conclusion drawn from this particular data set is that the state with the highest revenue -for the entire period the survey lasted- in Ohio, while the state with the most stores that participated in the survey is Texas, which is further

https://doi.org/10.26219/heal.aueb.6516

Table 5.5: Total sales per Brand

BRAND	CATEGORY	Total_units_sold	Total_revenue
PRIVATE LABEL	BAG SNACKS	1,455,159	1,974,512
FRITO LAY	BAG SNACKS	361,333	1,013,737
SHULTZ	BAG SNACKS	319,518	560,008
SNYDER S	BAG SNACKS	379,043	992,790
MKSL	BAG SNACKS	86,406	190,936
PRIVATE LABEL	COLD CEREAL	1,099,276	2,099,992
GENERAL MI	COLD CEREAL	1,689,338	4,866,484
QUAKER	COLD CEREAL	758,809	1,737,705
KELLOGG	COLD CEREAL	1,420,972	3,851,245
POST FOODS	COLD CEREAL	858,987	2,452,926
PRIVATE LABEL	FROZEN PIZZA	467,509	1,612,330
KING	FROZEN PIZZA	72,748	424,139
TOMBSTONE	FROZEN PIZZA	591,164	3,258,225
TONYS	FROZEN PIZZA	211,873	1,164,865
PRIVATE LABEL	ORAL HYGIENE PRODUCTS	184,687	297,429
COLGATE	ORAL HYGIENE PRODUCTS	1,833	13,581
P & G	ORAL HYGIENE PRODUCTS	128,478	507,046
CHATTEM	ORAL HYGIENE PRODUCTS	49,662	228,301
WARNER	ORAL HYGIENE PRODUCTS	156,559	681,475

Table 5.6: Frequency of occurrence of subcategories by category.

Sub_Categ-Categ.	Bag Snacks	Cold Cereal	Frozen Pizza	Oral Hygiene Prod.
adult cereal	0	2	0	0
all family cereal	0	4	0	0
kids cereal	0	3	0	0
rinses and sprays	0	0	0	3
mouthwashes (antiseptic)	0	0	0	4
pizza/premium	0	0	4	0
pretzels	5	0	0	0

south than the other states, as mentioned previously.

https://doi.org/10.26219/heal.aueb.6516

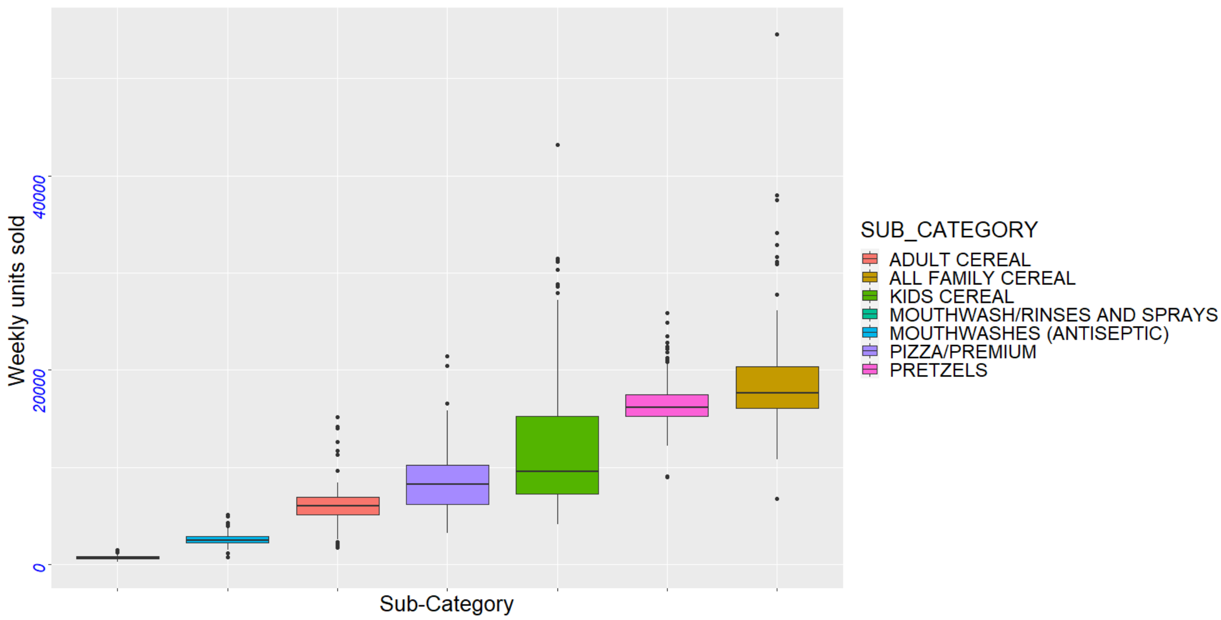

Figure 5.4: Box plots for all products sub-categories.

Table 5.7: Number of stores and cities per State

States	Texas	Ohio	Kentucky	Indiana
Stores:	41	31	4	1
Cities:	31	16	3	1

Table (5.8a) illustrates the total revenue per state, while Table (5.8b) shows the average revenue of each state per store. The values in the two tables indicate that the average revenue in a store in the northern states is higher than in a store in Texas. Table (5.9) refers to the weekly revenue for each state and the standard deviation of the states for the weekly revenues.

Figure (5.5) shows the weekly units sold per state/store. The pattern depicting the sales realized per state & store remains the same, while the sales of the three northern states have the same weekly average revenue, in contrast with Texas which has monthly sales reaching half compared to the rest.

As witnessed in Figure (5.1), data can be categorized hierarchically in products' categories. Additionally, there can be another hierarchical categorization related to geographical data. Those geographical data are divided into three ordered cate-

https://doi.org/10.26219/heal.aueb.6516

(a) Total revenue ($)

Texas	Ohio	Kentuky	Indianna
11.36M	14.35M	1.76M	0.45M

(b) Average revenue per store ($)

Texas	Ohio	Kentuky	Indianna
0.27M	0.46M	0.44M	0.45M

Table 5.8: Revenue per State

(a) Average weekly revenue ($)

Texas	Ohio	Kentuky	Indianna
73k	92k	11k	3k

(b) Standard deviation of weekly revenue ($)

Texas	Ohio	Kentuky	Indianna
10k	15k	2k	0.5k

Table 5.9: Statistical functions of weekly revenue by state

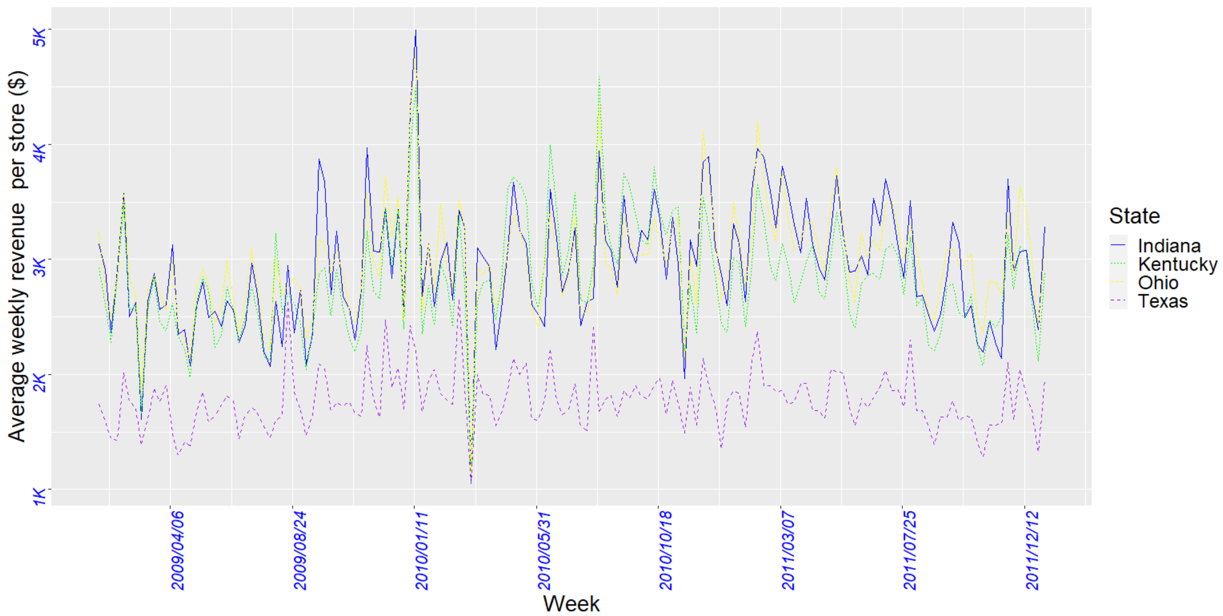

Figure 5.5: Average weekly revenue per store & state

gories, starting with the first category related to the states the research was realized, continuing with the cities of each state, and concluding with the stores participating in the research. The ordered structure described above is fully depicted in Figure (5.6).

In the data set, there were three product pairs with different product codes (UPC) but the same product description. This occurs as each product pair was the

https://doi.org/10.26219/heal.aueb.6516

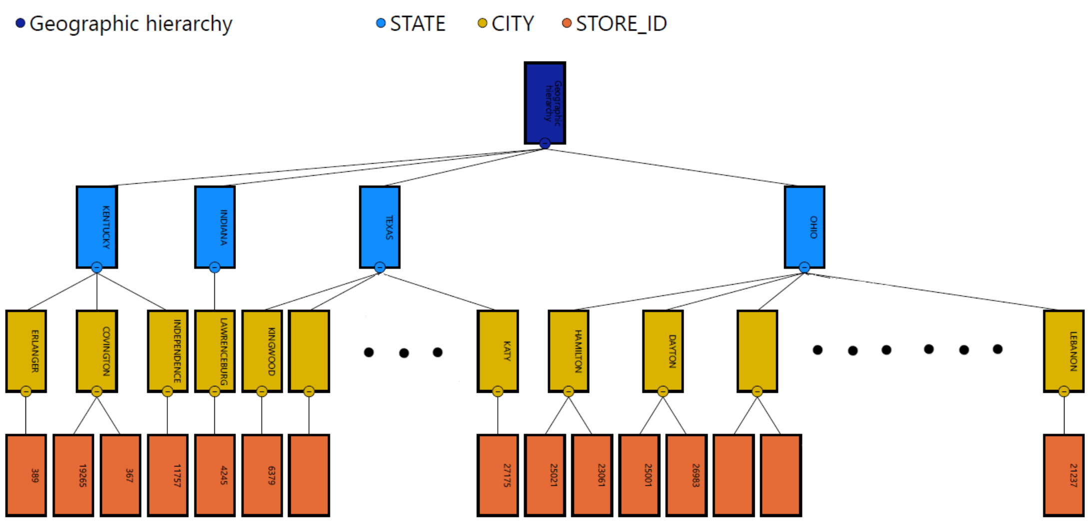

Figure 5.6: Location hierarchy

same product in a different quantity. In this case, it was vital to add a character to one product from each pair to have unique descriptions for each product. Grams (gr) for mass and milliliters (ml) for volume were also defined as units of measurement for the products. Also, the surface area of the stores is measured in square meters.

Finally, each of the 77 stores corresponds to a unique code. There are two codes with double values for all fields that describe each store except for the "SEG_VALUE_NAME" field. The two duplicate entries in the data set were deleted.

5.2 Fit Models - Hierarchy Data

Our upcoming data analysis will take place under the condition that we will only work for a portion of our initial data set as we can obtain in Figure (5.7).

The hierarchical dataset under consideration has bag snacks as the top level, resulting in manufacturers being positioned at the middle level, and the lowest level comprising the products they produce. A forecasting horizon of one month is to be considered, with a total of four forecasts to be conducted. In the first scenario, only the products will be consolidated, without taking into account the location where they were sold. Furthermore, models which have regressor (covariates) will not be employed. In other words, It will be assumed that the variance of the dependent variable is determined solely by the dependent variables of previous time points.

https://doi.org/10.26219/heal.aueb.6516

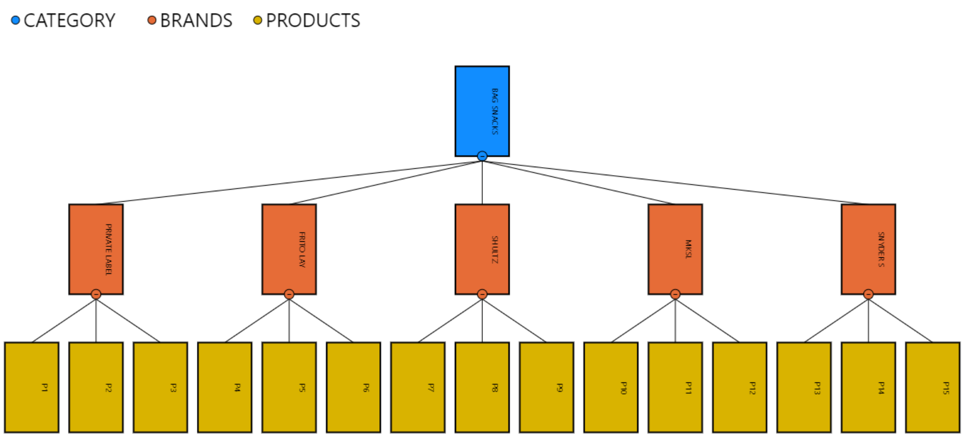

Figure 5.7: Product hierarchy

To develop our models, we exclusively utilized the *R* programming language and the fpp3 package. This statistical package is capable of selecting the most suitable model by identifying the one with the smallest AICc value. In specific, this process is followed for ARIMA and exponential smoothing (ETS) models. The algorithm implemented in this package is attributed to Hyndman and Khandakar (2008) and is recommended for automatic modeling. For instance, if we employ ARIMA models, the package will identify the optimal parameter values for each time series and hierarchy level that minimize the AICc value.

By leveraging the automatic modeling capabilities of the fpp3 package, predictive performance of various models, including ARIMA models, exponential smoothing (ETS), and simpler models such as Naive and SNaive. The reason for using these simple models is to identify and exclude them if they outperform our primary models.

Specifically, based on Figure (5.7), there are a total of fifteen time series at the low level, five at the middle level, and only one at the high level.

The performance is assessed for each model, strictly across the three hierarchical levels and the best model is chosen for each level. In specific, the mean model performance is evaluated for each time series in scope of the hierarchical level. The metrics in scope of this assessment are obtained by averaging the forecast error measures $(RMSE, MASE, RMSSE)$.

The results presented in Table (5.10) include the average metrics for each model, as discussed above. The table shows that in general, the error metrics are higher as we move up on the hierarchical level, as expected due to the higher number of

https://doi.org/10.26219/heal.aueb.6516

observations included for these levels. In addition, it is observed that for the first forecast period ARIMA models performed better than the other models, whilst for the remaining periods of the forecast, the best performance was observed in models which include seasonality, i.e. Snaive. This can be due to the fact that the forecasting periods include the last 3 weeks of the year and the first week of the next year, which typically exhibit strong seasonality in sales.

Table 5.10: Model Performance Across Hierarchical Level (Not Reconciliation Forecast)

LEVEL	h	MODEL	RMSE	MASE	RMSSE
Category	1	arima_search_all	1368.52	0.76	0.52
Brands	1	arima_search	240.14	0.52	0.39
Upc	1	arima_search	139.78	0.63	0.47
All_time-series	1	arima_search	233.06	0.62	0.46
Category	2	arima_search	547.28	0.30	0.21
Brands	2	arima_search_all	300.50	0.62	0.47
Upc	2	Snaive	218.80	0.72	0.50 t
All_time-series	2	Snaive	275.33	0.68	0.47
Category	3	Snaive	701.00	0.39	0.27
Brands	3	Snaive	374.20	0.63	0.44
Upc	3	Naive	196.80	0.97	0.71
All_time-series	3	Snaive	269.86	0.70	0.51
Category	4	ETS_search	650.76	0.36	0.25
Brands	4	ETS_search	185.57	0.30	0.22
Upc	4	ETS_search	99.84	0.47	0.35
All_time-series	4	ETS_search	146.49	0.43	0.32

The results presented in Table (5.11) include the average metrics strictly for the chosen model of each hierarchical level. The results show that in general ARIMA performed better than the other models, which is due to the fact that the assessment is made on an aggregate level of all forecasts. More details on the performance of each model per level can be found in Appendix A, Table (A.12).

It should be noted that in Tables (A.5)-(A.10) there are the correspondences of the parameters that have been selected for each time series model with the name

 https://doi.org/10.26219/heal.aueb.6516

of the model. For example, the model name Arima _stepwise in the time series at the top of the hierarchy corresponds to a model with parameters $ARIMA(1,1,1)(0,1,0)[52]$ (Table (A.9))

Table 5.11: Chosen Model Performance Strictly For each Hierarchical Level (Not Reconciliation Forecast)

LEVEL	MODEL	RMSE	MASE	RMSSE	DATA
Category	arima_search	1669.29	0.86	0.64	test
Brands	arima_search	451.84	0.63	0.51	test
Upc	arima_search	220.15	0.84	0.69	test
All_time-series	arima_search	344.32	0.79	0.64	test

In the case of reconciliation, there are a few differences in terms of the methodologies and models applied. In specific, three top-down approaches were utilized in different forms, meaning the weights applied to the observations. These approaches are a) Average Historical Proportions, b) Proportions of the Historical Average and c) Forecast Proportions. In addition, the middle-out approach is also utilized in the reconciliation where the middle level is Brands and the aforementioned three top-down methods are incorporated to reach the bottom-up approach.

As discussed in section (4.4), there is only one choice in order to reach top from middle-out approach, that is with bottom-up. Lastly, two $MinT$ optimal reconciliation approaches are utilized in the reconciliation, which effectively minimize the trace of Variance-Covariance matrix of actual and reconciliation forecasts in order to achieve an optimal solution. The top-down and middle-out methods are used in strictly hierarchical data.

The results presented in Table (5.12) include the average reconciliation metrics strictly for the chosen model of each hierarchical level. The results show that in general $MinT$ approach and ARIMA performed better than the other models, which is due to the fact that the assessment is made on an aggregate level of all forecasts. More details on the performance of each model are not presented due to dimensions of the data (i.e., 160rows).

The results presented in Table (5.13) include the average metrics for each model, as discussed above. The table shows that in general ARIMA family of models

https://doi.org/10.26219/heal.aueb.6516

Table 5.12: Chosen Model Performance Strictly For each Hierarchical Level (Reconciliation Forecast)

LEVEL	MODEL	Hier-Method	RMSE	MASE	RMSSE
Category	arima_search	MINT	1596.69	0.81	0.61
Brands	arima_search	-	451.84	0.63	0.51
Upc	arima_search	MINT	189.67	0.71	0.59
All_time-series	arima_search	MINT	323.69	0.71	0.59

is observed to perform better, across all hierarchical levels and forecast horizons. However, as in the case with no reconciliation, for the longest horizon (h=4) of the forecasts, ETC models are observed to have the best performance. In addition, *MinT* (Min Trace) approach leads to lower error metrics for across all hierarchical level for the first forecast horizon (h=1).

Table 5.13: Model Performance Across Hierarchical Level (Reconciliation Forecast)

LEVEL	h	MODEL	Hier-Method	RMSE	MASE	RMSSE
Category	1	arima_search_all	MINT	264.55	0.18	0.10
Brands	1	arima_search	MINT	128.24	0.34	0.26
Upc	1	arima_search_all	MINT	85.12	0.43	0.32
All_time-series	1	arima_search_all	MINT	110.19	0.41	0.31
Category	2	arima_search	-	547.28	0.30	0.21
Brands	2	arima_search	TD (ATH)	216.43	0.50	0.38
Upc	2	arima_neighb	MID OUT (ATH)	183.41	1.05	0.76
All_time-series	2	arima_search	TD (ATH)	213.51	0.89	0.65
Category	3	arima_search_all	MID OUT (ATH)	692.16	0.38	0.26
Brands	3	Snaive	BU	374.20	0.63	0.44
Upc	3	arima_search_all	MID OUT (ATH)	179.58	0.81	0.59
All_time-series	3	arima_search_all	MID OUT (ATH)	252.43	0.75	0.54
Category	4	ETS_search	-	650.76	0.36	0.25
Brands	4	ETS_search	TD (ATH)	138.69	0.24	0.18
Upc	4	ETS_search	TD (ATH)	70.18	0.38	0.29
All_time-series	4	ETS_search	TD (ATH)	114.14	0.35	0.26

According to Tables (5.12) and (5.13), it is reasonable to infer that the coherent forecasts exhibit smaller errors across the hierarchical levels and in the forecast horizon.

https://doi.org/10.26219/heal.aueb.6516

It should be noted that the above tables present the best model for each hierarchical level, whether or not reconciliation methods are used.

5.3 Dynamic Regression Models (Mid Level - Brands)

The dataset contains observations for product sales across 77 stores for each Manufacturer (Brand). A dynamic regression modeling approach is applied where 3 binary variables and Fourier term are introduced in the regression models at mid-level, based on the following logic:

$$\mathbf{X}_{B,i,t} = \begin{cases} 1, & \text{at least one product (out of three UPC) from Brand B at store } i\text{, at time } t \\ 0, & \text{otherwise} \end{cases}$$

The binary variables are specified in **Dunnhumby** data set and refer to product promotions/circulation/price reductions in each store. In specific, FEATURE binary variable is added to capture the effect of whether the product was in-store circular, DISPLAY binary variable is added to capture the effect of whether the product was a part of in-store promotional display and TPR_ONLY binary variable is added to capture the effect of whether the product had a temporary price reduction only (i.e. shelf tag only, not on display or in advertisement).

Given the above, the values of these binary variables are specified at a low level based on whether at least one store was observed to showcase one of the above promotions. Table (5.14) below summarizes the number of stores included in the analysis across mid-level.

In addition to the 3 binary variables mentioned above, the dynamic regression series for mid-level include Fourier terms. In specific, these terms are used in order to capture the seasonal patterns in the series. It is common, for a harmonic regression approach to be used where the seasonal pattern is modeled using Fourier terms with short-term time series dynamics handled by an ARMA error. The Fourier terms (sine and cosine pairs) are specified below:

https://doi.org/10.26219/heal.aueb.6516

$$\begin{aligned}
x_{1,t} &= \sin\left(\tfrac{2\pi t}{m}\right), x_{2,t} = \cos\left(\tfrac{2\pi t}{m}\right), \\
x_{3,t} &= \sin\left(\tfrac{4\pi t}{m}\right), x_{4,t} = \cos\left(\tfrac{4\pi t}{m}\right), \\
x_{5,t} &= \sin\left(\tfrac{6\pi t}{m}\right), x_{6,t} = \cos\left(\tfrac{6\pi t}{m}\right), \\
x_{7,t} &= \sin\left(\tfrac{8\pi t}{m}\right), x_{8,t} = \cos\left(\tfrac{8\pi t}{m}\right), \ldots.
\end{aligned}$$

It is noted that $k = m/2$ number of sine and cosine pairs are investigated in Fourier terms, as the seasonality of this study is 52 weeks ($m = 52$) and the literature suggests using this ratio.

Due to the nature of the time series, not all stores are included in this analysis. In specific, stores with at least one-time observation missing for a product are excluded because the statistical package used for the analysis cannot be applied when time points are missing in the time series. An alternative would be to fill in the missing data, however, this approach has not been followed in the current study and is out of scope.

As evident from Table (5.14) below, a large percentage of time series is included in the dynamic regression analysis. However, the percentage of time series included in the analysis is slightly lower for MKSL S brand but, the number of total stores for this brand is low overall and also in comparison with other brands.

Table 5.14: Summary of Time Series Included in Analysis for each Brand

BRANDS	TOTAL STORES	STORES INCLUDED	(%) OF INCLUDED STORES
Private label	77	76	98.7
Frito Lay	77	55	71.4
Shultz	36	30	83.3
Snyder S	77	58	75.3
Mksl S	35	19	54.3

In order to assess the performance of dynamic regression models at mid-level, the two approaches followed are based on a) AICc criteria (as the models can be compared) and b) Performance metrics of the models (rmse,mase,rmsse). It can be seen from Table (5.15) below that for approximately 10% (26) of the models both approaches are observed to lead to the same model choice. It is also observed

https://doi.org/10.26219/heal.aueb.6516

that in general, more than half of the models are specified with $K = 1$, meaning only 1 pair of cosine and sine is used in order to capture seasonality in the time series.

Table 5.15: Dynamic Regression Model Choice based on AIC & Performance Metrics

MANUFACTURER	STORE_NUM	MODEL	RMSE	MASE	RMSSE
Private Label	367	regr - K = 3	9.60	0.32	0.27
Private Label	15765	regr - K = 15	9.53	0.24	0.26
Private Label	25001	regr - K = 1	7.68	0.24	0.19
Private Label	26981	regr - K = 1	21.39	0.47	0.35
Frito Lay	2277	regr - K = 1	11.99	0.42	0.33
Frito Lay	6431	regr - K = 1	1.94	0.27	0.25
Frito Lay	15541	regr - K = 1	6.21	0.43	0.27
Frito Lay	21213	regr - K = 1	3.09	0.24	0.18
Frito Lay	21221	regr - K = 1	4.63	0.39	0.25
Frito Lay	21237	regr - K = 1	10.52	0.58	0.44
Frito Lay	23061	regr - K = 2	5.70	0.33	0.23
Frito Lay	25027	regr - K = 1	8.49	0.23	0.18
Frito Lay	25229	regr - K = 7	9.64	0.40	0.31
Frito Lay	26981	regr - K = 1	2.13	0.13	0.08
Shultz	8041	regr - K = 8	3.69	0.24	0.21
Shultz	23061	regr - K = 2	13.85	0.89	0.80
Shultz	25027	regr - K = 9	33.44	0.83	0.74
Shultz	26981	regr - K = 1	2.91	0.25	0.25
Snyder S	11757	regr - K = 2	3.84	0.31	0.25
Snyder S	15763	regr - K = 1	4.57	0.33	0.24
Snyder S	15765	regr - K = 1	9.87	0.76	0.68
Snyder S	23345	regr - K = 1	12.03	1.10	0.90
Snyder S	24991	regr - K = 1	31.84	1.06	0.82
Snyder S	25229	regr - K = 10	26.98	1.03	1.01
MKSL	21237	regr - K = 2	7.53	0.65	0.54

The assessment of model performance is made following the one-step approach, and the results for mid-level models are presented in Table (5.16) below. As evident, the model choices for all brands on the first and second forecasting horizons include a higher number of Fourier terms in comparison with the latter (2) forecasting horizons, where a lower number of Fourier terms are chosen. In addition, as stated in Appendix B, (B.5), the highest number of units sold is associated with private label which is also reflected in the error metrics calculated for the model forecasts.

Table 5.16: Mid Level - Dynamic Regression Model Performance For Each Forecasting Horizon

MANUFACTURER	h	MODEL	MEAN_RMSE	MEAN_MASE	MEAN_RMSSE
Private Label	1	regr - K = 20	22.48	0.74	0.53
Private Label	2	regr - K = 17	25.63	0.81	0.56
Private Label	3	regr - K = 3	28.58	0.87	0.62
Private Label	4	regr - K = 2	20.56	0.65	0.46
Frito Lay	1	regr - K = 4	7.54	0.49	0.34
Frito Lay	2	regr - K = 12	7.83	0.57	0.39
Frito Lay	3	regr - K = 3	6.04	0.41	0.29
Frito Lay	4	regr - K = 2	6.00	0.48	0.32
Shultz	1	regr - K = 7	10.45	0.54	0.44
Shultz	2	regr - K = 14	12.33	0.61	0.49
Shultz	3	regr - K = 7	15.66	0.86	0.69
Shultz	4	regr - K = 2	11.74	0.66	0.54
Snyder S	1	regr - K = 11	9.82	0.73	0.54
Snyder S	2	regr - K = 12	9.56	0.61	0.44
Snyder S	3	regr - K = 9	9.68	0.62	0.46
Snyder S	4	regr - K = 1	9.34	0.60	0.44
Mksl	1	regr - K = 14	6.36	0.65	0.51
Mksl	2	regr - K = 25	7.53	0.71	0.55
Mksl	3	regr - K = 10	6.19	0.67	0.54
Mksl	4	regr - K = 4	8.59	0.86	0.68

Table (5.17) presents the model choice for mid-level (brands) across all forecasting horizons. It is observed that in general, the error metrics are larger for Private label and Shultz, in line with table (5.16) above. In addition, it can be seen that a

https://doi.org/10.26219/heal.aueb.6516

larger number of Fourier terms are included in the model choice when all forecasting horizons are in the scope of the performance assessment.

Table 5.17: Mid Level - Dynamic Regression Model Performance Across All Forecasting Horizons

MANUFACTURER	MODEL	MEAN_RMSE	MEAN_MASE	MEAN_RMSSE
Private Label	regr - K = 11	32.82	0.83	0.68
Frito Lay	regr - K = 7	9.05	0.53	0.43
Shultz	regr - K = 12	15.55	0.70	0.66
Snyder S	regr - K = 9	11.99	0.66	0.58
Mksl	regr - K = 22	8.62	0.77	0.70

The analysis also includes a comparison of the same methodology applied to the family of models discussed and presented in section 5.2. More importantly, Table (5.18) presents the model choice for mid-level (brands) across all forecasting horizons when alternative (other) models are considered.

It is observed that similar to the Dynamic regression case, Private label and Schultz brands exhibit the largest errors in the forecast metrics, whilst the errors also exhibit a decreasing trend across all Brands for the latter horizons. Extending the forecasting horizon can provide further evidence on this however, exceeds the scope of this analysis. Furthermore, as shown in (5.10), the ETS_search model is consistently chosen across all Brands as the best in terms of forecasting, either when aggregating in terms of stores or without aggregation.

Table (5.19) presents the model choice for mid-level (brands) across all forecasting horizons. It is observed that in general the error metrics are larger for Private label and Schultz, in line with table (5.16) above.

It should be noted that Tables (5.18) and (5.19) present the best model for each hierarchical level and for each forecast horizon.

https://doi.org/10.26219/heal.aueb.6516

Table 5.18: Mid Level - Other Models Performance For Each Forecasting Horizon

MANUFACTURER	h	MODEL	MEAN_RMSE	MEAN_MASE	MEAN_RMSSE
Private Label	1	Naive	20.13	0.64	0.45
Private Label	2	Snaive	24.43	0.76	0.52
Private Label	3	Naive	26.33	0.75	0.53
Private Label	4	ETS_search	16.97	0.56	0.40
Frito Lay	1	Naive	6.23	0.48	0.33
Frito Lay	2	Naive	10.11	0.72	0.51
Frito Lay	3	arima_stepwise	6.01	0.44	0.30
Frito Lay	4	ETS_search	5.35	0.42	0.29
Shultz	1	regr - 1	14.81	0.75	0.60
Shultz	2	Snaive	21.73	1.06	0.85
Shultz	3	Naive	19.53	1.07	0.86
Shultz	4	arima_search_all	10.45	0.62	0.50
Snyder S	1	arima_stepwise	10.02	0.73	0.55
Snyder S	2	Drift	10.11	0.79	0.59
Snyder S	3	Mean	8.72	0.63	0.47
Snyder S	4	Mean	8.42	0.55	0.41
Mksl	1	Drift	7.08	0.72	0.56
Mksl	2	Drift	8.86	0.89	0.71
Mksl	3	arima_stepwise	6.64	0.73	0.59
Mksl	4	ETS_search	7.57	0.85	0.68

Table 5.19: Mid Level - Other Models Performance Across All Forecasting Horizons

MANUFACTURER	MODEL	MEAN_RMSE	MEAN_MASE	MEAN_RMSSE
Private Label	Naive	31.54	0.84	0.69
Frito Lay	ETS_search	8.81	0.52	0.44
Shultz	Naive	21.55	0.97	0.90
Snyder S	arima_search_all	11.59	0.69	0.60
Mksl	Drift	9.37	0.84	0.77

https://doi.org/10.26219/heal.aueb.6516

Chapter 6

Discussion on Future Research

This section aims to suggest potential avenues for future research that could contribute to the ongoing effort of improving forecasting accuracy and reliability with models utilizing hierarchical data. More specifically, the following approaches could be investigated in future research, including utilizing a geographical hierarchy, filling missing observations in product time series data, using forecast combinations, and exploring machine learning methods and Neural Networks.

- In future research, it may be beneficial to consider using a geographical hierarchy instead of a product hierarchy when making forecasts at the highest hierarchical level. This approach could potentially offer more accurate predictions for broader regions or entire countries, as it takes into account the unique economic, social, and cultural factors that can influence consumer behavior within each region. By focusing on the geographic context rather than specific product categories, it may also allow for more flexibility and adaptability in forecasting models as consumer preferences and market trends shift over time. Further research in this area could explore the feasibility and effectiveness of using a geographical hierarchy in forecasting models and compare its performance to existing methods based on product hierarchies.

- Another interesting approach to improve prediction accuracy would be to fill the missing observations in product time series data, which could then be compared to strict hierarchical models. By utilizing more complete data and time series, it may be possible to gain a more comprehensive understanding of product trends and consumer behavior, leading to better predictions. This approach could also help identify potential biases or inconsistencies in hierarchical models and provide insights into how to further refine or adapt these models. Future research could explore the feasibility and potential benefits

https://doi.org/10.26219/heal.aueb.6516

of filling missing observations in product time series data, and investigate the impact of incorporating this data into forecasting models.

- For future research, one possible approach to improving model forecasting is to utilize forecast combinations, such as combining ARIMA and ETS models. This approach may offer several advantages, such as leveraging the strengths of different models and reducing the impact of individual model weaknesses or biases. By combining forecasts from multiple models, it may also be possible to achieve more accurate and robust predictions, especially in scenarios where individual models may struggle to capture complex or non-linear relationships. Future research could explore the feasibility and effectiveness of using forecast combinations in various forecasting contexts, and investigate the impact of different combination techniques on prediction accuracy and reliability.
- For future research, machine learning methods and Neural Networks could be explored as potential approaches to improving forecasting models. These methods may offer advantages such as the ability to capture complex relationships and patterns in large and diverse datasets, as well as adaptability to changing environments and the potential for automated model selection and optimization. Future research could investigate the feasibility and effectiveness of using machine learning methods in forecasting, and compare their performance to traditional statistical models in various contexts. Additionally, research could explore the potential benefits and limitations of different machine learning methods, as well as the impact of different training data and feature selection techniques on model accuracy and generalizability.

https://doi.org/10.26219/heal.aueb.6516

Appendix A

Tables

https://doi.org/10.26219/heal.aueb.6516

Table A.1: Weekly sales per Brand ($)

BRAND	SALES(Average)	SALES(SD)	REVENUE(Average)	REVENUE(SD)
PRIVATE LABEL_Sn	9,328	1,766	12,657	2,015
FRITO LAY	2,316	832	6,498	1,506
SHULTZ	2,048	298	3,590	524
SNYDER S	2,430	629	6,364	1,399
MKSL	554	197	1,224	342
PRIVATE LABEL_Cereal	7,047	1,249	13,461	2,301
GENERAL MI	10,829	5,513	31,195	10,439
QUAKER	4,864	4,905	11,139	8,960
KELLOGG	9,109	6,398	24,687	12,638
POST FOODS	5,965	2,728	17,034	5,991
PRIVATE LABEL_Pizza	2,997	1,553	10,335	4,534
KING	742	415	4,328	1,914
TOMBSTONE	3,790	2,635	20,886	11,276
TONYS	1,358	1,505	7,467	6,314
PRIVATE LABEL_Hygiene	1,184	538	1,907	469
COLGATE	17	11	128	82
P & G	824	244	3,250	731
CHATTEM	318	68	1,463	269
WARNER	1,004	328	4,368	1,180

Table A.2: Weekly sales per Category

CATEGORY	BAG SNACKS	HYGIENE PROD.	COLD CEREAL	FR. PIZZA
Frequency	127462	116529	169689	111270
Number_of_weeks	156	156	156	156
Average_weekly_of_store	77	77	77	77
Average_weekly_of_cities	51	51	51	51
Total_units_sold	2601459	521219	5827382	1343294
Average_weekly_units_sold	16676	3341	37355	8611
Median_weekly_sales	16199	3206	35885	8286
SD_weekly_sales	2442	711	9312	3085
Max_weekly_sales	25857	6304	68999	21422
Min_weekly_sales	9032	1303	13995	3236
Total_revenue	4731983	1727831	15008351	6459558
Average_weekly_revenue	30333	11076	96207	41407
Median_weekly_revenue	30161	11083	94171	40626
SD_weekly_revenue	3838	1671	18984	12520
Max_weekly_revenue	43060	15835	153215	82337
Min_weekly_revenue	16376	4348	38602	16240

https://doi.org/10.26219/heal.aueb.6516

Table A.3: Weekly sales per State

State	Kentucky	Ohio	Texas	Indiana
Number_of_weeks	156	156	156	156
Total_sales	700,000	5,000,000	4,000,000	200,000
Average_weekly_sales	4,367	34,342	26,192	1,082
Median_weekly_sales	4,178	33,700	25,600	1,078
SD_weekly_sales	919	6,150	4,723	210
Max_weekly_sales	8,514	60,044	45,344	2,009
Min_weekly_sales	1,672	12,973	14,721	398
Total_revenue	1,761,697	14,352,266	11,355,084	458,676
Average_weekly_revenue	11,293	92,002	72,789	2,940
Median_weekly_revenue	11,146	91,387	71,943	2,910
SD_weekly_revenue	1,993	14,880	10,408	537
Max_weekly_revenue	18,371	148,252	108,954	4,996
Min_weekly_revenue	4,386	35,188	43,168	1,074

Table A.4: Maping ID with product

ID	PRODUCTS	ID	PRODUCTS
P1	PL MINI TWIST PRETZELS	P30	POST FRUITY PEBBLES
P2	PL PRETZEL STICKS	P31	PL SR CRUST SUPRM PIZZA
P3	PL TWIST PRETZELS	P32	PL SR CRUST 3 MEAT PIZZA
P4	RLDGLD BRAIDED HONEY WHT	P33	PL SR CRUST PEPPRN PIZZA
P5	RLDGLD TINY TWISTS PRTZL	P34	NWMN OWN PEPPERONI PIZZA
P6	RLDGLD PRETZEL STICKS	P35	NWMN OWN 4 CHEESE PIZZA
P7	SHURGD PRETZEL RODS	P36	NWMN OWN SUPREME PIZZA
P8	SHURGD MINI PRETZELS	P37	HMRN CLSC SAUSAGE PIZZA
P9	SHURGD PRETZEL STICKS	P38	HMRN CLSC SSG PEPP PIZZA
P10	MKSL MINI TWIST PRETZELS	P39	HMRN CLSC CHS PIZZA
P11	MKSL DUTCH PRETZELS	P40	DIGIORNO THREE MEAT
P12	MKSL PRETZEL STICKS	P41	DIGRN SUPREME PIZZA
P13	SNYDR PRETZEL RODS	P42	DIGRN PEPP PIZZA
P14	SNYDR SOURDOUGH NIBBLERS	P43	FRSC BRCK OVN ITL PEP PZ
P15	SNYDR FF MINI PRETZELS	P44	FRSC PEPPERONI PIZZA
P16	PL HONEY NUT TOASTD OATS	P45	FRSC 4 CHEESE PIZZA
P17	PL RAISIN BRAN	P46	PL BL MINT ANTSPTC RINSE
P18	PL BT SZ FRSTD SHRD WHT	P47	PL BL MINT ANTSPTC RINSE 2
P19	GM HONEY NUT CHEERIOS	P48	PL ANTSPTC SPG MNT MTHWS
P20	GM CHEERIOS	P49	COLG SPEARMINT WISP
P21	GM CHEERIOS 2	P50	CREST PH WHTG CLN MINT TP
P22	QKER LIFE ORIGINAL	P51	SCOPE ORIG MINT MOUTHWASH
P23	QKER CAP N CRUNCH BERRIES	P52	CREST PH CLN MINT RINSE
P24	QKER CAP N CRUNCH	P53	ACT MINT A/CAV FLUOR RNS
P25	KELL BITE SIZE MINI WHEAT	P54	ACT KIDS BBLGUM FLUOR RNS
P26	KELL FROSTED FLAKES	P55	ACT RSTR CL SPLSH MINT MW
P27	KELL FROOT LOOPS	P56	LSTRNE CL MINT ANTSPTC MW
P28	POST HNY BN OTS HNY RSTD	P57	LSTRNE CL MINT ANTSPTC MW 2
P29	POST FM SZ HNYBNCH OT ALM	P58	LSTRNE FRS BRST ANTSPC MW

https://doi.org/10.26219/heal.aueb.6516

Table A.5: Models Parameters chosen across low hierarchical level (part I)

UPC	ARIMA_STEPWISE	ARIMA_SEARCH
2840002333	ARIMA(1,1,1) w/ drift	ARIMA(1,1,1) w/ drift
2840004768	ARIMA(0,0,3) w/ mean	ARIMA(0,0,3) w/ mean
2840004770	ARIMA(0,0,3) w/ mean	ARIMA(0,0,3) w/ mean
7110410455	ARIMA(1,1,1)(0,1,0)[52]	ARIMA(4,1,0)(0,1,0)[52]
7110410470	ARIMA(1,1,1)	ARIMA(0,1,4)
7110410471	ARIMA(0,1,1)(0,1,0)[52]	ARIMA(0,1,4)(0,1,0)[52]
1111009477	ARIMA(1,1,1)(0,1,0)[52]	ARIMA(1,1,1)(0,1,0)[52]
1111009497	ARIMA(0,1,1)	ARIMA(0,1,1)
1111009507	ARIMA(3,1,0)	ARIMA(3,1,0)
7027312504	ARIMA(2,1,1)	ARIMA(2,1,1)
7027316204	ARIMA(4,1,2)	ARIMA(4,1,2)
7027316404	ARIMA(2,1,1)(0,1,0)[52]	ARIMA(5,1,1)(0,1,0)[52]
7797502248	ARIMA(1,0,0) w/ mean	ARIMA(0,0,2) w/ mean
7797508004	ARIMA(0,0,2) w/ mean	ARIMA(5,0,0) w/ mean
7797508006	ARIMA(0,1,0)	ARIMA(0,1,3)

Table A.6: Models Parameters chosen across low hierarchical level (part II)

UPC	ARIMA_NEIGHB	ARIMA_SEARCH_APR	ETS_SEARCH
2840002333	ARIMA(1,1,1) w/ drift	ARIMA(1,1,1) w/ drift	ETS(M,N,N)
2840004768	ARIMA(0,0,3) w/ mean	ARIMA(0,0,3) w/ mean	ETS(M,N,N)
2840004770	ARIMA(0,0,3) w/ mean	ARIMA(0,0,3) w/ mean	ETS(M,N,N)
7110410455	ARIMA(1,1,1)(0,1,0)[52]	ARIMA(1,1,1)(0,1,0)[52]	ETS(A,N,N)
7110410470	ARIMA(1,1,1)	ARIMA(1,1,1)	ETS(A,N,N)
7110410471	ARIMA(0,1,1)(0,1,0)[52]	ARIMA(0,1,4)(0,1,0)[52]	ETS(A,N,N)
1111009477	ARIMA(1,1,1)(0,1,0)[52]	ARIMA(0,1,2)(0,1,0)[52]	ETS(M,N,N)
1111009497	ARIMA(0,1,1)	ARIMA(0,1,1)	ETS(M,N,N)
1111009507	ARIMA(3,1,0)	ARIMA(0,1,2)	ETS(M,N,N)
7027312504	ARIMA(2,1,1)	ARIMA(0,1,2)	ETS(A,N,N)
7027316204	ARIMA(4,1,2)	ARIMA(1,1,1)	ETS(A,N,N)
7027316404	ARIMA(2,1,1)(0,1,0)[52]	ARIMA(5,1,1)(0,1,0)[52]	ETS(A,N,N)
7797502248	ARIMA(1,0,0) w/ mean	ARIMA(0,0,2) w/ mean	ETS(M,A,N)
7797508004	ARIMA(0,0,2) w/ mean	ARIMA(0,0,2) w/ mean	ETS(A,N,N)
7797508006	ARIMA(0,1,0)	ARIMA(2,1,1)	ETS(A,N,N)

https://doi.org/10.26219/heal.aueb.6516

Table A.7: Models Parameters chosen across middle hierarchical level (part I)

MANUFACTURER	ARIMA_STEPWISE	ARIMA_SEARCH
FRITO LAY	ARIMA(0,0,3) w/ mean¿	ARIMA(0,0,3) w/ mean¿
MKSL	ARIMA(1,1,1)(0,1,0)[52]	ARIMA(4,1,0)(0,1,0)[52]
PRIVATE LABEL	ARIMA(2,0,1)(0,1,0)[52] w/ drift	ARIMA(2,0,1)(0,1,0)[52]
SHULTZ	ARIMA(1,1,1)(0,1,0)[52]	ARIMA(5,1,1)(0,1,0)[52]
SNYDER S	ARIMA(1,0,0) w/ mean	ARIMA(1,0,0) w/ mean
TOP_LEVEL	ARIMA(1,1,1)(0,1,0)[52]	ARIMA(4,1,2)(0,1,0)[52]

Table A.8: Models Parameters chosen across middle hierarchical level (part II)

MANUFACTURER	ARIMA_NEIGHB	ARIMA_SEARCH_APR	ETS_SEARCH
FRITO LAY	ARIMA(0,0,3) w/ mean¿	ARIMA(0,0,3) w/ mean¿	ETS(M,N,N)
MKSL	ARIMA(1,1,1)(0,1,0)[52]	ARIMA(0,1,4)(0,1,0)[52]	ETS(A,N,N)
PRIVATE LABEL	ARIMA(2,0,1)(0,1,0)[52] w/ drift	ARIMA(1,0,1)(0,1,0)[52]	ETS(M,N,N)
SHULTZ	ARIMA(1,1,1)(0,1,0)[52]	ARIMA(1,1,1)(0,1,0)[52]	ETS(A,N,N)
SNYDER S	ARIMA(1,0,0) w/ mean	ARIMA(1,0,0) w/ mean	ETS(A,N,N)
TOP_LEVEL	ARIMA(1,1,1)(0,1,0)[52]	ARIMA(1,1,3)(0,1,0)[52]	ETS(M,N,N)

Table A.9: Models Parameters chosen across the top hierarchical level (part I)

CATEGORY	ARIMA_STEPWISE	ARIMA_SEARCH
TOP_LEVEL	ARIMA(1,1,1)(0,1,0)[52]	ARIMA(4,1,2)(0,1,0)[52]

Table A.10: Models Parameters chosen across middle hierarchical level (part II)

CATEGORY	ARIMA_NEIGHB	ARIMA_SEARCH_APR	ETS_SEARCH
TOP_LEVEL	ARIMA(1,1,1)(0,1,0)[52]	ARIMA(1,1,3)(0,1,0)[52]	ETS(M,N,N)

 https://doi.org/10.26219/heal.aueb.6516

Table A.11: Model Performance Evaluation across hierarchical levels (on train data)

MODEL	RMSE	MASE	RMSSE	DATA	LEVEL
ARIMA_search	1757.62	0.56	0.67	train	Category
ARIMA_search_all	1772.63	0.57	0.68	train	Category
ARIMA_search_apr	1772.63	0.57	0.68	train	Category
ARIMA_neighb	1859.98	0.57	0.71	train	Category
ARIMA_stepwise	1859.98	0.57	0.71	train	Category
ETS_search	1951.91	0.72	0.75	train	Category
Drift	2102.02	0.75	0.80	train	Category
Naive	2102.02	0.75	0.80	train	Category
Mean	2404.97	0.94	0.92	train	Category
Snaive	2617.63	1.00	1.00	train	Category
ARIMA_search	539.10	0.44	0.53	train	Brands
ARIMA_neighb	540.85	0.44	0.53	train	Brands
ARIMA_stepwise	540.85	0.44	0.53	train	Brands
ARIMA_search_all	541.06	0.44	0.53	train	Brands
ARIMA_search_apr	541.06	0.44	0.53	train	Brands
ETS_search	573.77	0.53	0.60	train	Brands
Drift	600.56	0.53	0.61	train	Brands
Naive	600.58	0.53	0.61	train	Brands
Mean	736.07	0.79	0.81	train	Brands
Snaive	951.33	1.00	1.00	train	Brands
ARIMA_search	203.25	0.47	0.53	train	Upc
ARIMA_search_all	203.41	0.47	0.54	train	Upc
ARIMA_search_apr	203.41	0.47	0.54	train	Upc
ARIMA_neighb	206.27	0.48	0.54	train	Upc
ARIMA_stepwise	206.27	0.48	0.54	train	Upc
ETS_search	222.78	0.53	0.60	train	Upc
Drift	230.27	0.52	0.62	train	Upc
Naive	230.28	0.52	0.62	train	Upc
Mean	288.79	0.81	0.81	train	Upc
Snaive	373.78	1.00	1.00	train	Upc
ARIMA_search	357.23	0.47	0.54	train	All_time-series
ARIMA_search_all	358.53	0.47	0.54	train	All_time-series
ARIMA_search_apr	358.53	0.47	0.54	train	All_time-series
ARIMA_neighb	364.68	0.47	0.55	train	All_time-series
ARIMA_stepwise	364.68	0.47	0.55	train	All_time-series
ETS_search	388.69	0.54	0.61	train	All_time-series
Drift	407.57	0.53	0.62	train	All_time-series
Naive	407.58	0.53	0.62	train	All_time-series
Mean	496.06	0.81	0.82	train	All_time-series
Snaive	618.14	1.00	1.00	train	All_time-series

https://doi.org/10.26219/heal.aueb.6516

Table A.12: Model Perfomance Evaluation across hierarchical levels (on test data)

MODEL	RMSE	MASE	RMSSE	DATA	LEVEL
arima_search	1669.30	0.86	0.64	test	Category
arima_search_all	1698.86	0.89	0.65	test	Category
arima_search_apr	1698.86	0.89	0.65	test	Category
arima_neighb	1701.29	0.90	0.65	test	Category
arima_stepwise	1701.29	0.90	0.65	test	Category
Snaive	2303.63	1.08	0.88	test	Category
Naive	2866.35	1.44	1.10	test	Category
Drift	2868.43	1.44	1.10	test	Category
Mean	3390.73	1.67	1.30	test	Category
ETS_search	3862.79	1.85	1.48	test	Category
arima_search	451.84	0.63	0.51	test	Brands
arima_neighb	457.39	0.67	0.55	test	Brands
arima_stepwise	457.39	0.67	0.55	test	Brands
arima_search_all	458.01	0.66	0.54	test	Brands
arima_search_apr	458.01	0.66	0.54	test	Brands
Snaive	571.85	0.79	0.58	test	Brands
Naive	633.70	0.90	0.74	test	Brands
Drift	635.32	0.90	0.73	test	Brands
ETS_search	755.55	0.99	0.82	test	Brands
Mean	867.86	1.60	1.29	test	Brands
arima_search	220.15	0.84	0.69	test	Upc
arima_search_all	224.39	0.86	0.71	test	Upc
arima_search_apr	224.39	0.86	0.71	test	Upc
arima_neighb	225.21	0.85	0.70	test	Upc
arima_stepwise	225.21	0.85	0.70	test	Upc
Naive	247.14	0.94	0.76	test	Upc
Drift	247.24	0.94	0.76	test	Upc
Snaive	247.81	0.80	0.66	test	Upc
ETS_search	292.30	1.02	0.83	test	Upc
Mean	349.57	1.58	1.27	test	Upc
arima_search	344.32	0.79	0.65	test	All_time-series
arima_search_all	350.23	0.82	0.66	test	All_time-series
arima_search_apr	350.23	0.82	0.66	test	All_time-series
arima_neighb	350.78	0.81	0.66	test	All_time-series
arima_stepwise	350.78	0.81	0.66	test	All_time-series
Snaive	422.86	0.81	0.66	test	All_time-series
Naive	463.90	0.95	0.77	test	All_time-series
Drift	464.46	0.95	0.77	test	All_time-series
ETS_search	572.62	1.05	0.86	test	All_time-series
Mean	617.79	1.59	1.28	test	All_time-series

https://doi.org/10.26219/heal.aueb.6516

Table A.13: RECONCILIATION TOP MODEL BY STEP FORECAST

LEVEL	MODEL	RMSE	MASE	RMSSE	DATA
Category	mint_arima_search	1596.69	0.81	0.61	test
Brands	arima_search	451.84	0.63	0.51	test
Upc	mint_arima_search	189.67	0.71	0.59	test
All_time-series	mint_arima_search	323.69	0.71	0.59	test

Table A.14: RECONCILIATION TOP MODEL BY LEVEL

LEVEL	MODEL	RMSE	MASE	RMSSE	DATA
Category	mint_arima_search	1596.6916	0.8101154	0.6099757	test
Brands	arima_search	451.8378	0.6273491	0.5135532	test
Upc	mint_arima_search	189.6689	0.7138086	0.5942371	test
All_time-series	mint_arima_search	323.6925	0.7112956	0.5926889	test

Table A.15: Forecast Measures in Train data

LEVEL	MODEL	RMSE	MASE	RMSSE	DATA
Category	Arima	1757.6193	0.5608310	0.6714541	train
Brands	Arima	539.0951	0.4400077	0.5275639	train
Upc	Arima	203.2479	0.4718660	0.5340280	train
All_time-series	Arima	357.2292	0.4685171	0.5390330	train

https://doi.org/10.26219/heal.aueb.6516

Table A.16: CV-BRANDS1

MODEL	MANUFACTURER	TYPE	RMSE	MASE	RMSSE
arima_neighb	SNYDER S	Test	576.56	0.61	0.64
arima_stepwise	SNYDER S	Test	576.56	0.61	0.64
arima_search_all	SNYDER S	Test	576.60	0.61	0.64
arima_search_apr	SNYDER S	Test	576.60	0.61	0.64
arima_search	SNYDER S	Test	577.12	0.61	0.64
ETS_search	SNYDER S	Test	607.01	0.65	0.68
Naive	SNYDER S	Test	625.15	0.58	0.70
Drift	SNYDER S	Test	639.76	0.60	0.71
Mean	SNYDER S	Test	643.22	0.71	0.72
Snaive	SNYDER S	Test	898.47	1.00	1.00
ETS_search	SHULTZ	Test	244.17	0.66	0.75
Naive	SHULTZ	Test	245.53	0.65	0.75
arima_search_all	SHULTZ	Test	246.70	0.70	0.76
arima_search_apr	SHULTZ	Test	246.70	0.70	0.76
arima_stepwise	SHULTZ	Test	247.06	0.70	0.76
arima_neighb	SHULTZ	Test	247.09	0.70	0.76
Drift	SHULTZ	Test	251.75	0.67	0.77
arima_search	SHULTZ	Test	264.67	0.73	0.81
Mean	SHULTZ	Test	303.67	0.89	0.93
Snaive	SHULTZ	Test	324.78	0.99	1.00
ETS_search	PRIVATE LABEL	Test	1350.87	0.70	0.67
Naive	PRIVATE LABEL	Test	1406.87	0.69	0.69
Drift	PRIVATE LABEL	Test	1425.29	0.70	0.70
arima_neighb	PRIVATE LABEL	Test	1534.53	0.78	0.76
arima_stepwise	PRIVATE LABEL	Test	1541.50	0.78	0.76
arima_search	PRIVATE LABEL	Test	1572.71	0.82	0.78
arima_search_all	PRIVATE LABEL	Test	1619.23	0.84	0.80
arima_search_apr	PRIVATE LABEL	Test	1619.23	0.84	0.80
Mean	PRIVATE LABEL	Test	1790.39	0.99	0.88
Snaive	PRIVATE LABEL	Test	2025.63	1.00	1.00

https://doi.org/10.26219/heal.aueb.6516

Table A.17: CV-BRANDS2

MODEL	MANUFACTURER	TYPE	RMSE	MASE	RMSSE
Naive	MKSL	Test	89.04	0.41	0.48
Drift	MKSL	Test	90.12	0.41	0.48
arima_search	MKSL	Test	94.47	0.47	0.51
ETS_search	MKSL	Test	95.79	0.44	0.51
arima_search_all	MKSL	Test	96.37	0.48	0.52
arima_search_apr	MKSL	Test	96.37	0.48	0.52
arima_neighb	MKSL	Test	97.98	0.48	0.53
arima_stepwise	MKSL	Test	97.98	0.48	0.53
Snaive	MKSL	Test	186.02	1.00	1.00
Mean	MKSL	Test	199.51	0.97	1.07
arima_search	FRITO LAY	Test	629.09	0.46	0.50
arima_stepwise	FRITO LAY	Test	631.09	0.46	0.50
arima_neighb	FRITO LAY	Test	632.24	0.46	0.50
arima_search_all	FRITO LAY	Test	634.46	0.47	0.50
arima_search_apr	FRITO LAY	Test	634.46	0.47	0.50
Naive	FRITO LAY	Test	639.79	0.38	0.51
Drift	FRITO LAY	Test	644.80	0.38	0.51
ETS_search	FRITO LAY	Test	670.22	0.41	0.53
Mean	FRITO LAY	Test	843.34	0.75	0.67
Snaive	FRITO LAY	Test	1263.66	1.00	1.00

Table A.18: BEST MODEL CV BRAND

MANUFACTURER	MODEL	TYPE	RMSE	MASE	RMSSE
SNYDER S	arima_neighb	Test	576.56	0.61	0.64
SHULTZ	ETS_search	Test	244.17	0.66	0.75
PRIVATE LABEL	ETS_search	Test	1350.87	0.70	0.67
MKSL	Naive	Test	89.04	0.41	0.48
FRITO LAY	arima_search	Test	629.09	0.46	0.50

https://doi.org/10.26219/heal.aueb.6516

Table A.19: CV-TOP

MODEL	TYPE	RMSE	MASE	RMSSE
Naive	Test	2129.62	0.77	0.82
Drift	Test	2156.34	0.78	0.83
ETS_search	Test	2162.05	0.82	0.83
arima_stepwise	Test	2168.84	0.81	0.83
arima_neighb	Test	2179.77	0.82	0.84
arima_search	Test	2189.15	0.85	0.84
arima_search_all	Test	2245.34	0.86	0.86
arima_search_apr	Test	2245.34	0.86	0.86
Mean	Test	2478.09	1.04	0.95
Snaive	Test	2606.25	1.01	1.00

Table A.20: BEST-UPC-CV

UPC	MANUFACTURER	MODEL	TYPE	RMSE	MASE	RMSSE
7797508006	SNYDER S	Naive	Test	264.47	0.49	0.62
7797508004	SNYDER S	Naive	Test	261.44	0.51	0.62
7797502248	SNYDER S	arima_search_all	Test	359.58	0.87	0.86
7027316204	SHULTZ	ETS_search	Test	97.24	0.61	0.67
7027312504	SHULTZ	Naive	Test	83.87	0.63	0.73
7027316404	SHULTZ	Naive	Test	93.83	0.83	0.87
1111009507	PRIVATE LABEL	arima_search_all	Test	271.21	0.33	0.44
1111009497	PRIVATE LABEL	ETS_search	Test	562.31	0.57	0.59
1111009477	PRIVATE LABEL	ETS_search	Test	622.94	0.71	0.71
7110410471	MKSL	Naive	Test	31.64	0.44	0.49
7110410455	MKSL	Naive	Test	44.20	0.46	0.54
7110410470	MKSL	ETS_search	Test	29.76	0.51	0.58
2840002333	FRITO LAY	arima_search_all	Test	89.82	0.37	0.46
2840004770	FRITO LAY	arima_search	Test	228.11	0.48	0.48
2840004768	FRITO LAY	arima_search	Test	319.41	0.50	0.52

https://doi.org/10.26219/heal.aueb.6516

Appendix B

Figures

Figure B.1: Time Series for the units sold per Category & Brands.

https://doi.org/10.26219/heal.aueb.6516

Figure B.2: Time Series for the units sold per Brand.

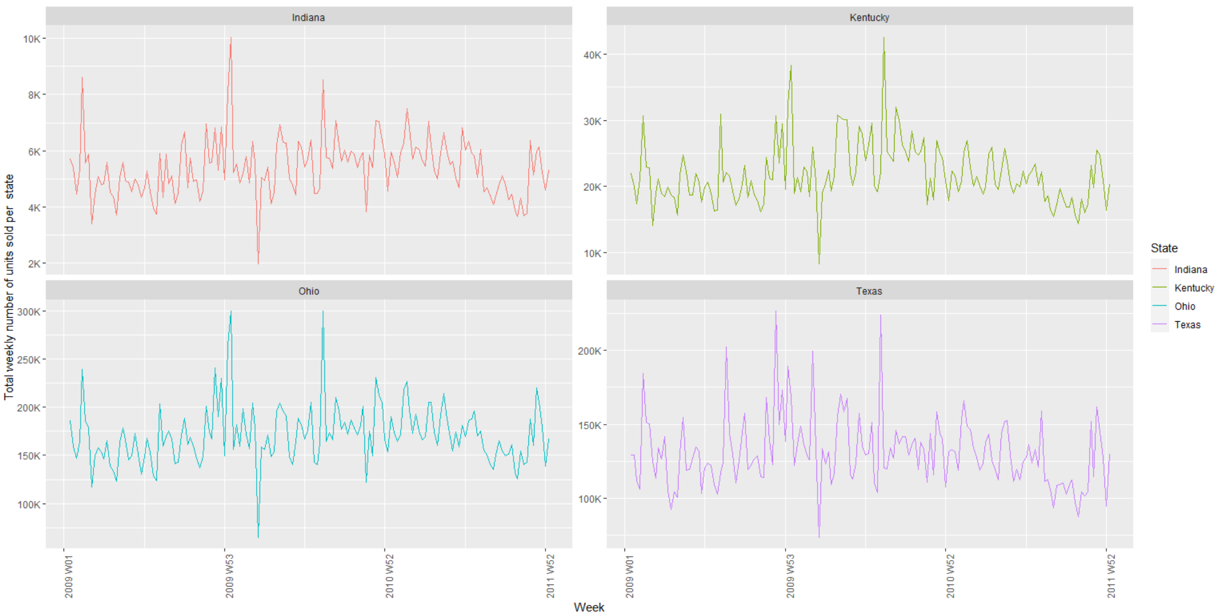

Figure B.3: Time Series for the units sold per State.

https://doi.org/10.26219/heal.aueb.6516

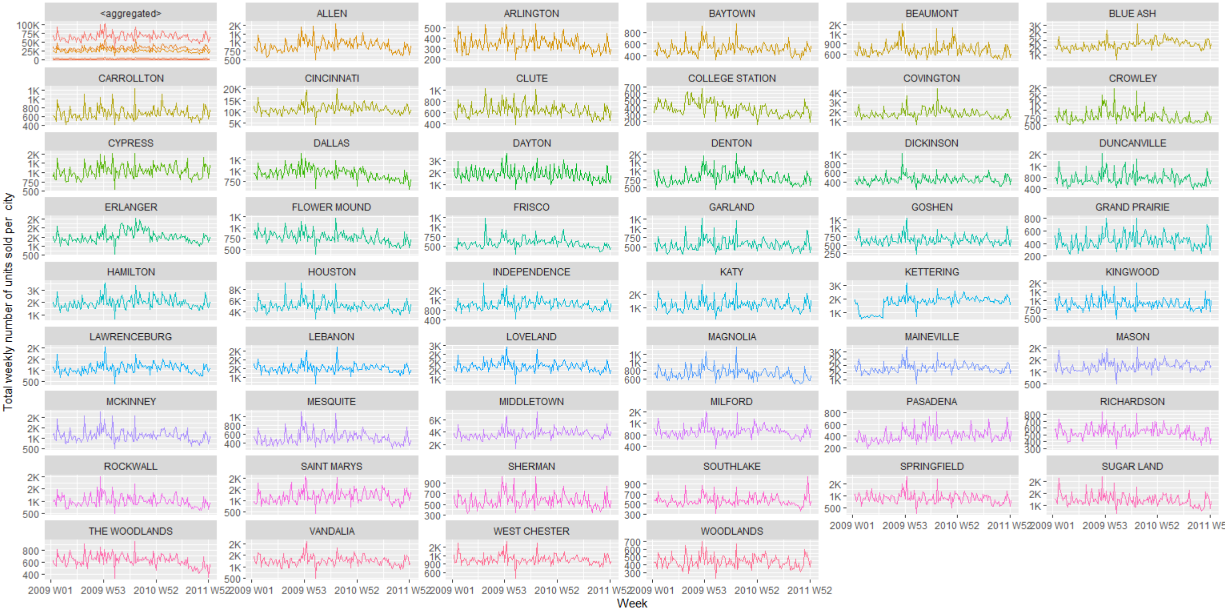

Figure B.4: Time Series for the units sold per City.

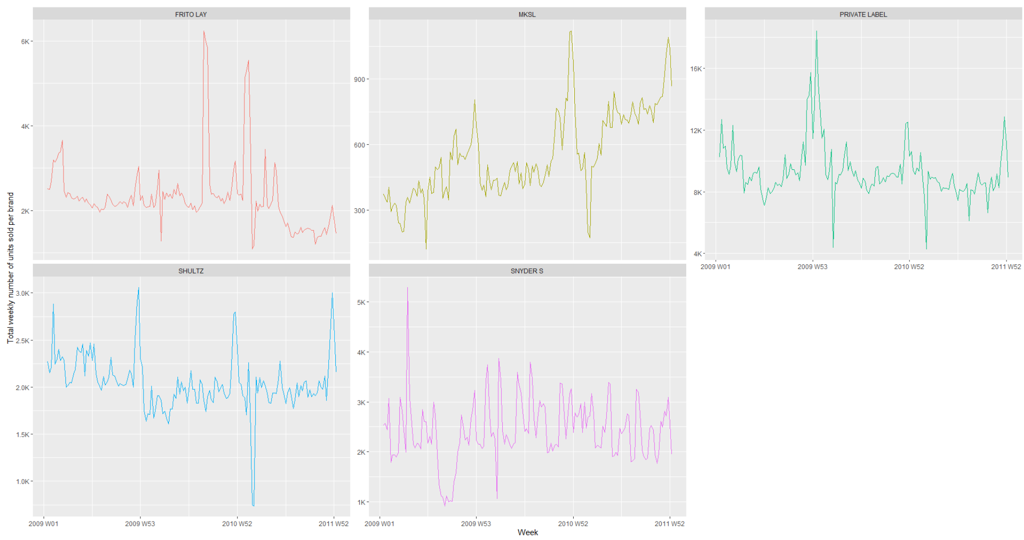

Figure B.5: Time Series for the units sold per Brand in Bag Snacks Category.

 https://doi.org/10.26219/heal.aueb.6516

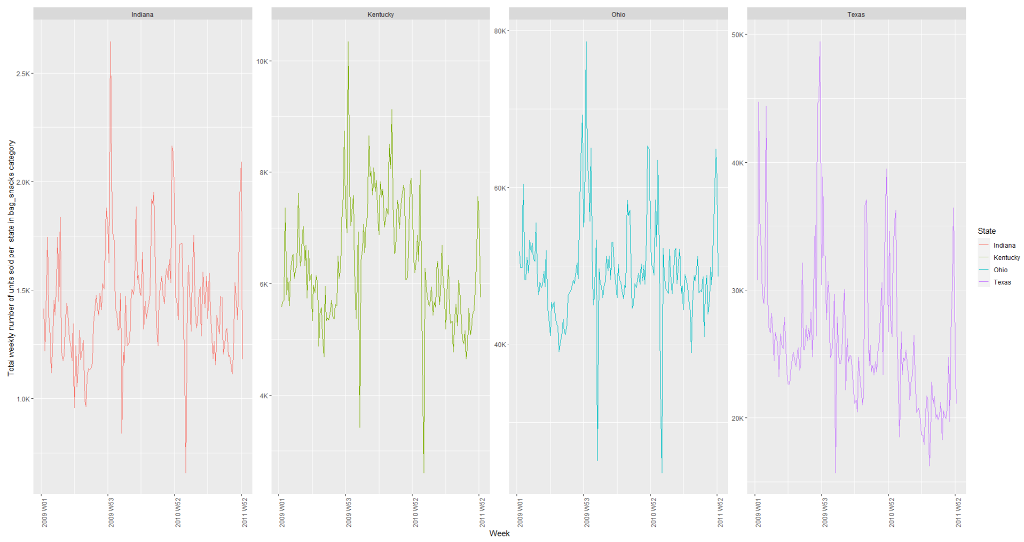

Figure B.6: Time Series for the units sold per State in Bag Snacks.

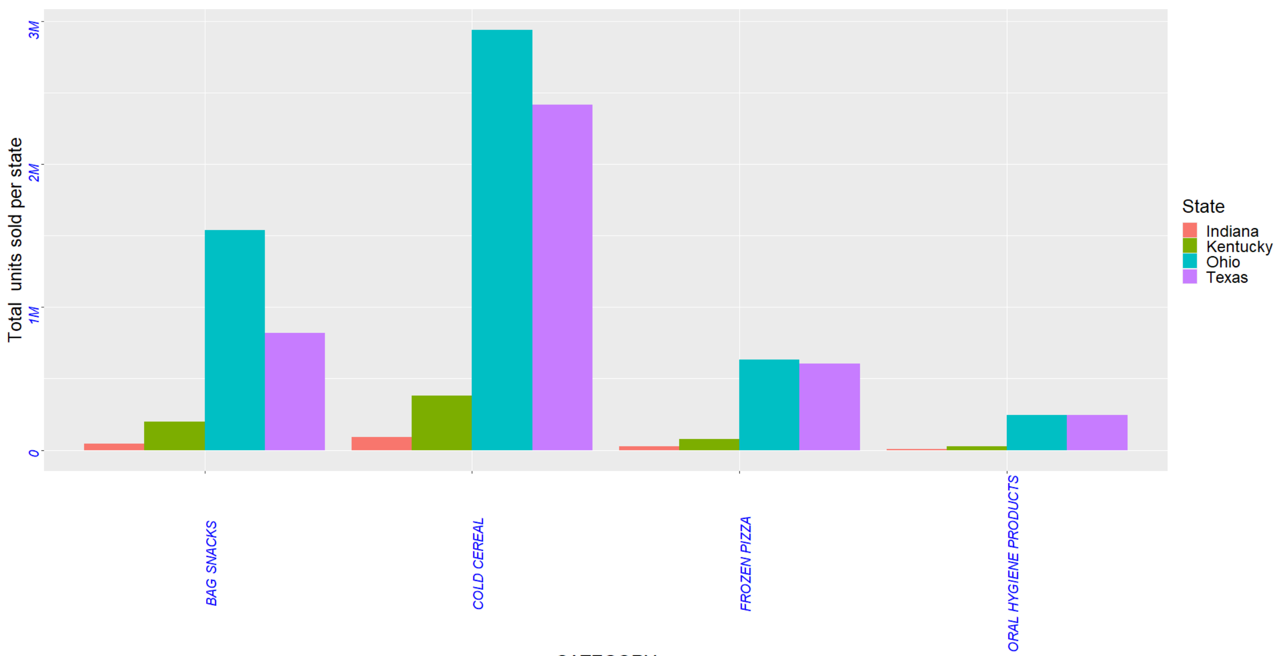

Figure B.7: Barplot per State & Category.

https://doi.org/10.26219/heal.aueb.6516

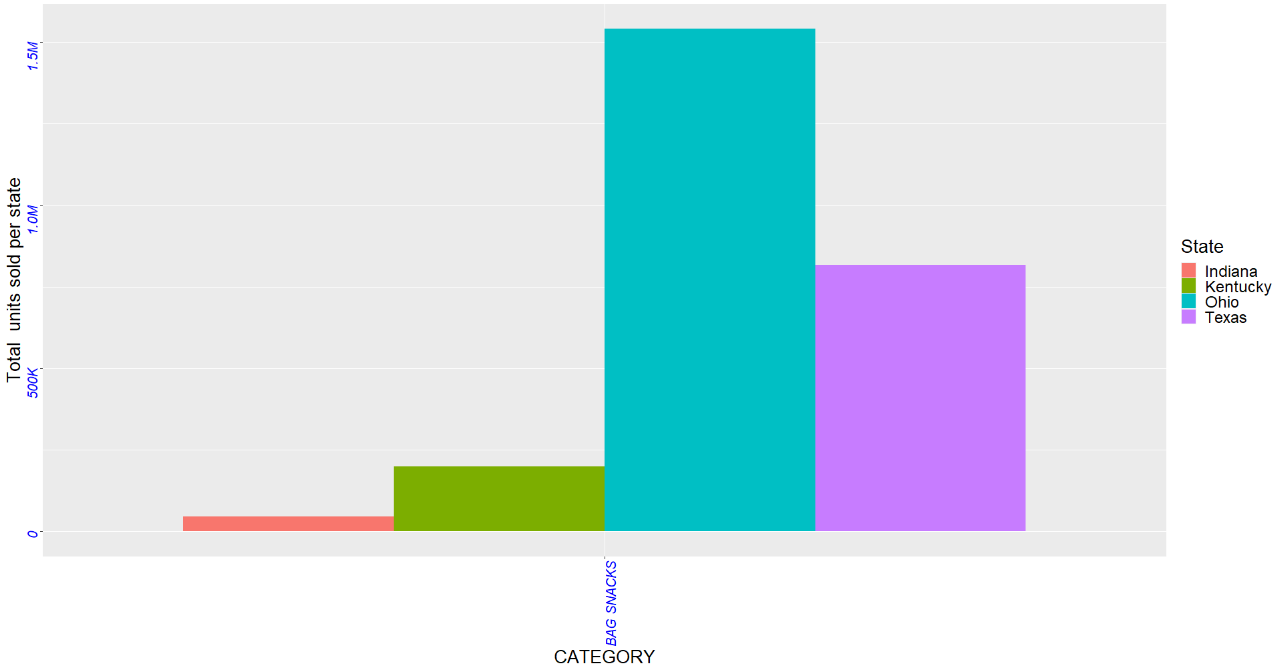

Figure B.8: Bag Snacks Barplot per States.

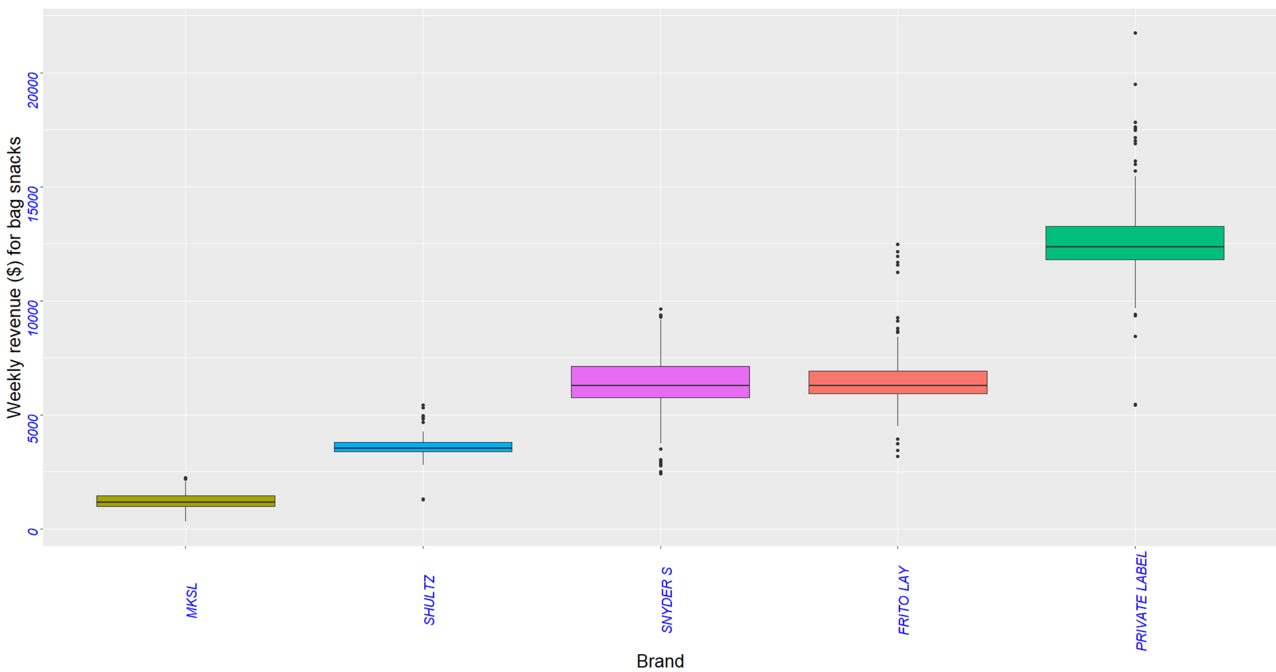

Figure B.9: Bag Snacks Box Plot per Brands .

https://doi.org/10.26219/heal.aueb.6516

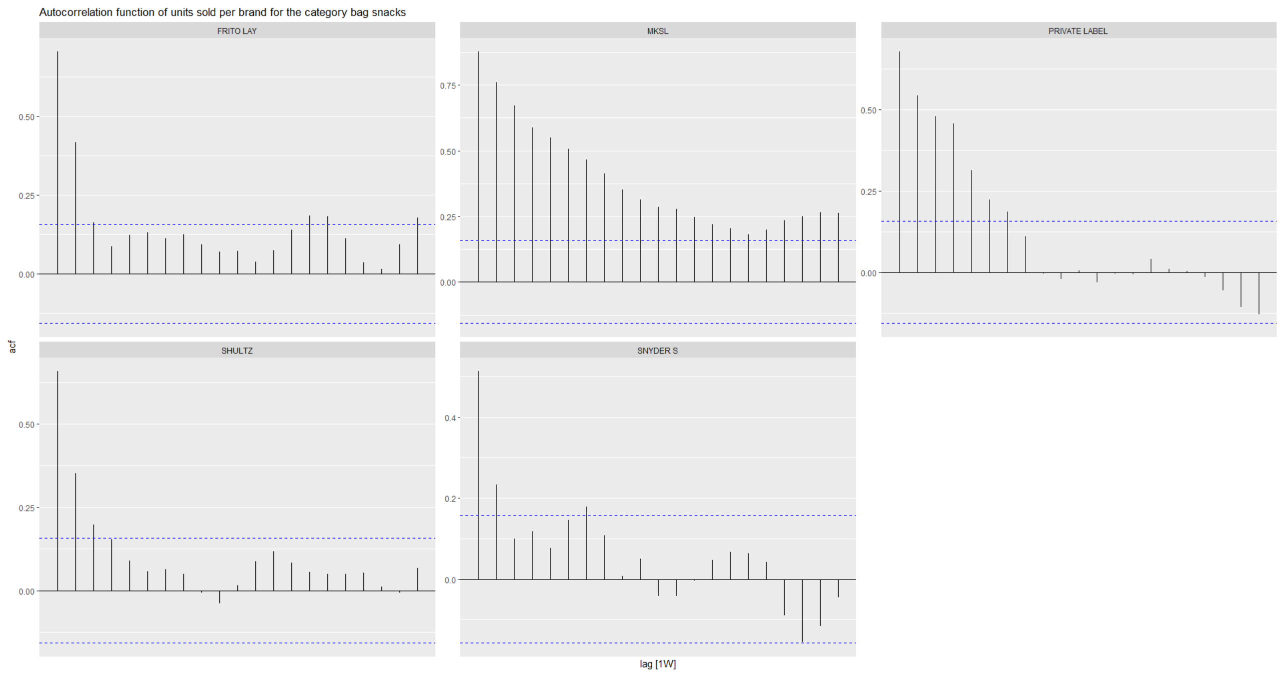

Figure B.10: Bag Snacks ACF Plot per Brands.

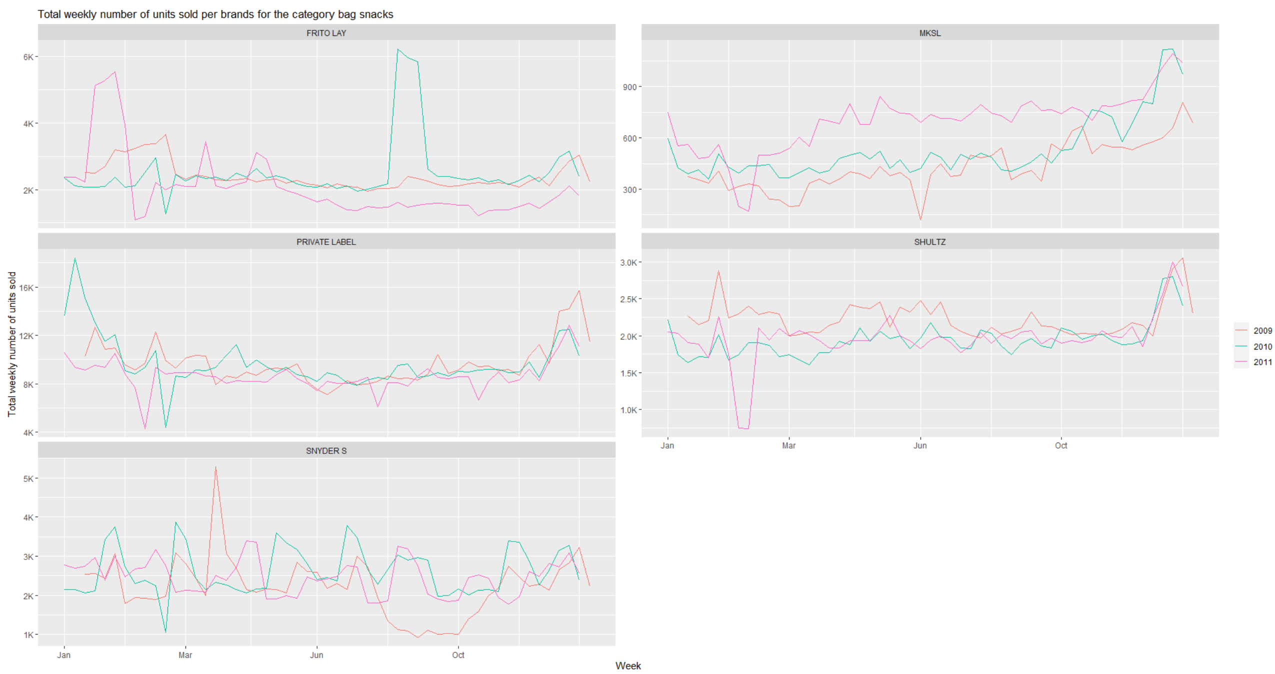

Figure B.11: Seasonal Plot per Brands in Bag Snacks Category.

https://doi.org/10.26219/heal.aueb.6516

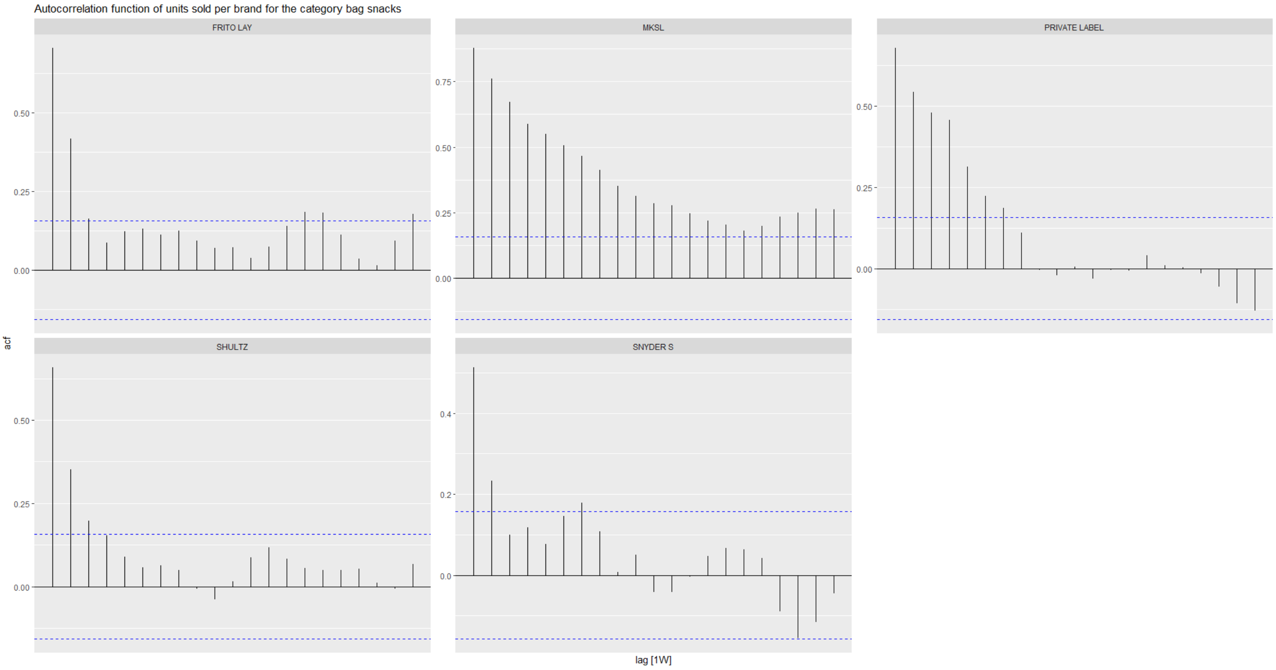

Figure B.12: ACF Plot per Brands in Bag Snacks Category.

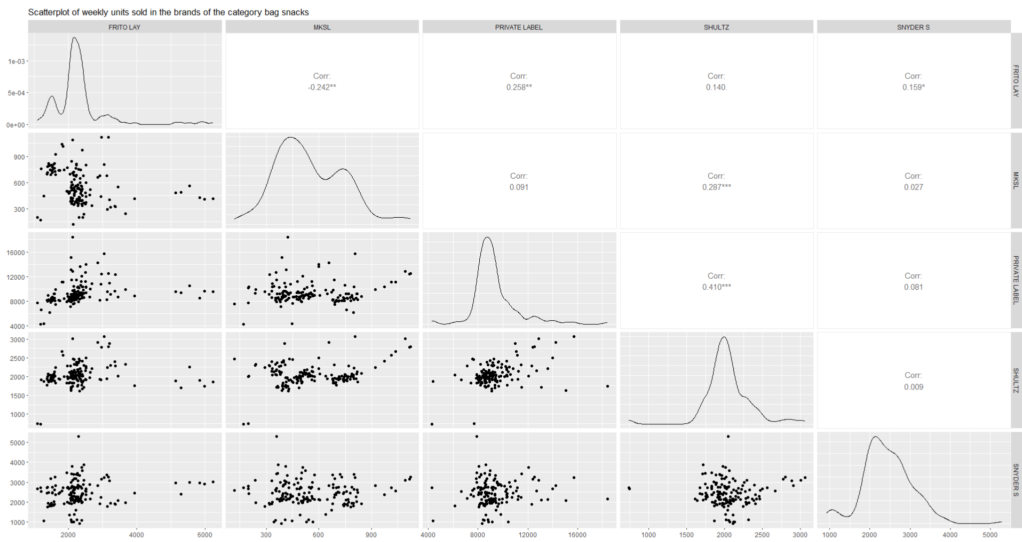

Figure B.13: Bag Snacks Scatter Plot per Brands in Bag Snacks Category.

https://doi.org/10.26219/heal.aueb.6516

https://doi.org/10.26219/heal.aueb.6516

Bibliography

J. S. Armstrong. *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Springer, 2001.

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

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

S. Banerjee, B. P. Carlin, and A. E. Gelfand. *Hierarchical modeling and analysis for spatial data*. CRC Press, 2015.

M. S. Bartlett. On the theoretical specification and sampling properties of autocorrelated time-series. *Supplement to the Journal of the Royal Statistical Society*, 8 (1):27–41, 1946. ISSN 14666162.

B. L. Bowerman, R. T. O'Connell, and A. B. Koehler. *Forecasting Time Series and Regression*. Cengage Learning, 2018.

G. E. P. Box and D. A. Pierce. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American Statistical Association*, 65(332):1509–1526, 1970.

P. J. Brockwell and R. A. Davis. *Introduction to Time Series and Forecasting*. Springer, 2016.

D. W. Bunn and B. Flyvbjerg. *Forecasting: An Essential Introduction*. John Wiley & Sons, 2018.

J. H. Chan, J. Yang, and X. Zhang. Forecasting chaotic time series with arima and stochastic time-change models. *Communications in Nonlinear Science and Numerical Simulation*, 58:102–115, 2018.

https://doi.org/10.26219/heal.aueb.6516

C. Chatfield and H. Xing. *The Analysis of Time Series: An Introduction with R*. Chapman and Hall/CRC Texts in Statistical Science Series. CRC Press, Taylor & Francis Group, 2019. ISBN 9781138066137.

M. P. Clements and D. F. Hendry, editors. *The Oxford Handbook of Economic Forecasting*. Oxford University Press, 2011.

J. D. Cryer and K.-S. Chan. *Time Series Analysis*. Springer, 2008.

M. Demetrescu and O. I. Moisescu. The use of arima models in forecasting the production of the romanian automotive industry. *Procedia Manufacturing*, 11: 2028–2035, 2017.

Dunnhumby. Breakfast at the frat: User guide. 2020.

J. Gareth, W. Daniela, H. Trevor, and T. Robert. *An introduction to statistical learning: with applications in R*. Spinger, 2013.

C. W. Gross and J. E. Sohl. Disaggregation methods to expedite product line forecasting. *Journal of Forecasting*, 9(3):233–254, 1990.

J. E. Hanke. *Applied Time Series Analysis for Managerial Forecasting*. John Wiley & Sons, 2003.

R. Hyndman, A. Lee, E. Wang, S. Wickramasuriya, and M. Wang. Package 'hts'. hierarchical and grouped time series. r package version 4.5. 2015, 2017a.

R. J. Hyndman and G. Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, 2018.

R. J. Hyndman and A. B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688, 2006.

R. J. Hyndman, A. B. Koehler, J. K. Ord, and R. D. Snyder. *Forecasting with exponential smoothing: The state space approach*. Springer Science & Business Media, 2008.

R. J. Hyndman, R. A. Ahmed, G. Athanasopoulos, and H. L. Shang. Optimal combination forecasts for hierarchical time series. *Computational Statistics Data Analysis*, 55(9):2579–2589, 2011. ISSN 0167-9473.

R. J. Hyndman, A. J. Lee, and E. Wang. Fast computation of reconciled forecasts for hierarchical and grouped time series. *Computational statistics & data analysis*, 97:16–32, 2016.

https://doi.org/10.26219/heal.aueb.6516

R. J. Hyndman, G. Athanasopoulos, S. Razbash, D. Schmidt, Z. Zhou, Y. Khan, and C. Bergmeir. forecast: Forecasting functions for time series and linear models, 2013, r package version 4.06. 2017b.

Y. Khandakar and R. J. Hyndman. Automatic time series forecasting: the forecast package for rj stat. *Soft.[Google Scholar]*, 2008.

D. Korobilis. Prior selection for vector autoregressions with informative steady-state priors. *Computational Statistics & Data Analysis*, 66:62–76, 2013.

J. K. Kuhn M. *Applied predictive modeling*. Spinger, 2018).

N. Lange and L. Ryan. Assessing normality in random effects models. *The Annals of Statistics*, pages 624–642, 1989.

G. M. LJUNG and G. E. P. BOX. On a measure of lack of fit in time series models. *Biometrika*, 65(2):297–303, 08 1978.

S. Makridakis, E. Spiliotis, and V. Assimakopoulos. M5 accuracy competition: Results, findings, and conclusions. *International Journal of Forecasting*, 38(4):1346–1364, 2022. Special Issue: M5 competition.

S. G. Makridakis, S. C. Wheelwright, and R. J. Hyndman. *Forecasting Methods and Applications*. John Wiley & Sons, 1998.

S. G. Makridakis, S. C. Wheelwright, and R. J. Hyndman, editors. *Handbook of Forecasting*. Elsevier, 2000.

S. G. Makridakis, S. C. Wheelwright, and R. J. Hyndman. *Forecasting methods and applications*. John Wiley & Sons, 2008.

S. G. Makridakis, S. C. Wheelwright, and R. J. Hyndman. *Handbook of forecasting*. Elsevier, 2013.

R. A. D. Peter J. Brockwell. *Introduction to Time Series and Forecasting*. Spinger, 2016).

F. Petropoulos and N. Kourentzes. *Robust Forecasting with Exponential and Holt–Winters Smoothing*. Springer, 2021.

R. H. Shumway and D. S. Stoffer. *Time Series Analysis and Its Applications: With R Examples*. Springer, 2017.

https://doi.org/10.26219/heal.aueb.6516

Q. Wang, J.-W. Lee, and A. G. Yeh. Forecasting time series with complex seasonal patterns using exponential smoothing and state space models: An empirical evaluation. *European Journal of Operational Research*, 281:527–545, 2020.

S. L. Wickramasuriya, G. Athanasopoulos, and R. J. Hyndman. Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. *Journal of the American Statistical Association*, 114(526):804–819, 2019.

https://doi.org/10.26219/heal.aueb.6516