Predictive Demand Models in the Food and Agriculture Sectors: An Analysis of the Current Models and Results of a Novel Approach using Machine Learning Techniques with Retail Scanner Data.

By

Aline Oliveira Pezente

B.A. Business and International Trade, Mackenzie, 2006
M.S. Corporate Finance, FIA, 2010

SUBMITTED TO THE MIT SLOAN SCHOOL OF MANAGEMENT IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE IN MANAGEMENT OF TECHNOLOGY
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 2018

Signature of Author: Signature redacted

MIT Sloan School of Management
May 11, 2018

Certified by: Signature redacted

Roberto Rigobon
Senior Lecturer, MIT Sloan School of Management
Thesis Supervisor

Accepted by: Signature redacted

Johanna Hising DiFabio
Director, Sloan Fellows and EMBA Programs
MIT Sloan School of Management

THIS PAGE WAS INTENTIONALLY LEFT BLANK

Predictive Demand Models in the Food and Agriculture Sectors: An Analysis of the Current Models and Results of a Novel Approach using Machine Learning Techniques with Retail Scanner Data.

By

Aline Oliveira Pezente

Submitted to MIT Sloan School of Management
on May 11, 2018 in Partial Fulfillment of the
requirements for the Degree of Master of Science in Management of Technology.

ABSTRACT

Agriculture commodities production and consumption are typically not aligned since the timing of commodity production with its pace of consumption is disjoint, once commodities are often produced periodically (with certain crops being harvested once a year) but with a continuous consumption throughout the year. The temporal mismatches in production and consumption require both commodities consumers (food industries) and producers (farmers) to predict future consumption based on limited unreliable information, about the future of demand and available historical data. Consequently, the lack of an appropriate understanding of what is the actual food consumption trend, lead's the producers in some cases to make wrong bets, which eventually causes food waste, price volatility and excess commodities stock. The commodities market has a good view of short-term supply fundamentals but still lacks powerful tools and frameworks to estimate long-term demand fundamentals, of which will drive the future supply. This thesis studies commodities demand forecasting using Nielsen's Retail Scanners data based on machine learning techniques to construct nonlinear parametric models of commodities consumption, using the U.S sugar cane as our use case. By combining Nielsen Retail Scanner data from January 2006 to December 2015 for a sample of 30% of U.S retail, wholesalers and small shops, considering a basket of products that has sugar as one of its main components, we were able to construct out-of-sample forecasts that significantly improve the prediction of sugar demand compared to classical base-line model approach of the historical moving average.

Thesis Supervisor: Roberto Rigobon
Title: Senior Lecture

THIS PAGE WAS INTENTIONALLY LEFT BLANK

Acknowledgments

I would like to sincerely thank Roberto Rigobon, my advisor, for his valuable guidance and time dedicated to directing the thesis process. Also, this research would not have been possible without the insights and guidance from Felipe Ferreira for the generous dedication of his time to edit and structuring the codes, which were detrimental for this thesis results. Finally, thank you to my husband, Fabricio Pezente, who supported and incentivized me during this challenging journey.

Cambridge, May 2018
Aline Oliveira Pezente

THIS PAGE WAS INTENTIONALLY LEFT BLANK

Contents

List of Figures

List of Tables

1. Introduction

1.1 Overview of Agriculture and Food Chain

Agriculture and food supply chains are the lifelines for human existence. Where these chains are located and the availability of food at the right time, appropriate quality, and right quantity are paramount.

The recent United Nations report "World Population Prospects: the 2017 Revision" projects that the population of the world will be 9.77 billion by 2050, and one of the biggest challenges will be to feed this growing population sustainably.

The series of processes, operations and actors that take and convert the food from its raw material state to the food in our plates is known as the food supply chain. It is not a single chain of individual entities and actors but a complicated web of interconnected entities and actors working to make food available.

Nearly all agriculture must undergo a variety of processes to transform it into food that can be consumed. These transformations can occur in three main categories: (i) transformations in space, (ii) transformations in time, and (iii) transformations in form.

The transformation in space involves the transportation of commodities from regions where they are produced (supply regions) to the places they are consumed (demand regions). Where commodities are produced are almost always located away from, and often far away from, the places where consumers reside. Transportation—the transformation in space—is fundamental to bring commodities from production areas to places where they are consumed.

Just as to the locations of commodity production and consumption usually do not align, the timing of production and consumption is often disjoint as well. Most of the agricultural commodities are produced periodically, with a crop being harvested once a year for some commodities, but consumed continuously throughout the year. (Eg. Corn, Soybean and coffee are an example).

These temporal mismatches in production and consumption create the need to engage in transformations in time, which consist of storing commodities and creation of inventories when supply is relatively high to demand. The storage is a form to create time transformation by smoothing out the effects of shocks on prices, production, and consumption. In conjunction to this, transformations in space and form require time to complete.

The actors involved in the three main categories of commodity and food transformation are shown in Figure 1, and the role of each of those actors is discussed briefly below.

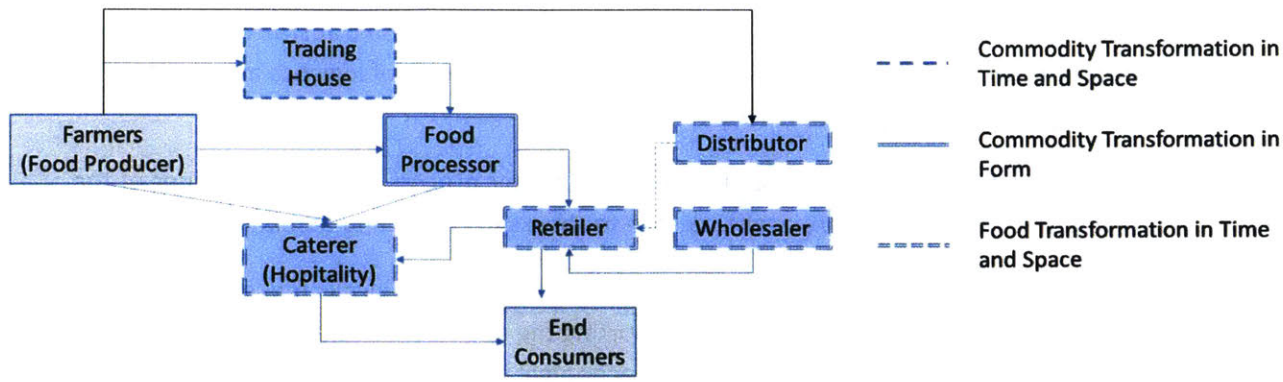


Figure 1 The Actors in Agriculture Commodities and Food Supply Chain

Producers

The food supply chain starts at the commodity producer end, which supplies food in its raw form for the agriculture and food industry system – grains, oilseeds, fruits, vegetables, meat, fish, poultry, and so on.

The producers are farmers whose businesses range from small family productive units to very large corporates. Every country needs an active and organized food production sector as it affects both food availability for the population and economic sustainability for the economy.

The farmers also have to deal with increasingly uncertain climatic weather patterns, scarcity of water, and land degradation, among other issues. As margins for producers within the commodity and food supply chain are getting smaller and smaller, farmers are growing what they can sell at a good price in order to have economic sustainability. Although this is fair, it impacts the availability of food products, food prices, and inflation levels of countries.

Trading Houses

The output from commodity producers moves downstream in two formats, either directly as fresh produce to the consumer (fresh fruit and vegetables, milk, grain, and so on) or in bulk as a raw material within the food processing industry.

The bulk purchase of raw commodities food is commonly done through partnerships with the farmers or through trading houses. Commodity trading houses are, in essence, engaged in the process of transforming commodities in space, time and sometimes in form.

They attempt to identify the most valuable transformations and undertake the transactions necessary to make these transformations. Those transactions consist of physical and operational actions necessary to deliver the bulk of commodities from producers to the food processing industry.

The value creation of a commodity trading house involves optimizing these transformations in space, time, and form through the usage of a complex interconnected network of logistics assets and information of food supply and demand chain.

The role of a commodity trading house is an intrinsically dynamic process because the values of the multiple possible transformations vary over time due to shocks to supply and demand forces.

For instance, a big harvest campaign of a commodity in one region will typically make it optimal to store additional quantities of that commodity since transporting the additional marginal output to consumption locations will not generate reasonable economic benefits.

An essential factor for the optimization process of commodity transformation is the cost of making the transformations, which includes transportation costs (transformations in space), storage costs (transformation in time), and processing/refining costs (transformations in form).

These costs are affected by constraints and bottlenecks in the transformation processes. Usually, considering all else equal, the tighter the constraints affecting a specific transformation process, the more expensive that transformation is.

Commodity trading houses define their role as finding and exploiting "arbitrage" opportunities in these transformational processes. Such arbitrage opportunities occur when the price of making the transformation exceeds the costs. (For example, when buying a commodity and hold it in the warehouse now is cheaper than buying in the future).

Food Processors

Food processors are companies or individuals in the food supply chain that transform the agriculture commodities in their form supplied by the farmers into products ready-to-eat format to be consumed by people, and that meet consumer requirements. The food-processing companies are diverse in nature and process products at different stages: for example, meat slaughtering and processing; packaging of fresh fruits and vegetables; milling of oilseeds and grains; manufacturing confectionery and bakery products; and many other types of food.

Food processing is essential, as it not only sustains the food sector economy by delivering to the demands and serving requirements of consumers but also helps to reduce waste and increase food availability by increasing the shelf life of raw food products that cannot be immediately consumed. Food processors need to work very closely with the downstream supply chain, which comprises the companies that are responsible for taking the processed food to the consumer.

Retailers and distributors

Once the food is processed, its transformation process in space and time stages is executed by two actors in the food supply chain: distributors and retailers. Food distributors are companies that act as the link between food processing industry, farmers (mainly in case of fruit and vegetables), and markets. The distributor's source either fresh products from farmers (eg. fruits, vegetables) or processed food from the processors and then distribute it through various channels to reach the final consumer. These channels are either retailing companies or other processing companies in the value chain (for example, restaurants) which supply the product to the consumer.

Distributors usually buy products in bulk and use an infrastructure of distribution centers and warehouses to deliver the products as and when required by consumers. Retailers generally buy from distributors or food processors (which can be integrated with distributors) and make it accessible to the final consumer. Retailers can comprehend local corner shops or large hypermarkets and supermarkets that manages hundreds of thousands of stock-keeping units (SKUs).

The retailers provide the consumer with the variety of products that the food sector has to offer. It is a highly competitive industry where food processors compete for shelf space in the retailer environments, and the retailers compete among themselves to attract more consumers through their doors. Consumer[1]s have a wide choice of retailers, retail channels and formats.[1]

Hospitality sector

The hospitality sector is a key actor in the food supply chain. Although not considered as a retailer or a distributor, it is also one of the critical links between the producer/processor and the consumer.

Foodservice companies, hotels, restaurants and takeaway places sources raw material from the producers or food processors and transform the food ready-to-eat to the final consumer.

These entities comprise millions of small and medium establishments or companies, sometimes one-person organizations [2].

1.2 The role of supply and demand (S&D) and the long term commodity consumption view

The challenge of feeding the world's population, which is estimated to exceed 9 billion by 2050[2], in the face of changing diets, increasing non-food demand for agricultural products, declining growth in agricultural productivity, and uncertainties derived from changing weather patterns has received a lot of attention lately, specially due to its effects in the global food industry [36].

Much of the research and disclosure in the market has focused on supply-side issues of declining productivity growth and sustainably increasing agricultural productivity. The understanding of the factors that govern future food supply capacity is an essential component of strategic planning for food and agriculture industries. However, it is equally important to improve the understanding of the demand-side drivers. The examination of food demand patterns over time and how these patterns

---

[1] Dani, Samir. (2015). Food Supply Chain Management and Logistics - From Farm to Fork - 1. Introduction to Food Supply Chains. (pp. 1-16). Kogan Page Publishers

adjust to rising incomes, changing prices and social-economic factors, enables better projection of food needs, provides an insight into the kinds of food consumers are likely to seek in the future, allows identification of new potential markets, and significantly improves predictions about how the food industry may be structured to meet the evolving trends in global demand. [36]

Equally to the benefit of a correct estimation of food and commodities demand, an incorrect assessment of demand also causes supply shocks and shocks to different kinds of transformation such as storage or processing which can cause financial losses, by the inefficient allocation of capital resources or processing capacity under-utilization for example, and food waste.

For instance, an increase in the demand for a commodity (e.g., the increase in the demand for wheat based products) increases the demand for logistical services provided by commodity trading firms, and simultaneously increases the demand for storage services and food industry processing capacity.

Moreover, the lower inventory due to an expected demand requires greater adjustments in production and consumption in response to demand and supply shocks and these adjustments are costly and a large demand shock primarily affects commodity prices.

These considerations highlight the danger of an inaccurate supply and demand estimation with the strategies that both commodities and food industries will assume. Although changes to underlying supply and demand for commodities affects demand for transformation services, the demand tends to be less volatile (especially when underlying demand and supply are highly inelastic), and there are frequently negative correlations between the demands for different types of transformations. [1]

The commodity chains deal in products such as palm oil, cocoa, coffee, sugar, cereals, grains and so on, and its supply chain model works with few buyers, mainly food industries, and many sellers. [1] The process works as a spot market, and the supply side fundamentals and price determines the movement of the product in the short-term (1 month or few months time horizon). However, the commodity systems keep information flow between trading parties to a minimum.

The commodity trader's intermediaries, when buying in bulk, utilize the information asymmetry to buy quickly, reduce costs by entering in hedging and maintain flexibility in product availability based on their view of future supply and demand dynamics.

Demand and supply volatility provides the necessary incentive to the futures trading between companies, and the interpretation of such volatility patterns and trends are myriad to the strategic orientation of actors in the food and agriculture supply chain.

Forasmuch as the purchasing between the food processor and commodity producer (farmer) does not happen directly but through many intermediaries that utilizes futures contract, the demand signals from consumers cannot be sent to the commodities producers (farmers) as there is a disconnect in the relationship.

Normally, the prices of major commodities are influenced by short term signals from climate change and uncertain weather patterns, variations in the reported global demand and supply (WASDE report)[2], and political processes such as trade agreements.

Therefore, the estimation of food supply and demand volatility, volumes and patterns are essential for an efficient allocation of capital and physical resources and expected profitability.

Commodities prices are determined by the understanding of supply and demand rather than government controls. It's this free market approach [35] that drives prices and consequently have an impact to household consumptions, companies' financial results and individual farmers efficient land utilization.

1.3 Problem statement and Thesis motivation

As described in the previous sections, to estimate accurately and in advance the demand of commodities is one the most attractive and powerful capabilities that players in the commodities and food value chain aims to have. Predicting the demand is very relevant for the efficient allocation of capital, human and physical resources.

Many research and studies that proved that food demand is more sensitive to long-run social-economic aspects and diet preference and ultimately to its costs relative to population income, than to climate change or bioenergy scenarios [37] . Such consumer's preference and associated social-economic aspects signals are more efficiently captures in historical and real-time retail consumption data than any other source of data.

The consideration of Retail-Scanner's data point to commodities demand forecast models could provide a powerful analysis and more accurate demand predictions for overall players in the industry and bridge the gap between the asymmetric information about demand pattern and responses across the players in in the commodities and food processing industry.

The agriculture market has succeeded so far to responded globally to increased food demand from population growth. The commodities and processed food supply has more than tripled since the 1960's [37] but prospects for the future are uncertain due to our challenge to keep feeding the world in a sustainable way, thus simulating possible scenarios of agriculture commodities futures requires analytical tools that can represent this world in a comprehensive way by reproducing the main structural drivers of supply and demand.

[2] The WASDE report is the World Agricultural Supply and Demand Estimates reports which contains fundamental market information such as crop production, ending stocks estimates, supply and demand estimates. The WASDE reports are released on a monthly basis and contain information for many different agriculture commodities.

And important component of such analytical tools is consumer demand modeling, which have different perspectives across modelers. The demand forecast exercise has been much more subjective to modelers different perspectives captured in their choice of behavioral parameters and how they see future food consumption might evolve rather than factual and quantitative data.

Therefore, the goal of this thesis is to investigate and propose a design for a adaptive system that forecasts the commodity demand by incorporating real-time frequent data from Nielsen's Retail Scanner's data applied to Machine Learning techniques and algorithms that accurately describes and forecast commodity demand.

This thesis will focus on U.S Sugar Cane demand estimation as our use case to build, validate and test the prediction models.

2. Literature Review

This chapter reviews the time series modelling in the aspects of stochastic models and machine learning approaches, as well as multi-step forecasting strategies which are useful for building a forecasting system for commodities demand.

In section 2.1, the basics of time series analysis are reviewed because forecasting commodity domestic demand is essentially a time series problem due to the seasonality of production and domestic deliveries associated with harvest season. In the past years, stochastic models and machine learning approaches were considered to be fundamental methodologies to solve the time series forecasting problems, as introduced in section 2.2 and 2.3.

2.1 The Time Series Analysis

There are many ways to make economic forecasts, including guessing, informal models, extrapolation, surveys, time-series models and econometric systems. When the assumption of constant parameters fails, the in-sample fit of a model may be a poor guide to ex-ante forecast performance (Clements, Michael 1999).

Usually, deterministic non-constant variables approach is one of the most important sources of forecast failures. (Clements, Michael 1999).

Time series is typically measured over successive times, representing a sequence of data points (Cochrane, 1997). The values measurements from an event in a time series are organized in proper chronological order.

Basically there are two types of time series:(i) Continuous and (ii) Discrete. Continuous time series observations are measured at every point in time, while a discrete time series encloses observations measured at discrete points of time.

In a discrete time series the successive observations are recorded at equally spaced time intervals such as hourly, daily, monthly or yearly time separations, which corresponds to the frequency of events and data observation. In general, to do further analysis, the data observed in a discrete time series is assumed to be as a continuous variable using the real number scale.

Time series analysis fits a time series into a correct model. The procedure of fitting a time series to a correct model is termed as Time Series Analysis. Practically, parameters of the model are estimated using the known data values, which comprises models that attempt to analyze and understand the nature of the series, and diagnose whether observed data values are most likely generated by stochastic or determinist dynamics. (Huffaker, Bittelli, Rosa 2017)

Time series models are useful for future simulation and forecasting after beings validated. A time series, in general, is assumed to be affected by four main components: trend, cyclical, seasonal and irregular components (Lee, 1999). These four components can be extracted and separated from the observed data.

Considering the effects of these four components, in general, addictive and multiplicative models are utilized for a time series decomposition. Additive models are substantiated on the assumption that the four components of a time series observation are independent of each other. Whereas, the multiplicative model, the four components can affect the others results and context and they are not necessarily independent.

Time series are often highly fluctuating, with random appearance of variables. The observed volatility is commonly attributed to exogenous random shocks to stable systems. Consequently, the reproduction of volatility is driven by a variety of linear-stochastic and probabilistic methods. (Huffaker, Bittelli, Rosa 2017).

2.2 Stochastics models

Stochastic models are tools for estimating probability distributions of potential outcomes, the application of such models initially started in physics and is now being applied in finance, engineering, social sciences, etc. The selection of a proper model is critical as it reflects the underlying structure of the series, and more importantly, to the fitted model which is essential for future forecasting.

There are two widely used linear time series models: Autoregressive (AR) and Moving Average (MA) models [8, 10]. The AR and MA models are widely analyzed and used so will not be introduced in details here.

An ARIMA(p, q) model is a combination of AR(p) and MA(q) models and is particularly suitable for univariate time series models [11]. However, the ARIMA models, described above can only be used for stationary time series data. In practice, many time series show nonstationary behavior pattern, such as those related to commodity and food business as well as those contain trend and seasonal patterns [13, 14]. Thus from the application point of view, ARIMA models are not useful and deficient to adequately describe non-stationary time series, which frequently encountered in commodities and food consumption data.

To evaluate and forecast time series, we always search for a statistical model capable of understanding the underlying processes of data and filtering the unwanted noise. This noise may have some pattern of autoregressive moving average i.e., ARIMA(p,q) processes. [12]

For this reason, the ARIMA model [5,11] is proposed, which generalizes ARIMA model to include the case of non-stationarity as well [8]. ARIMA model and its different variations are also generally known as the Box-Jenkins models for the reason that they are based on the famous Box-Jenkins principle [10, 15].

The popularity of the ARIMA model is due to its statistical properties as well as because of their relative simplicity in understanding it effects as well as implementation.

However, the major limitation of ARIMA model is the assumed linear form of the model. That is, a linear correlation structure is assumed among the time series values, and therefore, no nonlinear patterns can be captured by the ARIMA model. Although linearity is a useful assumption and a powerful tool in many areas, it became increasingly clear that the approximation of the linear models to complex real-world problem is not always satisfactory. [16]

For example, non-linear models are appropriate for predicting volatility changes in economic time series and have offered new insights into functional form, dependence, tail thickness, and nonstationary that are fundamental to the behavior of asset returns, as described by Linton and Yang in[17].

Various non-linear models have been suggested in the literature, such as the Threshold Autoregressive (TAR) model[16, 18], the Autoregressive Conditional Heteroscedasticity (ARCH) model and its variation like Generalized ARCH (GARCH) [17], the Non-linear Autoregressive (NAR) model [19] and so on.

According to Linton and Yang in [17], these nonlinear models are "model-driven approaches" in which we first identify the type of relationships among the variables (model selection) and afterwards estimate the selected model parameters.

2.3 Machine Learning approaches: Artificial Neural Networks

Machine learning models have been increasing its usage compared to classical statistical models in the area of forecasting in the last decade. Big advances have been made in this field in the past years, both in the quantity and variations of the models developed, equally in the theoretical understanding of the models.

Machine learning approaches are extensively being used for building more accurate forecasting models. Literature searches have found the combination of machine learning techniques with the econometrics models results in better forecasting accuracy in integrated models.

The results from different techniques are integrated to obtain better forecasting accuracy as demonstrated by Bajari, Nekipelov, Ryan, and Yang in [20].

Although the development of ANN was mainly biologically motivated, whereas its primary objective was to construct a model for simulating the intelligence of human brain into the machine, the Artificial neural networks (ANN) approach has been suggested as a widely used in various areas, especially for classification and forecasting purposes [22].

Similar to the dynamics of a human brain, ANN tries to recognize essential patterns and regularities in the input data, learn from experience and then provide generalized results based on the previous knowledge.

The ANN model is characterized by a network of three layers connected by acyclic links: input, hidden and an output layer [25]. In the class of ANN, as demonstrated in the Figure 1, the most widely used ones in forecasting problems are multi-layer perceptrons, which use a single hidden layer feed forward network[25,16].

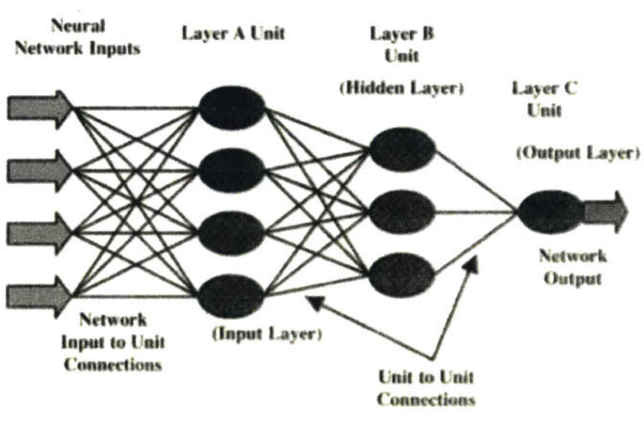


Figure 2 - Typically three-layer Feed-forward Neural Network topology.

Whereas many machine learning algorithms focus on solving classification problems, they can also be applied to regression problems, which is used in this thesis to forecast the U.S sugar domestic demand.

The recent growing research activities in Artificial Neural Network (ANN), as well as their numerous successful forecasting applications and results, suggest that they can also be utilized for time series forecasting [21, 22,23]. As opposed to the traditional model-based methods, ANN is a data-driven, self-adaptive, nonlinear, nonparametric statistical method in that there are few a priori assumptions about the models for problems under study [19,24].

ANN models can be useful for nonlinear processes that have an unknown functional relationship and as a result, are difficult to fit [26]. The process of constructing the relationship between the input and output variables is addressed by a certain general-purpose learning algorithm [16]. ANN modelling has been used as a new technique for estimation and forecasting in many fields of study in agriculture economics and statistics, since ANN's has freedom from restrictive assumption such as linearity that is often needed to make the traditional mathematical models tractable.

Most uses of ANN in economics have so far been in the financial market, in part because traditional approaches have had low explanatory power and in part, because the ANN approaches require abundant data, which is frequently available in financial sectors.

Notwithstanding, despite the increasing popularity and the sheer power of ANN models, the empirical forecasting performance of these models has been somewhat inconclusive [23].

Neural networks and traditional time series forecasting techniques have been studied and compared in several studies.

As some examples, Sharda and Patil in [27] use 75 out of the 111 time series from the well-known M-competition [23] as test cases and demonstrates that the neural networks perform as well as the automatic Box–Jenkins methodology.

In Kohzadi et al. study [28], it is demonstrated that the neural networks are superior to ARIMA methods for forecasting U.S monthly live cattle price and wheat cash prices.

In the context of economic data, Aliahmadi et al. [29] investigate the performance of neural network models compared with traditional regression methods to forecast crude oil exports, finding that they generally outperform traditional regression approaches even where there is no explicit nonlinearity.

In a paper from Crone et al. [30] it is demonstrated that neural networks are capable of handling complex data, including short and seasonal time series, beyond prior expectations. Nevertheless, most earlier studies of forecast accuracy for neural networks versus linear models do not consider pretesting for nonlinearity. [16]. In that context, Jha el al.[16] shows that execution of nonlinearity test provides a fairly a good indication to post-sample forecast accuracy for neural network models, which was measured by the RMSE and by the percentage of forecasts that correctly predict the direction of change , they demonstrate that nonlinearity tests of the series provide a reliable guide to post-sample forecast accuracy of the ARIMA and Time-Delay Neural Network models.

Besides, the literature suggests that the performance of nonlinear model should be evaluated on the basis of percentage of forecasts that correctly predict the direction of change instead of measures based on error terms.

The above facts indicate the importance of systematic investigation of nonlinearity on forecasting agricultural price, and volume series using the neural network models. Hence, in this thesis, an effort is made to evaluate the suitability of a time-delay neural network as a demand forecasting model in agriculture in comparison with the Box– Jenkins methodology using monthly demand volume series of sugar cane in U.S.

3. The Data

In previous chapters, the dynamics, importance of commodities forecast to the players within the agriculture and food industry have been explained, along with the challenges associated with the correct prediction.

This chapter we described the data structure, its sources and its challenges to obtain and process it and last we document some simple but profound empirical trends.

3.1 Data Sources

In this thesis, it is utilized a unique dataset consisting of Nielsen Scanner's data organized in a food composite portfolio for food consumption in units and price per upc (unit of product code). This data was obtained for a subset of U.S retail stores in all US markets which includes from approximately 55,000 groceries, drug, mass merchandiser, and other stores, in a weekly frequency of universe of sales and prices of every store by UPC.

Integrating Nielsen Scanners data, population growth, monthly sugar domestic deliveries, monthly sugar price data allows to compute and update measures of Food and Commodity equivalent consumption trend more frequently than the traditional slower-moving food consumption demand being utilized in the industry and governments.

The other dataset we utilize combined with Nielsen Retail Scanners Data are: Raw Sugar cane domestic consumption from USDA ERS, monthly SMD reports from 2005 to 2017. Sugar and Sweeteners Outlook sugar domestic delivery and stock changes. Sugar Price in the international market, NYBOT from 2006 to 2015.

The historical sugar domestic consumption data was assembled from ERS and FAS SMD3 monthly reports. We use the actual data and not the estimates since we have demonstrated that normally the SMD and ERS estimates are biased and under reported, as demonstrated in the next section 3.3 where we describe the ERS data.

In the section 5. it is demonstrated some trends in the sugar consumption demand, as outcomes from predicted models, compared with historical forecast over the same sample period done by industry.

Given the sensitive nature of Nielsen Scanners Data and data privacy, all individual identifications data elements to the stores and brand names were removed by Nielsen before sharing the data for the purpose of this study.

3.2 Collecting and Processing Retail Scanner's Data

The Nilsen Scanner's data was provided in tab-delimited format from Nielsen Kilts Data Center. A set of programs from Python and R programming languages were then used to pre-process the data, clean, homogenized and categorized to produce higher-level feature vectors and time-aggregated to weekly observations per product category and upc code.

Nilsen Kilts Data Center provided us with U.S consumer's transaction data in a subset of US retail and groceries stores, approximately to 55,000 stores across the country, in a UPC (Unit of Product Code, barcode) and store level, per week for the period of 2006 to 2015. The barcodes level information on prices and quantities purchased are approximately to 3,2 Million UPC codes, with 1075 produce modules (e.g Ice Cream, Carbonated – Soft Drinks), 125 product groups (e.g Health , Electronics, Eggs, Milk) and 10 departments (eg. Health & Beauty, Dairy , Deli, Frozen Foods).

The data represents approximately to 33% of total U.S expenditure shares. The store identities are anonymized, nevertheless 5-digit zip code for counties is provided linked to UPC, which allow us to make a geographical inference level for consumption. However, for the purpose of model construction and study objective the geographical categorization of data is not relevant and not covered in this research.

The sheer size and scope of dataset proved to be a significant challenge, from gathering and computing basic summary statistics. To develop basic intuition for the data we began by computing the total number of units sold and prices per week and upc code across all different categories.

Although Nielsen scanners data provide us a good vision about real food consumption compared to ERS estimations described in the Chapter 5, it has some limitation in terms of coverage. It does not cover all products in retailers, and particularly to the data set represented in this study, it represents approximately to 30% of U.S retail consumption not all. Therefore, one of the limitations of this results

[3] https://www.fsa.usda.gov/FSA/webapp?area=home&subject=ecpa&topic=dsa

relies on the data representation. Although we understand that 30% of total U.S retail consumption is quite relevant and representative we cannot assure if data sample is a fair representation of food consumption population.

One of the biggest challenges was the noise that this data set presents due to miss recording. The Figure [1] bellow demonstrates the data variability in a sample of Cookies consumption data set. This clearly is an example of wrong manual input which, if not correctly adjusted could lead to strong biases and misleading data interpretation for our algorithms.

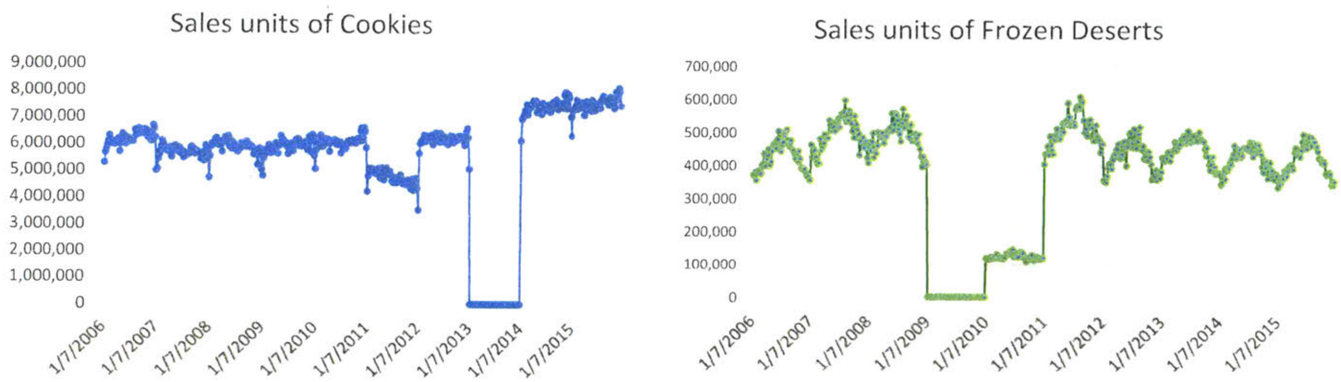


Figure 3 – Nielsen Scanner's Original Data set: Cookies and Frozen Deserts

To fix the data variability cause by wrongly data manipulation we have treated all data for call categories to exclude the top and bottom lowest 15% quartile in both units and sales price, before computing the total sales.

3.2.1 Nielsen Scanner's Data Sugar composite index

The Nielsen's food scanner's data presents strong seasonality in its consumption pattern for different product categories. Therefore, to minimize the seasonality effects and biased due to temporary promotions and discounts, we have considered the total sales index of a basket of products that possess sugar cane as one of its main components.

The food composites are kind of food that is composed of many ingredients in its form. Such as breakfast cereals, chocolates, carbonated beverage, cookies, which possess several commodity ingredients in its composition but among the main ones are sugar.

The estimation of ingredient component among every individual sku would be a great challenge, not only due to sheer size of skus and its different formula but also to the lack of granular data to allow us to do so.

These considerations suggest that we make use of a fixed base of a basket of food products to calculate our food composite. For this thesis, after evaluating many formulations and different products in different categories, we have selected the following product categories demonstrated in Table 1 to compose our sugar food composite basket.

Table 1 – Selected Food Categories for Sugar Retail Composite Index

Group	Category	
Candies	Candy Miniatures	Candy Non Chocolate Miniature
	Candy-Hard Rolled	Candy Non Chocolate
	Candy-Chocolate Miniatures	Candy Lollipops
	Candy-Chocolate	Marshmallows
	Candy Chocolate Special	
Jams, Jellies, Spreads	Jams	Marmalade
	Jelly	Peanut Butter
Juices	Fruit Juice Nectars	Fruit Drinks - Other Container
	Fruit Drinks Canned	
Breakfast Food	Pastires	Instant Breakfast - powdered
	Granola Bars	Breakfast bars
Cereal	Cereal - Ready to eat	
Desserts and Syrups	Syrup Specialty	Pudding -Seetened - mix
	Pudding/Pie Filling	Mixes-Ice Cream
	Syrup Chocolate	Desserts- Single servings - canned
	Toppings-Mixes	Syrup-Table
	Toppings - Liquid and Dry	Syrup-Shorghum and Sugar
Sugar, Sweeteners	Sugar Brown	Sugar Granulated
	Sugar Powdered	Sugar Remaining
Bread and Baked Goods	Bakery Bread, bunds, rolls , bage	Breakfast Cakes
	Muffins	Doughnuts
	Cakes	Pies
	Biscuits	Cheesecake
Carbonated beverages	Soft drinks - carbonated	Breakfast drinks - Powdered
	Soft drinks powdered	
Cookies	Cookies	Ice Cream Cones
Frozen Foods	Desserts	Pudding
	Ice Cream	

Following Diewert (1995,20) [40] analysis and considering the level of data disaggregation among the many different levels of skus it becomes necessary to define the average price and total quantity using a unilateral index number formula [40].

In the first stage of data aggregation, when we constructed the vectors of prices and quantities for the monthly annual periods, to be compared with ERS data, and insert these vectors into a bilateral index number formula, we aggregated the individual skus represented by upc's codes, per category in sort of average price and total quantities. This leads to unit value prices as being the natural prices at this first stage of data aggregation.

For our data length scope, we defined monthly yearly data, to be consistent with our data set from historical sugar domestic deliveries and prices. Such definition required a second level of data aggregation which was essentially the aggregation of the product of the aggregated price and quantity per category.

Also, price discounts lead to large fluctuations in quantity purchased and leads to large fluctuations in quantities purchased and therefore in the overall measures of price change and volume change, such fluctuations are not necessarily a demonstration of long term trend [41] and heavily discounted prices typically lead to a chain drift problem. By aggregating data and considering the composite sales, rather than units, minimizes such effects.

To formulate the sugar retail composite index, we have utilized the Laspeyres index where which can be written as follows:

$$Index_t = \sum_i w_{i0} \left(\frac{p_{it}}{p_{i0}}\right)$$

Equation (1)

where p_{i0} is the base period sales of item i, p_{i0} is the sale of item *i* in period *t,* for t = 1,...T and w_{i0} is food category i's share of total sales in the period 0. In practice, the sales are unit values for food category class i for each period *t,* of some pre-specified lenghth (e.g week, month or quarter).

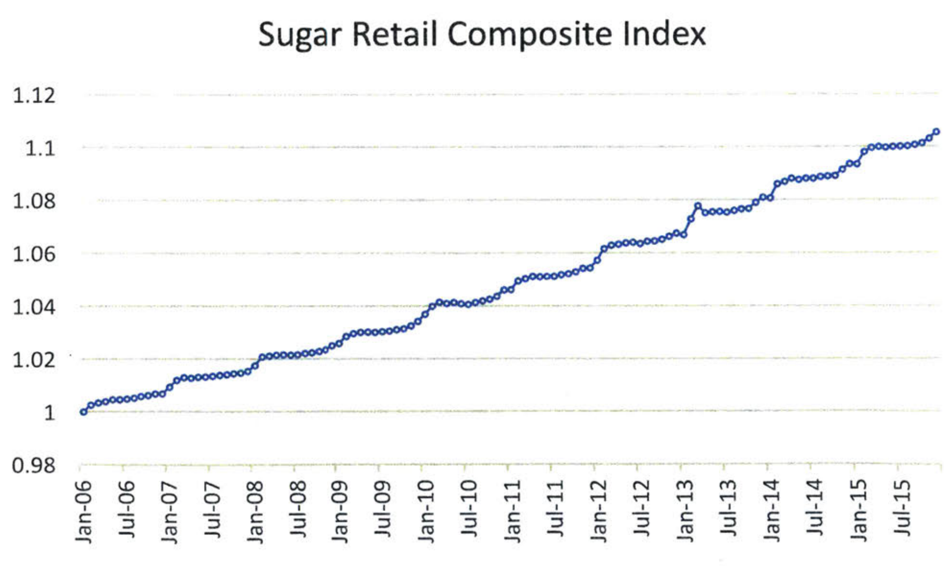


Figure 4- Sugar Retail Composite Index

The Figure 4 above represents the retail composite index calculated based on a selected basket of food products that contains sugar among its main component, which was described in the Table 1.

3.3 ERS and FAS Data historical S&D Sugar Analysis

The source of sugar domestic consumption is the data from actual sugar cane delivery for food consumption provided in the Sweetener and Market Data (SMD) report from the Farm Service Agency (FSA).

The data represents the quantity of raw foreign sugar purchased directly or not by food companies (SMD report or SMD non-reporter). The sugar represented in that data have already been physically cleared through U.S Customs and Border Protection for processing[4].

We use the actual computed results, consistent to the totals reported in the WASDE. The SMD report also reports estimates of sugar domestic however this analysis has been constantly being proved to be biased and underreported.

The USDA recognizes this in a report issued in April 2013, " Indeterminacy in Measuring U.S. Sugar Deliveries for Human Consumption" [42] where it is mentioned *"the analysis has shown that SMD estimates of sugar imports by SMD no reporters are biased and underreported. It is not clear what should be done about the problem, other than being aware of it. The SSO will continue its import analysis and use the implications for forecasting U.S. sugar demand. It is not clear what should be done about the problem, other than being aware of it."*

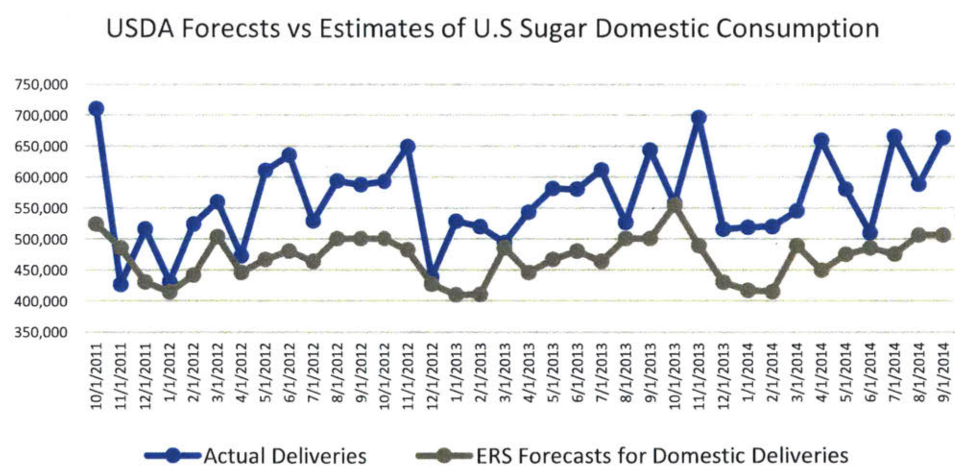


Figure 5 – USDA Domestic Consumption past prediction vs Actuals

The Figure 5 we demonstrate the 2012 and 2013 annual sugar cane domestic consumption projected in those years and its actual numbers in the same period. We later compare those numbers with the results from out out-of-sample predicted values.

---

[4] See http://fsa.usda.gov/Internet/FSA_File/sugar_data_user_manual.pdf - "CCC-835 On-LineReporting Instructions" for listing of cane refiners cane processors and beet processors who report to the SMD.

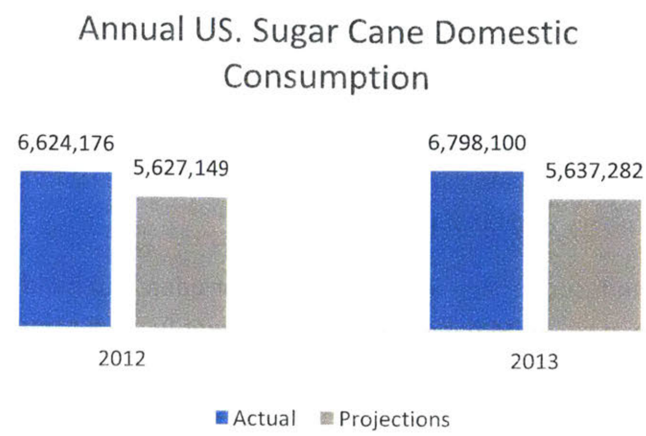


Figure 6 – U.S Annual Sugar Cane Domestic Consumption past prediction vs Actuals

4. Machine Learning Models applied to Commodity Demand forecast

Machine learning approaches are widely utilized for building accurate forecasting models.

Section 4.1 describes the features chosen to model the U.S sugar cane commodity demand. Some machine learning models are discussed in section 4.2, 4.3, 4.4, where we demonstrate the results of each considered model and its variations.

In this thesis, we do not dig into the details and variations of each model instead, we focus on the basic version of each models and discuss its pros and cons applying to U.S sugar forecasting problem by considering Nielsen's Retail Scanner Data as one of the estimator variables to predict sugar demand volume.

In this thesis we are going to apply three different methods to predict future values for the same time series of the U.S sugar cane demand based on Nielsen Retail Scanner's Data Sugar Composite Index, historical domestic consumption, stock levels and sugar price in the international market. Then we will make the comparison between the results of the three methods in order to determine which is better to use in similar situations and assess whether the Nielsen Retail Scanner's Data is a good predictor or not for equivalent agriculture commodity. Methods used are:

A) Multiple Linear Regression
B) Auto Regressive Model
C) Neural Nets

4.1 Features for commodities demand forecasting

Forecasts of agriculture commodity demand are useful for farmers, governments, food and agribusiness industries. The ability to accurately forecast the demand of agricultural commodities is, therefore, an important concern in both policy and business circles.

To forecast sugar demand in future, many features can be considered to build a good model. According to intuition and knowledge of sugar demand modeling, we identify the key features that are very likely to affect sugar demand, as summarized below:

- Retail Consumption: Equivalent consumption of food which contains sugar in its components. The retail consumption captures the volume, value and dietary preferences of consumers that might affect commodity demand.
- Sugar Commodity Price: Sugar Price that food industries paid to acquire the raw material to process and may influence industry search for substitutes or not.
- Previous Sugar Volume: Previous sugar volume purchased by food industry. The previous volume captures the pace of stocks build-up and historical consumption patter accordingly to seasonality.
- GDP per-capita: per capita expenditures on total food as well as individual food items due to income growth, in both urban and rural areas. The GDP per-capita is a relative stationary data where its effects is embedded to the Retail Consumption Index.
- Population: Food consumption is directly linked to population growth and size, equally to its purchase power. For the purpose of building vectors, this data if considered would have low significance considering its stationary profile in a monthly basis. Its long run effects it's also implied in Previous Sugar Volume consumption and Retail Consumption Index.

4.2 Multiple Simple Linear Regression

The Regression method is one of the most widely used statistical techniques and a straightforward approach for predicting a quantitative response to estimate the dependent variable by one or more independent variables [31].

In statistics, linear regression is an approach to modelling the relationship between scalars variables and one or more independent variables denoted X. The case of using linear regression to predict one independent variable is called simple regression, whilst more than one independent variables is called multiple regression.

In linear regression, the relationships are modelled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models.

To predict a quantitative response Y on the basis of a single predictor X, which assumes that there is a linear relationship between X and Y, mathematically, we can write this linear relationship as :

$$Y \approx \beta_0 + \beta_1 X$$

Equation (1)

Where β_0 and β_1 are two unknown constants that represents the intercept and slope in the linear model.

The linear regression has been around for a long time and was the first type of regression analysis to be studied rigorously and extensively used in practical applications [31]. One of the reasons that make linear regression practical use is because the models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters, besides the statistical properties of the resulting estimators and model results are easier to determine and interpret.

Most applications of linear regressions to commodities markets are forecasting, mainly price, consumption, export volumes, etc.

The Multiple regression analysis is a multivariate statistical technique used to examine the relationship between a single dependent variable and a set of independent variables. The purpose of multiple regression analysis is to use independent variables, whose values are known, to predict a single dependent variable. [32]

The effect of independent variables on the response is expressed mathematically be the regression or response function:

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon,$$

Equation (2)

Instead of fitting a separate simple linear regression model for each predictor, in multiple linear regression we extend the simple linear regression model (Equation 1) so that it can accommodate multiple predictors, giving each predictor a separate slope coefficient in a single model.

In our analysis, our system of dependent variables are Retail Composite Index, denoted by x_1, Sugar Price_Index , denoted by x_2 and our independent variable the Sugar Domestic Delivery Volume change, denoted by Y.

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon_1 + \epsilon_2,$$

Equation (3)

We define $(\epsilon_1 \epsilon_{,2})$ as the residuals of liner equation. We know that ϵ_1 and ϵ_2 are not quadratic since there are possibilities of interaction between the state variables i.e. X_1 can be affected by X_2 and vice-versa. This means that the model suffers from endogeneity.

The parameters β_0, β_1 and β_2 are unknown and estimated using the ordinary least squares approach (OLS), to fit the model.

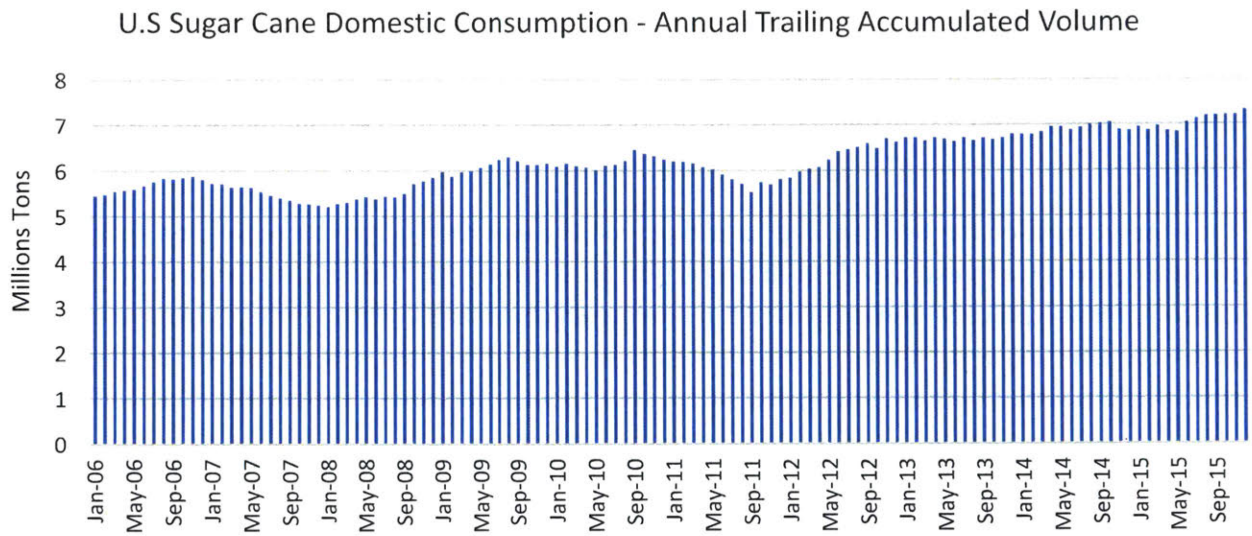


Figure 7 - U. S Monthly Sugar Cane Domestic Consumption by Food Industry

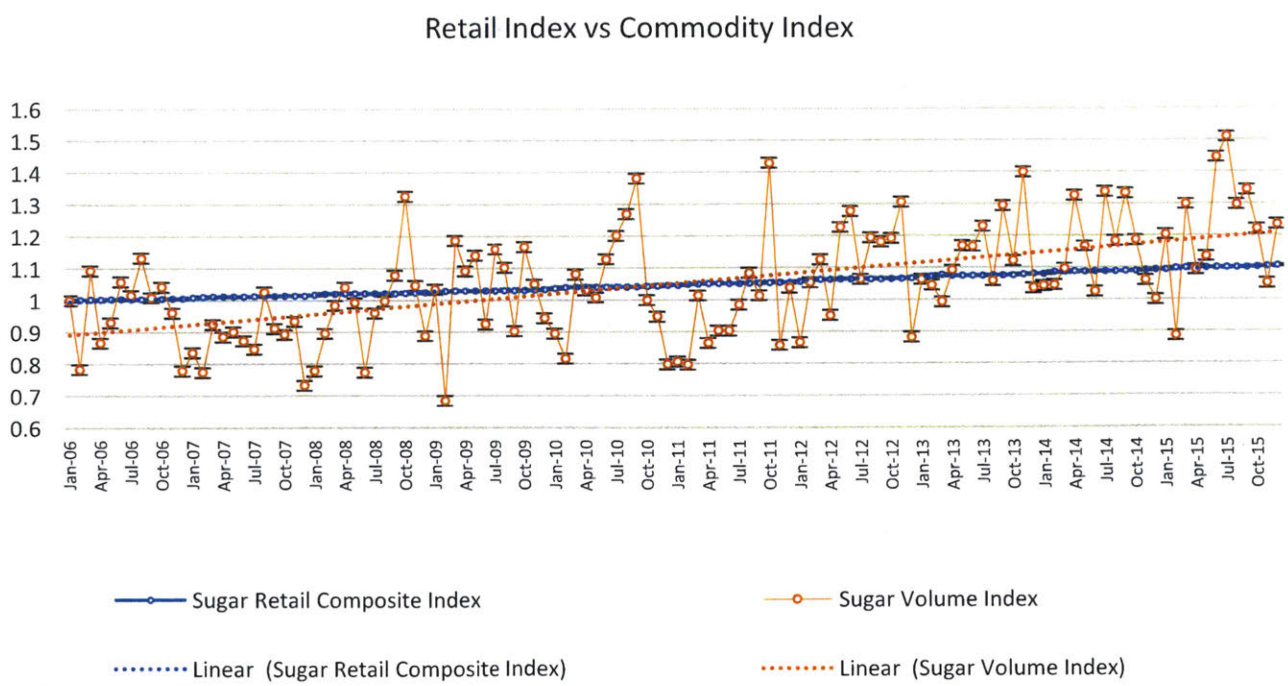


Figure 8 - Sugar Retail Composite Index vs Sugar Commodity Domestic Consumption Index

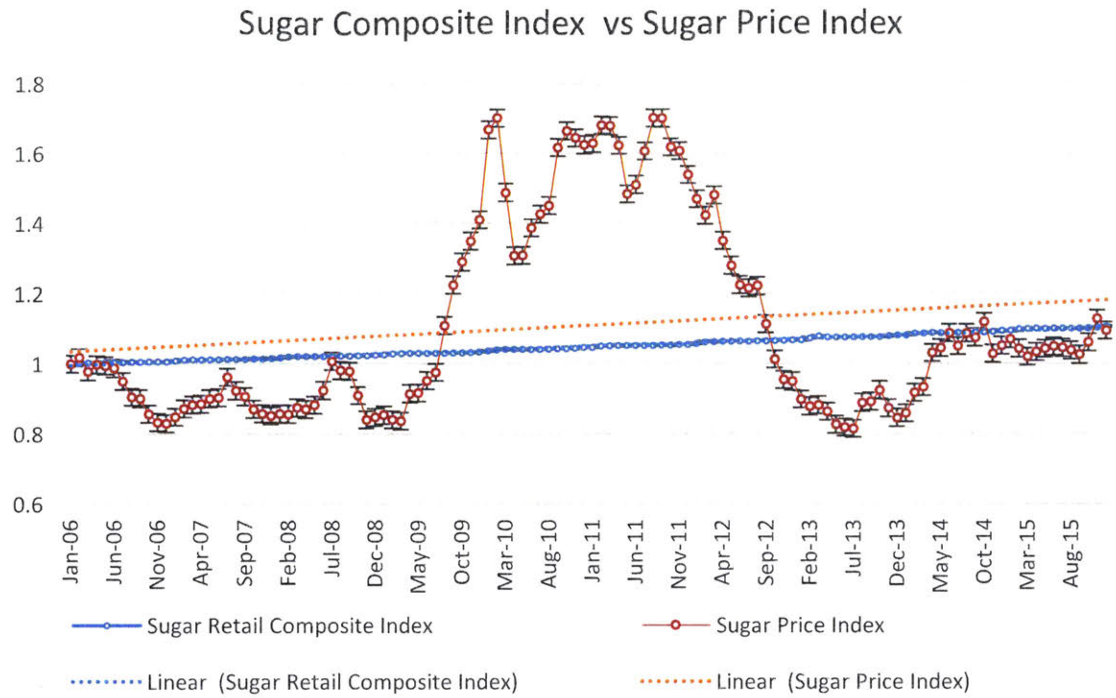


Figure 9 - Sugar Retail Composite Index vs Sugar Commodity[5] Price Index

Procedure:

- We use the linear regression model to regress the variable Sugar Volume (denoted by Y) against variables X_1 and X_2 .We regress both Monthly Series and 12-month trailing accumulated series.

- The 12-month trailing Sugar Volume (Y) is aggregated by year per month to smooth the effects of monthly seasonality embedded to Sugar Domestic Volumes Deliveries.

- We applied natural logarithmic transformation to the Sugar Monthly Volume data to stabilize the variance. A logarithmic transformation is used for data which can take on both small and large values and is characterized by an extended right-hand tail distribution. Logarithmic transformation is one of the data processing techniques which also convert multiplicative or ratio relationship to additive which is believed to simplify and improve neural network training. We note that the series is nonstationary even after logarithmic transformation and shows a stochastic upward trend.

- The first 7 years (from Jan 2006 to Dec 2012) is used to train the model, so that forecasting starts at the beginning of 2013.

---

[5] SB1 Contract – NY 11 Sugar Contract, Nearby futures traded in NY ICE

- Data from Jan 2013 to Dec 2015 is used as out-sample data in order to compared the accuracy of our prediction models.

- We have tested many trials with different variables to select combination of variables that would provide the highest statistically significance to the model.

For predictability, significant correlation between variables, in linear regression, is needed to make the model with a strong predictability capacity.

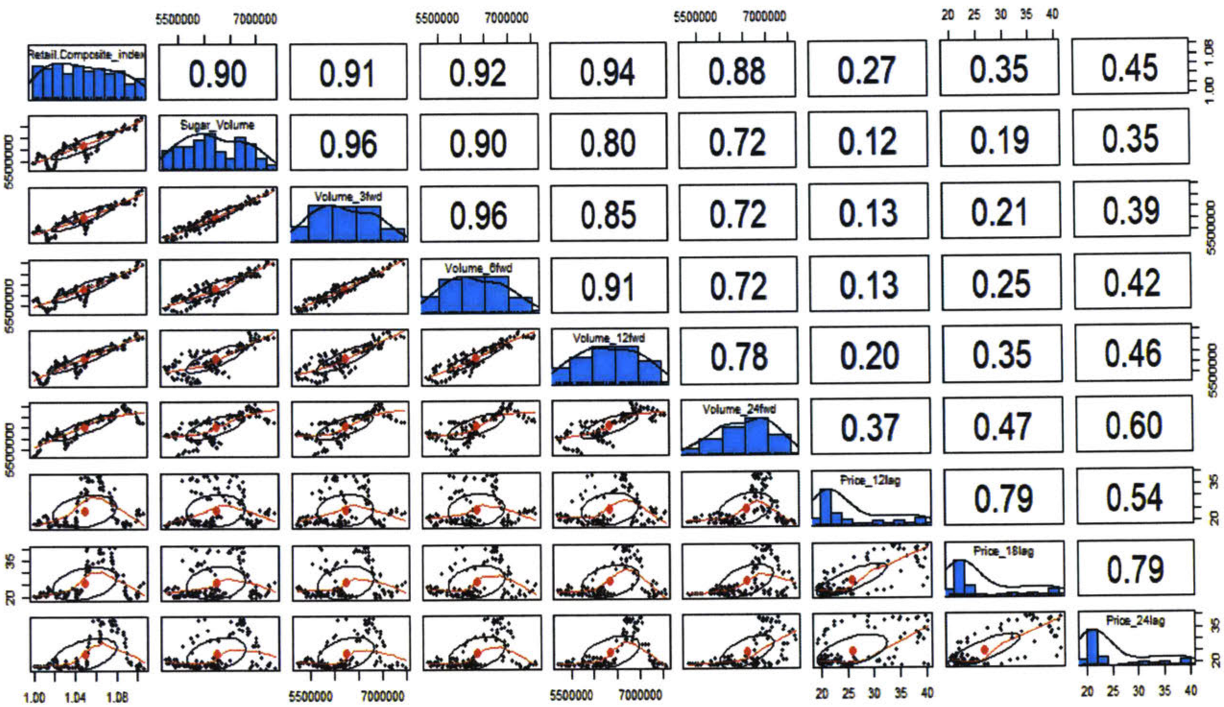


Figure 6 - Pair Panels of Bivariate Correlation and Pearson Correlation Coefficient (PCC) –Simple Linear Regression

For predictability, significant correlation between variables, in linear regression, is needed to make the model with a strong predictability capacity.

The Figure 6 above represents the distribution of pairs of variables used in the Simple Linear Regression Model (SLR), their linear regression results and estimated Pearson Correlation Coefficient[6]. We notice

[6] The correlation coefficient ranges from −1 to 1. A value of 1 implies that a linear equation describes the relationship between X and Y perfectly, with all data points lying on a line for which Y increases as X increases. A value of −1 implies that all data points lie on a line for which Y decreases as X increases. A value of 0 implies that there is no linear correlation between the variables. Source (https://en.wikipedia.org/wiki/Pearson_correlation_coefficient)

that the retail composite index has a strong PCC with sugar volume, 0.9 for Sugar 12 month accumulated trailing volume series and 0.54 for Sugar Volume series including in different lags and strong correlation with Sugar Price with 12 and 24 months' lag, which were the variables with highest significance. Such values are a strong indication of relationship of Sugar Retail Composite to predict future Sugar commodity volume consumption.

4.3 Time-series (ARIMA)

Demand forecasts are mostly made by using time series approaches. In time series modeling, past observations of the same variable are collected and analyzed to develop a model describing the underlying relationship.

The developed model is then used to extrapolate the time series into the future. In the last few decades, much effort has been devoted to the development and improvement of time series forecasting models. One of the most important and widely used time series models is the autoregressive integrated moving average (ARIMA) model. [16, 20]

The popularity of the ARIMA model is due to its statistical properties as well as the well-known Box–Jenkins methodology in the model building process. [16,25] However, a limitation of ARIMA model is the presumed linear form of the model. That is, a linear correlation structure is assumed among the time series values, and therefore, no nonlinear patterns can be captured by the ARIMA model.

In the classical statistical handling of time series, making predictions about the future is called extrapolation or time series forecasting. Some descriptive models can smooth or remove noise by utilizing time series.

The purpose of time series analysis is generally twofold: to understand or model the stochastic mechanisms that gives rise to an observed series and to predict or forecast the future values of a series based on the history of that series. [31]

In an autoregressive model (AR), time series variable is assumed to be a linear function of actual past values and random shocks. An AR model is mathematically defined by the following equation [3]:

$$Y \approx \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{T-2} + \cdots + \beta_p Y_{t-p} + \epsilon,$$

Where

$$\epsilon \sim WN(0, \sigma^2) \quad \text{and,}$$

$E(\epsilon_t Y_{t-j}) = 0\ , \forall j > 0, thus\ \beta_1$ is called as first-order autoregressive coefficient.

Equation (4)

As for autoregressive model (AR), in addition to use the above function to characterize the model the estimate of β_1 is done by the use of OLS equation:

$$\beta_{1=} \frac{\sum_{t=2}^{T}(Y_t - \bar{Y})(Y_{t-1} - \bar{Y})}{\sum_{t=2}^{T}(Y_t - \bar{Y})}$$

Equation (5)

The Y_{t+k} value after the *j* period is mathematically described as:

$$\begin{aligned} Y_{t+k} &= \beta_1 Y_{t+k-1} + \epsilon_{t+k}, \\ &= \beta_1[\beta_1 Y_{t+k-2} + \epsilon_{t+k-1}] + \epsilon_{t+k}, \\ &= \beta_1^k Y_t + \beta_1^{k-1}\epsilon_{t+1} + \dots + \beta_1 \epsilon_{t+k-1} + \epsilon_{t+k} \end{aligned}$$

Equation (6)

However, β1 is unknown parameter, hence the historical data of $\{Y_1, Y_2, \dots, Y_3\}$ must be first used to find out the estimate equation of β1 and construct the forecast equation of Y_{t+k}.

Considering the Loss Function Mean Square Error (MSE) , the information set $\theta_j = \{Y_{j,}, Y_{j-1}, Y_{j-2}\}$ can be used as conditional expected value.

There is an important time delay between retail food products demand responses to the commodity demand, which also suffers impact from sugar production and harvest seasonality.

Procedure:

- For Autoregressive model we use the 12-month trailing Sugar Volume (Y) as our independent variable to smooth the effects of monthly seasonality embedded to Sugar Domestic Volumes Deliveries. As demonstrated in previous exercise of simple linear regression, the 12-month trailing Sugar Volume provides more stability in time series and better results.

- Consistent with previous simple linear regression model, the first 7 years (from Jan 2006 to Dec 2012) is used to train the model, so that forecasting starts at the beginning of 2013 and the data from Jan 2013 to Dec 2015 is used as out-sample data in order to compared the accuracy of our prediction models.

- We ran several trials with many different combinations of independent variables to derive the best ARIMA (Auto Regressive Integrated Moving Average) model, which is demonstrated in Table 2 bellow. We defined ARIMA structure of differenced series, which was found based on the lowest AIC and RMSE values and autocorrelation function (ACF), partial autocorrelation function (PACF) and AIC information criterion.

- Similar to the SLR Model procedure, we have tested different variables for each independent variable, to fin the combination that would provide the highest significance to the model and lowest residuals. Such approach was important to define the variables that would provide the best predictability capacity and avoid model over-fitting. For the purpose of this thesis, our discussion will be limited to the best models identified only, although we have tested many different combinations of variables.

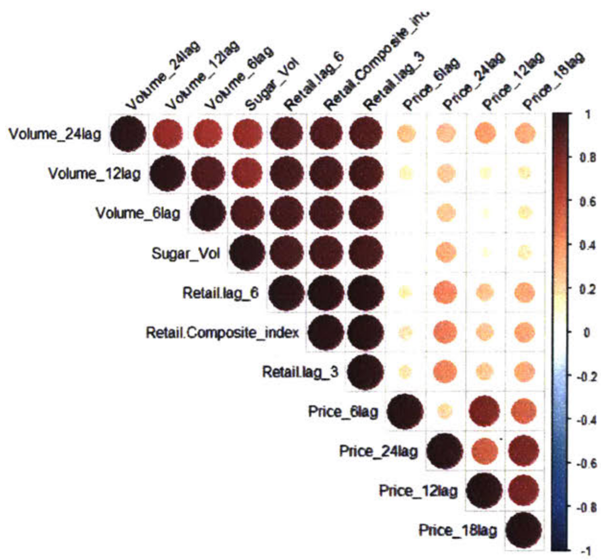


Figure 11 - Correlation Table Index– Autoregressive Regression Model

As explained in previous section, the Retail Composite Index captures the dollar value of a basket of food, which has inherent among its value components the pass-through cost of sugar commodity price. The high significance of variables in different time lags is understandable since the food industry usually stablish the commodity cost by trading the nearby futures contract of sugar and some players uses sophisticated derivatives mechanism around the same contract. This is a dynamic market and different players will form their cost in different periods of time, which ultimately is reflected in the overall food industry inventory cost level and consumption pace.

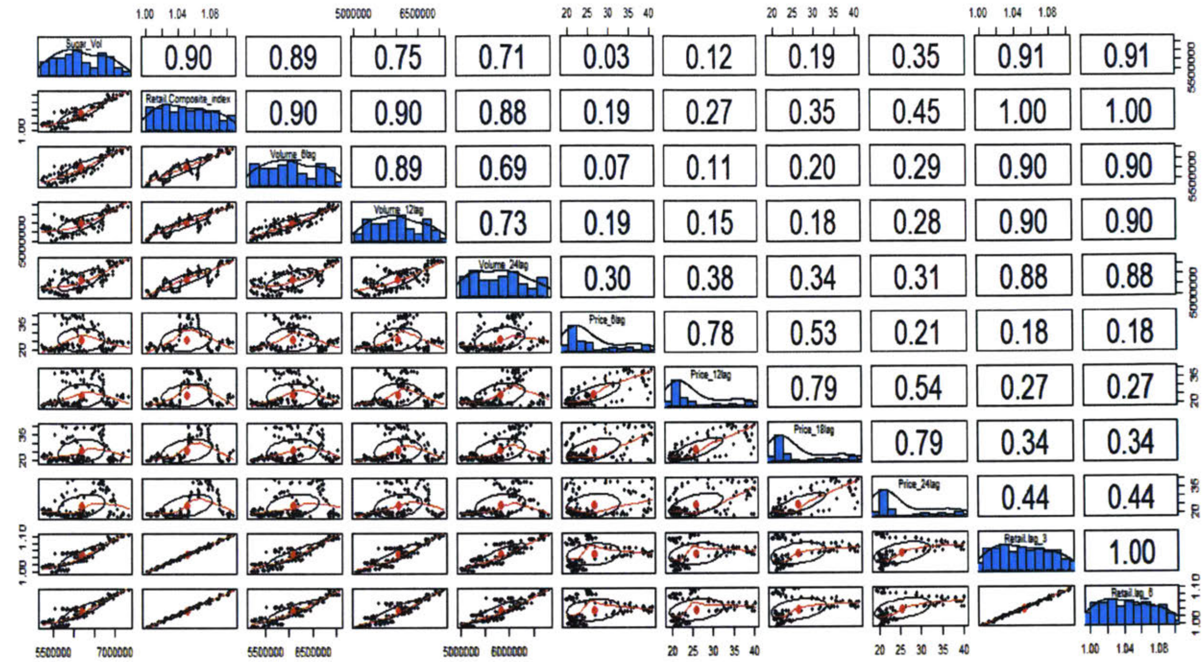


Figure 12 - Pair Panels of Bivariate Correlation and Pearson Correlation Coefficient (PCC) Autoregressive Model

Equally to previous section, we calculated the distribution of pairs of variables and their linear regression results and estimated Pearson Correlation Coefficient[7], demonstrated in Figure 12 above. The same pattern of strong PCC is observed in the retail composite index with sugar volume, including the lag time series.

Figure 12 is similar to that of the previous section, but pertaining to Autoregressive model and its statistical results and correlation between its variables. As we can see, we have high statistically significant coefficients however significantly less residuals than before. This is understandable as sugar commodities markets, in fact agriculture commodities market in general, are very close to random walks and are affected by previous information about consumption pattern, sugar price and volumes. There is a time lag between responses of consumer's responses in the shelf with the pace that sugar commodity is consumed to recompose food processing industry stocks and consequently retail stocks.

As such, the transmission of consumption in the supermarket to commodity industry takes at least 3 to 6 months between different categories of products. Because of limitations on our data sample size from

[7] The correlation coefficient ranges from −1 to 1. A value of 1 implies that a linear equation describes the relationship between X and Y perfectly, with all data points lying on a line for which Y increases as X increases. A value of −1 implies that all data points lie on a line for which Y decreases as X increases. A value of 0 implies that there is no linear correlation between the variables. Source (https://en.wikipedia.org/wiki/Pearson_correlation_coefficient)

Retail Consumers Index, we were not able to identify strong correlation or significance of retail Consumer Index lag above 6 months.

4.4 Artificial Neural Network

As explained in previous sections, the Neural Network models are computational methods that imitate the behavior of the human brain's central nervous system, considered as a class of generalized nonlinear, nonparametric, data-driven statistical method [16,24].

A general neural network architecture consists of an input layer that accepts external information, one or more hidden or middle layer that provides nonlinearity to the model and an output layer that provides the target value, as demonstrated in Figure 2 [16]. Each layer contains one or more nodes. All the layers in a multilayer neural network connect through an acyclic arc.

Time series data can be modeled using neural network in two possible ways. The first way is to explicitly represent time in the form of recurrent connections from output nodes to the preceding layer [15,16]. The second way is to provide the implicit representation of time, whereby a static neural network-like multilayer perceptron is bestowed with dynamic properties [14,16].

A neural network can be converted to a dynamic model by embedding long-term or short-term time memory, depending on the model attributes to be represented. A simple utilized in this study was through the use of time delay series into the input layer of Neural Network model.

The neural network structure for a particular problem in time series prediction includes determination of the number of layers and total number of nodes in each layer [16,34], which is usually determined through experimentation of the given data since there is no theoretical framework for assessing these parameters. It has been proved in many studies that neural networks with one hidden layer can approximate any nonlinear function given sufficient number of nodes at hidden layer and adequate data points for training [16,29,30,34]. After testing many different layers for our data set, the Neural Network model we present use 3 hidden layer.

In time series analysis, the determination of the number of input nodes which are lagged observations of the same variable plays a crucial role as it helps in modeling the autocorrelation structure of the data [34].

The general expression for a multilayer feed-forward time-delay neural network is given by the following function:

$$Y_{t+1} = g\left(\sum_{j=0}^{q} \infty_j \, f\left(\sum_{i=0}^{p} \beta_{ij}\, y_{t-i}\right)\right)$$

Equation (7)

where $y_{t+1} = \ln(y_{t+1}/y_t)$, f and g denote the activation function at hidden layer and output layer, p is number of input nodes (tapped delay), q is the number of hidden nodes , and β_{ij} is the weight attached to the connection between ith input node and the *j*th node of hidden layer, ∞_j is the weight attached to the connection from jth hidden node to the output node, and Y_{t+1} = is the ith input (lag) of the model. [16,34]

Each node of the hidden layer receives the weighted sum of all inputs including a bias term for which the value of input variable will always take a value one. This weighted sum of input variables is then transformed by each hidden node using the activation function f which is usually nonlinear sigmoid function. In a similar fashion, the output node also receives the weighted sum of the output of all hidden nodes and produces an output by transforming the weighted sum using its activation function g. In time series analysis, *f* is often chosen as logistic sigmoid function and *g* as an identity function, that always returns the same value that was used as its argument. [16,34]

For *p* delay nodes, *q* hidden nodes, one output node and biases at both hidden and output layers, the total number of parameters (weights) in a three-layer feed-forward neural network is $q(p + 2) + 1$.[16,24]

For a time series forecasting problem the past observation of variables are used as the input variables, therefore the Neural Network model will attempt to produce the following function:

$$y_{t+1} = f\left(y_t, y_{t-1} \dots y_{t-p+1}, w\right) + \epsilon_{t+1}$$

Equation (8)

where y_{t+1} refers to the observation at time $t + 1$, p is the number of lagged observation, w is the vector of network weights, and ϵ_{t+1} is the error term at time $t + 1$. Hence, the neural network acts like nonlinear autoregressive model. [16,34]

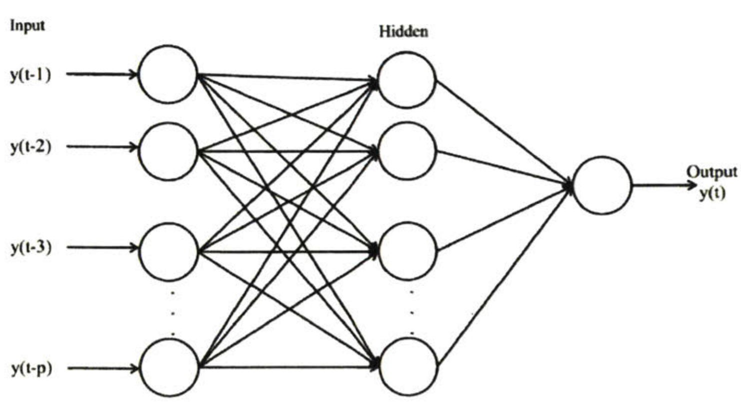


Figure 13 – Example of Neural Network with Time Delay series with one hidden layer mechanism.

Procedure:

- Equally to Autoregressive model we use the 12-month trailing Sugar Volume (Y) as our independent variable to smooth the effects of monthly seasonality embedded to Sugar Domestic Volumes Deliveries.

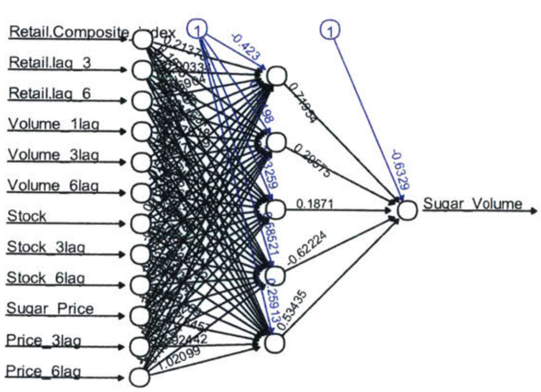
- For the purpose of this thesis, and in line with previous models, the first 7 years (from Jan 2006 to Dec 2012) is used to train the model, so that forecasting starts at the beginning of 2013 and the data from Jan 2013 to Dec 2015 is used as out-sample data in order to compared the accuracy of our prediction models.

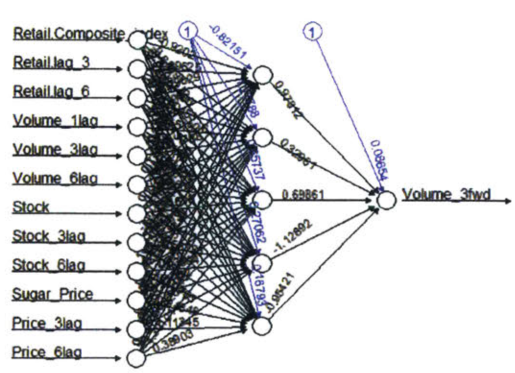
- The Neural Net model in this thesis adopts the resilient backpropagation with weight backtracking. The output layer with one layer having only 1 node is sugar volume forecasting.

- Many literatures demonstrate that the number of layers in hidden layer is often determined by trial and error method, so this thesis carries out trial and error method to select the number of layers in hidden layer by applying R software; according to the results of trial and error method, we have selected the model has the better results. Thus, this research Neural Net model adopts with 20 nodes; where the input layer with 12 layers, having 5 nodes and the input variables considered are demonstrated in Figure 14;

- Following the literature and others, the logistic and identity functions were used as activation function for the hidden nodes and output node, respectively [16,34]. We have focused primarily on the one-step-ahead forecasting.

- Like the number of hidden layers, the number of employed delays per each variable were determined through experimentation, therefore different numbers of neural network models were tried for each series before arriving at the final structure of the model, which results we discussion in the next section.





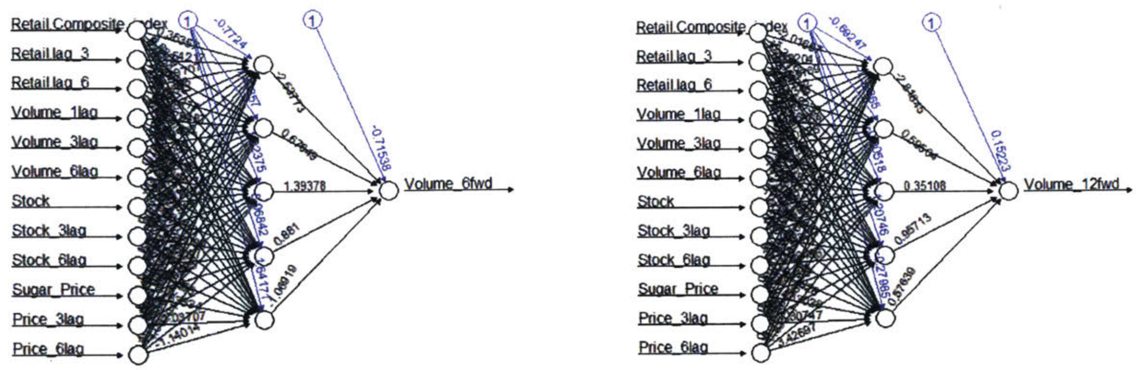


Figure 14 – Sugar Neural Network Models Structure.

4.5 Performance Measurement

The forecasting ability of the three models presented, and its variations to exclude the Retail Composite index, in the previous section is assessed with respect to three common performance measures: (i) the root mean squared error (RMSE), (ii) mean absolute deviation (MAD) and (iii) Correlation of test data predicted value with actual test data value.

The RMSE measures the overall performance of a model is given by:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y})^2}$$

Equation (8)

Where y_t is the actual value at time *t* and $\hat{y}$ is the predicted value for time *t,* where we have *n* as the total number of predicted values.

The MAD deviation, represents a measure of average error for each point predicted and the MAPE is the percentage representation of this error. It is mathematically given by:

$$MAD = \frac{1}{n}\sum_{t=1}^{n}[y_t - \hat{y}]$$

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{[y_t - \hat{y}]}{\hat{y}}$$

Equation (9)

Where y_t is the actual value at time t and $\hat{y}$ is the predicted value for time t, where we have n as the total number of predicted values.

The issue of finding a parsimonious model is taken into account while selecting the best model for each price series.

5. Key Findings

Table 2 – Overall Models Performance

	Prediction horizon (H)	Montly - 12 Month Trailing Values: Models BS	NN	ARMA	SLR	Year-end (December) Values: Models BS	NN	ARMA	SLR
RMSE	1	465,340	104,651	145,799	219,073	602,404	87,652	178,592	209,425
	3	522,261	198,630	380,602	206,240	694,745	233,231	439,541	204,308
	6	574,629	289,261	418,473	380,360	623,431	172,168	422,561	443,490
	12	624,108	251,616	141,322	394,046	691,775	270,559	58,086	466,415
MAE	1	431,230	81,586	129,599	171,198	585,588	81,417	158,245	458,527
	3	468,827	161,952	341,090	172,889	651,023	228,238	402,401	458,778
	6	504,016	236,294	332,729	338,434	592,728	160,284	358,683	903,935
	12	555,632	180,462	118,140	319,443	677,840	169,198	49,255	973,517
MAPE (%)	1	6.7%	1.2%	14.9%	2.5%	9.1%	1.2%	2.3%	2.2%
	3	7.2%	2.3%	15.5%	2.5%	10.1%	3.2%	6.2%	2.2%
	6	7.7%	3.3%	15.7%	4.6%	9.1%	2.2%	5.3%	4.2%
	12	8.5%	2.5%	15.4%	4.3%	10.5%	2.3%	0.7%	4.3%
OSR^2	1	0.79	0.99	0.98	0.95	(0.75)	0.99	0.96	0.97
	3	0.75	0.96	0.87	0.96	(1.16)	0.96	0.39	0.97
	6	0.72	0.93	0.85	0.88	(0.68)	0.98	0.65	0.91
	12	0.69	0.95	0.98	0.87	(2.18)	0.95	1.00	0.91

Table 3 – Table of Performance Comparison of Models – Training set and Out-of-Sample Data

	Prediction horizon (H)	Models: NN	ARMA	SLR
Training R2	1	0.97	0.967	0.541
	3	0.9649	0.8785	0.5633
	6	0.9889	0.8959	0.6758
	12	0.98	0.8906	0.755
Training OSR	1	0.99	0.98	0.95
	3	0.96	0.87	0.96
	6	0.93	0.85	0.88
	12	0.95	0.98	0.87

The Table 2 gives the comparative results for all the different models compared to the baseline with respect to the post-sample RMSE ratio. We can observe that for all different time series ratios, the smaller are forecast horizons of 1, 3, and 6 months.

With respect to the post-sample RMSE ratio, the models in Table 2 demonstrate better performance of NN over the other models.

However, we observe that the ARIMA model performs better than the NN model for the forecast horizon of 12 months as the RMSE ratio is lowest with respect to all series. A relevant aspect for the industry is to be able to capture signals of long-term consumption, mainly for the annual total consumption. The monthly data, although useful, are less relevant metric for long-term view, especially due to the variability of monthly deliveries associated with seasonality.

From Table 2 and 3 we infer that the NN model in general, including Retail Scanner's Data, performs better in all the cases. To demonstrate the predictability power of a model we consider the Out-of-sample R^2 (OSR^2) as one of our metrics. We observe that on average the out-of-sample models have performed better than base-line models, which does not consider Retail Scanners Data.

The Table 4 below demonstrates the significance of Retail Scanners Data and the other variables coefficients, and a parametric test (t-test).

Table 4 – Regression Coefficients of ARIMA and SLR

	ARIMA							
Prediction Horizon	1 months		3 months		6 months		12 months	
Variable	t-value	p-value	t-value	p-value	t-value	p-value	t-value	p-value
Intercept	-5.84	<0.001	-14.5	<0.001	-15.05	<0.001	-16.57	<0.001
Retail Composite Index	6.4	<0.001	18.37	<0.001	4.01	<0.001	19.1	<0.001
Sugar Volume lagged 1 month	17.17	<0.001	-	-	-	-	-9.911	<0.001
Sugar Volume lagged 3 month	-	-	-	-	-8.1	<0.001	-	-
Sugar Volume lagged 12 month	-7.5	<0.001	-13.46	<0.001	-14.6	<0.001	-10.46	<0.001
Sugar Volume lagged 24 month	-3.97	<0.001	-7.31	<0.001	-8.47	<0.001	-	-
Price lagged 6 month	3.69	<0.001	4.4	<0.001	4.25	<0.001	-	-
Price lagged 18 month	-4.21	<0.001	-5.7	<0.001	-5.24	<0.001	-	-

	LR							
Prediction Horizon	1 months		3 months		6 months		12 months	
	t-value	p-value	t-value	p-value	t-value	p-value	t-value	p-value
Intercept	-5.9	<0.001	-6.08	<0.001	-6.06	<0.001	-6.914	<0.001
Retail Composite Index	8.32	<0.001	8.4	<0.001	8.133	<0.001	8.995	<0.001
Price lagged 12 month	-	-	-	-	-4.87	<0.001	-4.1765	<0.001
Price lagged 18 month	-	-	-3.02	<0.01	-	-	-	-
Price lagged 24 month	-3.39	<0.001	-	-	4.31	<0.001	4.64	<0.001

As we can see the Retail Scanner Data has a strong significance to predict sugar volume consumption variables in many different time lags, which is consistent with the correlation matrix Figure 12 and 15.

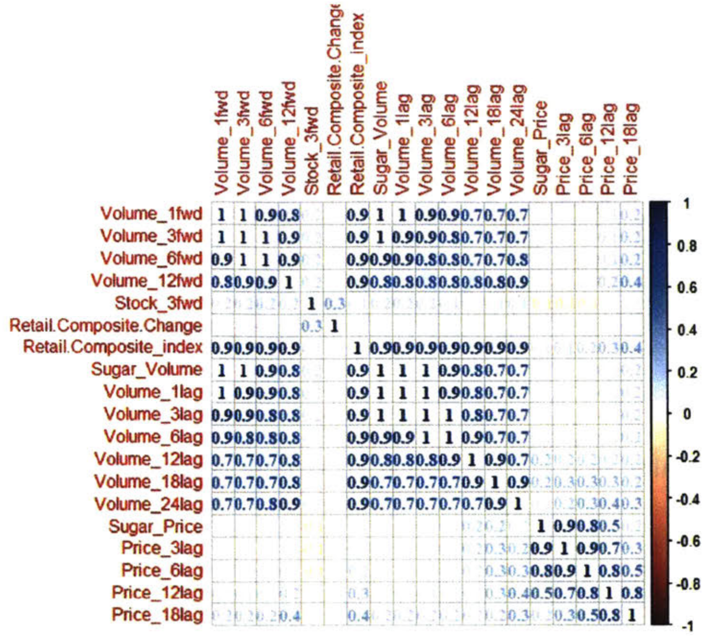


Figure 12 – Correlation Matrix of Sugar Model Variables

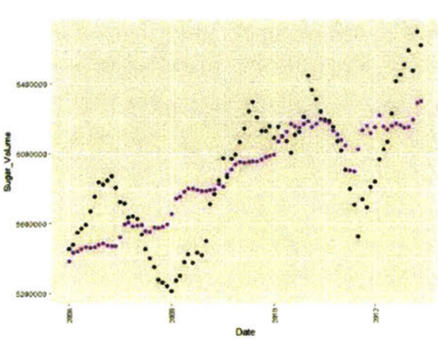
It is evident from the Table 2 and 3 that, the NN model performs better in all the cases, except in the case of forecasting 12 months ahead, while ARIMA models perform better for 12 months. At this juncture, it is worth mentioning that for all cases the best NN model in terms of the test RMSE is obtained for a forecasting horizon of 12 months, and the same model is used for other forecast horizons.

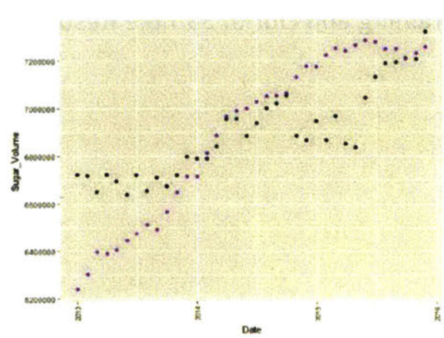
In this context, several researchers [15, 21] have recommended that a specific NN model should be selected for each forecast horizon, which implies that p and q may vary over forecast horizons [34,26]. However, one of the limitations of NN is its limited capacity for interpretability.

This will, in general, improve the performance of NN model with respect to each forecast horizon. Such approach is not of much advantage in case of ARIMA model [34].

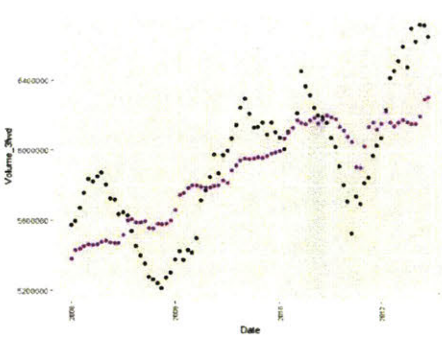
1month forward – Training Set **1 month forward – Out-of-sample**

**Black dots are Actual values and purple dots are predicted*

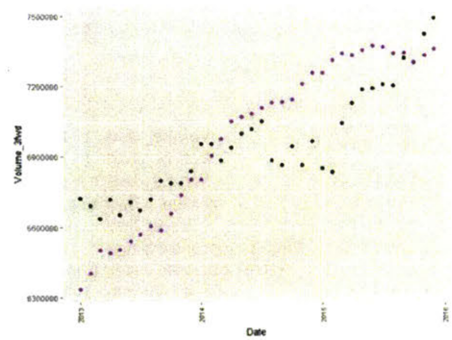




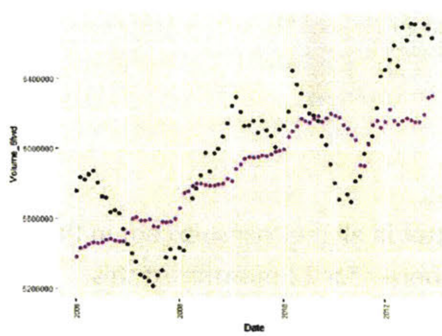
3 months forward – Training Set



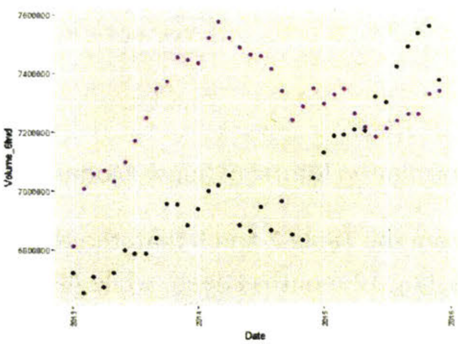
3 months forward – Out-of-sample



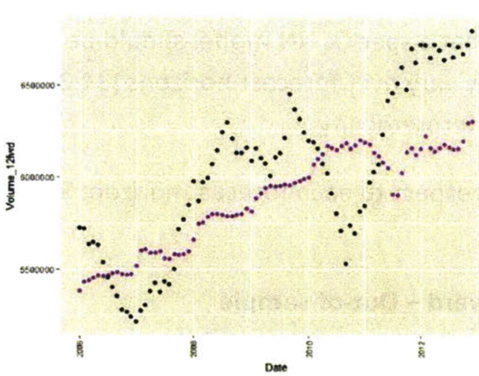
6 months forward – Training Set



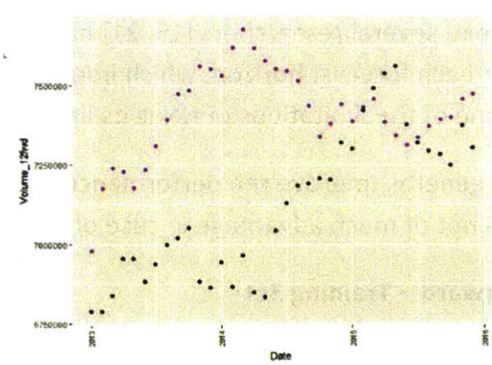
6 months forward – Out-of-sample



12 months forward – Training Set

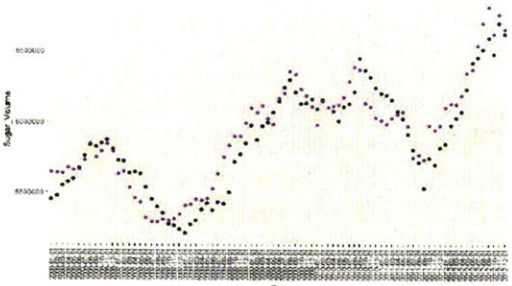


12 months forward – Out-of-sample

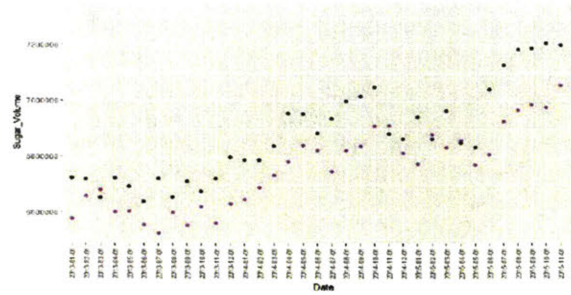


Figures 16 –SLR Training and Out of Sample Results

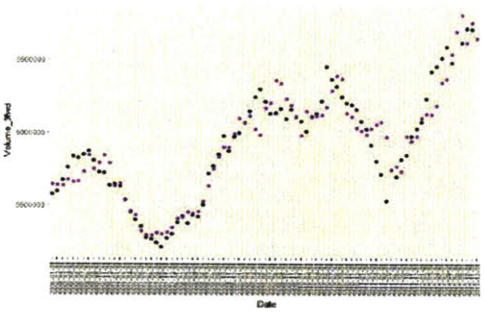
1month forward – Training Set



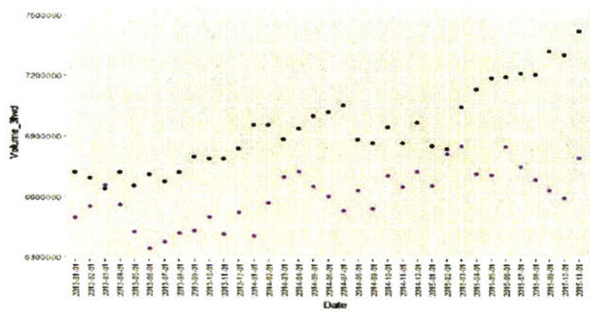
1 month forward – Out-of-sample



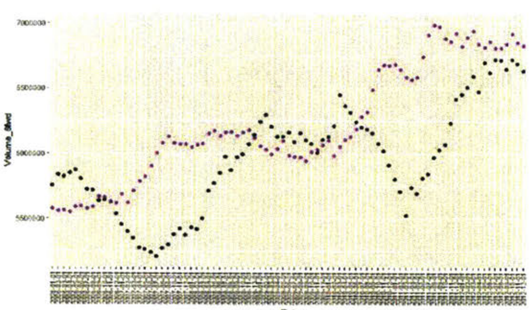
3 months forward – Training Set



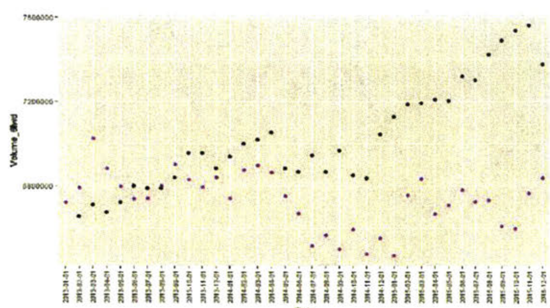
3 months forward – Out-of-sample



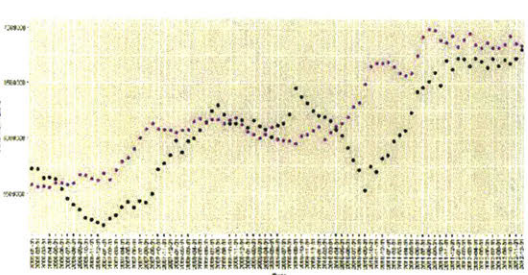
6 months forward – Training Set



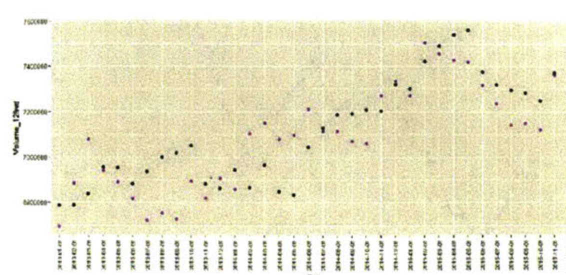
6 months forward – Out-of-sample



12 months forward – Training Set

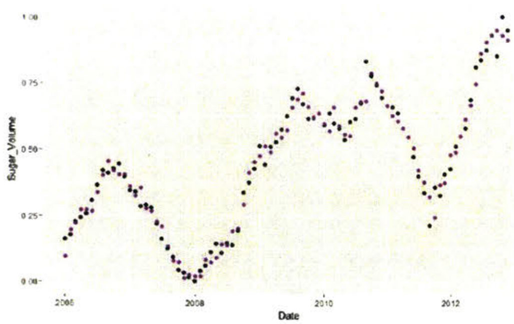


12 months forward – Out-of-sample

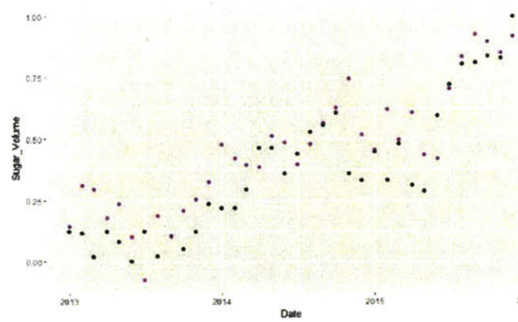


Figures 17 –ARIMA Training and Out of Sample Results

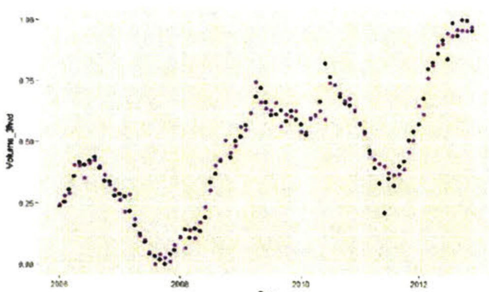
1month forward – Training Set



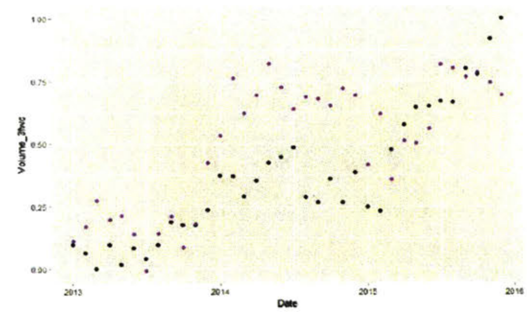
1 month forward – Out-of-sample



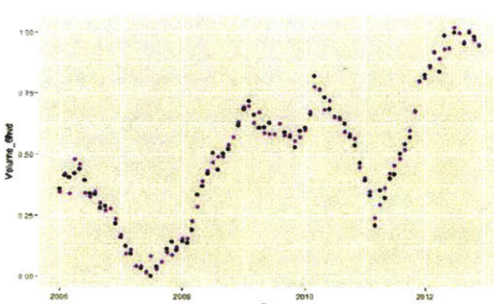
3 month forward – Training Set



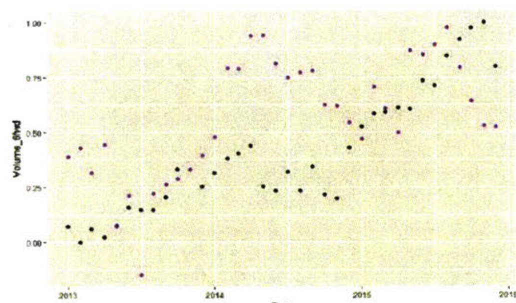
3 month forward – Out-of-sample



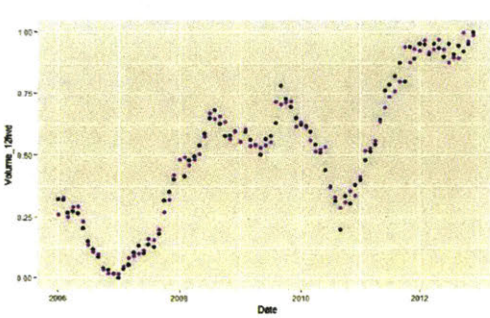
6 month forward – Training Set



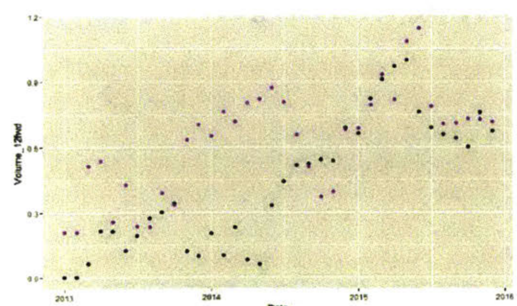
6 month forward – Out-of-sample



12 month forward – Training Set



12 month forward – Out-of-sample

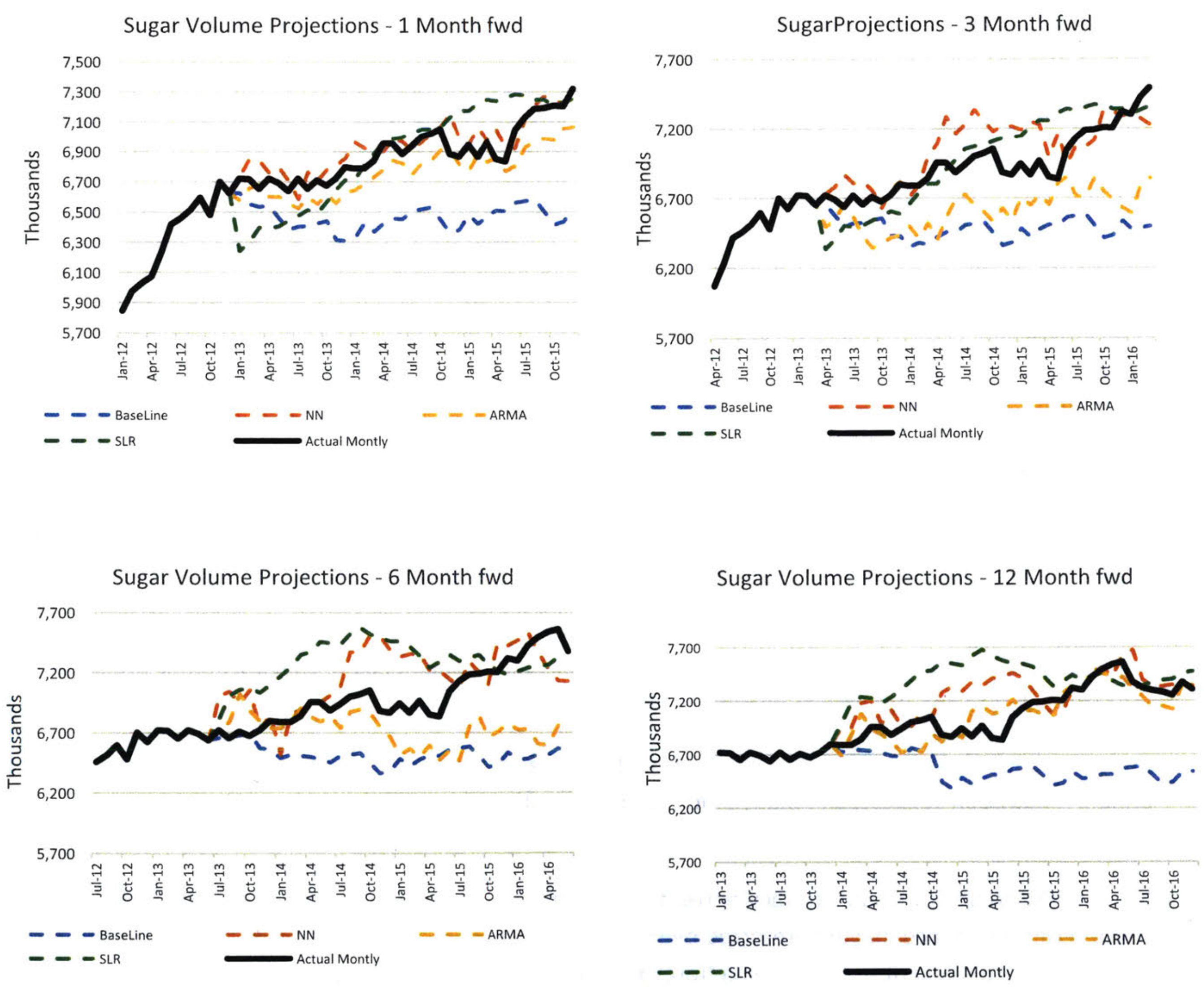


Figures 18 –NN Training and Out of Sample Results

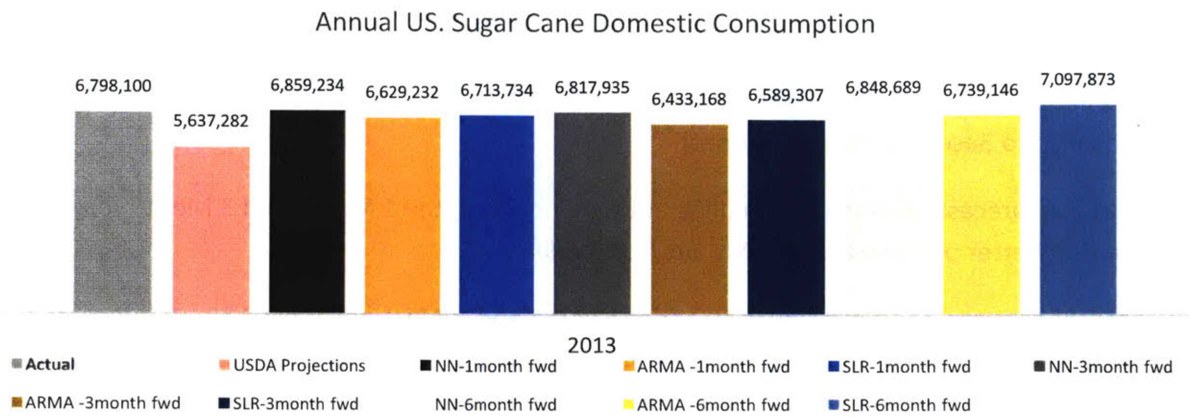
The observations of projected results compared to the different model perfromance, demontrated in Figure 19 bellow, shows a stronger performance of NN and ARIMA Models considering Retail Scanners Data, compared to Base Line model approach.

Concerning the forecast performance in different time horizons, the 1 Month and 3 Month forward time horizon have a better performance for NN and SLR models.

For 12 Months forward, the both Figure 19 and the RMSE and MAPE from Table 2 above suggests a better performance of ARIMA model over other methodologies for stronger non-linear series .



Figures 19 – Out-of-Sample projected results from different models



Figures 20 – Comparison of Machine Learning Model Results vs Actual and USDA estimates for 2013 U.S Sugar Cane Domestic Consumption

5.1 Summary of Results

This study has compared ARIMA , SLR and NN models, in terms of both modeling and forecasting using weekly Retail Scanners data and wholesale data of a equivalent basket of products that contains the underline commodity, which in this case was sugar.

The results demonstrated by this study show a stronger predictability performance of models by utilizing Retail Scanners Data associated with Machine Learning methodologies, than conventional industry Base Line methodologies.

This study suggests that before adopting any nonlinear model, one needs to check whether the series is indeed nonlinear and its correlation with volume consumption.

The NN model performance was better than ARIMA for all forecast horizons except for 12 months ahead. It has been suggested in many studies, that an optimum model should be selected for each forecast horizon.

NN and ARIMA perform substantially better than linear models in predicting the direction of change for these series and hence may be preferred over linear models in the context of predicting turning points, which are more relevant in case of price forecasting.

Many of the results of these studies align with conventional wisdom and traditional economic thinking. However, by producing very specific and detailed measures of food demand behavior, they offer superior forecasting potential, otherwise done by traditional methodologies used by companies in the industry. This is valuable information that contributes significantly to commodities and food companies production and investment strategy, food policies and market trade.

5.2 Model Extensions

There exist several potential extensions of this research. The proposed model to utilize Retail Scanner's Data as an important variable to predict commodities demand could be applied to other agricultural commodities.

The Retail Scanner's Data may be adequate for describing long-term trends of commodity consumption and could be used across many commodities markets. By extension, a hybrid model could be used to forecast the overall supply and demand effects and assessment of general logistics and productive industry capacity compared to future predicted demand.

Subsequently, it is recommended to evaluate the incorporation of sentiment analysis from web search trends. A large dataset of people's searches about food is posted online on Google Trends Website. This is an important untapped source of data that provides important insights about peoples behavior towards food consumption trends over the long run. Nevertheless, the collection and interpretation of this data is not trivial because its high correlation may not mean causation. Besides, historical long-term data, which would provide a substantial data set for historical inferences of people's search translating to actual food consumption, is not provided.

Nevertheless, the continuous and consistent collection of that data over time in the future could provide a powerful analysis of food consumption trends and, and make the decision making process of food and commodities industry players more efficient.

5.3 Conclusions

This thesis proposes a new forecasting system for commodities demand based on Machine Learning approaches and using the Retail Scanner's Data, as one of the data parameters. Rather than using single models, this thesis purposes ensemble methods to combine different methods and have them collaborate to build a comprehensive model.

This thesis compared NN, SLR and ARIMA models to forecast U.S sugar commodity annual trailing demand. Results show that forecasts from NN were considerably more accurate than those of traditional ARIMA, SLR and baseline models, which were used as a benchmark, except for predicting the 12 months time horizon, where the ARIMA model had better results.

The RMSE, MAD, and mean absolute percent error were all lower on average for the NN forecast than for the ARIMA. The means of the NN and ARIMA forecasts were also found to be statistically different.

The reason the NN model performed better than the ARIMA may be that the data contain non-linear behavior, which cannot be fully captured by the linear ARIMA model. The empirical results with the sugar data indicate that the combined models can be an effective way to achieve more forecasting accuracy than either of the models used independently.

Agriculture demand information is necessary for decision-making at all levels, and it is relevant information considering the globalization and high level of market integration. The utilization of Retail Scanner's Data and other alternative data points associated with Machine Learning can improve significantly the predictability of world sugar market demand. Future research should continue exploring the use of big data to predict the agricultural commodities supply and demand.

Bibliography

[1]Pirrong, Craig (2014). The Economics of Commodity Trading Firms . *White Paper.* https://www.trafigura.com/media/1364/economics-commodity-trading-firms.pdf

[2]Dani, Samir. (2015). Food Supply Chain Management and Logistics - From Farm to Fork - 1. Introduction to Food Supply Chains. (pp. 1-16). Kogan Page Publishers. Online version available at: https://app.knovel.com/hotlink/pdf/id:kt010YOY45/food-supply-chain-management/introduction-food-supply

[3]Alexandratos, N. and J. Bruinsma. 2012. World agriculture towards 2030/2050: the 2012 revision. ESA Working paper No. 12-03. Rome, FAO. http://www.fao.org/docrep/016/ap106e/ap106e.pdf

[4]United Nations, Department of Economic and Social Affairs, Population Division (2017). World Population Prospects: The 2017 Revision, Key Findings and Advance Tables. Working Paper No. ESA/P/WP/248. https://esa.un.org/unpd/wpp/Publications/Files/WPP2017_KeyFindings.pdf

[5] Clements, Michael P., and David F. Hendry. 1999. Forecasting Non-stationary Economic Time Series. Cambridge, Mass: MIT Press, 1999. eBook Collection (EBSCOhost), EBSCOhost (accessed April 15, 2018).

[6] Huffaker, Ray, Marco Bittelli, and Rodolfo Rosa. Nonlinear Time Series Analysis with R. n.p.: Oxford University Press, 2017. Oxford Scholarship Online, EBSCOhost (accessed April 15, 2018).

[7] John H. Cochrane. Time Series for Macroeconomics and Finance. 1997.

[8] K.W. Hipel and A.I. McLeod. Time Series Modelling of Water Resources and Environmental Systems. 1994.

[9] C.F. Lee, J.C. Lee, and A.C. Lee. Statistics for Business and Financial Economics. World Scientific Publishing Co. Pte. Ltd, 1999.

[10] Box, George E.P. ; Jenkins, Gwilym M. and Reinsel, Gregory C. ;Time series analysis : forecasting and control , Englewood Cliffs, N.J. : Prentice Hall, c 1994.

[11] R.O. Otieno J.M. Kihoro and C. Wafula. Seasonal time series forecasting: A comparative study of arima and ann models. African Journal of Science and Technology (AJST) Science and Engineering Series, 2004.

[12] Chaudhry, Akram M., and Irfan Ahmed. "Identification of Optimum Statistical Models for Time Series Analysis and Forecasting using Akaike's information Criterion and Akram Test Statistic: A Comparative Study." World Congress On Engineering 2007 (Volume 1) (January 2007): 956-960. Academic Search Complete, EBSCOhost (accessed April 15, 2018).

[13] C. Hamzacebi. Improving artificial neural netoworks' performance in seasonal time series forecasting. Information Sciences, 2008.

[14] J. Faraway and C. Chatfield. Time series forecasting with neural networks: a comparative study using the airline data. Applied Statistics, 1998.

[15] G.P. Zhang. Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, 2003.

[16] Jha, G.K. & Sinha, K. Time-delay neural networks for time series prediction: an application to the montly wholesale price of oilseeds in India, Neural Comput & Applic (2014) 24: 563. https://doi.org/10.1007/s00521-012-1264-z.

[17] Linton, Oliver B. and Yan, Yang, Semi-and Nonparametric ARCH Process, Journal of Probability and Statistics Volume 2011, Article ID 906212, 17 pages DOI:10.1155/2011/906212.

[18] H. Tong, Threshold Models in Non-Linear Time Series Analysis. Springer-Verlag, 1983.

[19] G.P. Zhang. A neural network ensemble method with jittered training data for time series forecasting. Information Sciences, 2007.

[20] Bajari, Patrick, Denis Nekipelov, Stephen P. Ryan, and Miaoyu Yang. 2015. "Machine Learning Methods for Demand Estimation." American Economic Review, 105 (5): 481-85. DOI: 10.1257/aer.p20151021.

[21] Kourentzes, Nikolaos ; Barrow ,Devon K. and Crone ; Sven F. ; Neural network ensemble operators for time series forecasting Expert Systems with Applications, Volume 41, Issue 9, 2014, pp. 4235-4244.

[22] Crone SF, Hibon M, Nikolopoulos K ; Advances in forecasting with neural networks? Empirical evidence from the NN3 competition on time series prediction., 2011 , Int J Forecast 27:635–660.

[23] Zhang G, Patuwo BE, Hu MY , Forecasting with artificial neural networks: the state of the art., 1998, Int J Forecast 14:35–62.

[24] Cheng B , Neural networks: a review from a statistical perspective. Statistical Science, 9:2–54 v.9 : no.1, 1994, p. 2.

[25] Agrawal, R.K and Adhikari, Ratnadip, An Introductory Study on Time Series Modeling and Forecasting, http://www.realtechsupport.org/UB/SR/time/Agrawal_TimeSeriesAnalysis.pdf,

[26] J. Cryer and K. Chan. Time Series Analysis with Applications. R. Springer, 2008.

[27] Sharda R, Patil RB ,Connectionist approach to time series prediction: an empirical test. J Intell Manuf 3:317–32, 1992.

[28] Kohzadi N, Boyd MS, Kermanshahi B, Kaastra I , A comparison of artificial neural network and time series models for forecasting commodity prices, 1995, Neurocomputing 10:169–181.

[29] Aliahmadi, A., Jafari-Eskandari, M., Mozafari, A., & Nozari, H. (2013). Comparing artificial neural networks and regression methods for predicting crude oil exports. International Journal of Information, Business and Management, 5(2), 40-58. Retrieved from https://search-proquest-com.libproxy.mit.edu/docview/1511436335?accountid=12492

[30] Crone SF, Hibon M, Nikolopoulos K , Advances in forecasting with neural networks? Empirical evidence from the NN3 competition on time series prediction, 2011, Int J Forecast 27:635–660

[31] James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. An Introduction to Statistical Learning: With Applications in R., 2013. Print.

[32] Lu, Z. Q. John. "The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edition." Journal Of The Royal Statistical Society Series A-Statistics In Society 173, (2010): 693-694.

[33] Moriya, N. (2008). Fengshan Yang, Noise-Related Multivariate Optimal Joint-Analysis in Longitudinal Stochastic Processes , Nova Science Publishers, Inc.: 223–260. ISBN 978-1-60021-976-4.

[34] Girish K. Jha and Kanchan, Sinhab , Agricultural Price Forecasting Using Neural Network Model: An Innovative Information Delivery System, Agricultural Economics Research Review Vol. 26 (No.2) July-December 2013 pp. 229-239.

[35] Morrison, K. (2008). Living in a material world : the commodity connection. Chichester, England ; Hoboken, NJ : Wiley, c2008.

[36] Regmi, Anita, and Birgit Meade. "Demand side drivers of global food security." Global Food Security 2, (September 1, 2013): 166-171. ScienceDirect.

[37] Valin, Hugo. "The Future of Food Demand: Understanding Differences in Global Economic Models." Agricultural Economics 45, no. 1 (January 2014): 51-67.

[38] Sugar and Sweeteners Outlook Special Article/SSS-M-296SA/April 16, 2013 Economic Research Service, USDA (http://usda.mannlib.cornell.edu/usda/ers/SSS//2010s/2013/SSS-04-16-2013.pdf)

[39] Balk, Bert. "A simple method for constructing price indices for seasonal commodities." Statistische Hefte: Internationale Zeitschrift Für Theorie Und Praxis 22, no. 1 (March 1981): 72.

[40] Diewert, W.E, 1995. Axiomatic and economic approaches to elementary price indexes. Discussion Paper No.95-01. Department of Economics, University of British Columbia.

[41] Ivancic, Lorraine, W. Erwin Diewert, and Kevin J. Fox. "Scanner data, time aggregation and the construction of price indexes." Journal Of Econometrics 161, no. Measurement with Theory (January 1, 2011): 24-35. Science Direct.

[42] USDA, ERS, Indeterminacy in Measuring U.S. Sugar Deliveries for Human Consumption, http://usda.mannlib.cornell.edu/usda/ers/SSS//2010s/2013/SSS-04-16-2013.pdf, April 2013.