**Validating Forecasting Models in the Energy Industry:**

**A Proposed Strategic Forecasting Framework**

by

**Manoj Dev Mahajan**

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

**Doctor of Philosophy**

in

**Technology, Policy, and Innovation**

Stony Brook University

**May 2020**

**Stony Brook University**

The Graduate School

**Manoj D. Mahajan**

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation.

**Robert J. Frey**
**Research Professor, Applied Mathematics & Statistics**

**Richard Chan**
**Associate Professor, College of Business**

**David Tonjes**
**Graduate Program Director, Department of Technology and Society**

**Shinjae Yoo**
**Adjunct Assistant Professor, Institute of Advanced Computational Science, Stony Brook University & Scientist, Computational Science Initiative, Brookhaven National Laboratory**

This dissertation is accepted by the Graduate School

Eric Wertheimer

Dean of the Graduate School

Abstract of the Dissertation

**Validating Forecasting Models in the Energy Industry:**

**A Proposed Strategic Forecasting Framework**

by

**Manoj D. Mahajan**

**Doctor of Philosophy**

in

**Technology, Policy, and Innovation**

Stony Brook University

**2020**

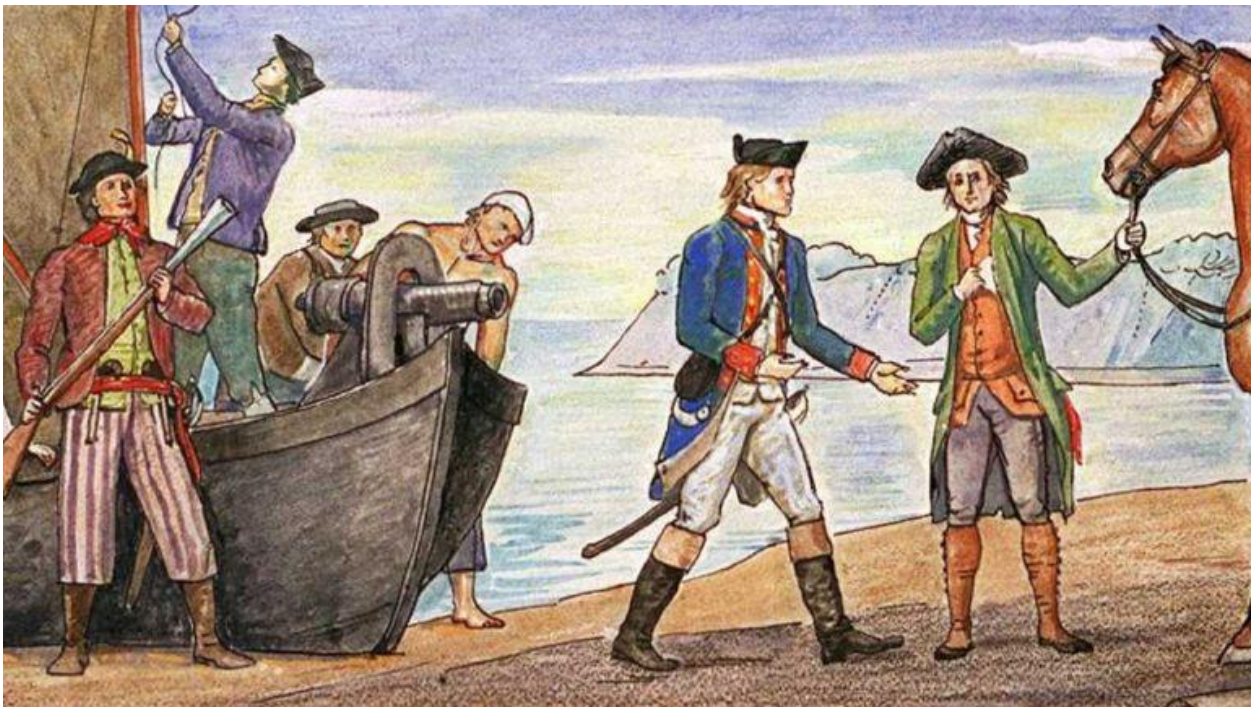
Many of the forecasts done today in finance, venture capital, and research & development institutions are becoming more complex and there is a need for a strategic validation process due to increased incertitude. This dissertation proposes a strategic multi-disciplinary forecasting validation framework to assist a data scientist or decision-maker in properly executing a forecasting model with their data set. The simulations shown conclude that there can be vast differences in forecast outputs from leaving parameters at default or 'out-of-the-box'. By changing simple and optional parameters, a decision-maker or data scientist can interpret, assess, and validate the proper outputs for the real-world answers they seek. The results show that even if the model is mathematically sound, other factors such as ETL (extract, transform, load), smoothing, or 'cleaning' of data can impact outputs. In addition, the order of magnitude, amount of data points, and visual displaying of the output forecast can also skew the results. The goal of this research is to ensure the decision-maker or data scientist can truly understand the lineage of their data as it changes through the process of data ingestion, preparation, and in the model to create the output. Through validation at each step, this dissertation intends to answer the question: How reliable are these models and how can they be validated to discover the best model for the data? The framework is editable, multi-disciplinary, and can work with basic and advanced forecasting approaches such as implementing neural networks. It can also assist a decision-maker in an implementation of the framework to create a validation-centric business process for forecasting to further data science quality in an organization and avoid the generic and default outputs that may be similar to their competitors.

## Dedication Page

This dissertation is dedicated to the late Dr. David L. Ferguson. He recruited me into the Ph.D. Program at SUNY Stony Brook upon my return from a deployment to Afghanistan during Operation Enduring Freedom and helped enable me to work on my degree over the years while working full-time. He was an upstanding, invaluable, sincere leader and mentor and will be missed by everyone he mentored and influenced.

It is also dedicated to my family and friends who have been supportive over the years.

**Frontispiece**



“A rendering by Allyn Cox, for a mosaic at Ward Melville High School, shows Austin Roe delivering a message to Caleb Brewster on the shore at Crane Neck Bend in Old Field. The message would be carried across the Long Island Sound by whaleboat to Fairfield Connecticut to be handed over to Maj. Benjamin Tallmadge.” Credit: Newsday / Newsday File Photo

# Table of Contents

## List of Figures

## List of Tables

## List of Equations

## List of Abbreviations

ANN – Artificial Neural Network

CNN – Convolutional Neural Network

CRNN – Convolutional Recurrent Neural Network

DARPA – Defense Advanced Research Projects Agency

ETL – Extract, Transform, Load

GAN – Generative Adversarial Network

NIST – National Institute of Standards and Technology

RNN – Recurrent Neural Network

SVM – Support Vector Machines

MLP – Multi-Layer Perceptron

ARIMA – Autoregressive Integrated Moving Average

LSTM – Long Short-Term Memory

DTS – Department of Technology and Society

SBU – Stony Brook University

## Preface

The decision to research this topic was due to having a career in the U.S. Government in multiple roles from analyst, venture capitalist, technologist, program manager, and research scientist realizing the difficulties of decision-making from the office to the battlefield in each of these positions. Using models to predict the next technology, natural disaster, terrorist attack, or forecast a research and development budget which can impact policy makers and American lives cannot be relied on by out-of-the-box forecasting models which are assumed correct.

The energy industry is used in the examples since it encompasses many global issues and policies from future technology investments to product-level prices. A developed, strategic, non-industry specific set of guidelines, or framework for everyday multi-disciplinary forecasting was not apparent in the literature. This dissertation provides a proposed validation framework to assist as a business process that can be easily applied in to any large corporation or small business framework to remind the data scientist or decision-maker of common failure points in hopes to discover the best model for their data set with the most realistic output to inform policy makers.

**Acknowledgments**

Dr. Robert Frey thank you for supporting me and motivating me to continuously learn and diversify my knowledge for the past 15+ years to be a better program director, leader, advisor, researcher, and of course being my advisor alongside the late Dr. David Ferguson.

Dr. Richard Chan thank you for helping me on a weekly basis to focus, write, organize, and ensure that I am relaying my goal of furthering science with my research as well as creating something beneficial to industry and business.

Dr. Shinjae Yoo thank you for being kind enough to share your world-class computer science expertise and always showing me the newest cutting-edge technologies being worked on to better understand architecture and implementation in large enterprises and acquisitions.

Dr. David Tonjes thank you for stepping up to help from the Department of Technology & Society after the loss of Dr. David Ferguson.

Kim Walls, Dale Walden, Dr. John Greer, Dr. Ralph B. James, Dr. Steve Musolino, Susan Pepper, Dr. Andrew Kalukin, Eric Salucci, and my family – Thank you all for supporting, enabling, helping, discussing, and making arrangements so I could complete this Ph.D. over the years back and forth from the Washington D.C. & Arlington, Virginia area to Setauket, Long Island, New York because you all realized that value of continued learning in STEM education fields and diversity of scientific practices.

# Chapter 1. Introduction

## I. Research Question

Forecasting models are used in technology and innovation investments to predict future technology markets and financial success. *How reliable are these models and how can they be validated to discover the best model for the data?*

Through answering the Heilmeier Catechism proposed by former Defense Advanced Research Projects Agency (DARPA) director Dr. George H. Heilmeier, to help evaluate and discern if a proposal is worthy of research (DARPA, 2018), this dissertation will look at validation of *forecasting models* in the following areas:

1. A literature review of historical, current, and future approaches to forecasting.
2. Propose a strategic framework for forecast validation.
3. A simulation using high-risk strategic investments in venture capital (energy sector) and technology forecasting approaches to better evaluate and validate appropriate forecast modeling.
4. A simulation using a specific energy industry subset as a data set – specific commodity chemical price to evaluate and validate forecasting models.
5. An understanding that models, though correct and generally accepted, must be validated at various steps and used appropriately based on the data set to prevent framing, flaws, and decision-making traps in data science. Furthermore, to prevent and mitigate these problems the current and future work of applying neural networks to forecast modeling and predictive analytics will be discussed.

## II. Motivation

Forecasting is used in many disciplines as a tool to attempt to predict the future. Some everyday examples can be a financial planner, intelligence analyst, economist, meteorologist, or supply-chain manager. At a more strategic level, forecasts can be one of the analyses utilized by a CEO, senior government official, policy maker, or organizational leader to influence their decision-making process. In government research and development as well as venture capital, forecasting and predictions of how to operate funding and invest in new technologies are imperative to deliver new technology or research in a timely fashion. This is important due to rapid and accelerating the innovation of new technologies that can render others obsolete. Venture capitalists may have an advantage of pivoting a company, or changing the product to a more appropriate one over time to meet a new market trend and similarly, government research and development needs to be able to adapt to constant changes of policy and technology requirements for the future. Planning for future technology areas such as satellite constellations, fighter jets, 'as-a-service' data applications, large AGILE software engineering deployments, efficient clean fuel engine technologies, and other disruptive technologies requires careful forecasting on the technology itself as well as the funding or budget. Many of the forecasts used commonly come from 'out-of-the-box' or 'off-the-shelf' default algorithms which are often haphazardly used by many data scientists and reach similar if not the same results. There are many new approaches to forecasting that can differentiate a forecasting result from the others with a more accurate and valid real-world output that can be used to act on by a decision-maker. However, there are many layers of complexity to forecasting that are often ignored since most literature deals with optimal model selection, model accuracy, and new hybrid or combinatorial approaches to forecasting that are industry or problem specific. There is a need for a strategic

forecasting validation framework that can help guide a junior or senior data scientist to a better understanding of the entire forecasting cycle from a raw data set in a specific industry, through the data preparation phase, to ingesting the data into the model, how the model processes the data, and how to display or interpret the results based on testing different parameters. Focusing on the lineage of the data set, a data scientist can better understand how their data passed through the forecasting model or process to help validate an output's real-world use.

## III. Forecasting Overview

Forecasting is used in a diverse number of fields such as: technology forecasting, finance, epidemiology, weather reporting, earthquake prediction, intelligence analysis, the gas, oil, and energy industry, and global currency markets among others. Research in applying and improving forecasting methods and prediction models in these fields is continuous since these models have many flaws, traps, and biases and they are used by decision-makers in most facets of society such as research and development, government, business, technology innovation, engineering, and finance. Forecasting and predictive analytics differ from regular modeling due to the estimation of the future using time series data which can come in various periods (such as quarterly or seasonally), which is why the need to validate the model, the data, and the relationship of the two is important. A definition of time series forecasting that is generally accepted in the literature is: "Time series forecasting is the projection of desired number of future values through some mathematical model. In [a] usual forecasting paradigm, an appropriate model is identified from a particular class and it is then used to generate the future values" (Adhikari, 2015). Furthermore, ensemble, hybrid, or combined forecasts apply other models to be used to create a more accurate forecast which can be found in the current literature. These hybrid approaches can be executed by creating a forecast based on multiple forecast models, using other statistical algorithms and existing models, preparing the data in a specific way, using machine learning and neural networks, or a combination or implementation of these methods to yield a more accurate or realistic result.

Today, many academic researchers and financial corporations use forecasting tool suites included in software packages, in-house developed forecasting tools and models, generally accepted industry-specific approaches for forecasting certain types of data set, or statistical software tools such as R, SPSS, and MATLAB to perform their forecasting. Many of these

forecasting models become a standard or create a business process that turns into a "this is the way we have always done forecasting" paradigm. This can be problematic as the scientific literature and progress of forecasting is continuously evolving. Moreover, there are issues with version control and forced-updating to some of these software packages that can cause small variations in outputs on the same data set. Currently, with the general acceptance of using cloud services, big data, neural networks, and other advances in computer science, much of the literature incorporates new techniques from other disciplines to improve forecasting models. Due to the advances in computing, many of these models can be implemented by a single user on a PC in a small business or from home rather than requiring a large information technology infrastructure such as a computing cluster. Moreover, for large scale forecasting models with complex data, researchers or financial corporations are now able to either maintain their own computing clusters or purchase cloud-based solutions from various service providers such as Microsoft Azure, Amazon AWS, Google Cloud AI, and IBM Watson. To reiterate, many of these providers include 'out-of-the-box' generic or generally accepted algorithms and protocols for forecasting that may have never been validated or assessed for a specific industry or discipline. Furthermore, special handling of the data being used with the model for a certain researcher or decision-maker may be required and an understanding of the data lineage, or how it is transformed, may be important in many instances.

There are strong applied mathematics and computer science components to modern day forecasting that not only includes the statistical and best practices of forecasting, but also the information technology infrastructure, and neural network framework choices which all require validation for specific forecasting outputs to a single research or financial problem on a case-by-case basis to maintain a level of accuracy for the output.

Forecasting has been a popular area of research since the 1950's but the results have been inconsistent and there is no theorem that can be used to assess a perfect model, or "[determine] the success or failure of a forecasting model" (Lemke & Gabrys, 2010). Historically, there have been traditional and classical ways to investigate time series data by fitting models based on judgement, experience, and recognition of patterns and trends (Lemke & Gabrys, 2010). The decision-maker or data scientist conducting the forecast should interpret the results and assess each stage of the validation process and can provide a valid qualitative assessment that can help verify the use of the results in the real world. An example of forecast model failure is after the financial crisis of 2008, where it was generally believed that emerging markets, such as China, Brazil, Russia, and India, would outpace the United States by 2013. However, the forecast was wrong, and in 2013, the average growth rate of these markets fell back to four percent (Sharma, 2014). This shows how forecasting can cause global economic policy changes and political positioning that is unnecessary due to failed forecasting. It also can lead to policy changes such as this example, which may impact economic or financial policy set by a government. Financial forecast failures are very similar to environmental prediction failures such as earthquakes, natural disasters, weather events, terrorist attacks (nuclear, power grid), and power-grid outages, which are difficult to predict and can cause abrupt policy changes, emergency management procedures, military action, or economic and budgetary adjustments.

These are examples of rare events, however, "black swans" are obvious prediction failures that forecasts will fail to predict, or as Nassim Taleb states, "As a forecasting period lengthens, prediction errors grow exponentially" (Taleb, 2007). These unpredicted and rare events can influence future policy decisions and make it imperative to attempt to discover the next event through forecasting. Furthermore, though many disciplines use forecasting, many of the basic

methods, statistical tools, and processes remain the same. Lemke and Gabrys (2010) created a table of popular and generally accepted time series model options and attempted to discern the 'meta-learning method' which included many of the current forecasting models from 1992 to 2009, most of which use similar methods such as exponential smoothing. Their conclusions were that, "neural networks, decision trees and support vector machines [and] building meta-models in a leave-one-out cross-validation methodology, [ … ] did not lead to convincing results," however their research did have positive results when using a ranking approach (Lemke & Gabrys, 2010). Even though their work is primarily for optimal model selection, it still makes apparent the need for understanding the specific data set and how it will work within the model given the context of the discipline, practice, or industry. For example, a model that works well with venture capital data may not work well in the gas, oil, and energy industry with a specific commodity chemical data. The data sets may have different time series periods, could be 'noisy', may require additional data preparation, or a host of other variables. At the simplest form, the order of magnitude or visual display choices of the data scientist may impact the results such as disregarding a forecast that is actually correct for a less accurate one due to limiting the forecasting period (this will be shown as an example in the simulations).

This dissertation does not focus on optimal model selection, but rather the validation of the chosen model as the data set flows through it to produce an outcome. This is to ensure that the model and data are being used properly within the scope of the data sets attributes (such as seasonality) or limitations (such as minimal data points and transformative issues in rounding, smoothing, and formatting) and properly meeting the model's requirements to produce a valid and useable result. The takeaway from this dissertation based on the conclusions is simple: attempt to change the model parameters, understand the changes to the data from preparation throughout

ingestion, and compare or analyze the forecast prediction periods and outputs to find the most accurate forecast that is the most realistic and useable for real world problems. However, the choice of the right forecasting results may still be a qualitative based approach of a seasoned data scientist and each data set should be treated individually on a case-by-case basis, especially if the product or industries are different than the last successful use of the same model.

## IV. Outline

The dissertation begins with a general literature review and history of forecasting to realize the continuous research and reasons for advancement in the field due to the complexities of how there is no perfect forecasting formula. The literature review will look at the modern history of forecasting, forecasting in the technology industry, and forecasting in the energy industry to understand what types of applications and reasons these industries use forecast modeling.

In Chapter 3, which is focused on methodologies, reviews of validation frameworks are discussed and the introduction of the 'proposed forecast validation framework'. The proposed forecast validation framework which is a multi-disciplinary approach to assist data scientists and decision-makers in conducting forecasts while being cognizant of data preparation, data modification, data flow, variables, parameters and how outputs can be impacted. The proposed forecast validation framework can serve as a template for conducting any forecast to assist the data scientist or decision-maker in validation of different data sets or forecasts to allow differentiation between 'out-of-the-box' and industry standard forecasting to help gain an edge in business and finance, or increase output accuracy for research and development.

The next sections, Chapter 4 and Chapter 5, include simulations using the proposed forecast validation framework for a strategic topic, energy investments in venture capital, and a product-level topic, methanol price in the energy industry. Note that methanol was the commodity chemical (over 97,000 million metric tons was consumed in 2019) selected for this study as an example (Methanol Institute, 2020), but easily could have been replaced with another energy industry product such as gas price, crude oil price, or product prices.

In Chapter 6, there is a general review of the complexities of neural networks and machine-learning when applied to forecasting models since forecast modeling is often mentioned in the literature using these newer techniques to form combinatorial, ensemble, and hybrid approaches which add complexity and require additional validation. Lastly, this dissertation concludes with a review of the takeaways from each chapter and results from the simulations to justify the need and use of the proposed forecasting validation framework. Moreover, it provides an understanding of some of the possible outcomes that may still need to be assessed by an experienced data scientist or decision-maker from a qualitative view. The proposed forecasting validation framework is to ensure the most accurate real-world execution of a forecasting model is chosen as the best result, rather than a default or mathematically correct, but not necessarily realistic result which may be wrong when applied to real-world prediction. It also reinforces the concept to not trust a generic forecast that may be presented in a slide of a business presentation or something that a data scientist created quickly as these simple predictions can become anchoring points based on an insufficiently validated forecast displaying the wrong output.

# Chapter 2. Literature Review

## I. Modern History of Forecasting

To better understand how complicated forecasting models have become based on the current research and literature, it is important to understand the history and notice that the basic forecasting model and concept have not changed significantly but are constantly being evaluated for improvement by additional testing, algorithms, and data preparation methods. These improvements to accuracy use new tools, algorithms, methods, and approaches from other disciplines such as statistics, computer science, and applied mathematics to better assist decision-makers in accurate real-world forecast results.

In the late 1960's, combination forecasts first appeared in mainstream literature in the publication titled "The Combination of Forecasts" by J. M. Bates and W. J. Granger from the University of Nottingham. This is generally accepted in the literature as the beginning of modern forecasting. They created a "composite set of forecasts" and their results were positive since they were able to "yield lower mean-square error than either of the original forecasts" (Bates & Granger, 1969). This combination or hybrid methodology is still used and examined today to improve forecasting results. In their research they were able to assess that when used alone, each forecast did not include all of the information and state: 1) "One forecast is based on variables or information that the other forecast has not considered" and 2) "the forecast makes a different assumption about the form of the relationship between the variables" (Bates & Granger, 1969). Their validation and assessment to test this theory used existing statistical tests and forecasting error analyses and further investigated the variances of each method. The conclusions of this early work were that certain forecasts, used alone, would not fully utilize all the information

contained in the original data set and that it was better to combine forecasts (which is what can be found in the abundance of literature today). The recognition of this loss of information and realization that combining forecasts creates complications and sometimes does not increase the accuracy of the forecast is a major reason of why it remains difficult to improve forecasting accuracy. However, it also introduces additional complexity to the forecast and process and therefore requires additional validation.

Today, many of these linear models are still used such as Box-Jenkins, Autoregressive (AR), and Moving Averages (MA) which are sometimes combined with neural networks and non-linear approaches since research has shown these methods ignore much of the non-linear time series data (Tealab, 2018). Moreover, C. W. J. Granger's previous literary work included a widely regarded publication titled, "Investigating Causal Relations by Econometric Models and Cross-spectral Methods," which set the groundwork for understanding relationships in causality and feedback in "… an explicit and testable fashion" (Granger, 1969). The "Granger Causality Test" will be discussed further in Chapter 5 since it is used in proposed model frameworks from literature in hybrid approaches to forecasting using social media data. In addition, it is used in combination with natural language processing (NLP). Newer literature includes attempts such as this to take qualitative data (social media data), convert it to quantitative, and apply it into a forecasting model to improve forecasting results.

This dissertation will show the uses and need for validation through the simulations conducted. The goal of this dissertation is to investigate forecasting models from various disciplines to better understand and validate forecasting in the energy sector using a strategic level example of venture capital (energy investments) data and a more specific tactical example of energy pricing (methanol price) data. The literature reviewed in this dissertation relates

directly to the dissertation proposal question of *how reliable are these models and how can they be validated to discover the best model for the data?* This question was derived from reviewing the topic of forecasting in academic literature where there is much debate and new tools and techniques being implemented into forecasting, but not much research done on an overall generic framework for a data scientist or decision-maker. Examples of comparisons can be seen in literature such as: univariate techniques conducted for the apparel industry by Min Li (Li, Wong, & Leung, 2014), neural networks to improve forecasting in the mid-1990's authored by W. Gentry (Gentry, M. Wiliamowski, & Weatherford, 2002), or in an electricity forecasting model comparison by Aman in 2014 (Aman, 2014). There are many industry specific publications that discuss improvements to forecasting but not many that can be used as a generic map of the major steps that require validation to better assist a data scientist or decision-maker in ensuring their results are the correct real-world results.

However, in 1971 a Harvard Business Review article was published titled, "How to Choose the Right Forecasting Technique" and discusses many of the factors and complexities of forecasting such as weighting, time periods/seasonality, and warns that forecast models should not "gold plate" working forecast models, or in other words, introduce extra unnecessary data, or non-instrumental data (Chambers, Mullick, & Smith, 1971). These publications are mostly topic specific and do not provide an overarching framework focused on validation. Some of the older literature's conclusions are out-of-date due to the advent of using computer science innovations in neural networks and technological innovations in graphics card processing to conduct advanced and sophisticated machine-learning. The focus of this dissertation is on validation of a forecasting model and ensuring it is being used properly and to its intended purpose with a chosen data set, which differs it from optimal model selection and a comparison of forecasting

models. The interest of this dissertation relies on forecasting frameworks and how to effectively convey the important steps when conducting a forecast to better prepare a data scientist or decision-maker in outputting the best result.

Due to the recent innovations in computer science in the practices of neural networks, deep learning, big data, machine learning, and artificial intelligence, many algorithms are being applied to create advanced forecasting models and hybrid forecasting models which present new traps, flaws, biases, and complexity. Many of the historical models are statistical and lack complexity, which is why current literature is focusing on more non-numerical approaches such as described by Dohnal & Doubravsky (2016). These efforts have been in hybrid models implementing computer science techniques such as natural language processing on social media data or using big data algorithms and database scrapers for topics such as discovering publication rates for a trending innovative technology or stock movement prediction (Nguyen, Shirai, & Velcin, 2015) and including this data in a forecasting model. Much of this conversion from qualitative to quantitative data can be difficult or set by the tool or algorithm used to mine or gather the data. Moreover, artificial intelligence can help solve some of these numerical quantification issues and factors that are difficult to quantify through models such as a Prohibitively Complex Forecast (PCF) (Dohnal & Doubravsky, 2016). These approaches are an attempt to bring in additional qualitative and quantitative variables and their data sets and add new parameters to increase the accuracy of forecast results. They are complex, subjective, and multidimensional as noted by the researchers, however, they are able to include social, ecological, political, and macroeconomic variables in a forecast (Dohnal & Doubravsky, 2016). Preferably, these hybrid approaches would be efficient and easy to validate but they may cause unknown explanations and require further investigation as to how they were able to increase the accuracy. This is shown where they can take

verbal knowledge such as a signaling statement and convert it to quantitative values that impact the outcome. An example of comparing neural network methods was published by IEEE in 1999 discussing a comparison of neural network forecast techniques using river flow data and time series to compare various neural network forecasting models (Atiya, El-Shoura, Shaheen, & El-Sherif, 1999). These newer methods which rely heavily on qualitative data being converted into quantitative data, such as in the case of social media analysis data preparation can have a significant impact on the outcomes. For example, social media data can be influenced by marketing and media efforts and skew the data points in a data set. This type of statistical trap is not necessarily the fault of the data scientist, but the data being collected can be inherently flawed. In addition, as Kaneman and Lovallo (2003) stated, most people are optimistic and exaggerate their talents so there is a tendency to "take credit for positive outcomes" while using external factors as an excuse for negative outcomes, so when attempting to gather social media data, or analyze a speech from a CEO or government official it can be difficult to quantify how this will impact a financial forecast. This kind of positive-negative verbal statement can easily disrupt a data set used in an advanced forecasting model, or a data set using a PCF model. When harvesting a new data set using sentiment and social media analysis algorithms there are many possibilities for incorrect data points due to socio-technological factors such as negative news being reversed into positive news. There is evidence of this in executive speeches and annual corporate reports where this concept holds true (Lovallo & Kaneman, 2003) which can cause statistical traps and biases such as signaling or influence a forecasting model. In addition, research claims that entrepreneurs and decision-makers themselves can be highly biased (Lovallo & Kaneman, 2003). For a decision-maker using forecasting models or glancing at a forecast from a slideshow every month from their finance department, they should be aware of not only in their own decision-making process but

the methods and process used by others they are investing with, or the company or project they are investing in to. Further research by Kahneman and Lovallo (2003) link this to anchoring (a common decision-making trap) and competitor neglect. Moreover, there can be political pressure that causes the entrepreneur (or decision-maker) to possibly downplay negative news (Lovallo & Kaneman, 2003), which can all impact social media and sentiment analyses, and other new big data collected information. These are some social (qualitative) factors, aside from the mathematical formulas that can impact the output of these models, which is why it is important to recognize that forecast modeling has many steps starting with the choice of data, the data preparation, the model parameters, the variables kept as default, the modified equations, and the hybridization of multiple models which all impact the accuracy of outputs that a decision-maker may rely on. However, some social scientists argue that "every tweet counts" since they concluded that social media data with sentiment analysis to create a complex forecasting model holds a correlation to mass survey data (Ceron, Curini, Iacus, & Porro, 2014). This argument can be compared similarly to the decision to use extract, transform, or load (ETL) methods on data, or 'cleaning the data' by conducting removal of data points that are generally considered anomalous before ingesting the data into a model. Due to the hybrid and combination approaches these forecasting models as whole can become extremely complex driving the need to mitigate incertitude.

The literature review in this dissertation will look at current research on forecasting methods and forecasting frameworks with a focus on validation across many disciplines to discover flaws, traps, and biases, whether they are in the model itself, the data quality needed for the model to output the intended results, the misuse of a model for an output that is incorrect but accepted, and the understanding that sometimes there is not a one-size-fits-all model. Chapter 3,

which is based on Methodologies will also provide a 'Proposed Forecast Validation Framework' to assist a decision-maker or data scientist in the process of validation from the data set, model, and the output. Many of these traps, flaws, and biases can be avoided through proper validation and exploratory data analysis (EDA) which will be used in the simulations in this dissertation in Chapter 4 and Chapter 5. Having accurate and reliable forecasting models or a combination of forecasting models may allow for higher quality prediction results for decision-makers and when making a strategic or executive level decision, a simple graph showing a forecast can hold significant weight to a decision-maker. Therefore, validation of forecasting models is imperative – using the right model for the data to produce the appropriate output.

## II. Technology, Innovation, and Research Forecasting

Technology forecasting is a field where a decision-maker or R&D manager must decide if a technology is worth innovating regarding cost or initial investment, return on investment, and competitive advantage over other similar technologies. For a decision-maker in a role to decide which new technology to assign research funds or a program manager who must decide which technologies to advance, improve, or engineer, having a strong understanding of forecast modeling is invaluable, especially when handling a large monetary budget. This is discussed by Porter (1985) and Brettel, Mauer, Engelen, & Küpper (2011) regarding corporate effectuation and causation (prediction) in research and development projects. These projects can be as small as a software company deploying cloud-based software for small businesses that are applying for venture capital funding or a government's large-scale spaced-based satellite program or fighter jet program. Regardless of size, there is an investment of time, money, and resources. It is critical to be able to mitigate risk and uncertainty, especially in a competitive environment where innovation is constantly redefining and replacing obsolete technology. Christianson (1997) describes how new technologies can cause firms to fail and why it is important to mitigate financial risk in technology investments, whereas newer approaches (Zhang & Tang, 2017) look at how risk and failure can be mitigated and included in to a model by investing in R&D internally within a firm rather than outsourcing. Approaches such as these can work on a case-by-case basis and being able to financially forecast budget requirements as well as forecast the investment in the technology sector should be taken into consideration.

In regard to statistical traps, flaws, and biases, Derbyshire & Giovannetti (2017) argue that when attempting to mitigate risk and uncertainty, forecasting techniques should not be used exclusively and there are "critical thresholds" for these models – which is where forecasting can

fail or output an incorrect real-world result. This indicates the need for innovation and improving the accuracy of these forecasting methods when used for technology forecasting. In regard to real-world applications of technology forecasting, Bildosola, Gonzalez, & Moral (2017) conducted a case study on cloud computing technologies to find the sub-technology that was the most active (their conclusion was: privacy & security), as an indicator for potential growing market and investment. Being able to discover a need or gap in a technology sector and focus on improving or delivering that technology can be successful or fail due to reaching the market late. Similarly, Adomavicius, Bockstedt, Gupta, & Kauffman (2008) used the wireless technology industry to look at interdependent technologies to identify complementary technologies that benefit from the innovation of Wi-Fi technology. Other researchers such as Kyebambe, Cheng, Huang, He, & Zhang (2017) were able to use academic literature itself (publications, patents, journal articles) as a data set to feed their forecasting model to try and attempt to identify upcoming disruptive technologies. They did this by using publication totals and rates over time of specific technologies as a form of signaling to discover where innovation is occurring. However, these analyses need to be appropriately used to avoid falling into decision-making traps. An example could be the publication rates of certain scientific practices versus others where significant publications (those that truly further the specific science as a whole) take a longer time to research, are more detailed, and may not be published as often as an popular mainstream industry backed-scientific topic. For example, research in venture capital has been conducted on patents in the technology industry and shows that, “Patents reflect improvements in innovation and can contribute positively to the performance of firms and their market value” (Hoenen, Kolympiris, & Schoenmakers, 2014). Therefore, the concept of venture capital firms being attracted to start-up companies with patents holds true, however the research concluded that after the initial deal and investment, additional

patents or filings for patents do not have a large impact during further rounds of financing as they do initially (Hoenen, Kolympiris, & Schoenmakers, 2014). This can cause errors in using patent publications as a data set and is an example of a statistical flaw or socio-technological trap in the data set and data preparation that could be discovered during validation stages. Moreover, this also shows that certain advanced technologies may only require a few patents and that improvement via publication and additional patents may not be beneficial or assist in the technology transition or adoption of the technology by the time it is finished and deployed into the market.

At a strategic level, Dohnal & Doubravsky (2016) discuss how these new approaches are trying to handle complex forecasting and are focused on solving quantification problems by applying a qualitative framework and reasoning. Heuristics such as equal weighting, fast and frugal trees mixed with machine learning, and regression models were used in research conducted by German and Swiss researchers to show the impact of venture capital learning, statistical properties of the investment environment itself, and the information in the business plan resulted that an equal weighting strategy is the most robust (Woike, Hoffrage, & Petty, 2014). However, what works in a scientifically derived industry such as pharmaceuticals may not work as well in an industry such as gas and oil (energy) or the automotive industry (manufacturing) without rigorous validation. This research also concluded that the decision strategies differed in respect to the impact of additional information on the outcomes (Woike, Hoffrage, & Petty, 2014). This is because most decision-makers use similar methods and tools and employ common out-of-the-box forecasting models. Moreover, Woike, Hoffrage, and Petty (2014) also discovered that decision-makers can be completely unaware of their own underlying strategies. This can cause problems downstream, especially if the business model changes, or the product line or market is re-aligned or shifted in a minor or drastic way. These are some arguments to why a seasoned and experienced

data scientist familiar with forecasting and validation of forecasting models is imperative to ensure the results are as accurate to real-world outcomes as possible.

Since there is no agreement for a single forecasting model or 'out-of-the-box' forecasting model for technology or innovation forecasting, many approaches are currently being revised or newly developed in the literature in various fields. The venture capital industry is particularly interested in finding "the next big thing" by their ability to exploit extreme events, or exceptional opportunities through the relationship of expensive capacity and negative externalities from high utilization (de Treville, Petty, & Stefan, 2014). Having a strong forecasting validation framework that can meet the requirements at a strategic level as a checklist can greatly assist the venture capital industry in terms of forecasting different products and services in various sectors at the market (strategic) and product (tactical) levels. The data used in financial forecasts can be very complicated especially during the data preparation due to the type of financial data points. For example, it is difficult for a forecasting model using historical time series data to handle and understand data points that there were manual corrections or financial functions such as splits, recalls, roll overs etc. which emphasize the need for a data scientist to validate these aspects of the data. This also touches upon the need for understanding data lineage and how data changes throughout a forecasting process from raw data to output result.

## III. Industry Price Forecasting

The oil, gas, electricity, and energy industry heavily rely on forecasting models for price points on crude oil, methanol, ethanol, other fuels and additives, which is the same as the need for forecasting models in technology disciplines. Moreover, electric price forecasting for power companies using a large portfolio of alternative energy sources are specifically interested in being able to maintain cost of maximum and minimum power load requirements during multiple times during the day. Much of the data for energy industry price forecasting works well in time series and seasonality modeling, both of which are predominantly used in the models found in the literature. The 1957 report by Charles C. Holt to the Office of Naval Research (ONR) is of literary importance when discussing forecasting as it provided a "systematic development of the forecasting expressions for exponential moving averages," and researched non-seasonal, seasonal, and error structures (Holt, 2004). This work describes the framework for forecasting a linear trend by identifying seasonality and adjusting for a percentage trend and a seasonal trend for future outputs (Holt, 2004). In addition, in 1959, Robert Goodwell Brown published a book discussing forecasting for industry, specifically in inventory control, which used short-range forecasting based on mathematical methods of the time to improve efficiencies in inventory control systems (Brown, 1959). He discusses exponential smoothing, averages with optimum weights, moving averages, trends, abrupt changes in the market, and seasonality, much of what is still used and researched to improve in today's forecasting literature in electricity price (Brown, 1959). Further research into methods for forecasting are conducted in Peter R. Winters' publication in 1960 where he discusses exponentially weighted moving averages for sales and the need for forecasts due to the growing use of computers and products for businesses at the time. In the 1960's much of the forecasting research was focused on how forecasting can be implemented as a tool to assist

inventory control systems and production planning. Winters' recognized then, that these models would have to support "hundreds of thousands" of data points and that the forecast may be conducted, "monthly or weekly, on a routine basis," which implied early on that time series and seasonality approaches may be important parameters and the concept of 'big data' would be necessary for these models to handle in the future (Winters, 1960). Over 20 years later, in 1983, the U.S. Navy published research on how to automatically monitor forecast errors due to the issues with forecasting models. Their evaluation looked at discovering biased forecasting errors and concluded, "A tracking signal based on the autocorrelation in errors is recommended for forecasting models other than exponential smoothing, with one exception. If the time series has a constant variance, the backward [cumulative sum control chart] should give better results" (Gardner, 1983). Over 20 years later as well as today there is still significant reason to investigate the basic algorithms, methods, and implementation of forecasting models.

In more recent history there are several continuous improvements and new approaches to time series modeling. According to a 2003 paper authored by Zhang mentions, "moving average, exponential smoothing, and ARIMA," which are all linear predictions as being popular and widely accepted forecasting methods still used today. In addition, Zhang also describes nonlinear time series approaches such as "bilinear, threshold regressive (TAR) modeling, and autoregressive conditional heteroscedastic (ARCH) modeling," which he concludes are just as important and can serve to create better results depending on the forecasting data set (Zhang G. P., 2003). Inherently, each of these models, whether linear or non-linear, need to be validated to verify if these methods and their relationship with the data set can be used to create better forecasting outputs.

Currently, most approaches are a combination or hybrid approach and many of them implement new technology from other disciplines. For example, in 2016, Li, Xu, Yu, & Tang

proposed a sentiment analysis and implemented natural language processing which is a dictionary-based approach to discover trends using social media data to inform their forecasting model. In 2018, Zhao, Wang Guoa, & Zeng implemented a moving average trend analysis, Autoregressive approach, and Support Vector Regression (SVR) as a hybrid model to discover trends which resulted in less than 4% percent error of the fitted oil price. This is an example of using hybrid forecasting models which implement innovations in the computer science field. Another new approach was modeled by Wang in 2018 using artificial intelligence and an integrated data fluctuation network with a forecasting model to predict crude oil prices. This method also used a preprocessing approach taking advantage of the time series crude oil data set, which increases the need for validation due to extract, transform, and load (ETL) changes to the initial data. These older literary works and publications set the standards for many of the forecasting models used today. The difference now is the implementation of new computer science technologies to be able to forecast extremely large data sets, and the internet, which allows for access to more qualitative and quantitative data that a data scientist can attempt to quantify that is readily available. An example of qualitative data that can be obtained from technologies such as the internet is social media data. This social media data should not be ignored as there is value in incorporating it into a forecasting model, however, it can be costly to create, maintain, and validate as it is not naturally created and vetted scientific data from a laboratory or corporation with process standards. Regardless, social media data can be used to improve forecast and prediction models through methods such as sentiment analysis. For example, a case study was conducted on political elections which demonstrated that sentiment analysis, which is similar to mass survey data, can be a strong data source but only if the citizens (in their use case) are consistent by the time of an election and an increasing number share their opinion (Ceron, Curini, Iacus, & Porro, 2014).

Similarly, researchers believe that sentiment data harvested from social media platforms such as Twitter (www.twitter.com) can be utilized in forecasts in order to improve the output (Asur & Huberman, 2010). However, quantifying social media data using a standard off-the-shelf forecasting model created for time-series data or specific financial prediction data sets may require additional validation and testing to ensure the data is properly being processed. It also adds to the complexity in the sense that it could provide useless or non-instrumental information, or valuable information that actually does not impact the outputs for the amount of effort and time required to procure, mine, transform, and ingest it into the forecasting model.

New approaches to forecasting in the energy sector include methods such as 'adaptive learning forecasting', first coined in the International Journal of Forecasting in 2019 (Kyriazi, Thomakos, & B., 2019). Researchers from the US, Greece, Italy, and UK recently proposed a methodology for forecasting that "allows for both forecasting averaging and forecast error learning" where they were able to discover the learning rate is shown to be non-linear if based on the past forecast errors (Kyriazi, Thomakos, & B., 2019). This method was used with agricultural data with success and with real GDP from various countries where they were able to assess the "irregular cyclicality" of the agricultural products can be directly linked to economic performance of the countries using this model. It should be noted, their smoothing method for the data uses a simple exponential smoothing forecast to discover the forecasting error. However, they describe using "a recursive structure [...] that it looks at one model/method only and learns only from that model/method" by using forecasting errors with a "learning parameter" (Kyriazi, Thomakos, & B., 2019). The conclusions from their work argue that with certain data sets and conditions, using *a priori* information to lower the forecast error with a "learning parameter" can improve overall performance. However this method cannot be used with "perfect/efficient forecast needs" since

there is no forecasting error parameter to use, which can be seen as a limitation (Kyriazi, Thomakos, & B., 2019). This example also shows how the forecasting literature is continually developing and there is no perfect model.

## IV. Neural Network Approaches to Forecasting

A 'Neural Network' is a specific group of algorithms (based on neuroscience) that can be used with machine learning to help execute the complexities of the input to output. Neural networks have been used in various disciplines, however, due to computing advances, they are being more commonly implemented in many disciplines to increase performance. Recently, they are becoming even more common in forecasting and to supplement existing forecasting models to improve accuracy, precision, and data handling. Inherently, by using neural network approaches to forecast models it exacerbates the complexity of attempting to validate the model due to the additional error within the neural network, which should be noted. A general understanding of neural networks and their use cases will be discussed in Chapter 6 to understand how and why certain frameworks and methods are used in combination with forecasting models.

Neural networks being developed and tested in computer systems began in the late 80's after what Cowan and Sharp call the "Age of Neoconnectionism" in 1990 (Eberhart & Dobbins, 1990). It is important to note the timeline, as neural networks have become a common term in current news, business, and computing as if it is a new concept, but now artificial intelligence is mainstream due to advances in computer science and graphics processing that are capable of handling large amounts of data and higher throughout. Furthermore, the article states, "Significant work is also occurring in areas ranging from the prediction of protein folding using supercomputers to formulation of new network learning algorithms and neurode transfer functions" (Eberhart & Dobbins, 1990) which shows that the realization of other applications for training neural networks and applying them in different scientific and academic fields was a long-term goal, mostly stalled by computer technology of the time not being able to handle the data processing. Researchers from other disciplines heavily contributed to early neural network research such as Donald O. Hebb,

who first defined a method to update 'Hebbian' (synaptic) weights which introduced the 'neuro-physiological postulate' discussing perceptron regarding metabolic firing rates (Eberhart & Dobbins, 1990). The reason these biological and historical anecdotes are relevant is because these are the types of developed algorithms and models that are currently being applied to forecasting models to produce more accurate outputs. It is vital that a data scientist conducting a forecast utilizing neural networks is well-versed in the basic concepts of neural network architecture to better perform validation and mitigate risk and incertitude.

To better understand neural networks, it is imperative to understand perceptrons. Perceptrons use backpropagation where the input is calculated in the neuron, and then the output is created. Early work on neural networks 1958 introduced the term 'Perceptron' by Frank Rosenblatt to understand how the physical world is interpreted and sensed by a biological system, how information is handled and stored, and how it impacts the biological system's behavior due to being stored (and organized) in its memory (Rosenblatt, 1958). In his work, he mentions what we now use as classifiers, a way to distinguish "dog" or "not dog" which calls a stimulus class in the brain, and that "Given an ideal set of binary characteristics (such as dark, light; tall, short; straight, curved; etc.), 100 stimulus classes could be distinguished by the proper configuration of only seven response pairs. In a further modification of the system, a single response is capable of denoting by its activity or inactivity the presence or absence of each binary characteristic" (Rosenblatt, 1958). However, his work is generally binary, so there were many limitations until neural network research became more prevalent. The original models were based on eyesight depicting the retina receiving inputs, however, this "classic perceptron" can be mathematically described which furthered the understanding of eyesight and how biological (optical) sensors interpreted and sensed the world. Similarly, this model is used to show how the neural network

will ‘understand,’ ‘store,’ ‘correlate,’ and ‘learn’ from the provided time series data in a forecasting model to produce an output. Figure X shows the classic perceptron model based on Rosenblatt’s work from 1958, which is currently used to describe the operations of a single perceptron in a neural network.

**Perceptron Model**

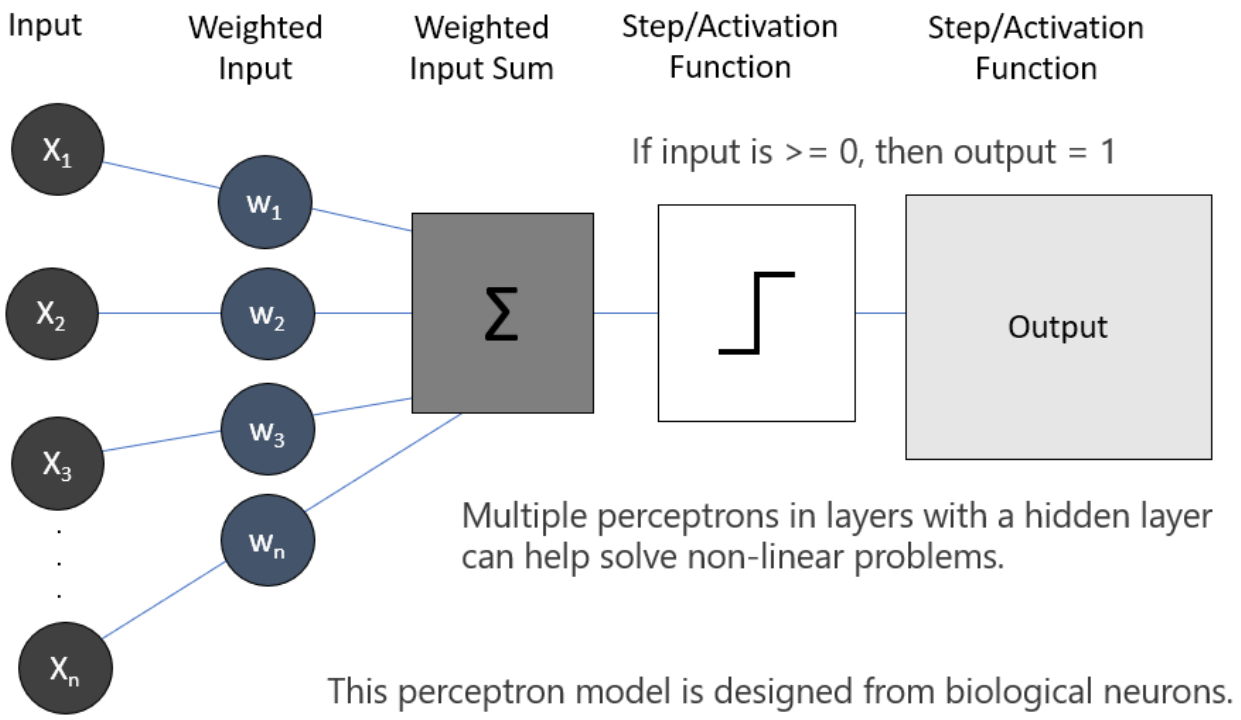



*Figure 1. Perceptron Model (Colorized) based on "The Classic Perceptron" (Rojas, 1996)*

However, artificial neural networks use an architecture with many perceptrons to handle non-linear data sets made up of multi-layer perceptrons, which is shown in Figure 1. This architecture diagram shows how the arrangement consists of an input layer, hidden layer, and output layer. The perceptron model, designed from biological neurons, are included within the input and hidden layers. This is a basic flowchart of how neural networks that are commonly combined with forecasting algorithms are used today. This model depicts 'feed forward neural networks' which are popular and commonly found architectures, especially in early artificial neural network literature for forecasting. The input layer is where the data set is ingested into a specific type of formatting for use in the neural network model, the hidden layer correlates the data between itself to discover relationships, and the output layer can be used to identify the answers that meet the criterion that the data scientist is trying to discover in the data set, or the output set by an algorithm. This type of neural network is used in many disciplines due to its robustness for data processing (Rojas, 1996).

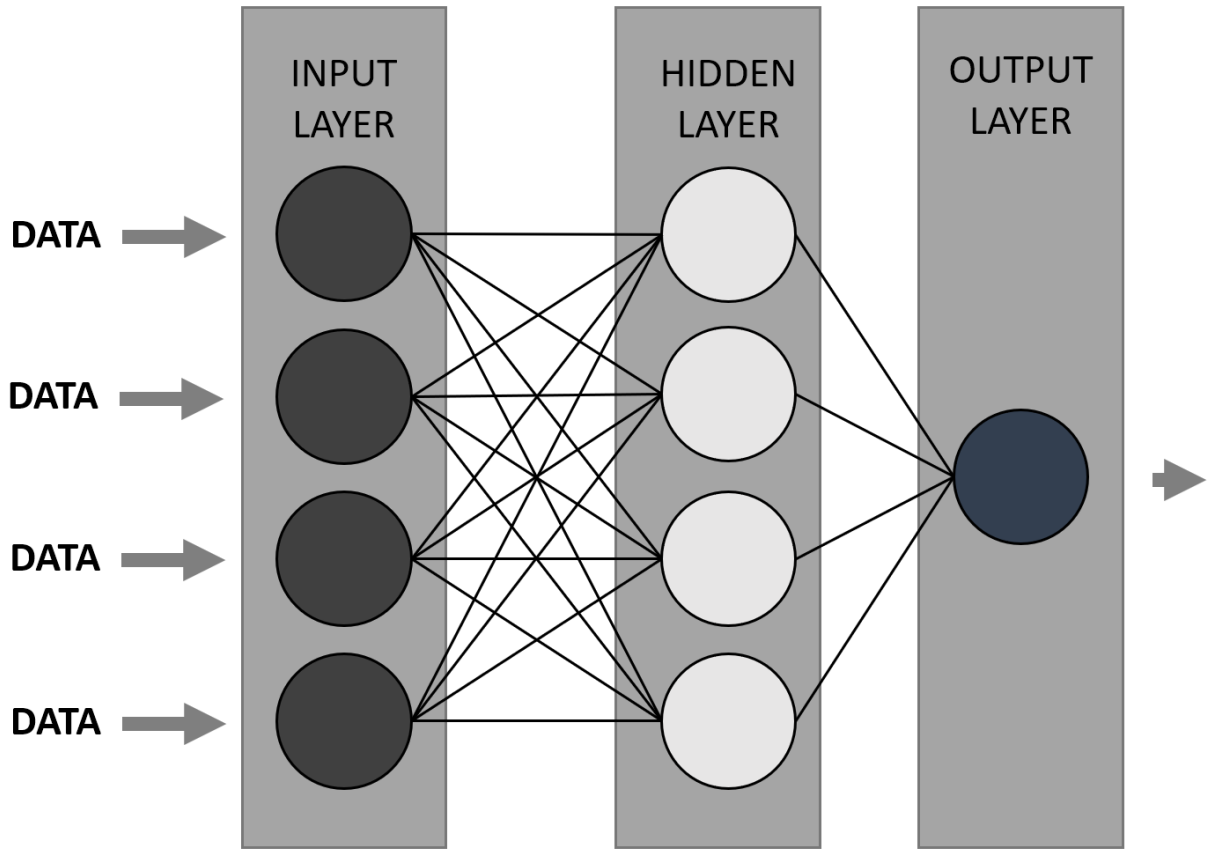



*Figure 2. Traditional Two-Layer Feed-Forward Neural Network Diagram (Colorized)*

*(Nguyen & Le, 2012)*

The traditional two-layer feed-forward neural network can be found in many early neural network publications, especially in late 1990's where it was applied to not just studying the human brain and biology, but also in sciences such as chemistry for topics such as potential energy surface fitting, spectroscopy, chemical reaction, process examinations, and electrostatic potentials due to the fact that it can assist with pattern identification (Nguyen & Le, 2012). Neural networks are complex and need validation regardless, whether they are being used with a forecasting model or not, and they are not required for forecasting, but just another tool and method of improving forecast outputs and accuracy. A data scientist should recognize them as a tool, algorithm, or option for improving their real-world results, not as a requirement when conducting a forecast. However, when using a forecast model including a neural network, rigorous and complex validation may be required. A traditional two-layer feed-forward neural network is shown in Figure 2. Note that this builds upon the perceptrons shown in Figure 1. It should be noted that when used or combined in a full forecasting model, there is a lot of additional complexity by using a neural network.

Innovative advancement in neural networks in computer science is impacting new approaches to forecasting today through the neural networks' ability to utilize deep learning and pattern recognition of extremely large data sets. For example, Haseltine & Eman use a neural network with a binary classification to improve forecasting to attempt to predict power grid failures (Haseltine & Eman, 2017). However, it should be mentioned that there is literature on the subject of applying neural networks to forecasting models since the early 1990's, though the GPU-driven architecture relying on graphics cards and languages such as NVidia's CUDA was not developed as it is today. An artificial neural network (ANN) approach was used by Park, El-Sharkawi, and Marks II in 1991 (El-Sharkawi, Marks II, & Weerasooriya, 1991) and by Lu, Wu, and Vermuri in 1993 (Lu, Wu, & Vermuri, 1993) to learn the relationships of temperatures and loads to better

forecast electrical loads in a 24-hour load profile. By the early 2000's, a study by G. Peter Zhang mentions that the past three decades a popular forecasting method was using autoregressive integrated moving average (ARIMA), and by combining this method with an artificial neural network would take advantage of the linear and nonlinear modeling strengths of both (Zhang G. P., 2003). These energy industry-related forecast models are due to the large amount of uncertainty in previous forecasting models and methods that were used and is an attempt to increase their accuracy and results by using a neural network forecasting approach. However, this also opens a new environment for flaws, traps, and biases incorporating similar problems that big data analytics and neural networks have in the computer science field. It seems that there are many variables that cause flaws, traps, and biases in forecasting models for decision-making at different levels: the data set, data preparation, the amount of data, the model itself, the thresholds of the model, and the advanced approaches all have inherent traps that must be avoided which can cause confusion and lead to bad decision-making. These flaws, traps and biases are not a new concept as E. Michael Azoff discusses in "Neural Network Time Series Forecasting of Financial Markets" in 1994, where he stated that the "proper use of a neural network involves spending time understanding and cleaning the data: removing errors, preprocessing and post-processing" (Azoff, 1994) which is what many data scientists and companies are forced to spend resources on. Having to efficiently prepare data sets is still one of the major risks in applying neural networks to forecasting models today due to the need for validation and data lineage analyses. Moreover, policy decisions regarding the electricity price changes and controls rely on forecasting models and therefore validation is necessary. An example is in Spain, where the electricity sector from 1988 until 1997 was regulated by 'Stable Legal Framework,' known as "Marco Legal Estable" however after that changed, the market did not know how to react as it had never had any competition. When the

liberalization of the market occurred, it created risk and failure to be able to find an appropriate model to forecast the electricity price as they never had before in Spain (Ortiz, Ukar, Azevedo, & Múgica, 2016).

These statistical traps, flaws, policy failures, and biases at various steps of the raw data to forecast output are reasons for the need to rigorously validate the forecasting processes used. Neural networking approaches and hybrid forecasting models significantly increase risk and complexity since each step has requirements such as preprocessing, postprocessing, and other data transformations to make the data suitable for the model.

To illustrate the complexity of ensemble or hybrid approaches combining neural networks, proposed frameworks for validation of neural networks and new methods in other disciplines are shown in the literature in publications such as (Humphrey, Maier R., Wu, Mount, & Dandy, 2017); (Chen & Shi, 2019); (Sun & Trevor, 2018); (Gonzalez-Carrasco & Garcia-Crespo, 2012); " (Wang & Hongguang, 2018)"; and others. The processes, conclusions, frameworks, and validation methods will be discussed in Chapter 6 which focuses on forecasting with neural networks. These publications are from various disciplines, yet they all use forecast models combined with neural networks to increase the accuracy of results for real-world applications. Many of them include measures for validation, but most are focused on the actual forecast model combination rather than on the validation itself. To reiterate, there is a need for a strategic forecast validation framework, especially since forecast modeling is used heavily in scientific, engineering, and mathematical disciplines and being able to provide a framework, flowchart, or checklist that covers the general start to finish forecast modeling cycle is beneficial for the academic community.

# Chapter 3: Methodologies

## I. Applying Validation in Forecasting

Validation of forecasting models in this research will be used to understand how the model works and how slight variable, parameter, or data preprocessing modifications can easily influence the output. This argues that even though the mathematical formula of the model is correct, the output may be inaccurate, yet perceived or accepted as correct. The reason for understanding the term 'Validation' is due to ambiguous assumptions that are often made.

For validating data sets, exploratory data analysis (EDA) can be one of the techniques used. EDA as defined by Tukey (1962) are "procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data" (Tukey, 1962). Using EDA, Tukey notes that using this analysis should be implemented alongside, rather than replacing significant statistical confidences. Furthermore, using this as an analytic data technique allows understanding that there are decision functions that influence data sets (Tukey, 1962). This is also a strong tool for a seasoned data scientist to employ when they are conducting a forecast. Since the math is correct regardless, sometimes having an experienced data scientist interpret the output and how the forecast was conducted is the best way to create and choose a reliable real-world output that can be acted on by a decision-maker.

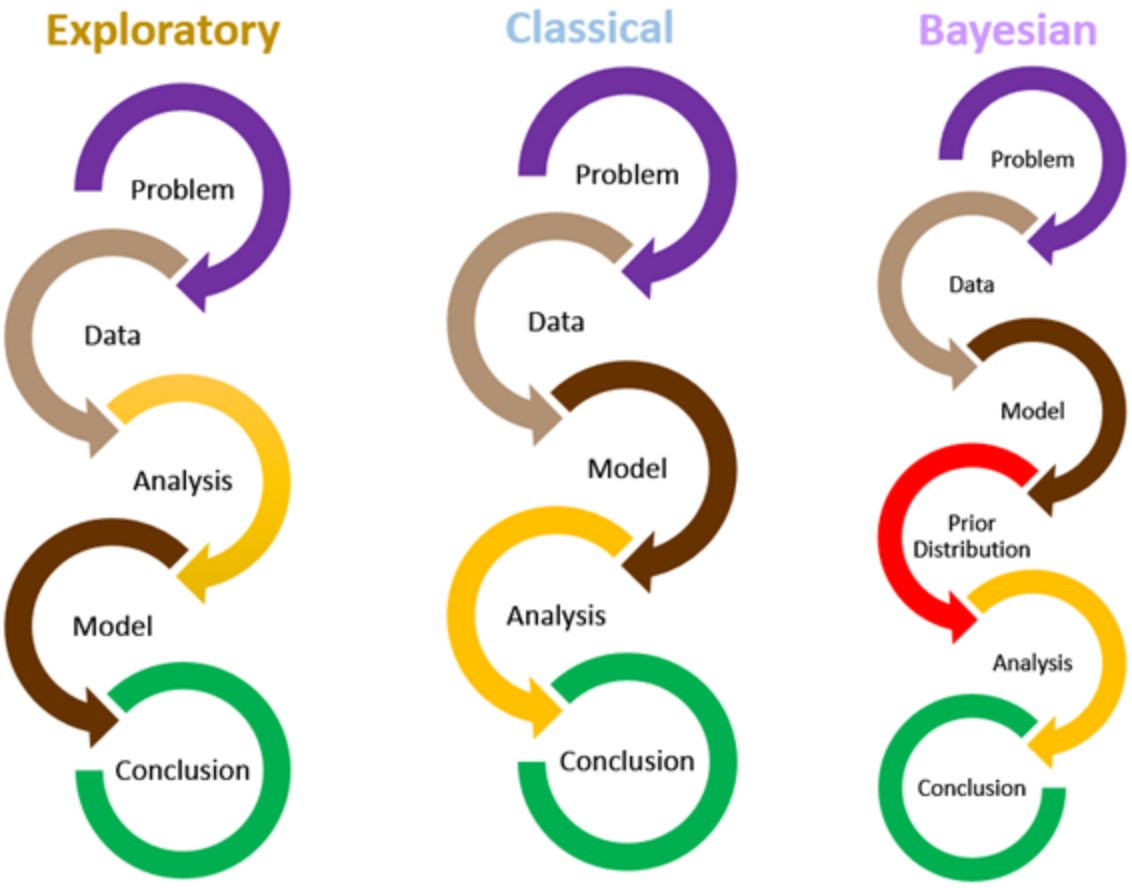



*Figure 3. Comparison of Data Science Schools of Thought*

Exploratory Data Analysis requires a careful visualization or “look” at the data, which may cause potential biases and flaws, however, this vetting process should help in the validation process. In addition, “[this] approach does not impose deterministic or probabilistic models on the data” (NIST, Exploratory Data Analysis, 2015). An example could be if a data set visually “looked” like a previously analyzed data set where visual cues such as the shape of the distribution caused the data scientist to skip other analyses and rush to use an inappropriate model. Whereas the ‘Classical Approach’ may not have this occur since the analysis would take place after deterministic and probabilistic models have already been chosen and outputs can be compared. Examples of common classical approaches are regression models and analysis of variance (ANOVA). However, while using this method for validation, it may become apparent earlier by the data scientist that there are validation issues or that they are on the right track, which can qualitatively come from experienced judgment of the data scientist with a strong history of forecasting specific data in a certain discipline. If successful, the data scientist may be able to glean information through visualization techniques that speed up their business process and move on to validating the next steps. In addition, constructing histograms and "looking" at the data to see its distribution can be a useful tool for a data scientist. Often the histogram will include a line that depicts what the shape would look like if the distribution were truly normal, and the data scientist can visually “see” how much the actual distribution deviates from this line. Herein lies the statistical trap or flaw in both at the strategic level – if something was assumed to be correct at first glance using only EDA for validation, it could cause major issues throughout the other steps of the validation since the data is inherently flawed. In addition, a Bayesian approach may be useful, but needs to be used appropriately “by imposing a data-independent distribution on the parameters of the selected model” and “formally combining both the prior

distribution on the parameters and the collected data to jointly make inferences and/or test assumptions about the model parameters," (NIST, Exploratory Data Analysis, 2015) which can complicate the validation process and could be avoided.

The goal is to truly understand how the data is interacting with the forecasting model (or neural net at later steps) as best as possible, based on experience as well as statistical testing and experience with similar data sets to ensure each step of the framework is valid, vetted, relevant, mathematically sound and is hopefully the best way to move forward to produce the best output.

Popular methodologies used throughout forecasting literature include moving average, exponential smoothing, polynomial regression, theta-model, Box-Jenkins approaches, time series models, and neural networks (Lemke & Gabrys, 2010). These methods are usually tested or graded singularly using the performance averaged per data set and standard deviation which has produced different models outperforming one another throughout the years according to the results from the Makridakis Competitions due to improvements in areas such as neural network computing (Crone, 2010). Moreover, many of these are used as a hybrid, ensemble, or combination approach.

When converting qualitative to quantitative, the data should be investigated and assessed to ensure that the data scientist conducting the forecast needs to include this data, and that it will increase the accuracy of the forecast. In addition, it can help to simply plot the data set and use simple statistical methods (histograms, standard deviation graphs, and/or box plots) to understand the data better as well as visually identify if there is seasonality, enough data points, missing data (where histograms can be misleading), or anomalous data points which can all impact the quality of the data set. An example of anomalous data points with financial data could be indications of previous market crashes and potential bubbles. This may also allow for identification of non-

instrumental information, or a data set that is unnecessary to include in a forecast but may seem relevant. Non-instrumental information may seem relevant to the decision-maker or data scientist, but actually has no impact on the end result of the decision (Bastardi & Shafir, 1998) or has a negative impact (time, memory, computing requirements increased to process). Kfir Eliaz and Andrew Shotter (2010) further affirmed this concept, named, the "confidence effect," by discovering that when making a "decision under risk, [decision-makers] are willing to pay for information on the likelihood that this decision is ex-post optimal, even if [the information or data set] will not affect their decision" (Eliaz & Schotter, 2010). To avoid this trap, it is imperative that the data scientist conducting a forecast validates the data sets involved to ensure they are not only correct but are necessary to be included in the model. In addition, this risk mitigation concept should follow throughout each step of the validation framework.

When validating the data preparation before the data set is ingested into the forecasting model, the decision-maker or data scientist must ensure that they are preparing the data in a way that does not disturb the output significantly. An example could be the auxiliary regression, which Franses (1994) proposed and considers a simple method for trend curves, however, he warns "we have to interpret this figure with caution," in regards to potential traps and biases (Franses, 1994). He recognizes that the Gompertz process is asymmetric while the logistic curve is symmetric, and because of this, there can be a "substantial impact on forecasting" and it is important to know which to use depending on the data set (Franses, 1994). In addition, Franses & Haldrup (1994) also recognized that for time series data, one of the main traits is that there are non-stationary patterns and this can be valuable for testing effects of additive outliers (Franses & Haldrup, 1994).

For validating forecast models, a method that can be used is cross-validation to attempt to statistically estimate and understand the error in forecast for data sets. This is not a new concept

and there are many publications that use, discuss, or improve on cross-validation methods such as Stone (1974) with research for the Deleted Cross Validation Statistic, which "provides a robust estimate of forecast error variance" (Stone, 1974). One of the most popular is K-fold cross-validation for model evaluation. However, there are inherent flaws when applying cross validation to time series data, and there are specific models developed for the testing and validation of these types of data sets (Bergmeir, J.Hyndman, & Koo, 2018). Their approach also researches machine-learning and concludes that k-fold cross-validation works better on time series data than out-of-sample and non-dependent cross-validation. Furthermore, Ailon, Jaiswal, and Monteleoni (2009) describe a new cross validation method for streaming k-means approximation with two approaches: " … a derivation of an extremely simple pseudo-approximation batch algorithm for k-means (based on the recent k-means++), in which the algorithm is allowed to output more than k centers, and a streaming clustering algorithm in which batch clustering algorithms are performed on small inputs (fitting in memory) … " (Ailon, Jaiswal, & Monteleoni, 2009).

For the simulations conducted in Chapter 3, the 'forecast' package in "R" will be used since it is a commonly used, popular, and generally accepted way to conduct statistical forecasting models. By default, the R package uses Exponential Smoothing State Space Model (ETS) on the data which chooses a model such as Akaike's Information Criterion (AIC) (Hyndman, 2018). Many users of the 'forecast' package in "R" may not be aware of the underlying transformations on the data that are inherent when using the functions included in the package so it is imperative to read the documentation and not accept a model, package, or function at face-value. The ETS model has been used since the 1950's and can include exponential smoothing methods (which are parameters that can be added or modified in "R" when writing forecast code.) The main three trend components in an ETS model are either none (default), additive, or multiplicative. In addition,

there are ‘damped’ versions of additive and multiplicative. These smoothing methods essentially will allow older observations, or data points, to carry less weight based on time, in a time series model. The simulations will show some of the impacts and the importance of choosing the right parameters in subsequent chapters. Furthermore, the ETS method additionally will output prediction intervals, which the data scientist should confirm using exploratory data analysis methods to begin a validation.

The model for ETS is:

$$yt = w(xt - 1) + r(xt - 1)\,\varepsilon t$$

$$xt = f(xt - 1) + g(xt - 1)\,\varepsilon t$$

*Equation 1. Exponential Smoothing State Space Model (ETS) Model*

*(Frausto-Solís, Chi-Chim, & Sheremetov, 2015)*

In the ETS Figure 1 (above), $yt$ is denoted as an observation at $t$ and $x_t$ representing the state vector for the unobserved components (trend, seasonality, and white noise) which are unlike an ARIMA approach, and it is noted in the literature that this can cause forecasting results to be “sensitive to initial states and parameter settings” (Frausto-Solís, Chi-Chim, & Sheremetov, 2015).

The Holt-Winters method was created by C.C. Holt and Peter Winters in 1960 and is an exponential smoothing approach used by many companies on their sales data. Paul Goodwin’s research column titled, “The Holt-Winters Approach to Exponential Smoothing: 50 Years Old and Going Strong” wrote, [Holt-Winters] produce short-term demand forecasts when their sales data contain a trend and seasonal pattern” (Goodwin, 2018). Holt-Winters is designed specifically to

handle data that is seasonal, and other researchers such as James Taylor have been able to modify Holt-Winters to handle additional seasons for other applications such as forecasting hourly electricity rates (Taylor, 2009).

## II. Forecasting Methodology – Assessing Existing Validation Framework

In this section, a validation framework will be proposed for decision-makers and data scientists to attempt to assist in critical thinking and helping to discover the best forecasting model for the data set. The literature referenced in this section will focus on existing forecasting frameworks and forecasting methodologies to better justify the need for a validation framework that considers many of the variables, steps, problems, and issues to be aware of when conducting a forecast. There are several methods that are covered in the literature discussed in this research, however, it is important to show that the flaws, traps, and biases that can cause forecast failures and unintended results can be seen in the most basic models. The simulations conducted in this research will show how easily forecasting models can have very different outputs that can be counter-productive to the decision-making process, which reinforces the idea that when a decision-maker, R&D manager, government policy maker, venture capitalist, or financial institution must use a forecasting model, they should understand that some models will be more accurate than others depending on the type of data, quality of data, amount of data, industry, or other variables. The reason for conducting these simulations is to demonstrate how the trap of overconfidence can be avoided by selecting one forecast model for everything a decision-maker does. It will also attempt to prove that model validation and data preparation are important when conducting predictive analytics and show that there can be significant differences due to the data set and minor changes in the model, even though the math, model, and inputs remain mathematically sound in each simulation.

Validation frameworks discovered in the literature are usually not as specific or applicable in practice with dynamic data sets – especially for forecast modeling. For example, in the publication, "A proposed best practice model validation framework for banks," (note the title

is not ‘for finance’ or ‘generic model’ but specifically  for the banking industry) the researchers acknowledge forecasting models, but their framework is a strategic one that can be used for other forecasting models. Their goal is to demonstrate best practices in model validation due to the financial crisis in 2008-2009 by focusing on model development to mitigate risk. The focus of their work was creating a scorecard approach “[comprised] of three principal elements: model validation governance, policy and process” (de Jongh, 2017). This framework by de Jongh (2018) realizes the potential for human error, flawed model construction, assumptions, and inappropriate model applications while recognizing the issues described in this dissertation. This should not be confused with optimal model selection, the goal is to better choose the right forecasting model for the data set being used – not just the forecast that seems to perform the best and continuously using it on new sets of data with a specific set of optimal parameters that were decided at the time of selection. This is described by de Jongh by stating: “even sound models which generate accurate outputs may exhibit high model risk if they are misapplied.” To reaffirm, de Jongh’s two principle risk reasons are: 1) flawed modeling and assumptions and 2) “models used outside the environment for which they were designed” (de Jongh, 2017). This shows the issues shown in this dissertation regarding forecasting models, though de Jongh is speaking to any model at a strategic level.

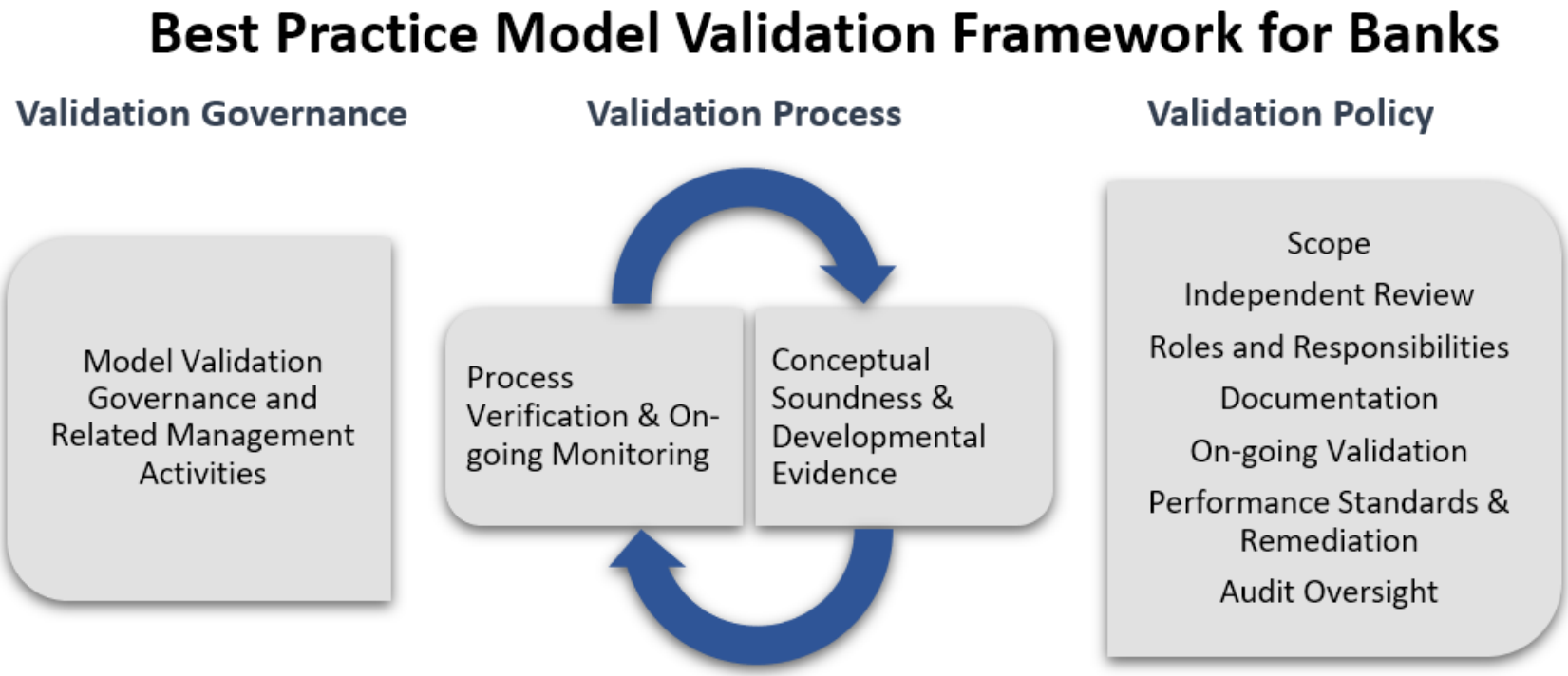



*Figure 4. Model Validation Framework for Banks*

*(de Jongh, 2017)*

The validation process proposed by de Jongh (2018) uses a continuous validation with ongoing monitoring once the 'validation governance' is defined which leads to a validation policy that includes criterion such as: performance standards, review, and scope. Essentially, in regards to forecasting models, the researchers refer to this continuous validation by stating, "A comparison of model outputs against corresponding actual outputs should be conducted regularly, including the assessment of forecast accuracy, appropriateness of statistical tests, expert judgement of outputs produced and confirmation that outputs make business sense" (de Jongh, 2017). This is where the proposed validation framework in this dissertation can be applied. Continuous, rigorous testing, assessment, verification, and validation are important when attempting to find the best and most accurate outputs of a forecasting model. By using the same model or not comparing or working through a validation framework, differences such as those shown in the simulations in Chapter 3 and Chapter 4 of this dissertation can occur. It should also be noted that the proposed validation framework can be used with industry specific validation framework models and modified with the steps of an industry specific forecasting framework.

An example of a framework used in industry and published by Bridgei2i, a digital transformation company using AI accelerators, describes an "ensemble" (or hybrid) forecasting method. Their model deconstructs the data by de-seasonalizing it, then running it in parallel through common statistical and forecast methods such as ARIMA, exponential smoothing, and moving averages, afterwards, it seasonalizes the data, and places it in a time series object to discover the error. This type of framework can be used in a general sense, however, there is no accountability or mention of validation other than 'error correction using [time series]", which may not account for the data transformations taking place within each method. However, this

framework may work well for a specific product or industry, but it should be decomposed more and certain steps, such as “de-seasonalizing data” should be validated or flagged of interest to the data scientist or decision-maker based on their experience, the decision-makers goals, and the value and impact of the outputs. A positive use of this framework is its focus on the data itself, and the transformations and algorithms that may change the data. Many forecasting frameworks are heavily focused on the model or the outputs and ignore the data changes or severity of certain smoothing techniques.

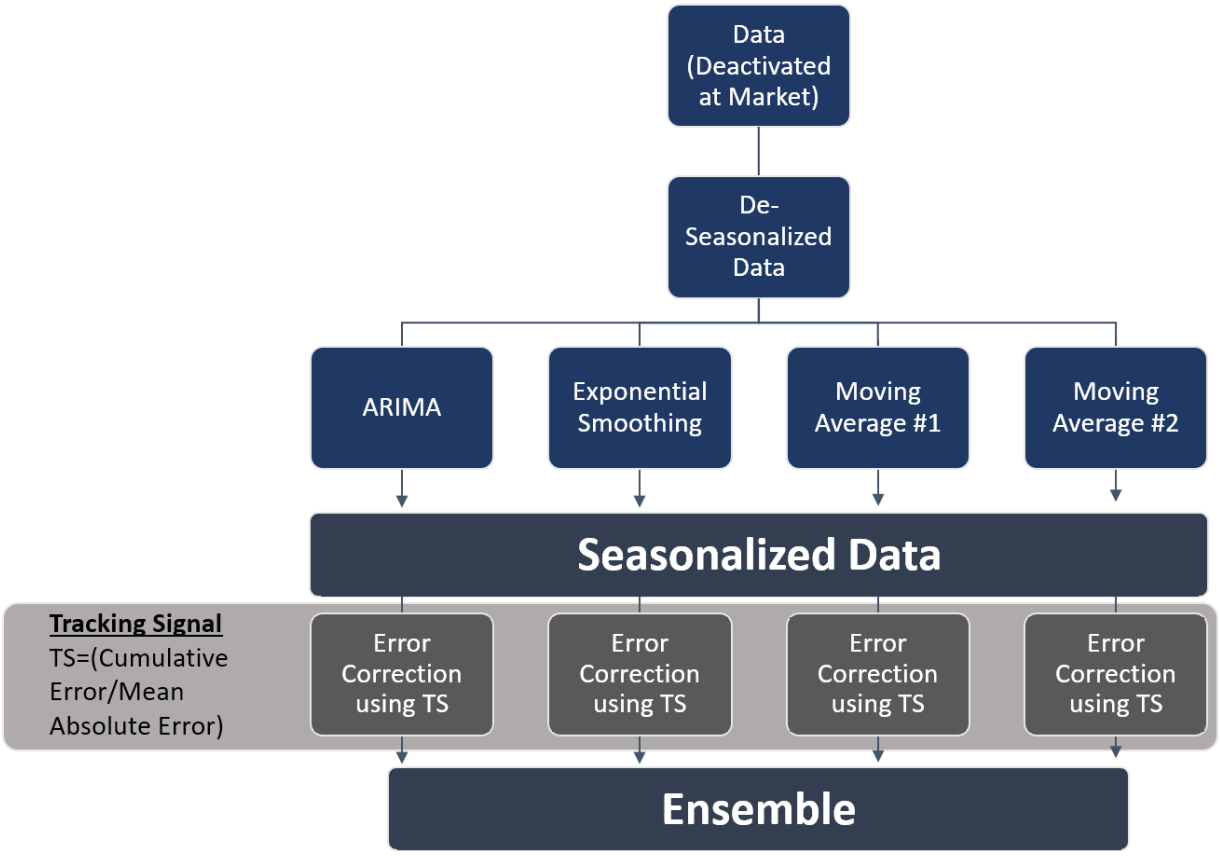



*Figure 5. Schematic diagram of Forecasting Method*

*(Laha & Roy, 2018)*

An example of a forecasting validation framework proposed in 2001 recognizes the decision-maker or data scientist may have 'analogous data' – or converted, changed, transformed data, and that the analysis should be tailored to the decision-maker (Armstrong, 2001). The researchers state that "forecast[s] that are too low might lead to a loss of customers, which might have more serious consequences than forecasts that are too high. Leave this concern to the planners and decision makers." And that the data scientist should, "provide unbiased forecasts and good assessments of prediction intervals" (Armstrong, 2001). These two points are important because the data scientist must remain unbiased regarding the data and outputs, but still be aware and draw on experience when performing a validation.

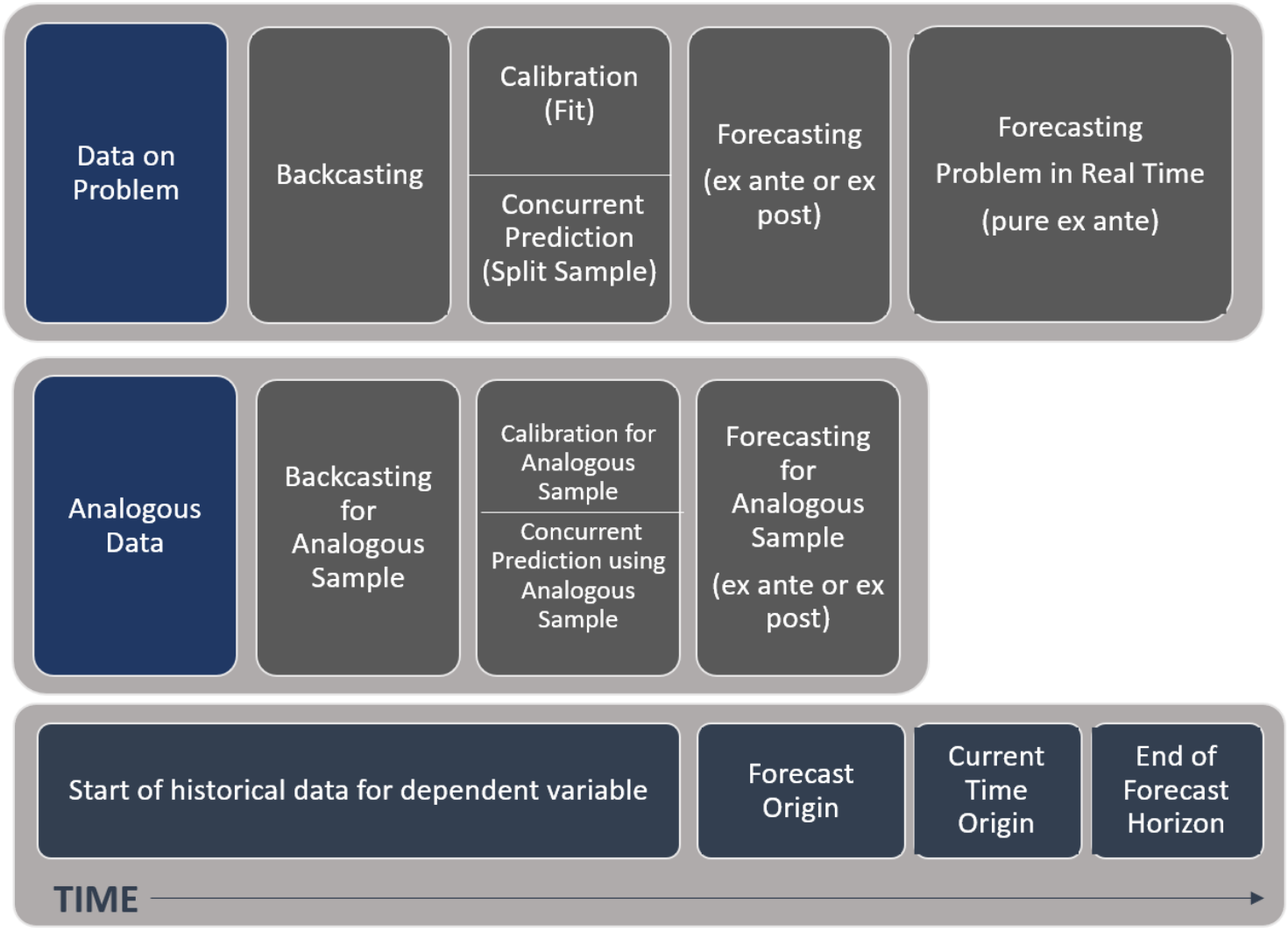



*Figure 6. Validation Matrix (Framework) for Forecast Validity*

*(Armstrong, 2001)*

Some of the criterion that is not noted in their criterion in the 'Validation Matrix' that are of interest to conducting a forecast validation are: 1) is there a potential source of bias by forecaster? 2) Assess the data reliability and validity, 3) Do not use R-square to compare forecasting models. A larger criterion is shown is described in the paper by Armstrong (2001) with many relevant validation tools and concepts to keep in mind when conducting a forecast.

One of the takeaways from the validation framework proposed by Armstrong (2001) is the recommendation and researcher's avoidance of using R-square or ($R^2$) to compare forecasting models. Armstrong (2001) states that, "$R^2$ should not be used for time series forecasts, not even on a forecast-validity sample. For one thing, it overlooks bias in forecasts. A model can have a perfect $R^2$, yet the values of the forecasts could be substantially different from the values for all forecasts" which is justified with an example in his work (Armstrong, 2001). This is an example of something to keep in mind for a data scientist when performing validation to test a forecasting model. Armstrong (2001) goes on to discuss how it can be misleading and cites other works of these statistical flaws, traps, and biases. An example used is a data set titled, "Historical Statistics of the United States," (Ames & Reiter, 1961) and Armstrong's (1970) data on sales in 31 countries of a specific product. The "assessing outputs" section in their 'Evaluation Principles Checklist' can be applied to validation of forecasting today, however, due to the heavy use of neural networks and hybrid approaches, there is room for modification, additions, and adjustment to the criterion. Still, it can be used as a starting point for learning best practices when conducting a forecast. There is not much of a focus on the data and transformations in this framework, it is assumed the data is in the best condition or 'perfect' and this is more focused on the model itself. In addition, neural networks add new layers of

complexity that are not mentioned by the author, which are some of the reasons this validation framework could be considered out of date.

## III. Proposed Forecast Validation Framework

After reviewing forecast validation frameworks, model validation frameworks, methodologies, and validation methods, as well as having data science experience in handling forecasting data, the following figure was created to help organize common scenarios that occur often when conducting a forecast. To better create a strategic level framework, it is important to understand the items in Figure 7 since they are the potential outputs, problems, issues, traps, flaws, and biases that the data scientist and decision-maker are trying to avoid. As shown in Part II of this chapter, a strategic validation framework is difficult to find, yet all the validation frameworks discussed for unique industries and practices share many common terms, problems, and steps. In order to propose a newly created framework, it is imperative to incorporate as many of these common problems together in order to inform the data scientist or decision-maker of areas of concern. Some of the frameworks focus on the data, others focus on the model or statistical tests. Most of these models do not account for the implementation of advanced artificial intelligence practices such as machine learning  or neural networks, or the potential subsets of machine learning such as deep learning models because they become extremely complex, however, they should be included as optional in an actual strategic framework but will require additional validation.

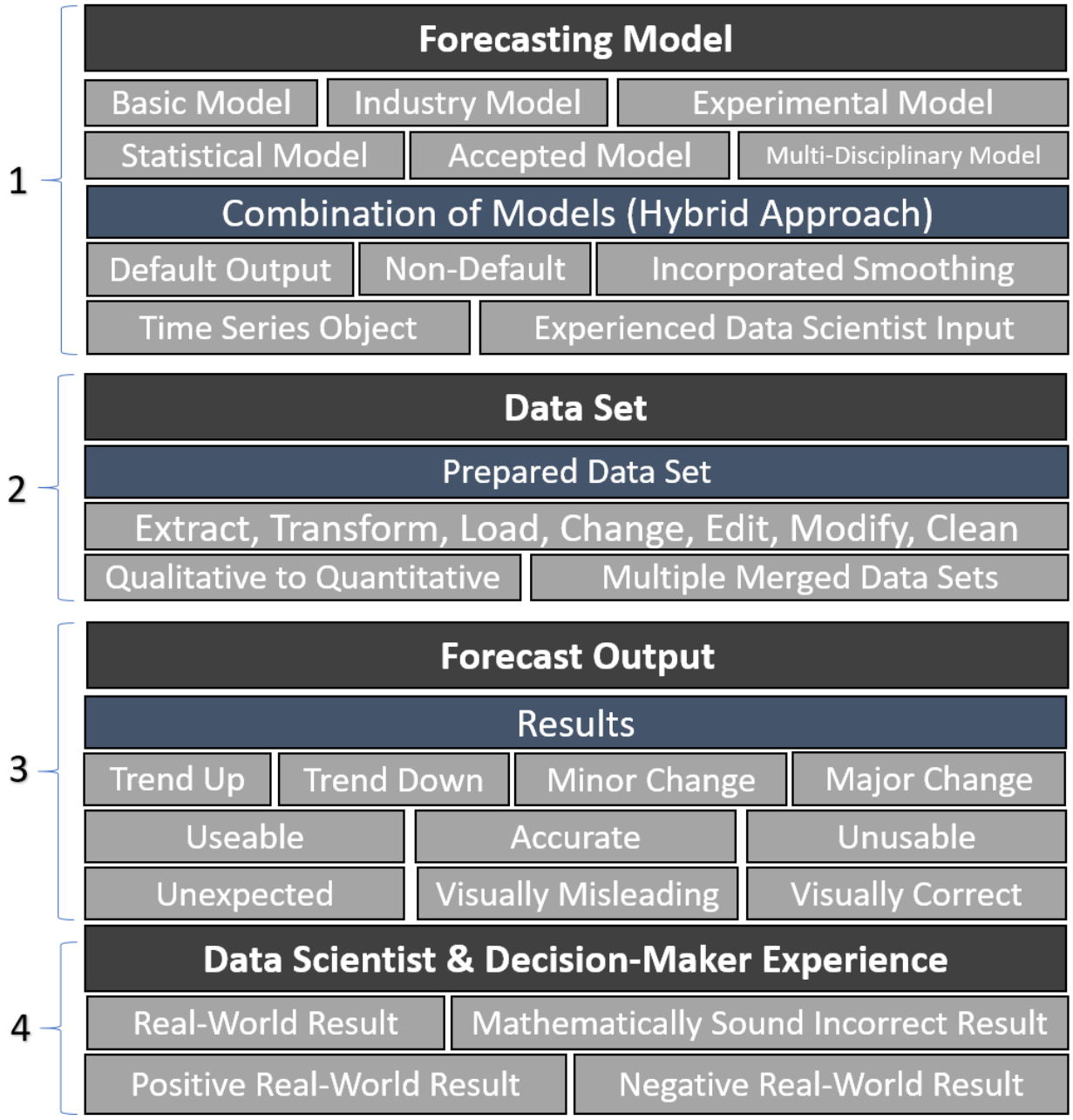



*Figure 7. Forecast Validation Elements*

Figure 7 represents various examples of potential variables, issues, decision-points, and outcomes in 4 different stages to help discern a path forward in concert with the proposed forecast validation framework in Figure 8. Figure 7, titled, "Forecast Validation Elements" was created by reviewing various frameworks from the literature to characterize common elements or potential outcomes. By following these guidelines, the data scientist and decision-maker should be better able to create proper data to model relationship understanding and produce more understandable outcomes. Since this is a multi-disciplinary approach, editing the proposed forecast validation framework and applying to a specific industry, product, program, financial area, or business process should be done by an experienced data scientist. The simple takeaway from this dissertation is to truly understand the data, the model, and how the data interacts with the model throughout the additional complexities added at each step. In addition, it is to understand how complicated forecasting can become when adding newer approaches such as neural networks, machine learning algorithms, natural language processing, and sentimental analyses. Many times, in venture capital and financial industry, there is no standard or discussion regarding validation, but there are many commonly used algorithms which give similar results. In order to differentiate and create a more accurate prediction or forecast, new forecasting approaches are constantly being developed.

A proposed forecast validation framework is shown in Figure 8 which strategically represents an approach to validate at three different layers – the Data layer where the data set that will be actually ingested in to the model is created, the Data Preparation layer where the data sets are modified (joined, extracted, transformed) or prepared using smoothing algorithms, seasonalized, de-seasonalized, etc. and the Forecasting Model layer, where the various parameters and additional forecasting models and algorithms are combined or used to produce the final

forecast model output. Therefore, a decision-maker implementing forecasting techniques and tools into their analysis to make strategic decisions must avoid the one-size-fits-all approach and the out-of-the-box forecasting model approach to forecasting. Acknowledging that each forecast can be different based on numerous variables from amount of data points to visual cues qualitatively discerned by an experienced data scientist is imperative.

Other frameworks will be discussed when implementing neural networks in Chapter 6 due to the additional complexities added by ensemble and hybrid approaches combining forecasting models with neural networks. For the simulations conducted in Chapter 4 and Chapter 5, the focus is on basic forecasting techniques without complications and implications of neural networks to show how even in a more simple form, there are many issues that can arise before adding the complexities of a neural network and advanced ensemble forecasting models.

# Proposed Forecast Validation Framework

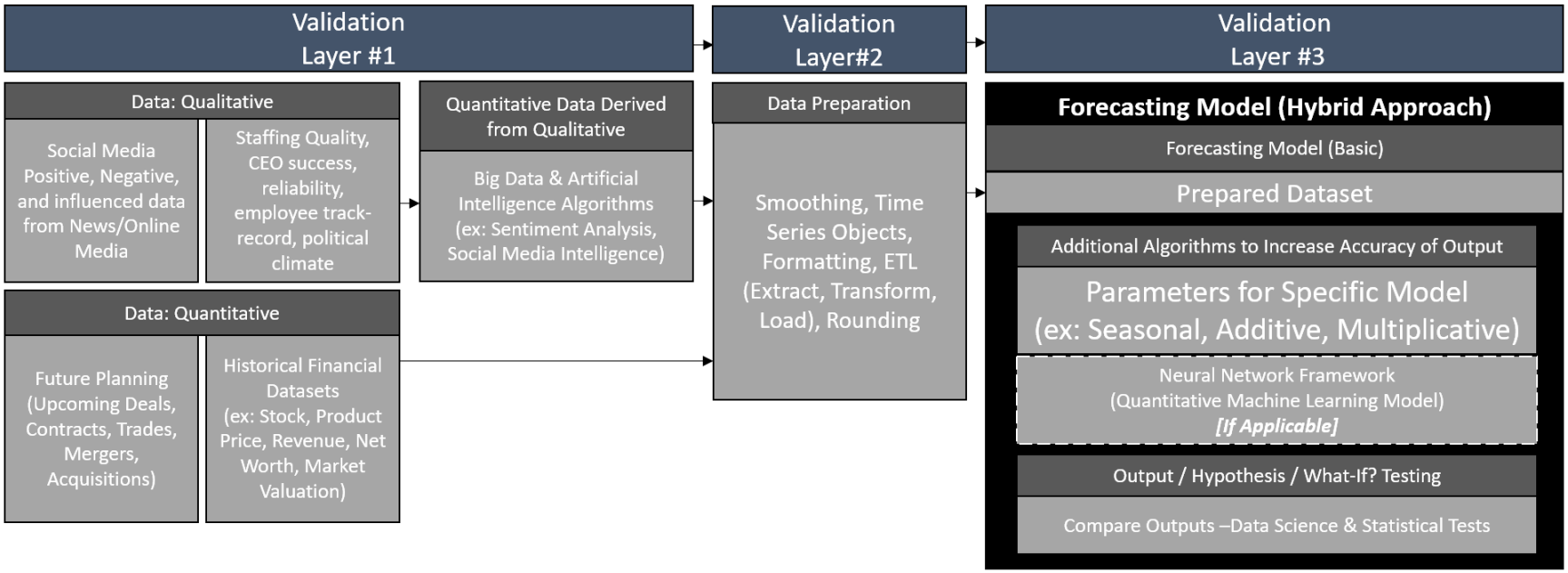



*Figure 8. Proposed Forecast Validation Framework*

It should also be noted that the optional box located in the "Forecasting Model (Hybrid Approach)" section of "Validation Layer #3" includes an optional box for "Neural Network Framework." This is because the strategic framework that is proposed acknowledges the use of advanced forecasting techniques incorporating machine learning. However, validation of neural networks, machine learning and deep learning models is a daunting task that requires computer science validation methods that go beyond statistical and simple model validation described in this dissertation. These complex methods will be discussed in Chapter 6; however, the proposed forecast validation framework can be used beforehand to prepare the data before going into the neural network model so the data set is properly validated.

## IV. Technology Forecasting: Venture Capital Use-case

For the first simulation in Chapter 4, a forecast will be conducted using the 'forecast' package in "R", a generally accepted and popular software and package, followed by methods to validate the model using a venture capital investment data set. The data set that will be used is from the National Venture Capital Association (NVCA) and will attempt to prove that even when the statistical calculations and data are correct, forecast outputs can be drastically different from slight variable changes or smoothing of the data, making it difficult to decide which forecast model is the correct for the decision-maker. The output of the NVCA data set will be focused on the energy industry investments in venture capital and the model will predict future data points. The basic model's parameters will be changed to show the variations in results, which are all correct, and it is up to the data scientist to decide which fits reality for that data set.

In Chapter 4, a simulation will be conducted using data from the venture capital industry. A strategic industry framework that includes validation is one proposed by Steve Blank by creating the Investment Readiness Level (IRL) thermometer. Steve Blank, who is "recognized for developing the customer development method that launched the lean startup movement" and "Lean Launchpad" (Wikipedia, 2019), was able to gauge a startup's level of investment readiness using this 'thermometer' that he derived from the technology readiness level thermometer used by the Department of Defense and NASA for Program Managers (Griffith, 2013). To understand the use-case of venture capital and how a forecast would fit in, it is important to understand the validation process within the industry.

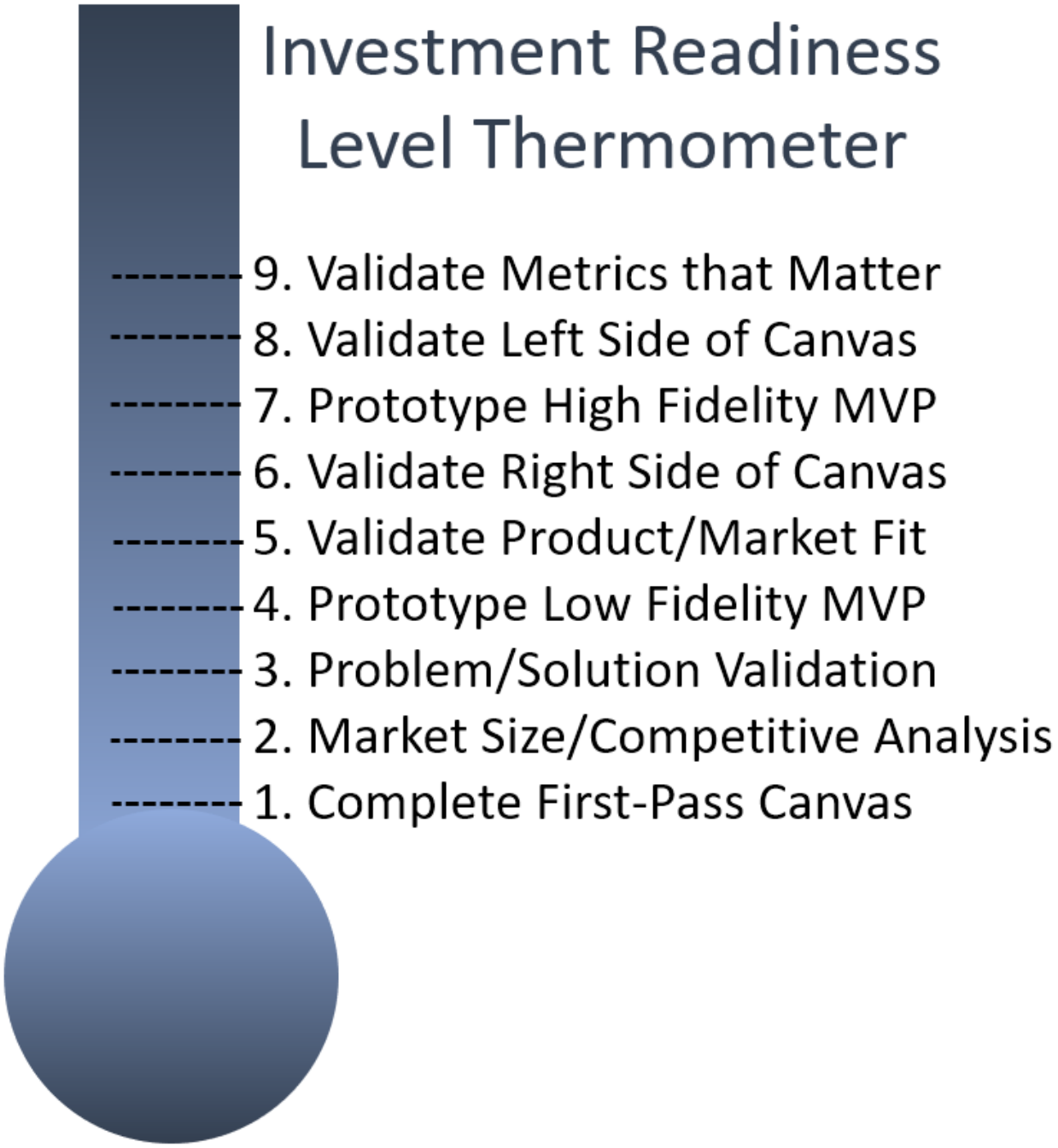



*Figure 9. Investment Readiness Level Thermometer*

*(Griffith, 2013)*

This is an example of using a quantitative approach in venture capital assessment, however, the terms can be considered vague and interpreted incorrectly or force-fit by the venture capitalist. The focus for forecasting can be attributed to #9 in the investment readiness level thermometer, "Validate Metrics that Matter." The following simulation would take place during level #9 to assist with level #4 and level #2. Forecasting in financial disciplines is necessary and commonplace however this 'validation' can be subjective. Validation in these levels may assume experience of the venture capitalist, data scientist, or decision-maker, and rely on qualitative inputs which may differ from firm to firm in terms of standards and weight of the entire toolset. However, by using the proposed forecast validation framework in this dissertation, one can use this as a guide for validation of financial forecasting methods. The venture capital firm, decision-maker, or data scientist must acknowledge and realize the placement and usefulness of this tool in their valuation process. It is designed to help; however, it could be discovered that it hinders their decision-making if not used appropriately and only as a process tool that becomes routine without proper validation of the forecasted outputs, which could cause failure in the investment readiness level thermometer as a metric for investment.

In order to decompose the IRL thermometer, the word "Validate" is used in levels #9, #8, #6, and #5. In level #5 validating the product/market fit may have a different meaning to a different company for all products, or a different meaning for each product, especially if one technology is, for example a hardware item in the energy industry, and another is a software in the information technology industry which requires an enterprise infrastructure to function properly. Similarly, this is why a data set used for forecasting in the finance and venture capital industry may react differently or on a different order of magnitude in a reliably used and

accepted forecasting model than when that same model is used with methanol or energy industry data. It is the data scientist or decision-maker's responsibility to recognize how to move forward with the appropriate level of how they define the term "validate" in many of these steps. The takeaway from the investment level readiness thermometer is that if a venture capital firm is using this as a method, they should use it as a guide and define the terms comfortably within their own company – which leads to the potential of a very different criterion, weighting, and effectiveness compared to another firm or venture capitalist's valuation method. Moreover, it at least makes the data scientist or decision-maker that is constantly validating or using a framework such as the 'Proposed Forecast Validation Framework' in Figure 8, have an industry criterion to assist them in a strategic investment process. The "Investment Readiness Level Thermometer" as proposed by Blank can be used in parallel with the "Proposed Forecast Validation Framework".

## V. Energy Industry Forecasting: Methanol Price Use-case

For the second simulation in Chapter 5, a similar forecast will be conducted using the 'forecast' package in "R", and methods of validating and analyzing the outputs will be discussed. The data that will be used is from the Methanex Corporation, which "supplies, distributes and markets methanol worldwide" which includes monthly pricing data from May 2002 until May 2017 (Methanex Corporation, 2018). This data set represents a tactical or specific product or subset (methanol) in the gas and oil industry of the energy sector. The model will predict future data points using the same models; however, it differs from the strategic level data set for technology forecasting since this is a more tactical and specific subset of data. The goal is to show that forecasting models are not one-size-fits-all, even for data that may seem similar – in this case – financial data from different areas and levels of the energy sector. Energy industry data such as oil price, gas price, methanol price, and others are commonly forecasted due to the need for financial prediction of global markets. It is important for these energy prices to be properly predicted as they have severe global impacts on policy decisions of entire countries.

# Chapter 4: Forecast Validation using Venture Capital Data for Energy Sector

***Executive Summary:*** *The goal of this chapter is to discover data science flaws in forecasting models using venture capital data from the energy sector. Venture capital in the energy sector represents a strategic level, dynamic, and high-risk technology and innovation investment space. Using forecasting as a statistical analysis tool is generally accepted, and this chapter will focus on a literature review including current approaches to forecasting technology and of financial time series data. The simulations (conducted in R) will attempt to prove that even when the statistical calculations and data are correct, forecast outputs can be drastically different from slight variable changes or smoothing of data sets, which require validation by the data scientist. The approaches will be reviewed and validated on the venture capital data set from the energy sector to show limitations and manipulation of present approaches. If successful, by comparing the different outputs from forecasting models in R, this will show that forecasts can be framed or misused which can cause inherent decision-making traps. This impacts policy and decision-makers in venture capital, R&D investing, government policy, and non-profit research centers. The data used in the analysis will be quarterly data from 1995 to 2015 from the National Venture Capital Association (NVCA).*

| Data Sample: Venture Capital Energy Industry Investments | | |
|---|---|---|
| **Year** | Quarter | Total Investment (Millions USD, $) |
| 2006 | 3 | 530 |
| 2006 | 4 | 443 |
| 2007 | 1 | 431 |
| 2007 | 2 | 577 |
| 2007 | 3 | 1034 |
| 2007 | 4 | 934 |
| 2008 | 1 | 1269 |
| 2008 | 2 | 1228 |
| 2008 | 3 | 1270 |
| 2008 | 4 | 875 |

*Table 1. Venture Capital Energy Industry Sample Data*

(National Venture Capital Association, 2018)

## I. Venture Capital Industry

Venture capital in the energy sector represents a strategic level, dynamic, and high-risk technology and innovation investment space. It also is a data set that could be used by a financial firm or government R&D directorate to attempt to identify if it is worth investing in technologies in a specific sector. For the simulation, energy was chosen as it is a largely researched and heavily invested in sector including start-ups, major banks, and government agencies today. In 2008 alone, over $4.1 billion was invested in clean tech, however it declined from 2011 to 2017, with the top 12 clean tech venture capital firms and funds attempting to regain the initial 2008 traction in the marketspace (Pitchbook, 2018). The simulation conducted in this chapter will attempt to prove that even if the model and data are correct, forecast outputs can be drastically different from slight changes or smoothing of data sets. At the minimum this proves the need for validation and that the framework is a valid first step. If successful, comparing the different outputs will show that forecasts can be framed or misused which can cause inherent decision-making traps. This differs from the next chapter's simulation using methanol price data as this is for a strategic industry and sector.

## II. Foundations and Approaches to Technology Forecasting

In order to best forecast future research and development (R&D) and financial investment in new technologies it is important to understand the various approaches to technology forecasting in current academic literature. In this section, some of the new and popular techniques will be reviewed. Many of the approaches are all attempts to add or modify the baseline forecasting formula by adding new variables and applying new methods to increase the accuracy of the output. Many of these variables are specific to the data set or type of industry the forecasting model will be applied to. Technologies such as crowdsourcing, neural networks, and big data analytics have come more mature in recent years and can be applied towards developing new and better forecasting models depending on the reliability and strengths of the data and the data of the newly implemented variables. The standard formula for forecasting from the National Institute of Standards and Technology (NIST) is shown in Equation 2 and is described as, "the new forecast is the old one plus an adjustment for error that occurred in the last forecast" (NIST).

Forecasting the next point:

$$St + 1 = \alpha yt + (1 - \alpha)St, 0 < \alpha \leq 1, t > 0$$

New forecast is previous forecast plus an error adjustment:

$$St + 1 = St + \alpha \epsilon t$$

$$where\ \epsilon t\ is\ the\ forecast\ error\ for\ period\ t.$$

*Equation 2. Forecast Model with Exponential Smoothing*

*(NIST, Forecasting with Single Exponential Smoothing, 2018)*

Technology forecasting and entrepreneurial innovation are related in the current literature since innovative products can be seen as a source of competitive advantage (Porter, 1985) and more recent studies on corporate effectuation and causation (prediction) recognize that R&D projects require a high level of innovation, which is usually attributed as a responsibility to the R&D manager (Brettel, Mauer, Engelen, & Küpper, 2011). Christianson discusses disruptive technologies and the sustainment of technology in "The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail" (Christensen, 1997). This is why it is important to mitigate risk in finance - venture capital backing technological innovation in the tech industry, R&D projects, and innovation in firms – which create new disruptive technologies. These firms can become "unicorn" venture capital portfolio companies, start-ups that reach $1 billion market value through their disruptive technology. Moreover, new strategies in the literature include internal R&D collaboration and support from management within firms and companies (Zhang & Tang, 2017). The promotion of internal R&D collaboration resulted in a positive outcome with employees for long term competitiveness. These are the factors that improved technology forecast models used as variables, and why some technology may be easier to forecast than others (Zhang & Tang, 2017). There are inherent flaws and traps in multiple steps in the forecasting models and they need to be accounted for especially in technology forecasting, since using bibliometrics as a data variable can inform forecast models to be able to predict emerging technologies based on topics such as research activity.

While reviewing the literature regarding forecasting emerging technology much of the research is to discover how companies, start-ups, firms, and venture capitalists can better identify success. Uncertainty in a new product or start-up is epistemic and ontic or "real" in nature and therefore certain researchers argue that forecasting techniques cannot mitigate uncertainty

exclusively (Derbyshire & Giovannetti, 2017). In their model, they use simple forecasting techniques to identify "the 'critical threshold' point where diffusion becomes self-reinforcing" (Derbyshire & Giovannetti, 2017). The technology diffusion point they are attempting to identify is through social network threshold model for innovation proposed previously by combining basic technology lifecycle adopter categories - early adopters, early majority, late majority, laggards (Valente, 1996).

Following this framework, researchers have discovered that Privacy and Security has been the most active sub-technology and they forecast this will continue for the future (Bildosola, Gonzalez, & Moral, 2017). Their study focused on cloud computing research activity since cloud computing heavily impacts information technology enterprises and can be used to gauge the entire market or technology-space at a strategic level. The method proposed by Bildosola, Gonzalez, & Moral is shown Figure 10.

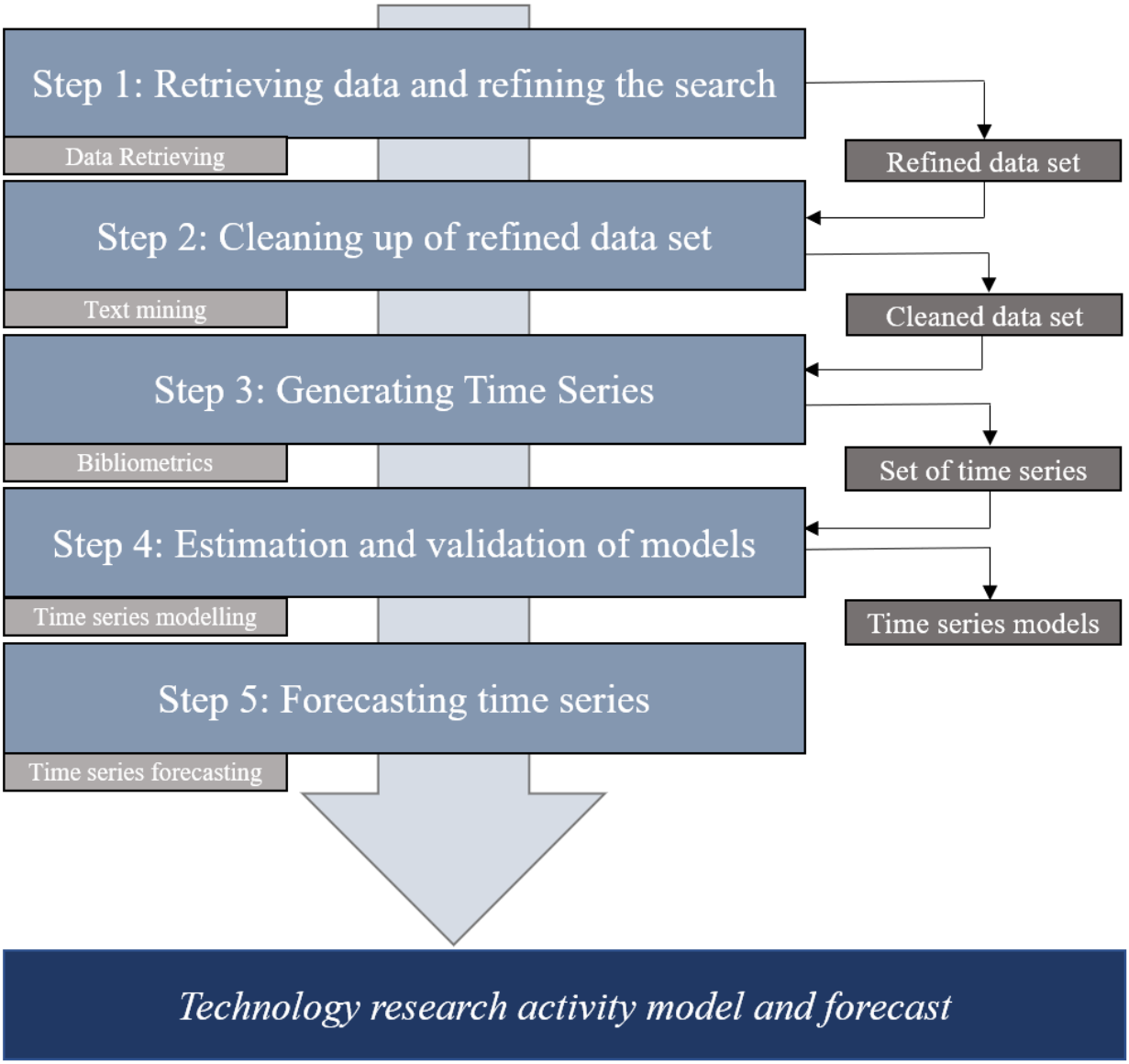



*Figure 10. Research approach flow diagram methodology*

*(Bildosola, Gonzalez, & Moral, 2017)*

This approach uses many transformations of the original data, which each may change, alter, or disrupt the actual data set. Validation should be required within the first few steps of their process to ensure that the refining, cleaning, and creation of the data in a time series format is correctly executed. There could also be significant differences in outputs using this approach due to the data retrieval and search parameters entered. Statistical testing and testing of different search parameters may be of value to a decision-maker or data scientist to perform validation on each of these steps. Step 1 in this approach begins with 'Data Retrieving' to create a 'Refined data set' and Step 2 includes 'Cleaning up of refined data set' which can alter the raw data. For an experienced data scientist, they should note that other methods could be used to 'refine' the data at this step before moving on to Step 3, "Generating time series." In this case, 'Bibliometrics' are being used to generate the time series data sets, and validation is recognized. However, validation should be recommended at each step due to the transformation of the data set. By Step 5 a forecast is conducted which concludes the research approach workflow for using this model (Bildosola, Gonzalez, & Moral, 2017).

Another example of technology forecasting is by using process theory with a design science research approach to understand the relationships between the information technology components, products, and infrastructure, while monitoring a specific technology evolution. Using this method, researchers conducted interviews with industry experts to better understand real-world utility and motivation regarding the components, products, and infrastructure following design science framework (Adomavicius, Bockstedt, Gupta, & Kauffman, 2008). To demonstrate this approach, Wi-Fi technology was used as a real-world example since each generation is certified by IEEE. This showed that there were interdependencies in the technologies and that, "all technology forecasting methods have inherent assumptions, and the accuracy of these assumptions

influences the predictive accuracy of the forecast" (Adomavicius, Bockstedt, Gupta, & Kauffman, 2008). The results of this study were to identify technology evolution by recognizing patterns over time for use in technology forecasting models and using this "…approach may complement existing technology forecasting methods" (Adomavicius, Bockstedt, Gupta, & Kauffman, 2008).

A study on discovering disruptive technologies that uses a supervised learning approach to forecasting for R&D planning concluded that the algorithm can "retrieve as high as 70% of emerging technologies in a given year with high precision" (Kyebambe, Cheng, Huang, He, & Zhang, 2017). In this model, the researchers used patents as their data set, but looked at specific variables such as: number of claims, number of citations, Number of citations made to non-patent literature, Cited Technology Similarity Index (CTSI) to measure the impact outside its own field, and others. They then processed the patent clusters to train the prediction model by creating a Lucene index to create feature vectors and used K-means for the final cluster.

**Cluster-Labeling Algorithm**

T <- start year

**while** T < end year

G <- Set of patent groups granted in year T + 1 and grouped by main class

C <- Set of patents clusters granted in year T and clustered basing on PFV

**foreach** cluster of patents *c* **in** *C*

**foreach** group of patents *g* **in** *G*

$BCS_{gc}$ <- Bibliographic Coupling Similarity between *c* and *g*

**end foreach**

link *c* to *g* **if** their values of $BCS_{gc}$ is the highest

**if** the *main class* of *g* to which *c* has been linked was established in year $T + 1$ label *c* as *ET cluster* **else** label *c* as *NET cluster*

**end foreach**

T <- T + 1

**end while**

*Figure 11. Cluster-labeling algorithm (colorized)*

*(Kyebambe, Cheng, Huang, He, & Zhang, 2017)*

The proposed cluster-labeling algorithm used decides whether the clusters are linked to a classification that emerged in the following year. Major flaws in this model include the changes to the classifications schemes of the patents after 2015, the implementation of social media/sentiment analysis for technology trends via Twitter and Facebook, and some patent databases that do not include full citation lists. These are similar problems that other forecasting models have in common with a supervised approach (Kyebambe, Cheng, Huang, He, & Zhang, 2017). These can also be potential flaws, traps, and biases that the data scientist should acknowledge. When looking at what the actual cluster-labeling algorithm does, it is apparent that it can be edited to change the 'Bibliographic Coupling' and other parameters which could change the outputs severely in attempts to increase or accidently decrease accuracy. The data set or the 'clusters of patents' could also become an issue depending on many variables such as the rate of patent publication, the amount of patents in a certain field of technology, or the impact of many significant patents versus a few but extremely significant and disruptive patents.

Technology forecasting can be used as a strategic level forecasting approach for the energy technology market, however, there are many qualitative data entry and decision-making points which must be accounted for and therefore, validated. In addition, venture capital investments in the energy sector and the financial forecasting models can also be an indicator for future technology investment. In order to analyze the aspect of traps in mathematical and statistical modeling in the decision-making process of venture capital firms, it first must be realized that the environment is dynamic and there are alternative methods that should be compared or used appropriately on a case-by-case basis.

## III. Methodology: Forecasting Venture Capital Data

This section will use forecasting models such as ETS and Holt-Winters, which are generally accepted smoothing techniques used often in forecasting literature and other popular forecasting approaches implemented by R code. The simulation will use venture capital data from the National Venture Capital Association to show statistical traps, flaws, and biases in the output results. This will prove how even if the math is correct, the output can be manipulated or accidently accepted as correct forecast results for real-world decision-making. The difficulty is in the interpretation of the output once the variables have been modified – since the math is correct, and therefore, the forecast is correct, it is up to the data analyst and decision-maker to decide whether or not the forecast results are worth including in their decision-making process. This relies on proper validation and preferably, the experience and knowledge of the data scientist conducting the forecast to use proper judgement when choosing the answer that will impact a real-world decision. The data set includes quarterly price points from 1995 until 2015 which allows for the time series and forecasting packages in R to be used. The time series model is used due to the potential for detecting patterns and seasonality in the data set.

The simulation will output the venture capital energy industry investments data set using the 'forecast' package to predict the next four data points (quarters). The R package will use ETS (Exponential Smoothing State Space Model) on the data and Akaike's Information Criterion (AIC). The output shows an upward trend increasing in dollar value over the next four data points, or quarters. However, there are many smoothing and normalization techniques, and to reiterate, the Holt-Winters Filtering method which is used in the 'R' programming language package for handling time series data will also be used.

# Validation Framework for Simulation #1: Venture Capital Data

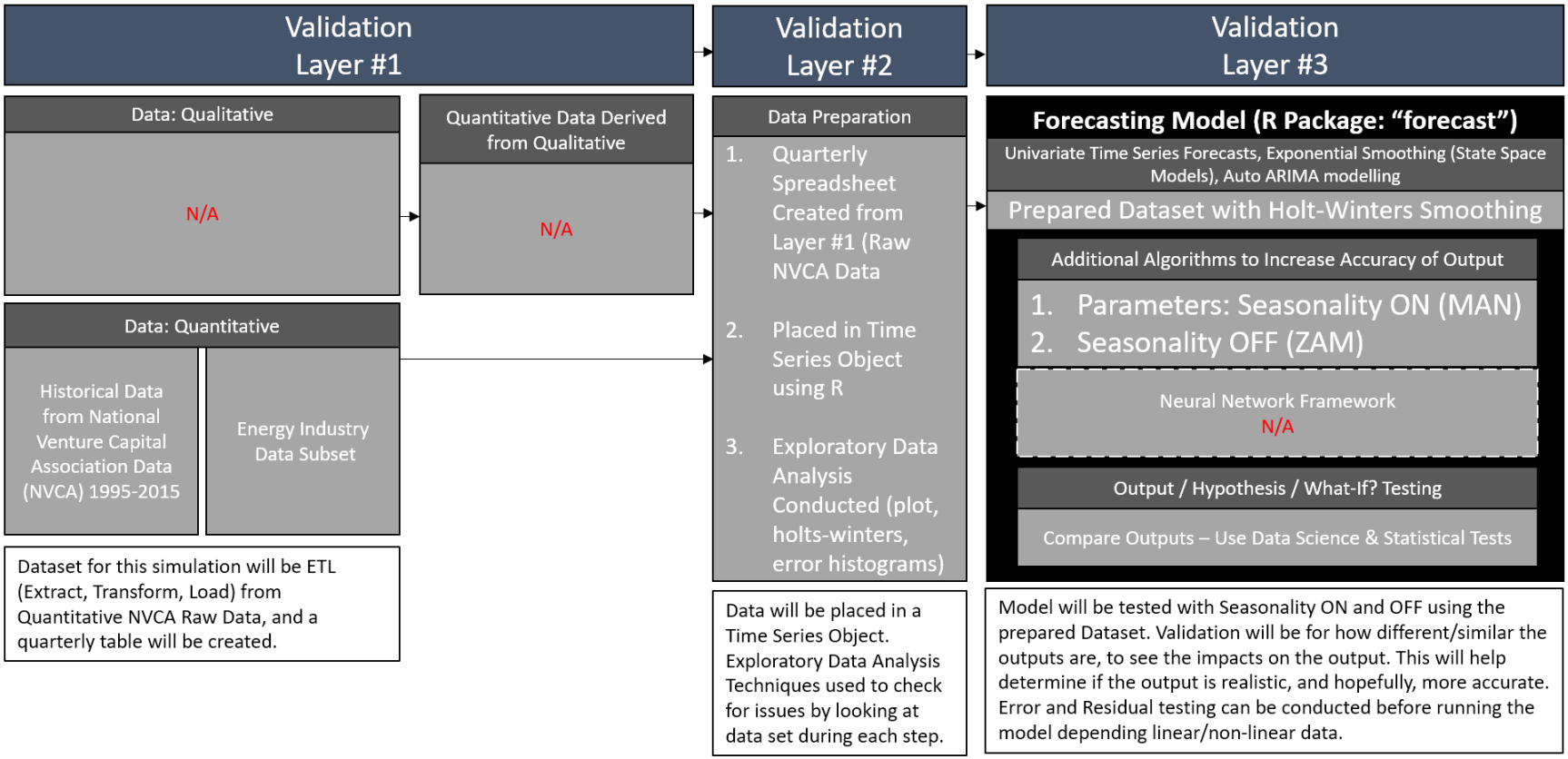



*Figure 12. Proposed Forecast Validation Framework Applied to Venture Capital Data*

## IV. Analysis of Simulation: Forecasting Venture Capital Data using R

The results of this simulation in the next section (Section V) show that there are various outputs using the same data with the same forecasting model package in R for statistics. Through minor parameter changes, the data outputs can vary significantly to a decision-maker. Therefore, the data scientist needs to truly understand how the data is working within the model. In order to better understand that, these outputs are analyzed from the basic forecasting function (Figure 13) to decomposition.

Figure 15 is an example of using the venture capital energy industry investment data set that can easily be misinterpreted or used in a biased or framed way. Figure 15 demonstrates that the smoothing (such as exponential smoothing) as a way of normalizing or preparing a data set and then using it in a forecasting model can cause some suspicious outputs. Therefore, the proposed forecast validation framework presents a layered approach to crafting the model to output results which, in this case, may be favorable to the decision-maker – such as a venture capital firm or start-up company. Figure 15 is a simple graph created in R, using the venture capital industry data set with the popular 'forecast package' in R. Since the 'forecast package' includes various functions used and accepted by R programmers and statisticians, a data scientist easily could believe that they have normalized their data and this is the one and only output, therefore accepting it at face-value. The forecasting formula is shown in Equation 3. The Holt-Winters data set of venture capital investments in the energy industry are outputted in a visually "averaged" looking forecast for the next four data points (quarters) in the graph shown in Figure 15. This smoothing technique causes a significantly different interpretation of the output. Therefore, it is important to realize how smoothing or normalization impacts the data set throughout the analytic process since it can produce an output that a decision-maker may rely on with a high level of magnitude from

other indicators, analyses, reports, and conclusions that prove otherwise. In addition, an erroneous understanding of the output could fortify a negative decision or analysis.

**Forecasting Formula (NIST, 2016)**

Forecasting the Next Point

$$S_{t+1} = \alpha y_t + (1-\alpha)S_t,\ 0 < \alpha \leq 1,\ t > 0$$

New Forecast is the Previous Forecast, plus Error Adjustment

$$S_{t+1} = S_t + \alpha\epsilon_t$$

"In other words, the new forecast is the old one plus an adjustment for the error that occurred in the last forecast." –NIST 2016

*Equation 3. NIST Forecasting Formula*

The difference between Holt-Winters and Exponential Smoothing State Space Model (ETS) is that Holt-Winters estimates smoothing parameters by optimizing the MSE (mean squared error and, using the heuristic values for the initial states beforehand. However, ETS estimates the smoothing parameters and initial states by optimizing the likelihood function (Hyndman, 2011). These methods are used and noted in the Proposed Forecast Validation Framework for these simulation outputs in this research.

The red colored line of Holt-Winters Filtering in the data set represents the new Holt-Winters filtered data set. When this data set is used with the same forecast method, it outputs the following graph shown in Figure 15, which is significantly different than the forecast from the original data set. This method uses the default Holt-Winters filtering that is used in the 'R' package 'forecast'. Without further investigation into the exponential smoothing methods and variables of the function itself, a venture capitalist can easily make an incorrect assumption and create the wrong output. However, since the "math is correct" a decision-maker or R&D manager may not realize this due to the complex multi-stage decision-making process. This is why it is best to understand each ETL (extract, transform, load) conducted on the data, and to realize that some are inherently done within the model, software, or package itself.

Rather than using default parameters in the forecasting model, providing an optional parameter of model="MAN" the result is shown in Figure 16. This is a visually "linear" approach.

The forecast in Figure 17 is conducted using model="ZAM" rather than leaving defaults, where seasonality is turned on and produces what looks to be a 'more accurate' predictive forecast for the next 3 years. This is due to the seasonality influencing the data set and the automatic approach in the parameter.

Figure 18 shows a decomposition function's output on the data set. This is a strong exploratory data analysis technique that should be used as a validation tool. The trend gives a more 'smoothed' look at the data to discover large high and low points, the seasonality can be verified here, and ensures the data has a legitimate seasonal pattern. The output of a decomposition breaks the data set (which is stored in a time series object) into components for further analysis on the y-axis.

Another exploratory data analysis technique is to discover the residuals in the data set, which can often be ignored by linear approaches. This can show the need for non-linear approaches or that some data is being ignored, therefore validating that the output may not be as accurate as it can be and another model, approach, or data preparation method should be used as shown in Figure 14, Figure 15, and Figure 16.

To continue analyzing the data set, using the ACF function can display the estimates of the autocovariance or autocorrelation function. For this function, a 'Lag' of 50 was used, however, this number should be adjusted by the data scientist to attempt to find the best fit. (R-core@R-project.org, 2019). This is shown in the output graphs Figure 17, Figure 18, and Figure 19.

Creating a function to display the histogram of forecast errors can provide a deeper look into the data set to discover outliers or best fit compared to others as shown in Figure 25, Figure 26, and Figure 27. It should be noted in these histograms to recognize or reconfigure the axis to scale properly for comparison as to not visually mislead the data scientist or decision-maker.

## V: Simulation Results: Forecasting Venture Capital Data

Plot **DEFAULT** Model (next 4 Quarters)

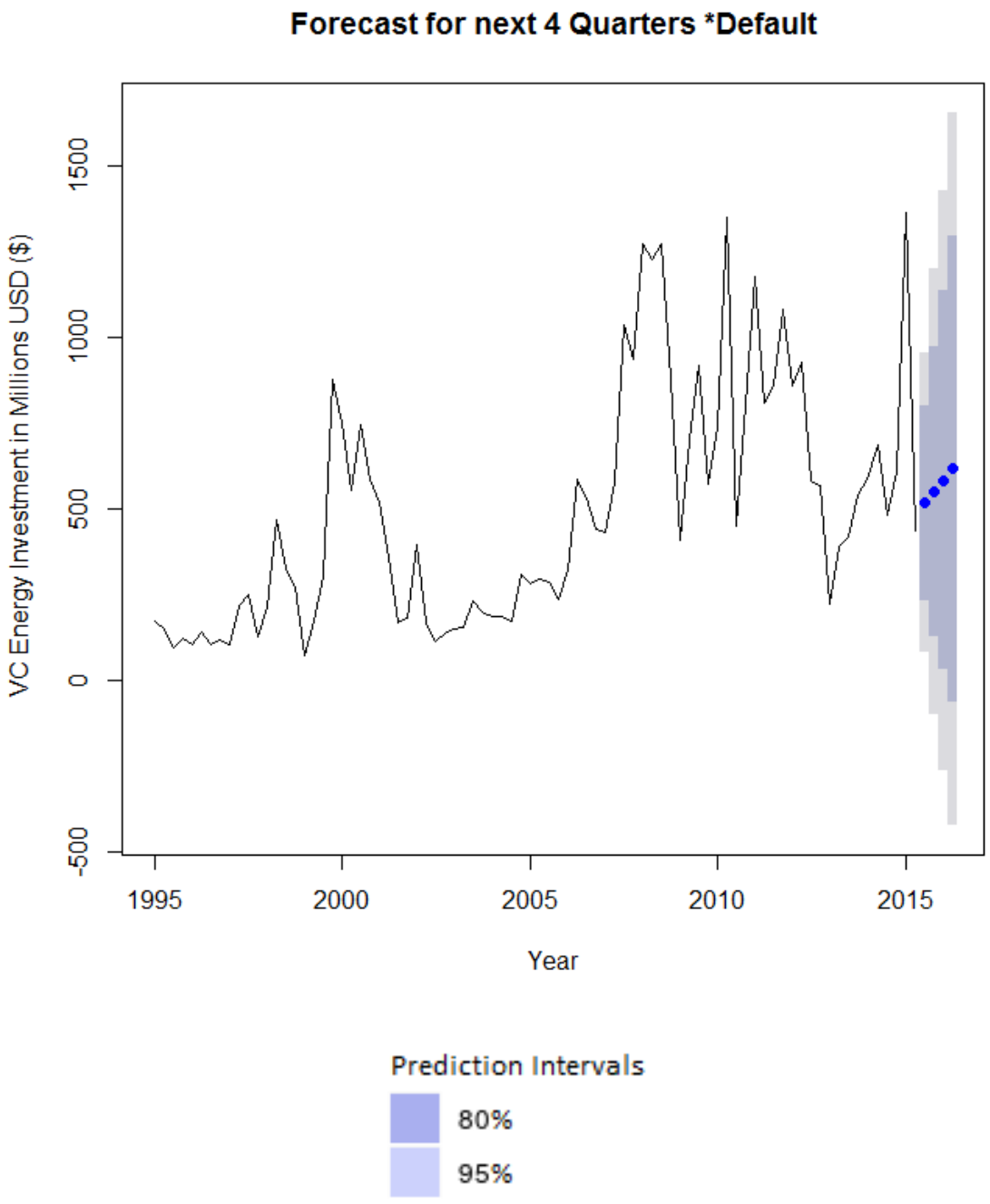



*Figure 13. Forecast for next four Quarters using Venture Capital Data (1995-2015) (Default)*

Plot **Holt-Winters**

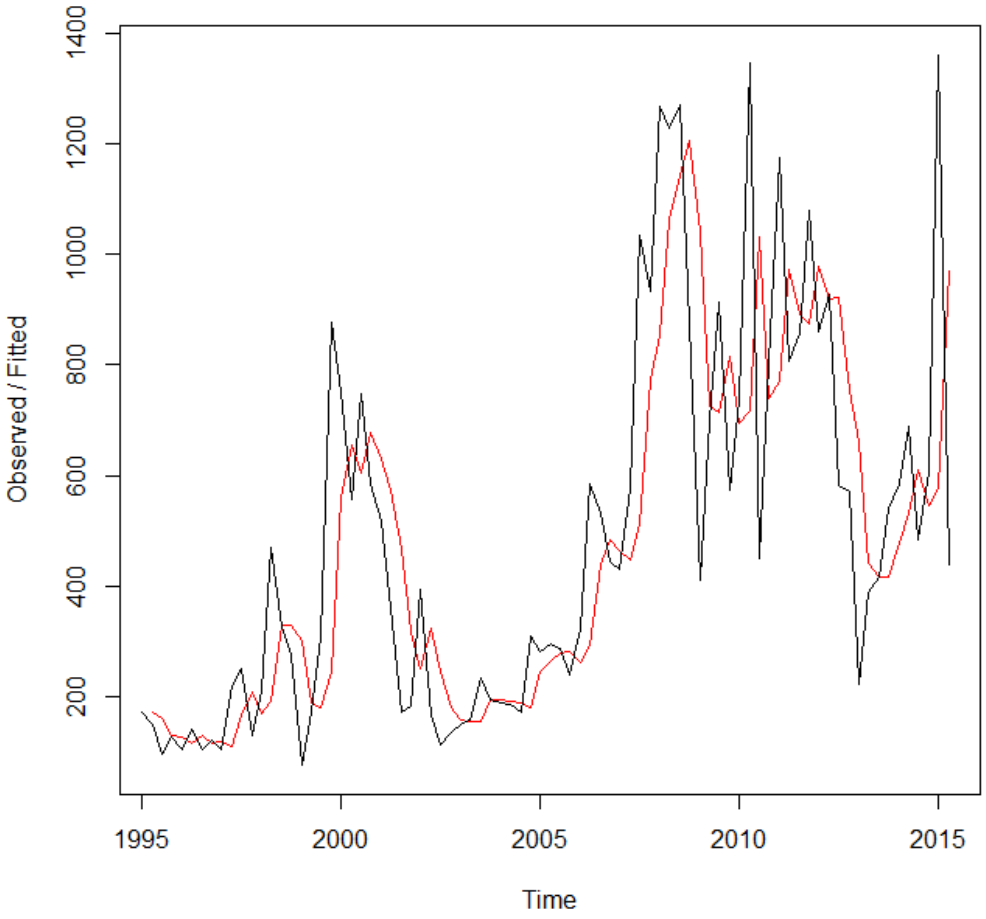



*Figure 14. Venture Capital Data (1995 – 2015) with Holt-Winters Filtering (Red)*

Plot **Holt-Winters** Forecast Model (next 4 Quarters)

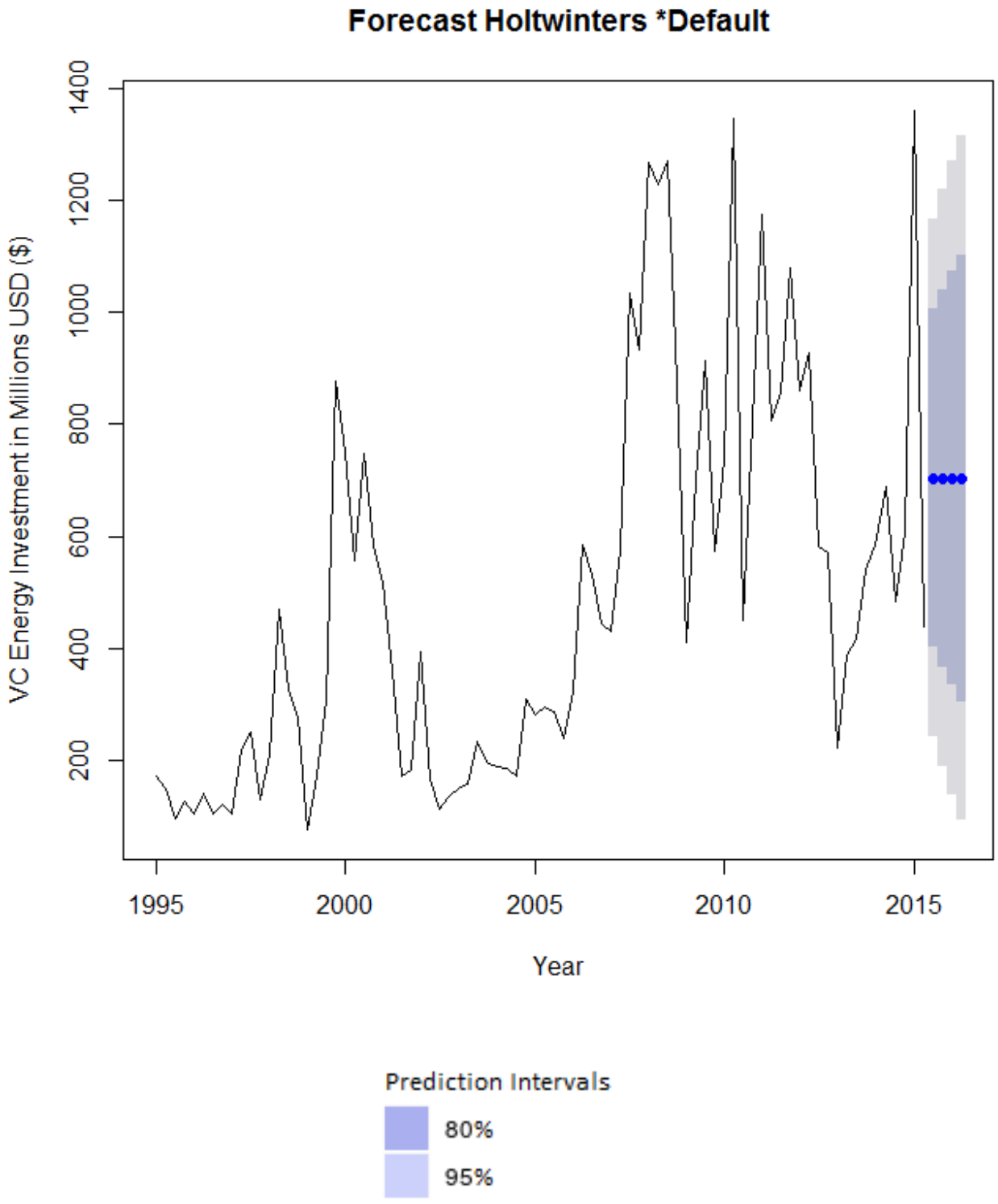



*Figure 15. Forecast for four Quarters (Default) after Holt-Winters Preparation*

Plot **Holt-Winters** Forecast Model (next 3 years)

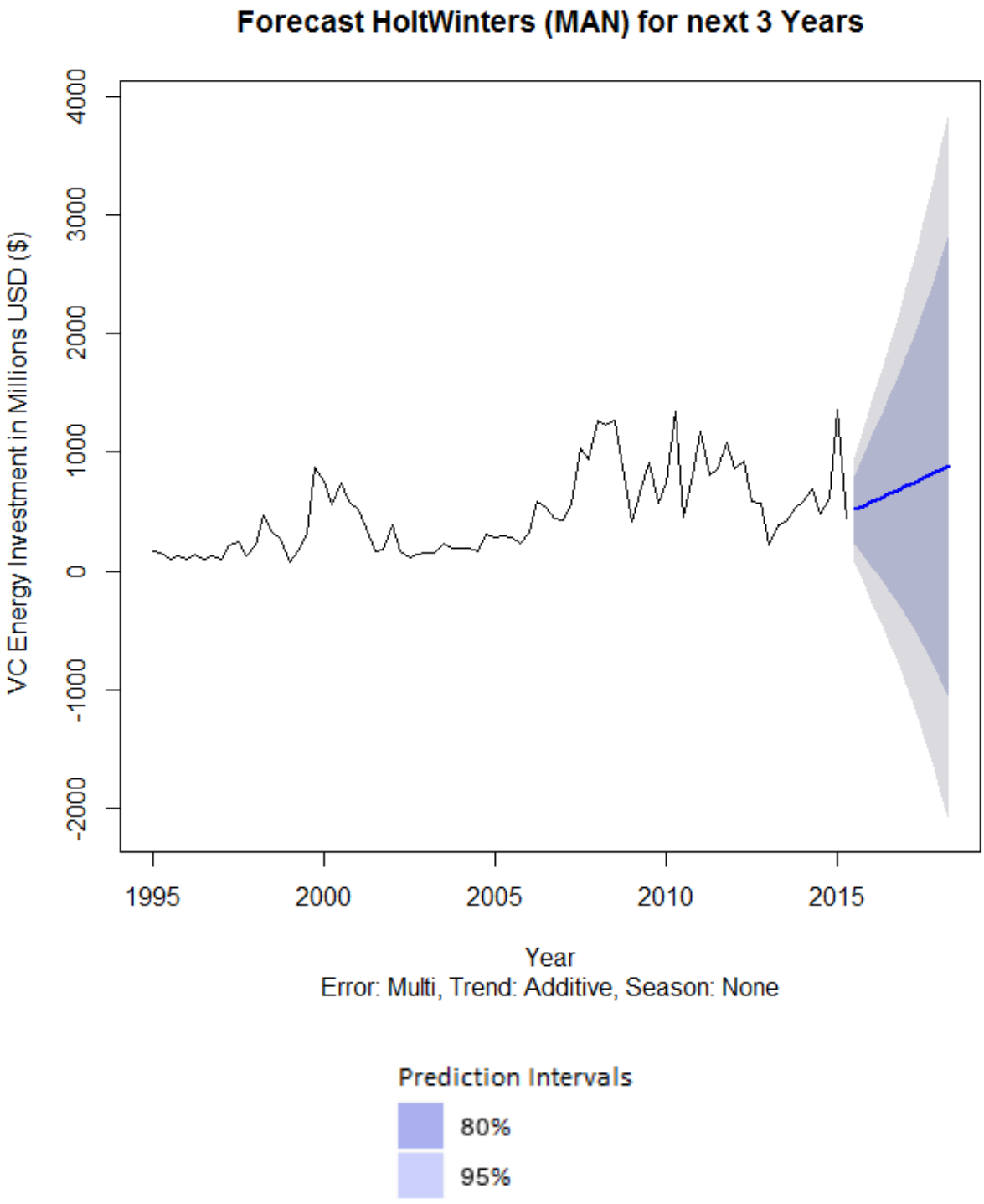



*Figure 16. 3 Year Forecast using Holt-Winters and MAN Model Parameters*

Plot **ZAM** Model Forecast (Error: Auto, Trend: Additive, Season: Multi)

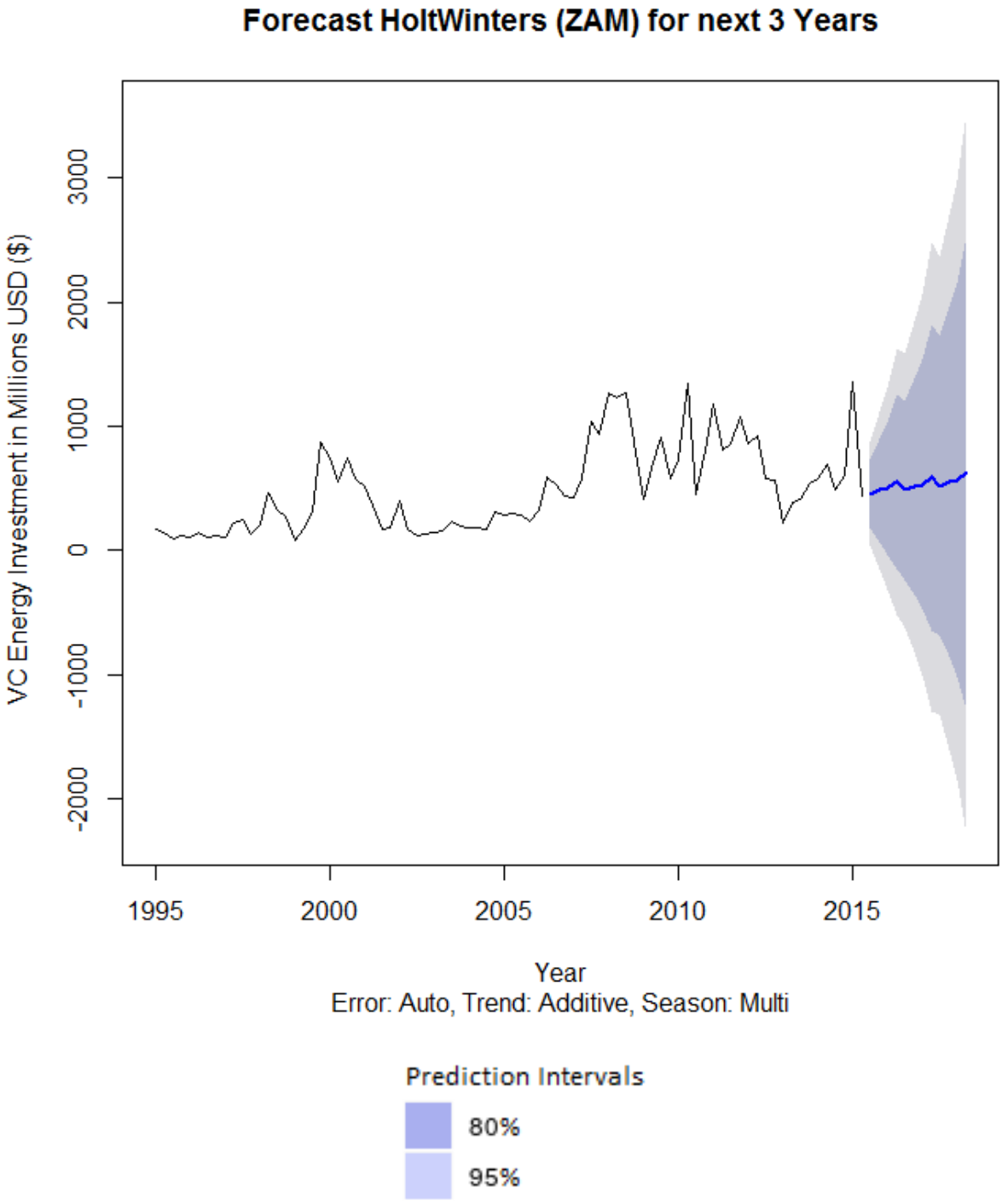



*Figure 17. 3 Year Forecast using Holt-Winters and ZAM Model Parameters*

**Decompose** Venture Capital Data set Time Series Object (*confirm seasonality*)

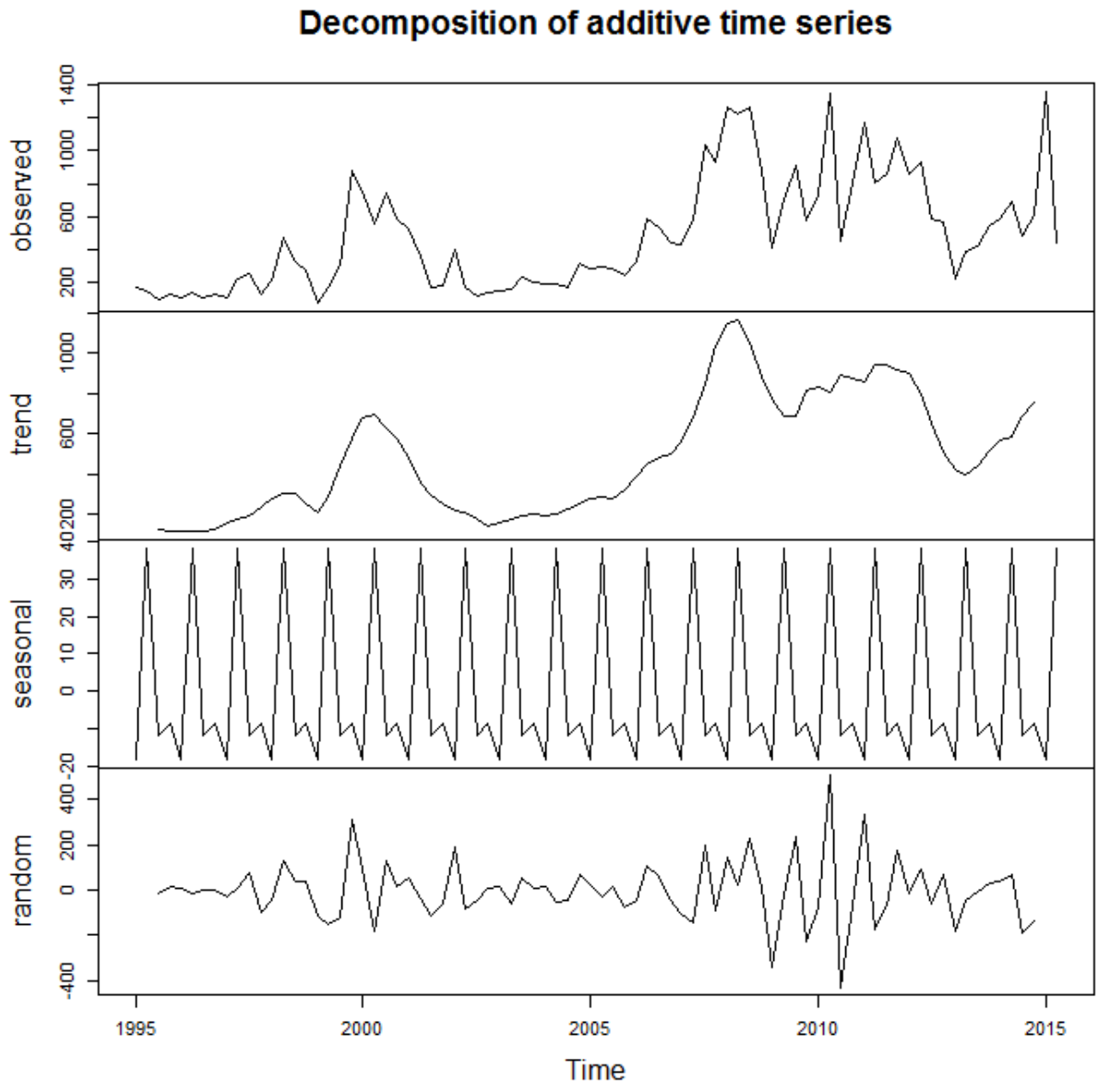



*Figure 18. Time Series Decomposition of Venture Capital Energy Data*

Plot **Holt-Winters** Model Forecast **Residuals**

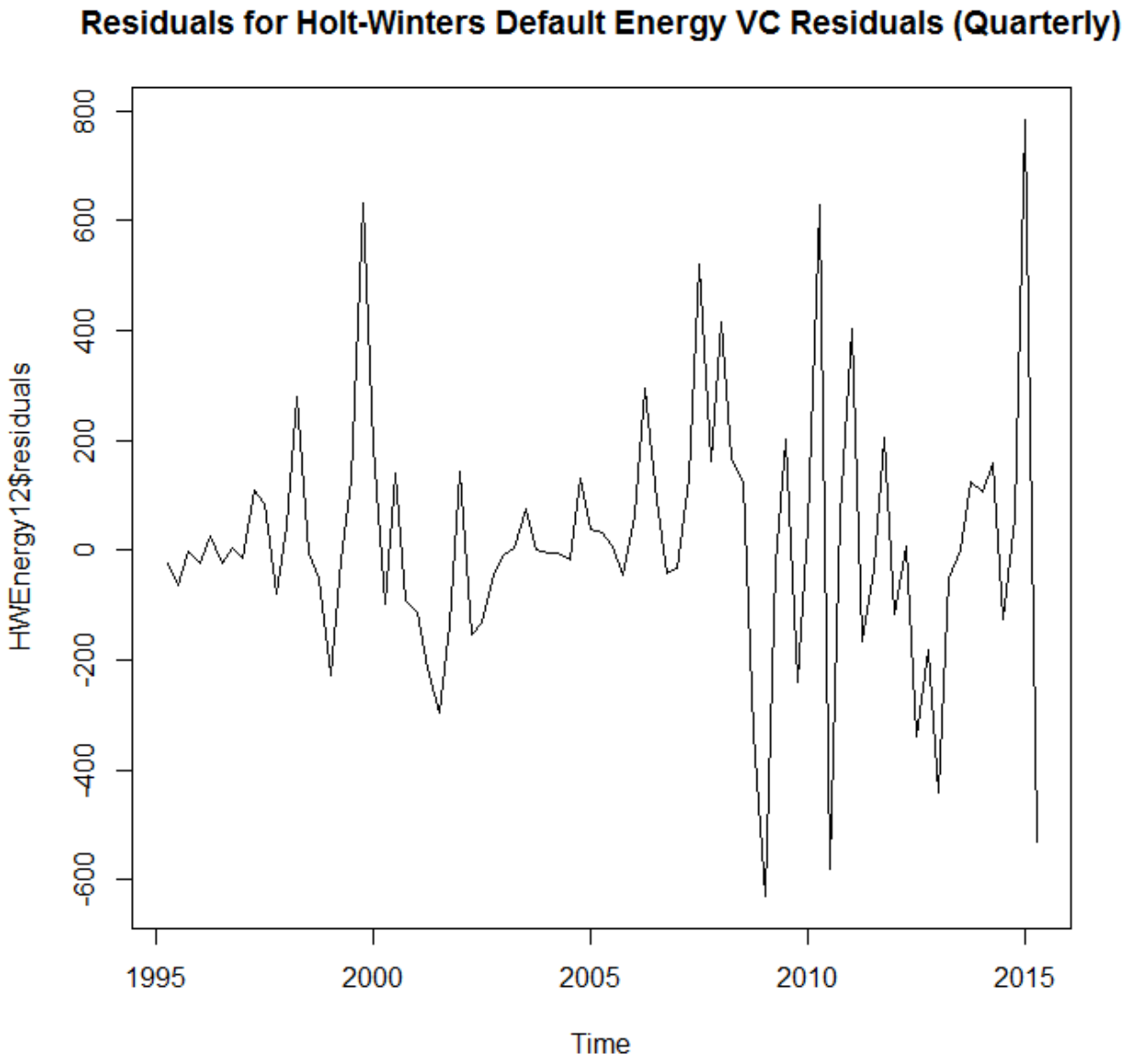



*Figure 19. Residuals of Quarterly Venture Capital Data using Holt-Winters Smoothing*

Plot **MAN** Model Forecast **Residuals**

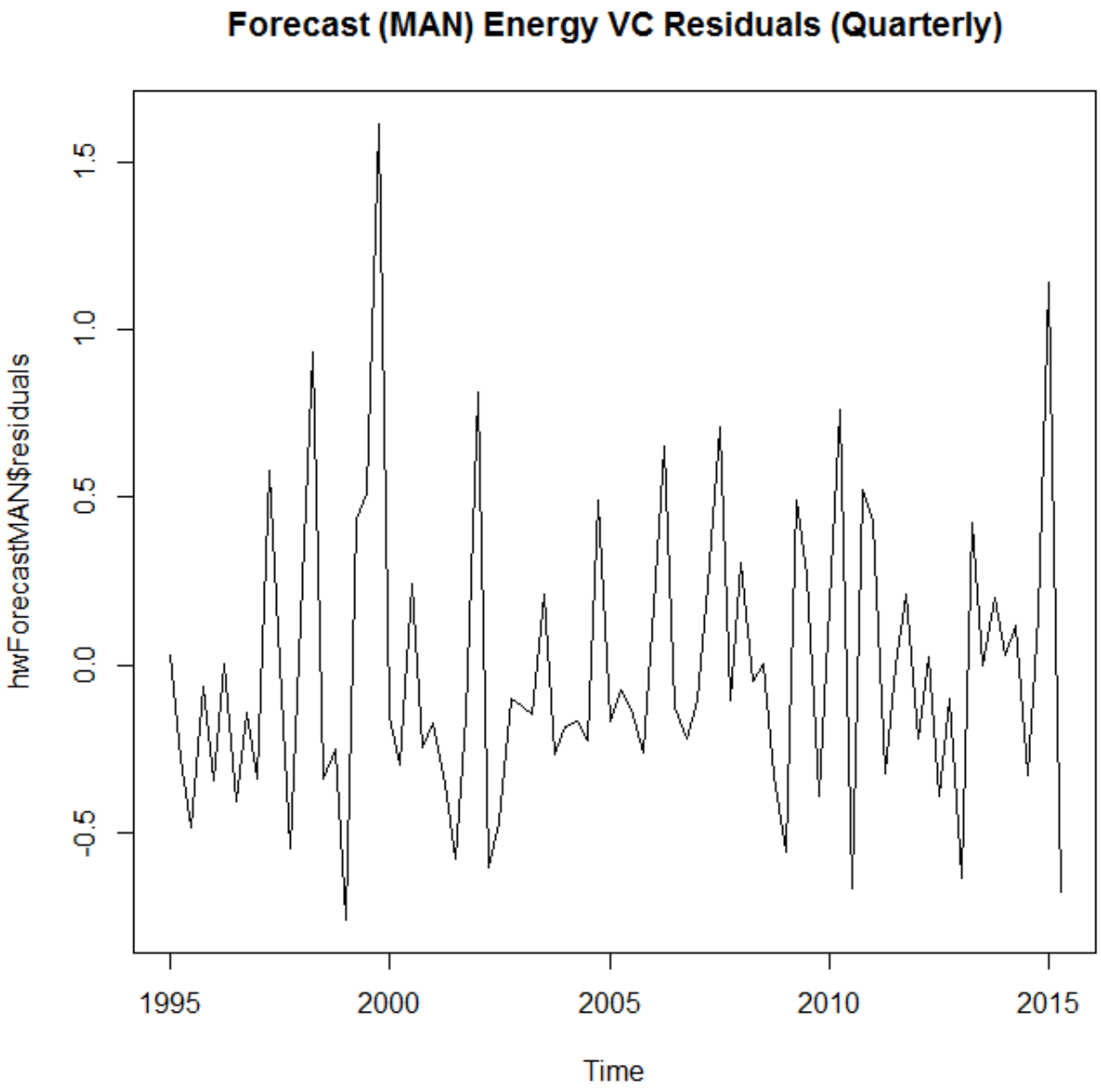



*Figure 20. Residuals of Quarterly Venture Capital Data using Forecast (MAN)*

Plot **ZAM** Model Forecast **Residuals**

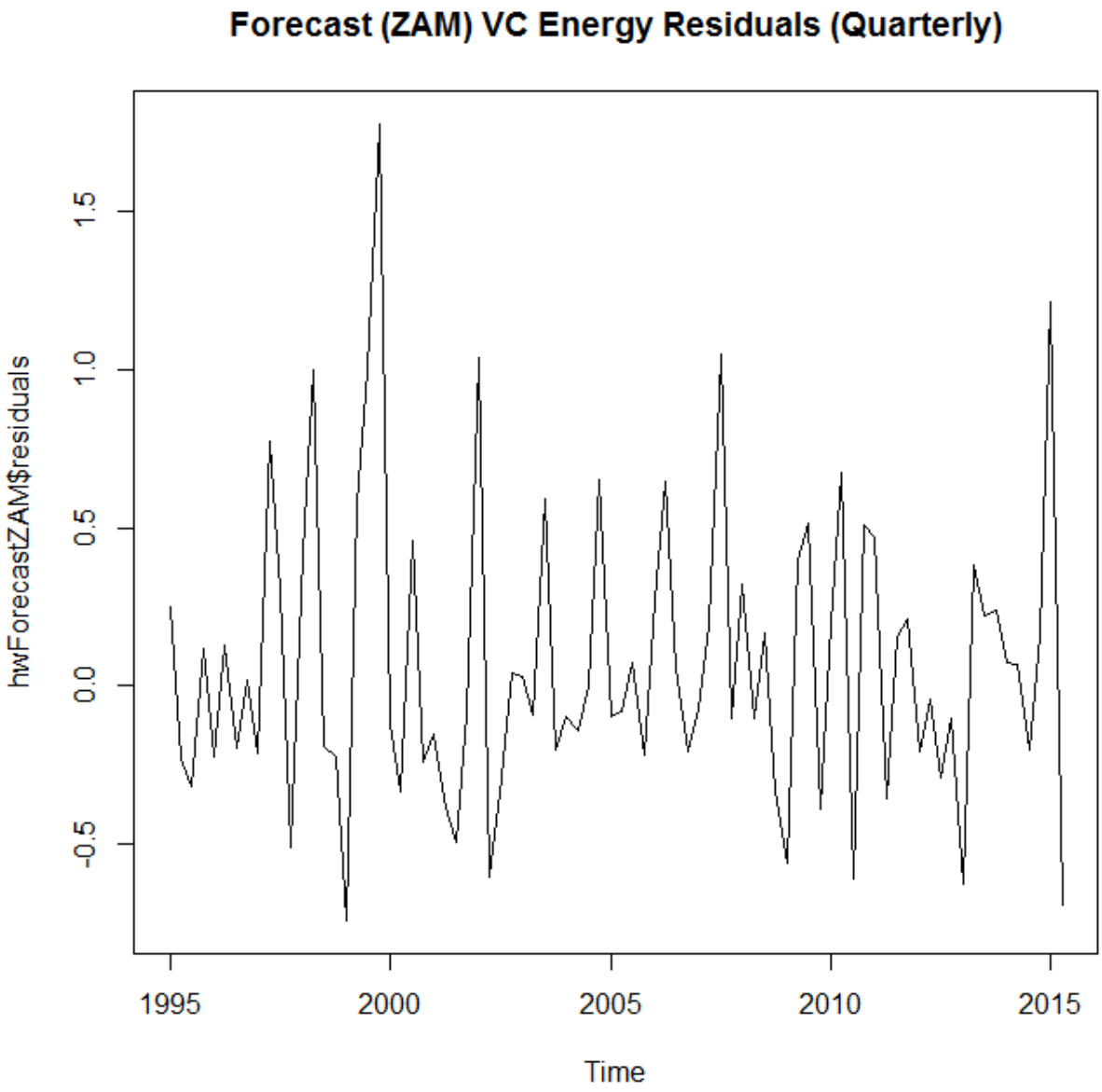



*Figure 21. Residuals of Quarterly Venture Capital Data using Forecast (ZAM)*

Plot **ACF** for **Holt-Winters** Forecast Model

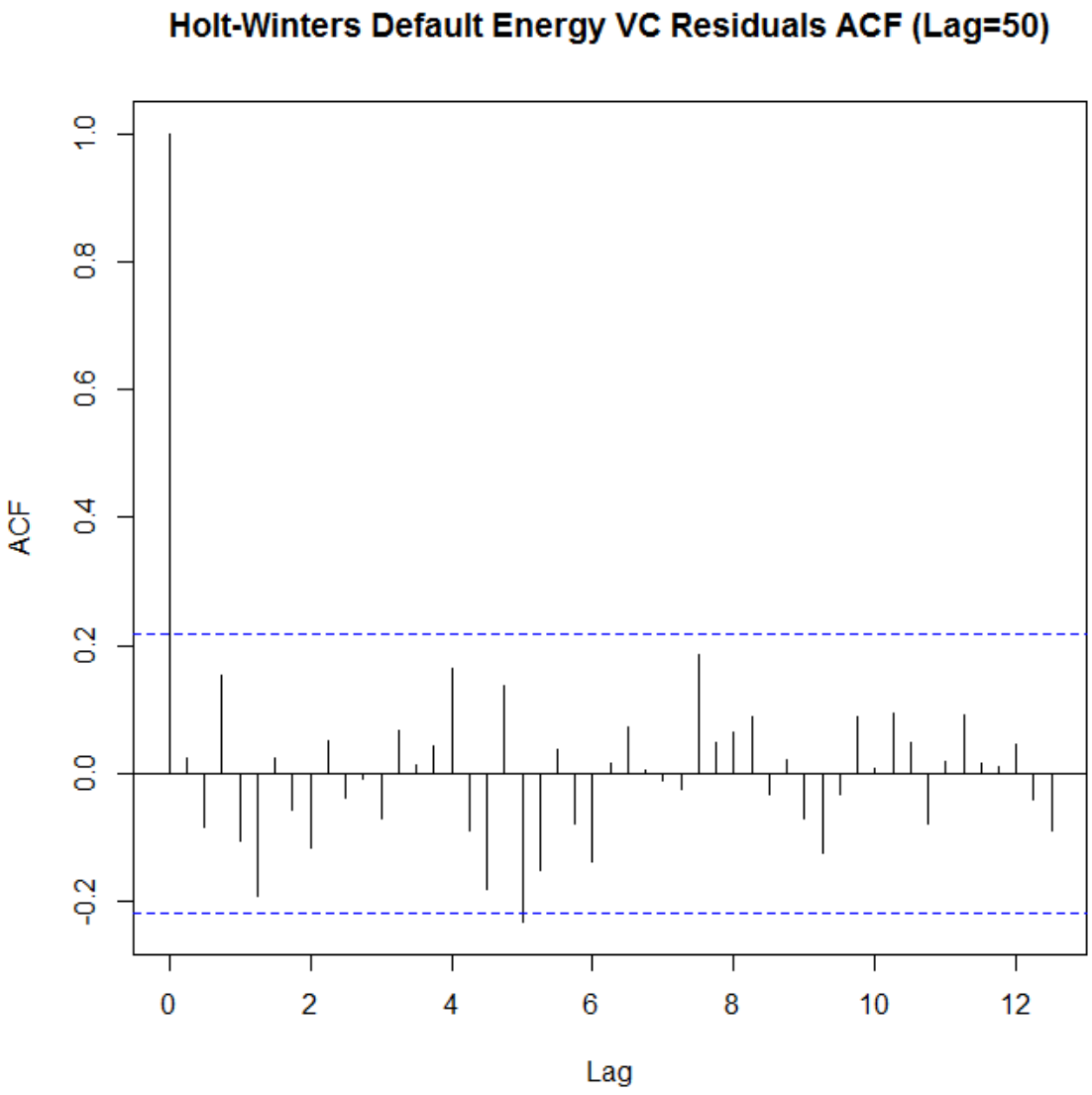



*Figure 22. ACF for Energy Venture Capital Data Set Residuals (Default)*

Plot **ACF** for **MAN** Forecast Model

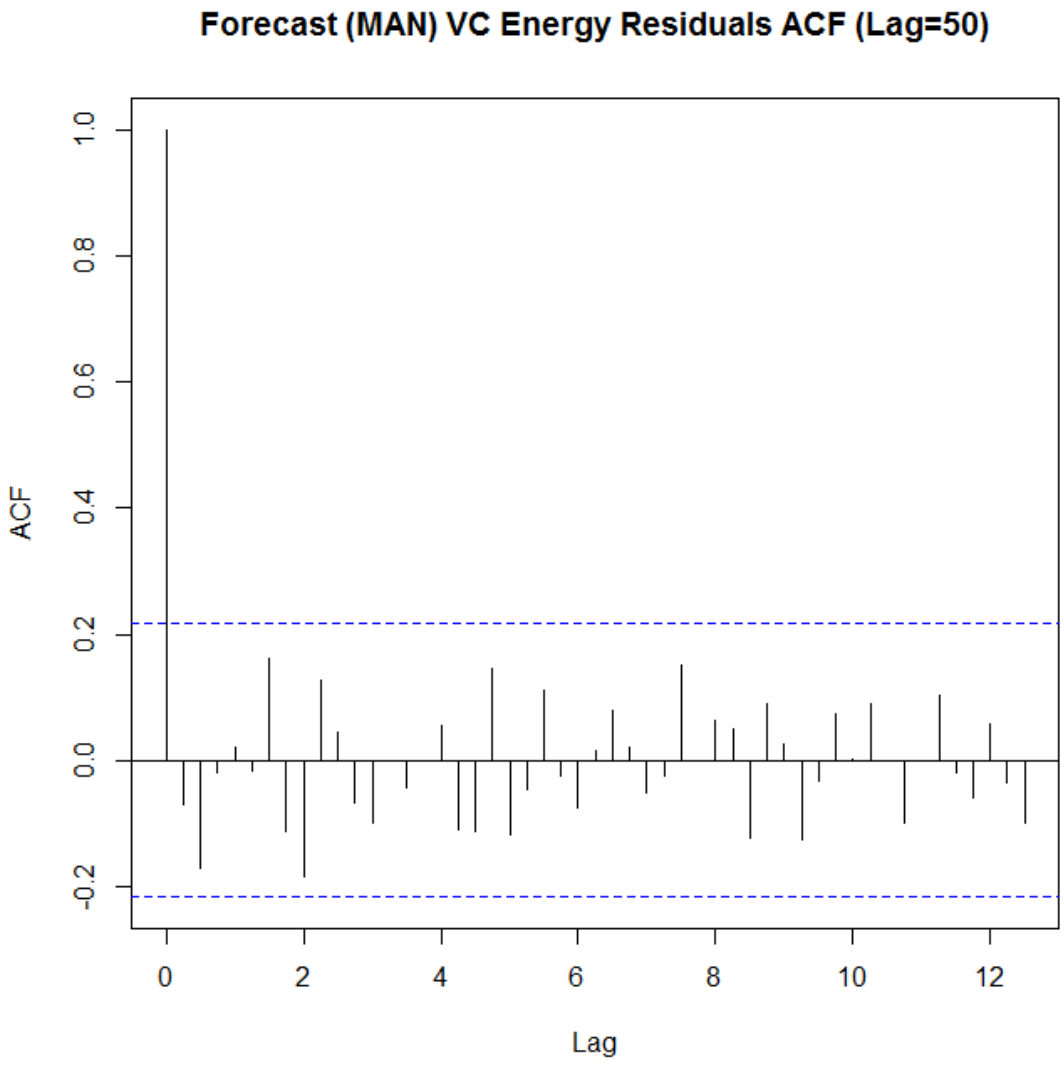



*Figure 23. ACF for Energy Venture Capital Data Set Residuals (Model=MAN)*

Plot **ACF** for **ZAM** Forecast Model

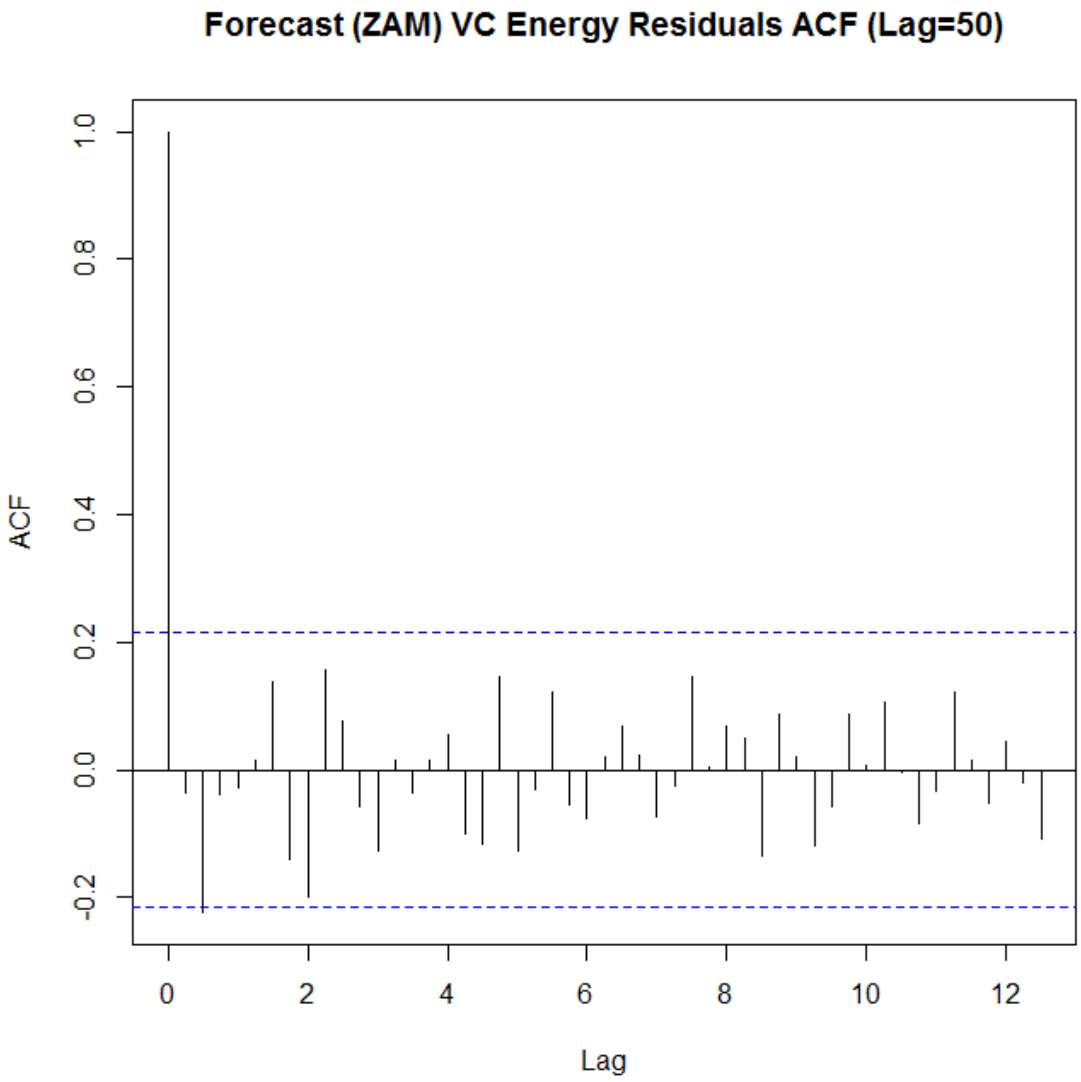



*Figure 24. ACF for Energy Venture Capital Data Set Residuals (Model=ZAM)*

Generate **ForecastErrors Histogram** on previous **Holt-Winters** Forecast output data

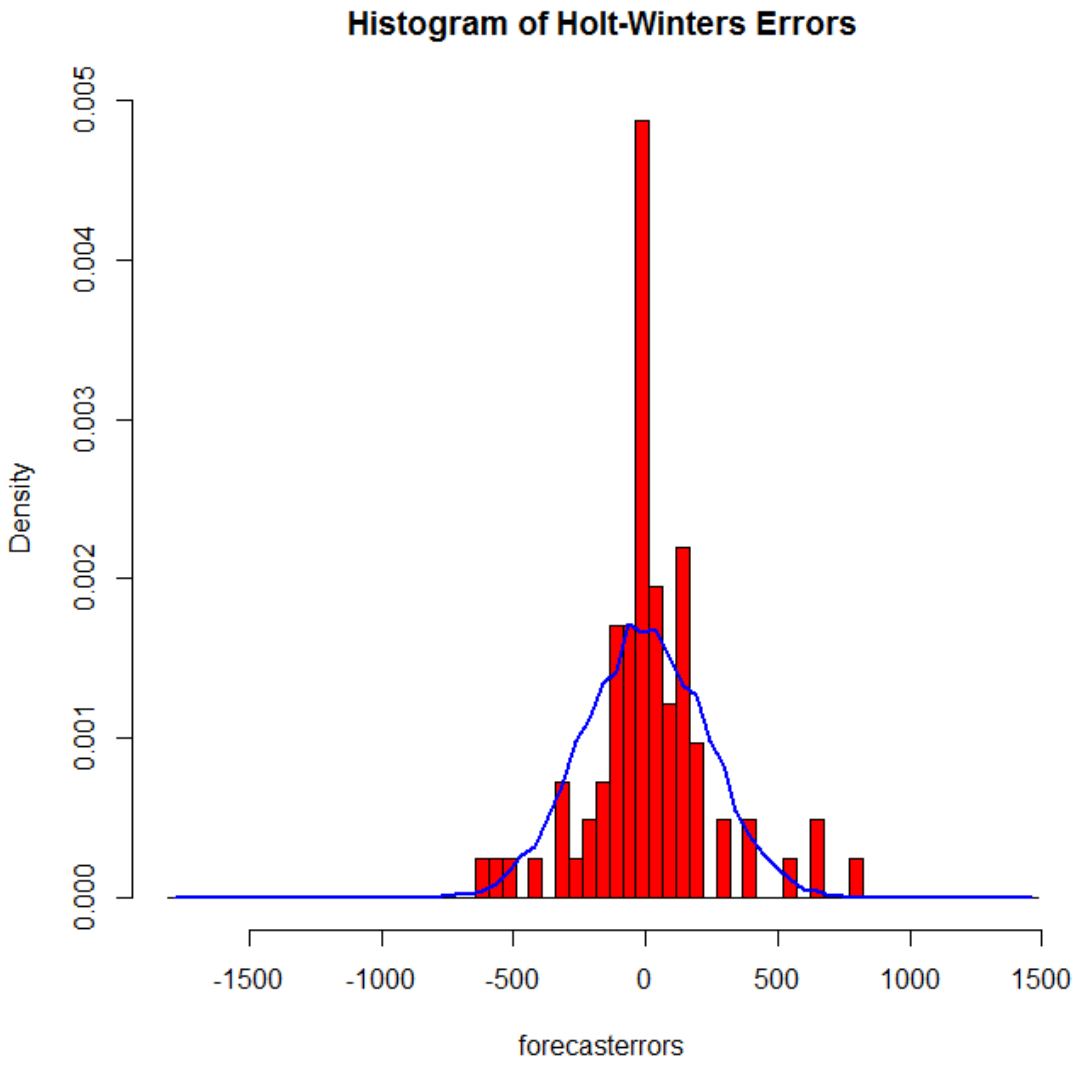



*Figure 25. Histogram of Errors: Energy Venture Capital Data, Holt-Winters*

Generate **ForecastErrors Histogram** on previous **MAN** Model Forecast output data

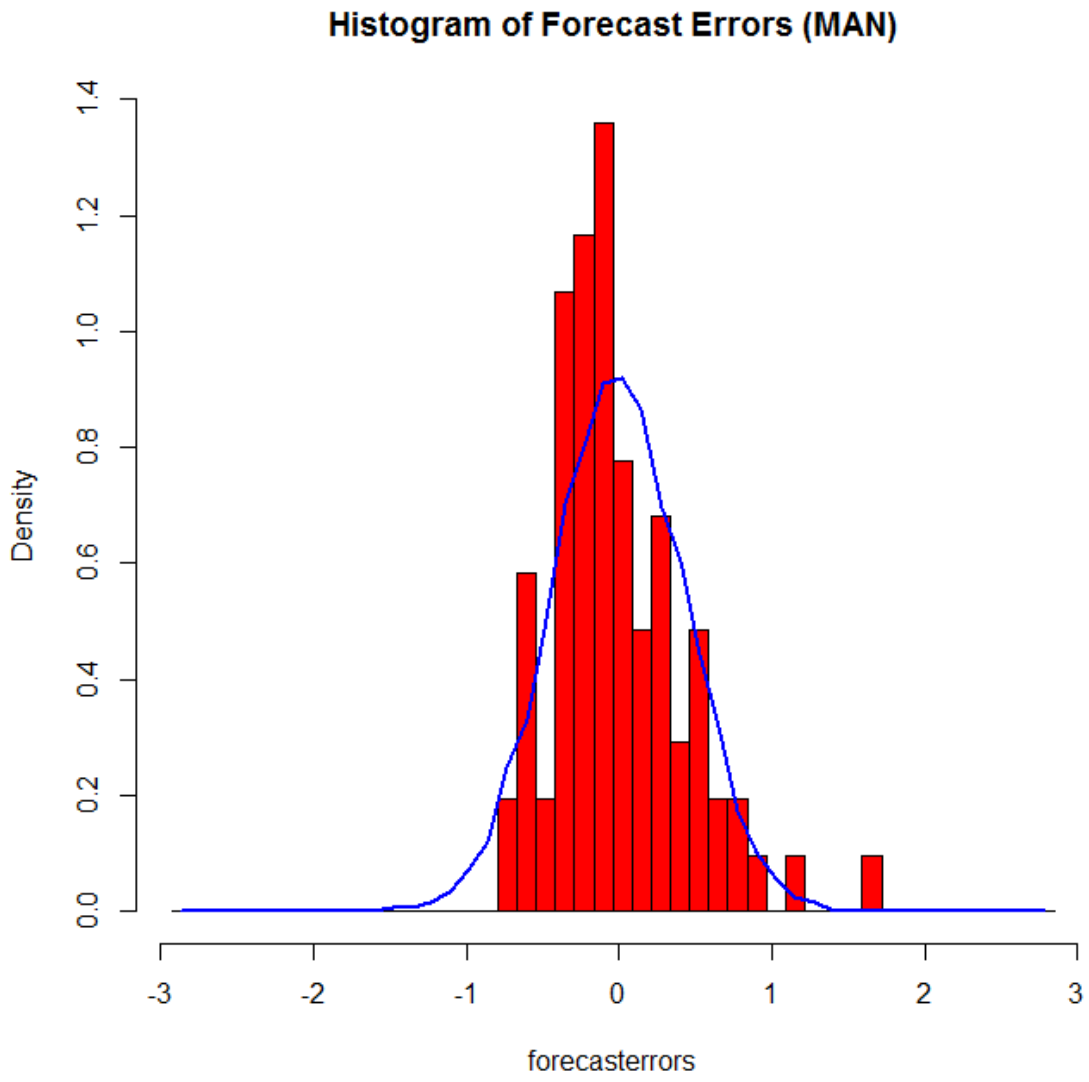



*Figure 26. Histogram of Errors: Energy Venture Capital Data, Model (MAN)*

Generate **ForecastErrors Histogram** on previous **ZAM** Model Forecast output data

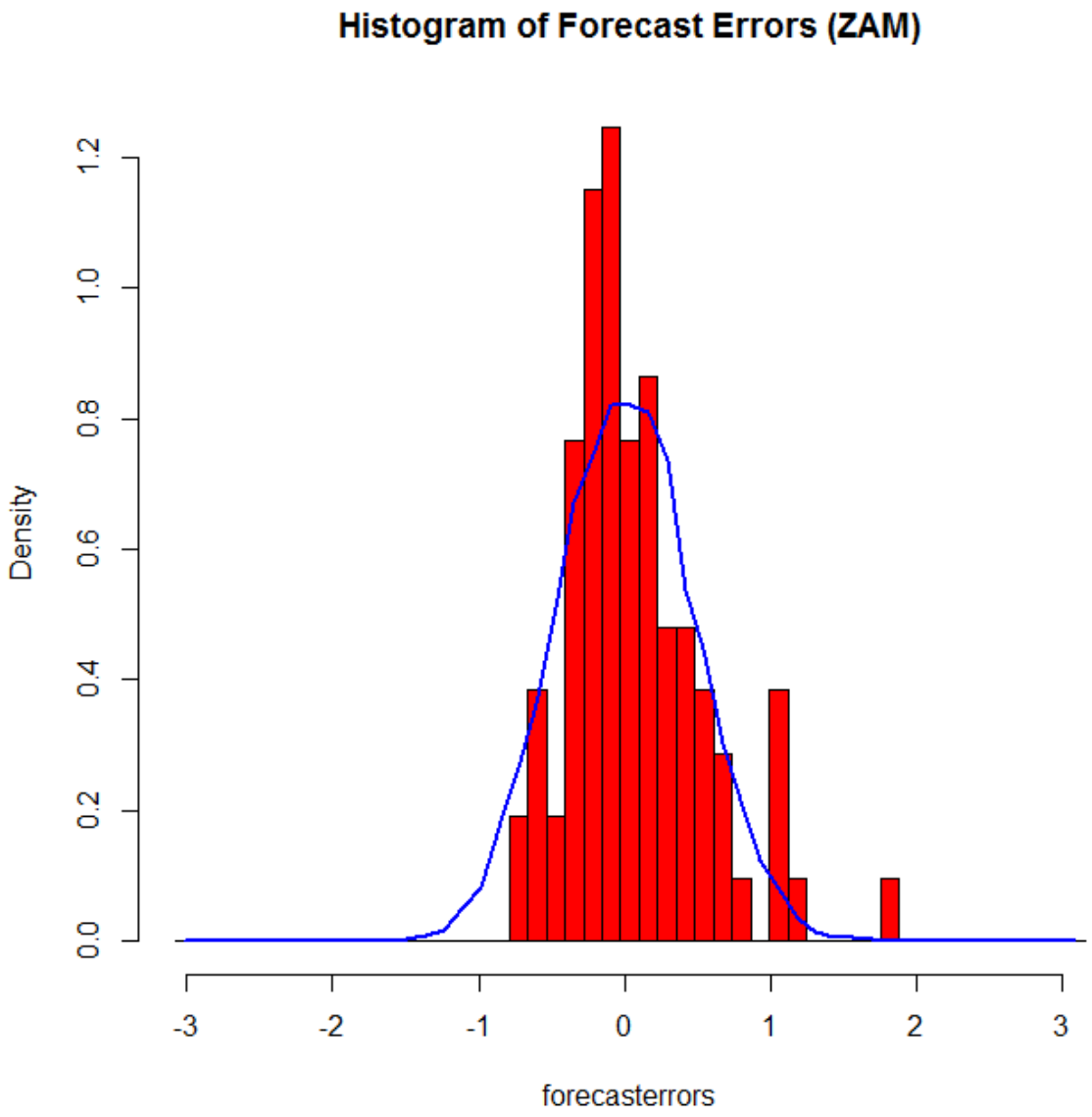



*Figure 27. Histogram of Errors: Energy Venture Capital Data, Model (ZAM)*

## VI. Simulation Takeaways

In Section V various forecasts were conducted using the same forecast function in the 'forecast' package in R with smoothing, taking advantage of common parameters, to attempt to discover differences in outputs. Figure 13 is the output from the default forecast on the venture capital energy industry data set. The results show a positive trend for the next four quarters (or one year), which could be interpreted by a decision-maker as a positive trend, and therefore influence their decision-making process. This could be the considered a common output that anyone using R with the same data set would come to as a 'correct' conclusion. In some cases, this graph could be used in a business presentation to influence or convince an audience of the intended goals based on this positive trend line. However, in the later simulations, such as Figure 15, the same forecast is conducted after a Holt-Winters smoothing is applied which gives a very different output that may not be as convincing to include in a presentation. The goal of these simulations is to show the process of how to validate a forecasting model for use in the real-world and recommend a few output graphs to create forecasts for following the Proposed Forecast Validation Framework: Validation Layer #3 shown in Figure 12.

Another takeaway from this simulation is to watch out for visual traps, such as in the error histograms scaling in Figure 25, Figure 26, and Figure 27. Be mindful of the Y-axis units and ensure that each graph is compared properly to avoid visual biases. This is something that can occur in multiple output graphs, many times this is due to leaving default parameters and not setting the specific labels for the X-axis and Y-axis.

The results in this section show how different outputs can be based on the choice to use optional parameters. A data scientist should also be cognizant of the size of the data set, amount of data points, time series spacing (for example: quarterly, monthly, yearly), and if possible,

another validation point could be to attempt to run the same analyses on another data set possibly from another industry's venture capital data that complements or has a similar data plot.

# Chapter 5: Forecasting Validation in Energy Markets: Methanol

***Executive Summary:*** *The goal of this chapter is to review the literature in a specific energy industry – methanol – and utilize a data set of monthly pricing data of methanol to conduct statistical predictive analyses using forecasting models to evaluate how accurate the forecast results are against known data. By portioning the methanol price data set, one can discover if there is over-fitting and identify other statistical traps by changing tuning parameters as a method of validation. This chapter will use current approaches to forecasting financial pricing data and if successful, attempt to discover best-fit forecasts and prove that not all forecasting outputs are suited for all data by simple trial and error testing. Understanding these analytic tools impact policy and decision-makers in science & technology planning and investing and directly applies to the energy, gas, and oil industry. The data that will be used is from the Methanex Corporation, which "supplies, distributes and markets methanol worldwide" which includes monthly pricing data from May 2001 until May 2017* (Methanex Corporation, 2018). *However, oil price, or other energy industry product prices could easily be applicable for these simulations.*

| Data Sample: Methanol Price (Full Data includes May 2001 – May 2018) | |
|---|---|
| **Date** | Price per Gallon (in USD, $) |
| **May 2018** | 1.23 |
| **June 2018** | 1.16 |
| **July 2018** | 1.13 |
| **August 2018** | 1.13 |
| **September 2018** | 1.16 |
| **October 2018** | 1.19 |
| **November 2018** | 1.16 |
| **December 2018** | 1.24 |

*Table 2. Data Sample: Methanol Price (May 2001 - May 2018)*

*(Methanex Corporation, 2018)*

## I. Methanol Industry

Methanol is a large volume global commodity chemical. Worldwide, there are over 90 plants with a combined annual capacity of 36.6 billion gallons. Since 2015, the demand for methanol has grown by over 50%, in part due to its growing use as a transportation fuel. The reason for choosing methanol data for this dissertation and its simulations are due to the growing demand, along with price data availability from the Methanol Institute and Methanex, the largest producer of methanol located in Canada (Methanol Institute, 2019). The simulation conducted in this chapter will attempt to prove that even if the model and data are correct, forecast outputs can be drastically different from slight changes or smoothing of data sets, as shown in the previous chapter, using venture capital data from the energy industry. At the minimum this will also prove there is a need for validation and that the framework is worthy of using to self-check an expert in the field. If successful, comparing the different outputs will show that forecasts can be framed or misused which can cause inherent decision-making traps and possibly cause errors in financial and global predictions for a specific type of fuel or chemical. This differs from the venture capital energy data as this is for a specific product at a tactical level.

## II. Foundations and New Approaches to Financial Energy Pricing Data

A review of models and approaches to forecasting using financial data will be discussed in this section. Many of these approaches incorporate new technologies and algorithms developed in data science and computer science practices which have produced algorithms and architectures to attempt to engineer more accurate forecast outputs with the help of big data, GPU implementations of neural networks (deep learning), and processing capabilities.

Researchers have created forecasting models using sentiment analysis and implementing natural language processing (NLP) to improve forecasting of oil price trends. News releases, which can include new policy changes, technology strategies, and signaling from technology in government, military, and the technology industry – such as Silicon Valley, are discovered while conducting due diligence. Since big data technologies are increasingly mature, sentiment analysis can be used as a variable to increase the accuracy of forecasts. Below is the methodology using sentiment analysis and natural language processing as a dictionary-based approach to find positive and negative words combined with a Granger causality test to discover relationships and correlations with oil price as time series data. This is a method that can be used to better inform the forecasting model to create a more accurate oil price trend, or forecast output (Li, Xu, Yu, & Tang, 2016).

$$S_t = \frac{w_p(A_t) - w_n(A_t)}{w_p(A_t) + w_n(A_t)}$$

*Equation 4. Sentiment Analysis Model*

*(Li, Xu, Yu, & Tang, 2016)*

In the sentiment analysis model, $At$ is the available news at time, $t$, and the other variables refer to the positive and negative words to discover $St$, the sentiment. This model itself is inherently flawed due to the subjective matter of defining what a 'positive' word is and what a 'negative' word is. In addition, some other obvious flaws can be: if there is any translation of foreign news, or not including foreign news, the magnitude of the amount of news articles from specific news, magazines, or periodicals compared to industry specific newswires. At the minimum, validation needs to take place on the news data set to ensure negative news is not only due to financial loss of a certain actor, country, location, or news firm itself, and/or a financial gain as it may cause massive error in resulting positive or negative data points. It is up to an experienced data scientist to ensure that the "news" data set being used is properly vetted and validated. Deep learning could improve this model, however, once additional languages and dialects are added, it could increase in complexity exponentially, and will likely add additional risk.

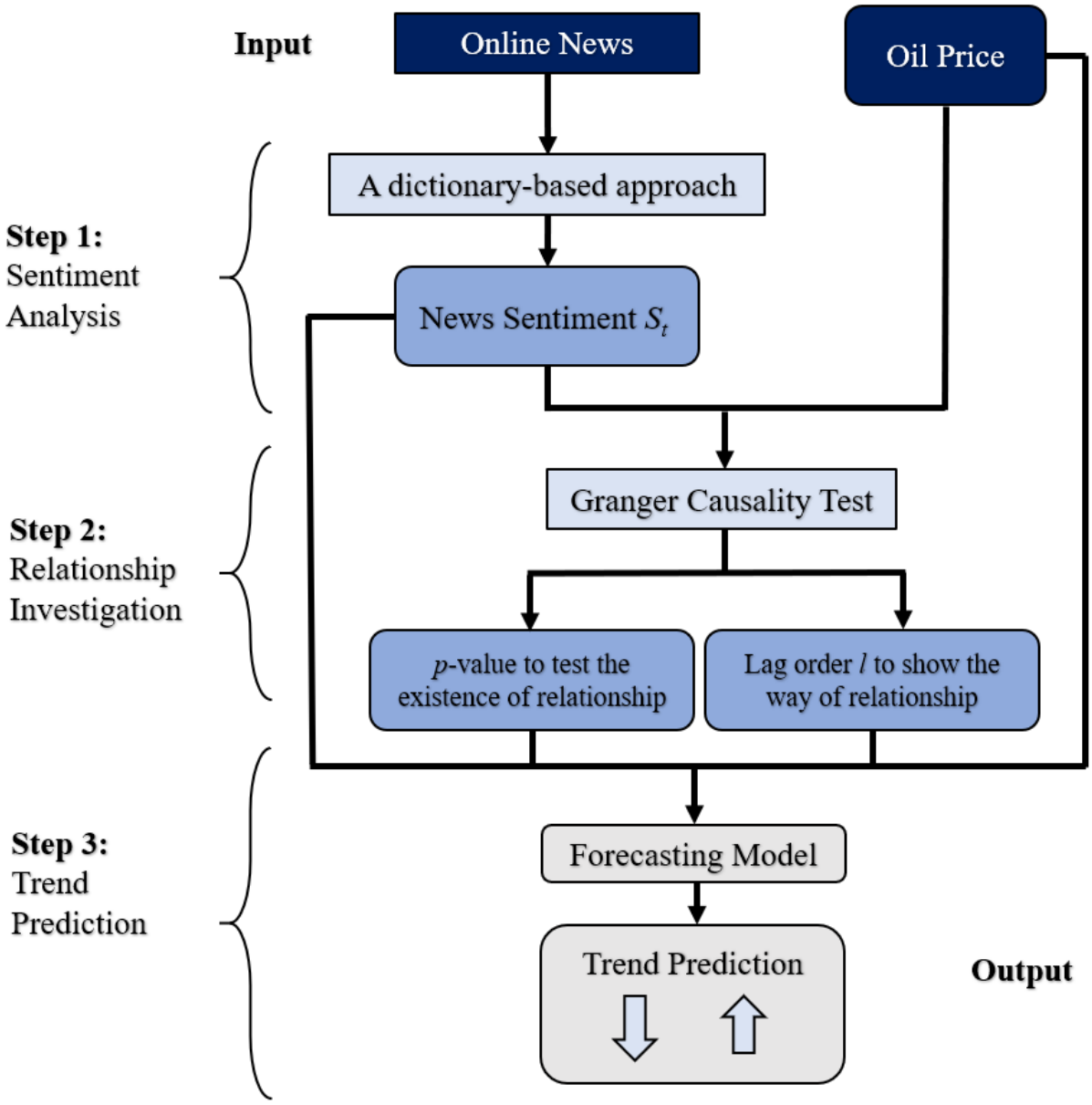



*Figure 28. Sentiment Analysis Framework – Simplified*

(Li, Xu, Yu, & Tang, 2016)

The new unique approach of this method shown in Figure 28 is the use of the Granger causality, which is shown in Equation 5. This model uses two time series data sets ($x$ and $y$) with a lag length vector ($n$) combined with trend prediction. This statistically confirms that "rich online data" can help forecasting price trends in the oil industry (Li, Xu, Yu, & Tang, 2016). However, keep in mind that this "rich online data" can also be corrupted or manipulated to trigger negative or inaccurate impacts if this was used as a common or flagship formula for decision-makers in large governments or firms such as oil companies. This reinforces the necessity of validation at the data level in this framework. Figure 28 shows the "Sentiment Analysis Framework" used to conduct such a forecast. The first step seems to be the highest risk and requires the most validation due to the qualitative news data set being used and converted into a quantitative value to be combined with the quantitative time series oil price data.


*Figure 29. Granger Causality Test (from Figure 28)*

$$\Pr(x_t \mid I_{t-1}) = \Pr(x_t \mid I_{t-1} - Y_{t-n}^{n})), (t = 1, 2, \dots, T)$$

*Equation 5. Granger Causality of Two Stationary Time Series*

(Li, Xu, Yu, & Tang, 2016)

Another approach uses vector trend forecasting method that resulted in less than 4% percent error of the fitted oil price (Zhao, Wang, Guoa, & Zeng, 2018). This approach is derived from current literature on forecasting including methods such as moving average trend analysis, Autoregressive approaches such as ARIMA, and Support Vector Regression (SVR) to discover trends as they are considered common methods for this application (Zhao, Wang, Guoa, & Zeng, 2018). Their model is successful in assisting with ultra-short term to long-term decisions according to their conclusions and uses a simple fitting function. They also discovered, "… that the monotonic function (or exponential and linear functions) were better than periodic functions with regards trend-forecasting" and they recognize the policy impacts that the petroleum industry has on economics and daily life. For a general or basic forecaster such as a decision-maker, their method could be extremely complex to validate and may require a strong background in other exponential and linear functions to know when to apply it.

Much of the literature in this area focuses on the oil industry within the energy sector due to the volatile price and its impact on a nation's national security and economic stability. An example of an approach to improve forecasting of crude oil is by using a hybrid model that uses integrated data fluctuation network and artificial intelligence algorithms which is called the DFN-AI model (Wang, et al., 2018). It uses a complex network time series analysis technique to preprocess the data set. The need for improving forecasting on crude oil predictions is because it is very nonlinear with high degrees of incertitude. This approach differs from others since it attempts to add artificial intelligence methods. To begin, the researchers map the time series into a data fluctuation network (DFN) and use a back propagation neural network (BPNN) which is commonly used for classification and prediction with neural networks because it is a multi-regression analysis. The backpropagation formula used is displayed in Figure 7.

$$W(n) = W(n-1) - \Delta W(n),$$

$$where$$

$$\Delta W(n) = \eta \frac{\partial E}{\partial W}(n-1) + \gamma \Delta W(n-1)$$

*Equation 6. Back Propagation Formula*

*(Wang, et al., 2018)*

However, the researchers note there is a flaw in this model since it uses a gradient method, "the learning convergent velocity is slow and a convergence to the local minimum always occurs. In addition, the selection of the learning and inertial factors affects the convergence …" Therefore, this modeling attempt brings new flaws and traps inherently within the formulas included (Wang, et al., 2018).

## III. Methodology: Methanol Price Data and Validating the Model

In this section data from the Methanex Corporation will be used to demonstrate traps, flaws, and biases in basic forecasting models. The forecasts in this section could be used by decision-makers and R&D managers to better understand the future of the methanol or energy industry. To begin, basic forecasting models will be used with slight variable changes to notice discrepancies as well as discover a forecast output that fits the data well. The results will be analyzed and compared in a similar way to the results in Chapter 4. The data includes monthly pricing data from May 2001 until May 2017 (Methanex Corporation, 2018). For the simulation, the 'forecast' package will be used in 'R' and the data set will be stored in a time series object using the 'ts' function. The data set fits well in to a time series object since there is no missing data, and the data is monthly from 2001 – 2018, which equates to 18 years of data with 12 data points each (one per month), totaling 216 data points. Since this data is sampled at equally spaced points in time, a time series object is the proper way to store it to forecast seasonality (R-Project.org, 2018). To validate using k-means, current methods in the literature will be used. In addition, exploratory data analysis (EDA) methods will be used to try to confirm or assist with validation.

# Validation Framework for Simulation #2: Methanol Data

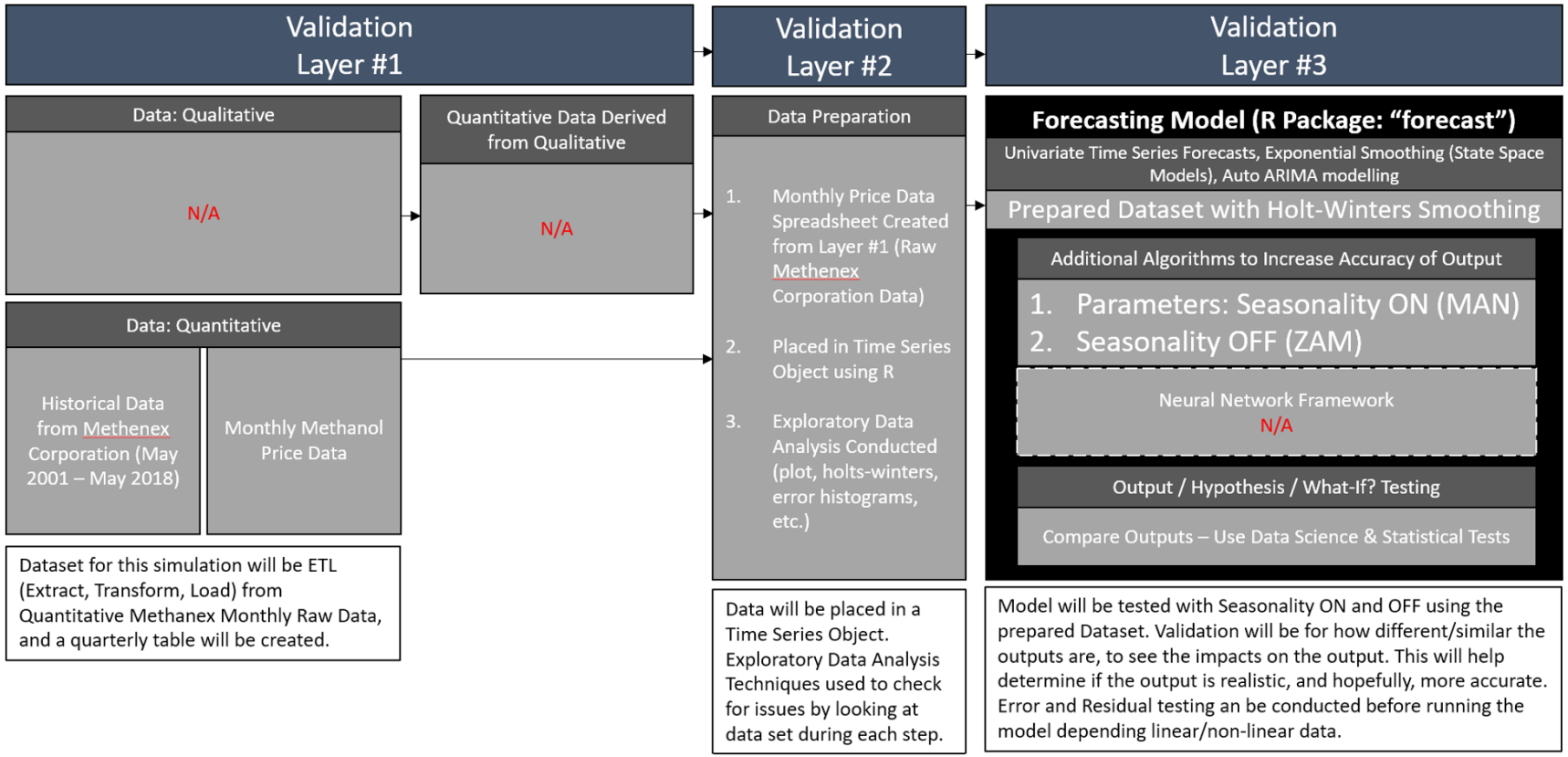



*Figure 30. Proposed Forecast Validation Framework Applied to Methanol Data*

## IV. Analysis of Simulation: Forecasting Methanol Price Data using R

The results of this simulation in the next section (Section V) are discussed in this section. The simulations were conducted using the same forecast function in the 'forecast' package in R for each output graph. The different graphs were created to discover differences in outputs on the methanol data set and to identify common visual traps. Figure 31 is a graph of the methanol data set for visual exploratory data analysis. The forecast result in Figure 32 is the output from the default forecast on the methanol data set which shows a positive trend for the next 12 months (or one year), which could be interpreted by a decision-maker as a positive trend, and therefore influence their decision-making process and Figure 33 shows a Holt-Winters filtering output graph. Figure 32, the default forecast output for 1 year, could be considered a common output that anyone using R with the same data set would come to as a 'correct' conclusion. In some cases, this graph could be used in a business presentation to influence or convince an audience of the intended goals based on this positive trend line. However, in the later simulations, such as Figure 34, the same forecast is conducted with different optional parameters and gives a steeper and more positive forecast output. Deciding which is correct and other visual traps are investigated in this section. The goal of these simulations is to show the process of how to validate a forecasting model for use in the real-world and recommend a few output graphs to create forecasts for following the Proposed Forecast Validation Framework: Validation Layer #3 shown in Figure 12. Moreover, Figure 35 shows a different set of parameters with what visually may be a more accurate forecast. Figure 35 has the optional parameter for multiplicative chosen, which makes the output forecast portion of the graph less linear and more in line with the variation of the time series and seasonal pattern.

The decomposition function can be used at the minimum to validate that the data set shows significant seasonality. Figure 39 shows a decomposition function's output on the data set and from a qualitative approach such as exploratory data analysis a data scientist can validate multiple aspects of the data set. The trend and seasonality can be verified, and the data scientist can discover if the data has a legitimate seasonal pattern, at least from a qualitative standpoint. The output of a decomposition breaks the data set (which is stored in a time series object) into components for further analysis on the y-axis.

By plotting the residuals on the graphs shown in Figure 36 (Holt-Winters), Figure 37 ('MAN' parameters), and Figure 38 ('ZAM' parameters), a data scientist can identify the variance and fluctuations. In this case, the variance is reasonable (or, what a data scientist with experience would expect), and it is easier to discover where the highest fluctuations are, Year 2006 to Year 2009.

In Figure 43, Figure 44, and Figure 45, a generally accepted error function is used to discover the forecast errors. If this graph is interpreted by the data scientist as too noisy, they may want to reassess the model they are using on the data set. The histogram results are reasonable and expected showing no severe issues or statistically significant anomalies with the data.

The graph in Figure 46 shows the forecasted data (in blue) for the time between 2013 and 2014 against the actual data from 2013 (continuous black) to 2014. At first glance, this may seem "off" however, using other tools such as linear regression at this stage, may be appropriate for justifying or discovering which output model is best. However, when we extend the forecast out using the real data from 2013-2015 (24 months out, rather than 12 used in the first graph), the variance due to 'seasonality' and the forecast begins to become more inaccurate in Figure 47.

Figure 47 was a test to see if it is worth investigating a larger forecast period. Since it seemed worth analyzing, Figure 48 was produced as an output. Figure 48 and Figure 49 are the same graph, except Figure 48 is visually stretched (or Figure 49 is visually shrunk) for exploratory data analysis reasons. These graphs are real data vs. 5 year forecasted data. There is an apparent seasonality that can be observed, and overall, the forecast line (in blue) seems realistic to the observer – and relatively acceptable. This shows how common stretching and output traps can visually distract or cause issues for a data scientist.

Working off the outputs from the simulation, Figure 50 shows a comparison of methanol pricing forecasts. The first (or top) forecast in Figure 50, "Methanol Price Forecast 2013-2014" shows the forecast output in blue for next 4 quarters (points) starting from 2013 to 2014 (1 year). The actual methanol price data is represented in black for comparison. At first glance using qualitative visual methods or exploratory data analysis on this output, a data scientist or decision-maker may notice this forecast may seem inaccurate, visually, as it does not trend similar to the real data points and begins with a negative trend for the first few quarters. This may seem unappealing to a decision-maker compared to a forecast that trended upwards such as the ones in Figure 31, Figure 33, or Figure 34. This initial visual assessment may cause the data scientist to dismiss applying this forecasting method due to the magnitude or variable of time. This is due to not validating properly and editing parameters such as how many quarters (data points) to forecast forward. A social or qualitative trap of "one year" being 4 quarters requires the understanding and realization of the data scientist when assessing the output and that additional testing or changing of simple parameters (additional forecasting points) may give a more accurate output. In addition, if the real data was not known, this forecast may look severely inaccurate compared to simple forecasts that trend upwards.

## V. Simulation: Forecasting Methanol Price Data using R

```
> methanolPRICE <- ts(v[,"PRICE"], start=c(2001,5),frequency=12)
> methanolPRICE
      Jan  Feb  Mar  Apr  May  Jun  Jul  Aug  Sep  Oct  Nov  Dec
2001                     0.77 0.67 0.57 0.51 0.42 0.37 0.40 0.40
2002 0.38 0.36 0.38 0.42 0.50 0.56 0.62 0.62 0.62 0.62 0.62 0.62
2003 0.69 0.79 0.82 0.82 0.82 0.82 0.78 0.72 0.70 0.68 0.68 0.68
2004 0.75 0.75 0.75 0.75 0.75 0.81 0.84 0.84 0.84 0.84 0.90 0.95
2005 0.95 0.95 0.95 0.95 0.95 0.95 0.90 0.90 0.90 0.96 0.96 1.02
2006 1.02 1.07 1.07 1.07 1.03 1.03 1.00 1.03 1.33 1.80 1.80 1.80
2007 1.80 1.65 1.55 1.01 1.01 1.01 0.93 0.93 0.96 1.70 2.00 2.50
2008 2.50 2.10 1.90 1.60 1.50 1.58 1.58 1.58 1.58 1.50 1.40 1.00
2009 0.70 0.70 0.65 0.60 0.60 0.60 0.68 0.72 0.84 0.95 1.00 1.10
2010 1.10 1.10 1.10 1.10 1.00 1.05 1.05 1.05 1.08 1.08 1.33 1.38
2011 1.35 1.28 1.28 1.28 1.28 1.28 1.28 1.38 1.38 1.38 1.38 1.38
2012 1.34 1.34 1.34 1.34 1.38 1.38 1.32 1.32 1.32 1.32 1.45 1.45
2013 1.45 1.45 1.55 1.55 1.60 1.60 1.60 1.60 1.60 1.65 1.80 1.90
2014 1.90 1.90 1.90 1.80 1.70 1.60 1.45 1.45 1.45 1.45 1.50 1.43
2015 1.35 1.25 1.25 1.25 1.33 1.33 1.33 1.25 1.10 1.10 1.05 1.05
2016 0.90 0.75 0.75 0.75 0.75 0.80 0.80 0.80 0.83 0.88 0.96 1.10
2017 1.25 1.25 1.50 1.33 1.23 1.16 1.13 1.13 1.16 1.19 1.16 1.24
2018 1.44 1.52 1.49 1.49 1.49
```

*Table 3. Methanol Price Data set May 2001 – May 2018*

Plot **Data set** for visual **Exploratory Data Analysis**

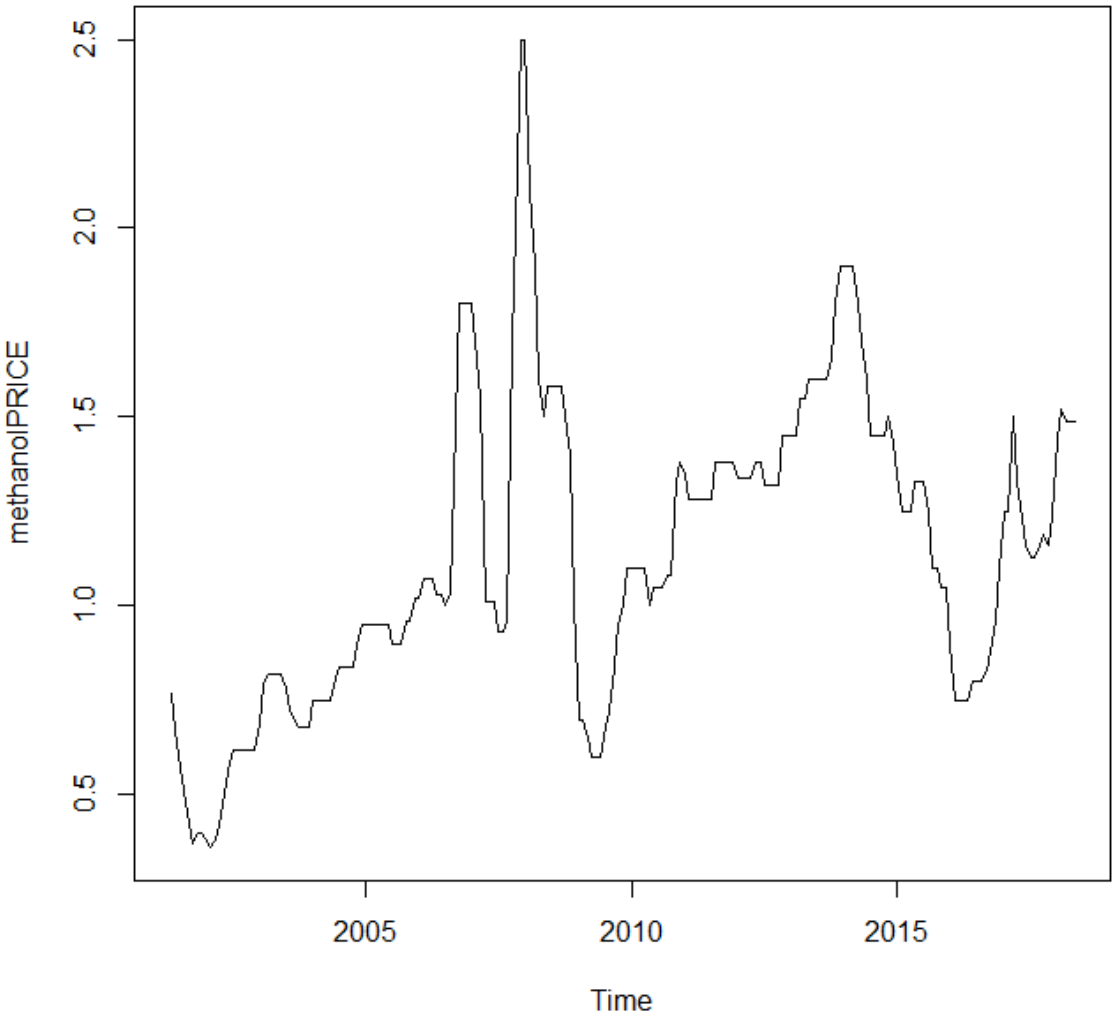



*Figure 31. Methanol Data Plot*

Plot **DEFAULT** Model Default (12-Month Forecast)

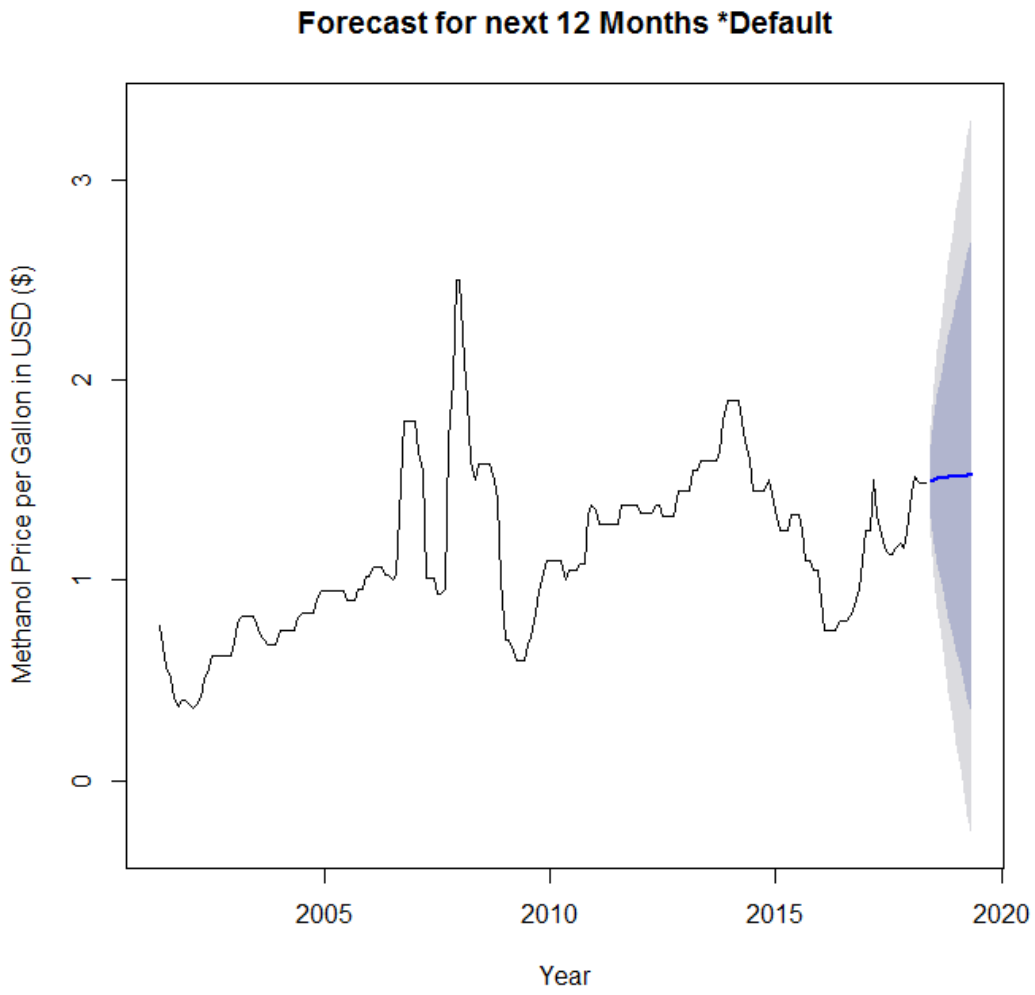



*Figure 32. 12 Month Forecast for Methanol Data (Default)*

Plot **Holt-Winters**

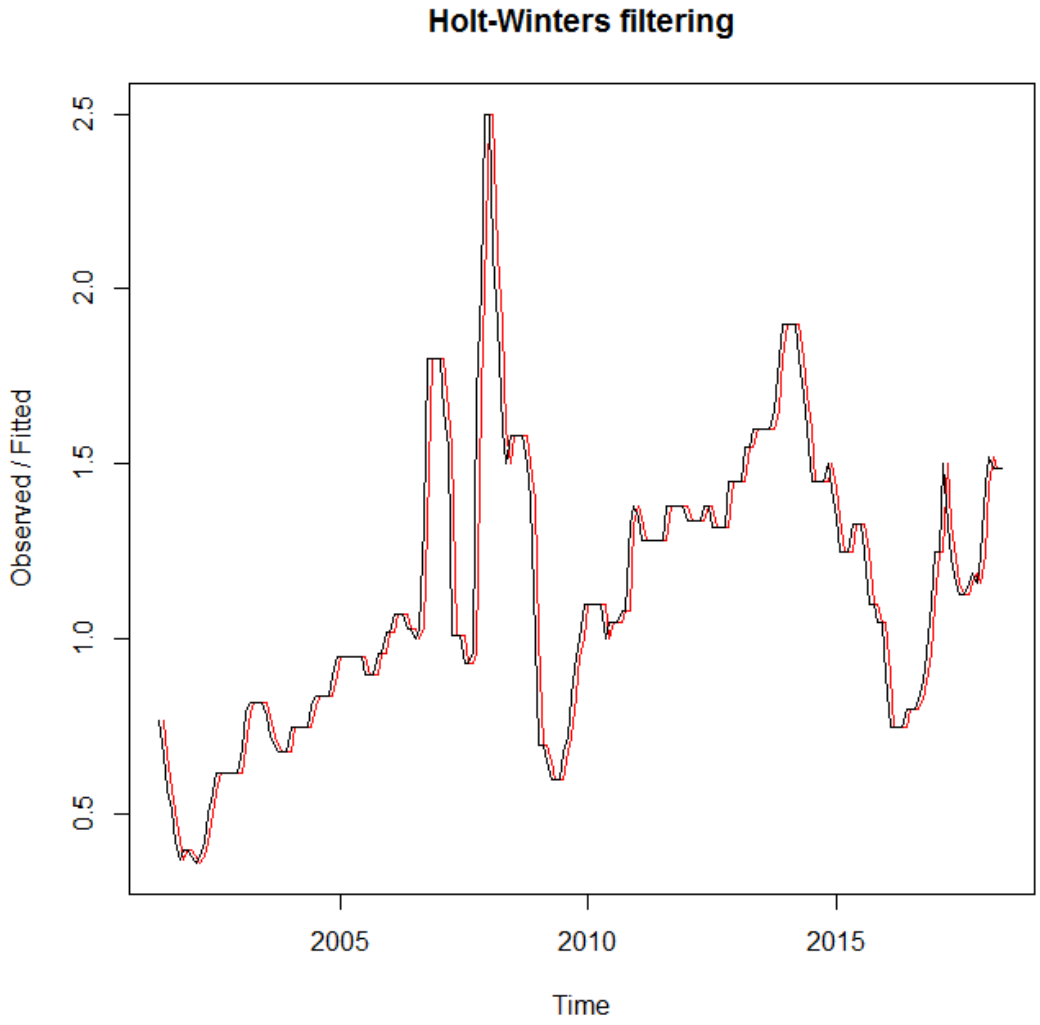



*Figure 33. 12 Month Forecast for Methanol Price (Holt-Winters)*

Plot **MAN** Model Forecast (Error: Multi, Trend: Auto, Season: None)

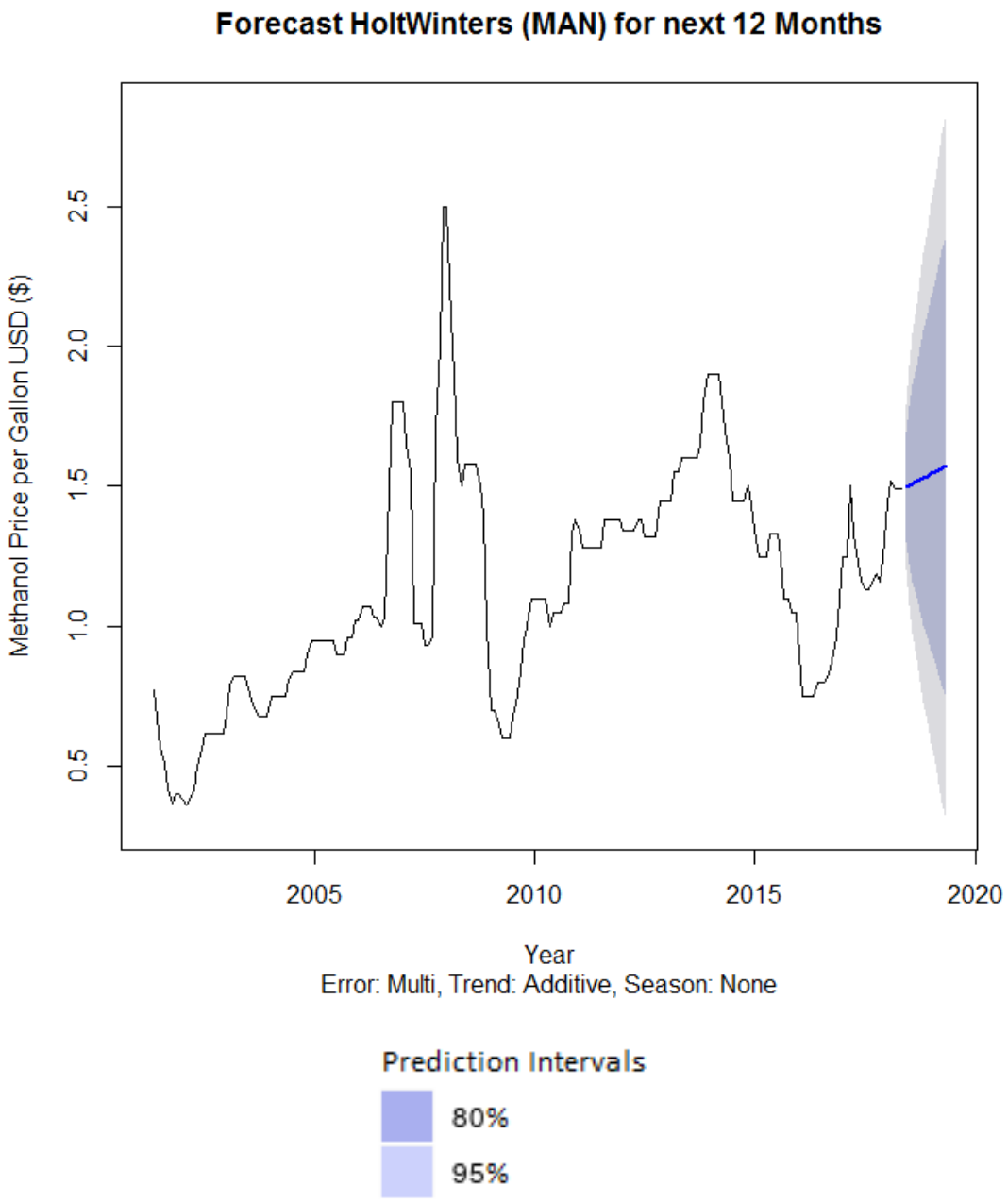



*Figure 34. 12 Month Forecast for Methanol Price (Model=MAN)*

Plot **ZAM** Model Forecast (Error: Auto, Trend: Additive, Season: Multi)

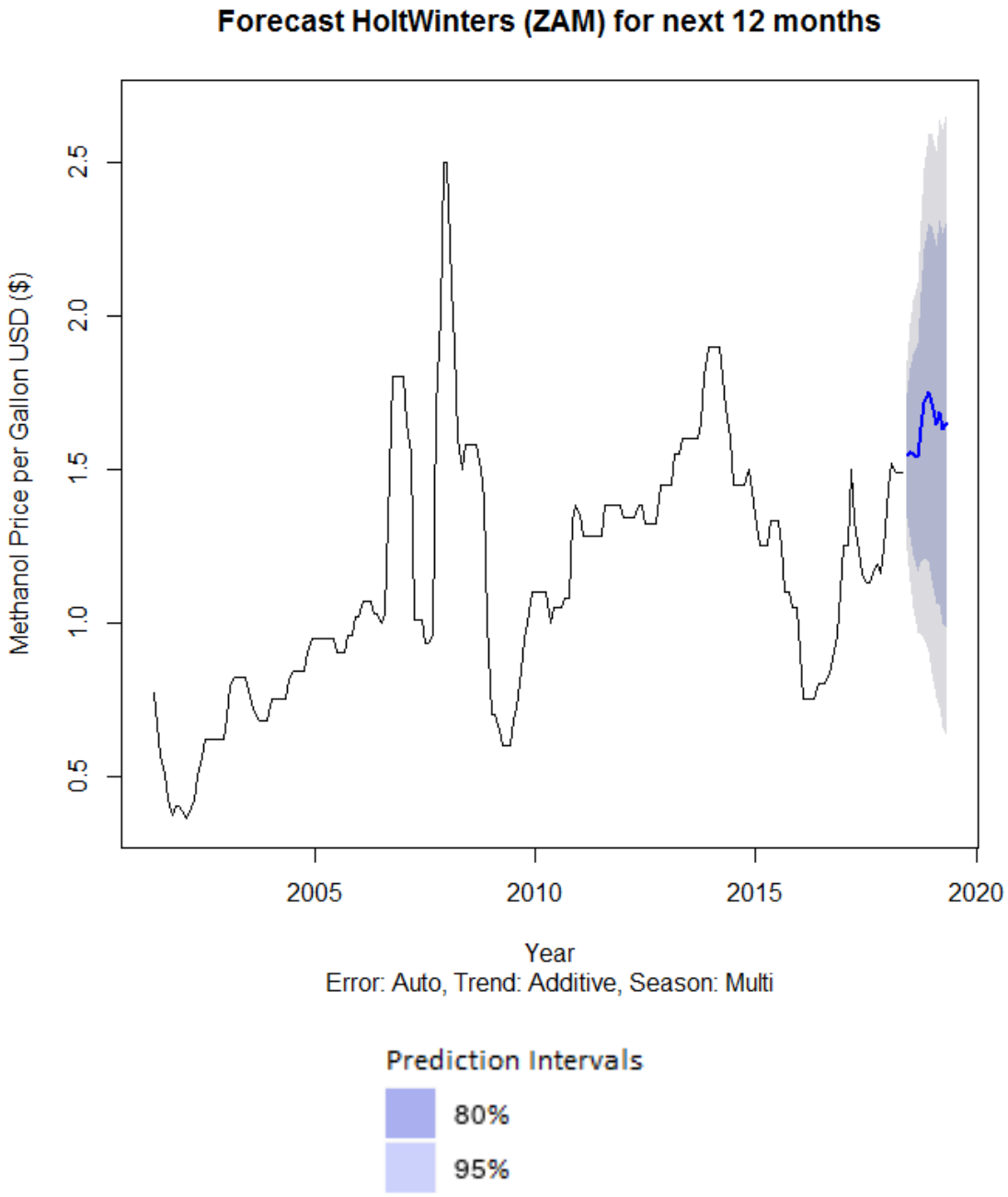



*Figure 35. 12 Month Forecast for Methanol Price (Model=ZAM)*

Plot **Holt-Winters** Model **Residuals**

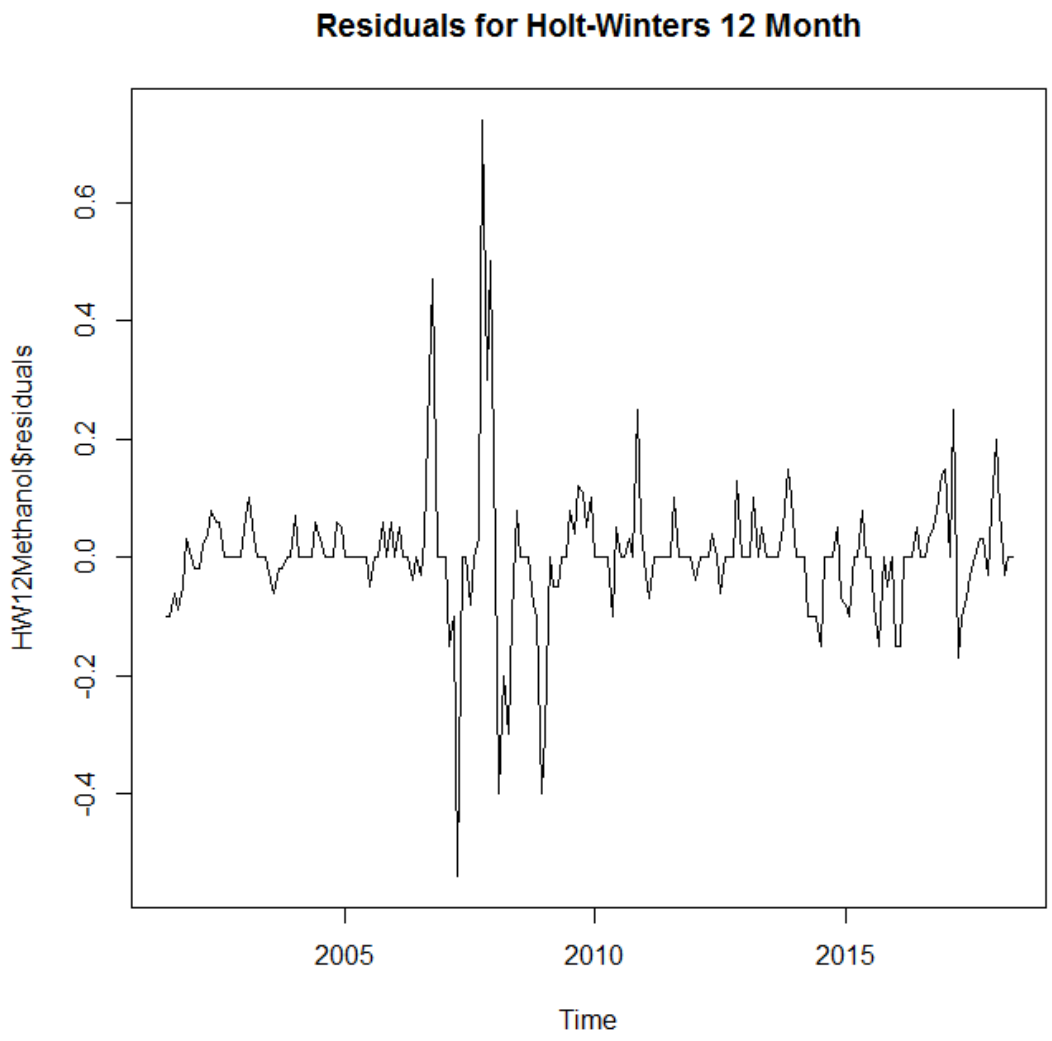



*Figure 36. 12 Month Forecast Residuals for Methanol Price (Holt-Winters)*

Plot **MAN** Model Forecast **Residuals**

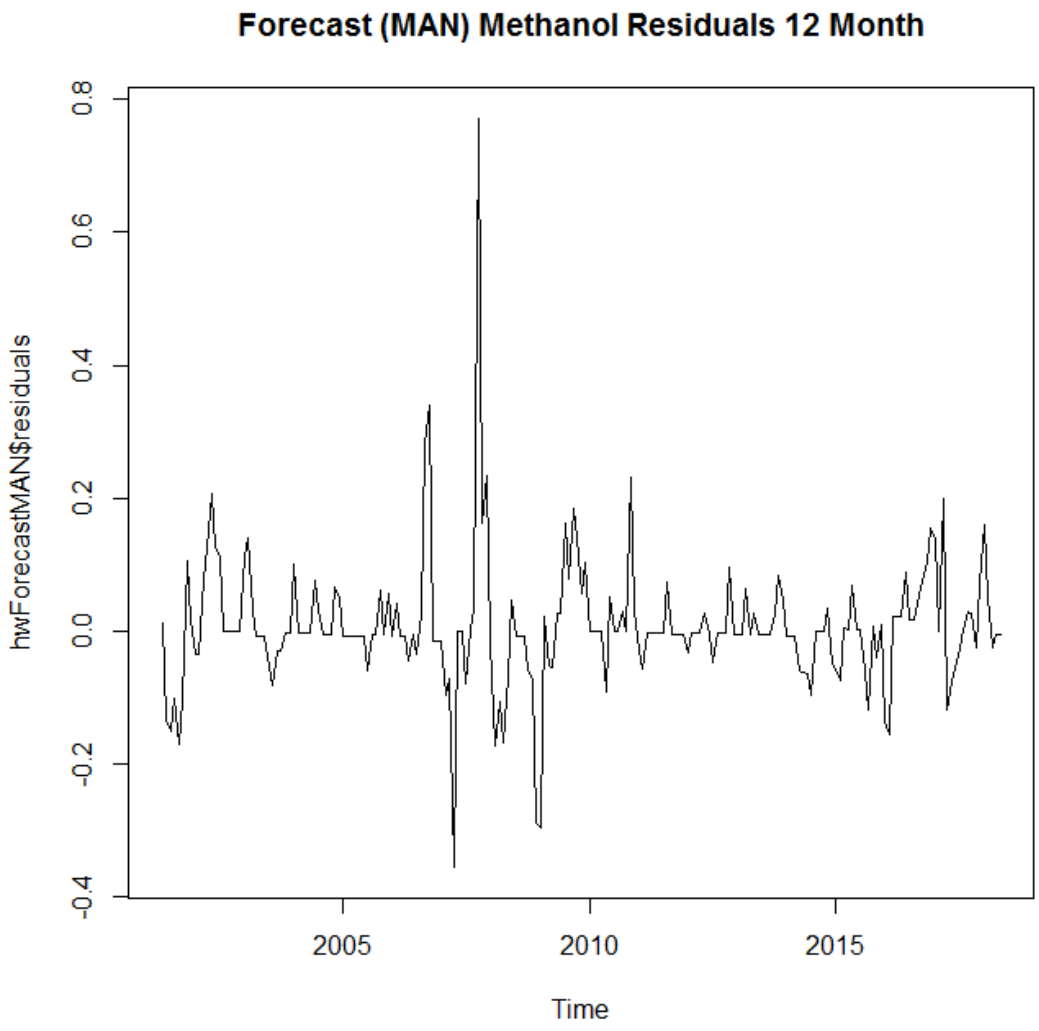



*Figure 37. 12 Month Forecast Residuals for Methanol Price (Model=MAN)*

Plot **ZAM** Model Forecast **Residuals**

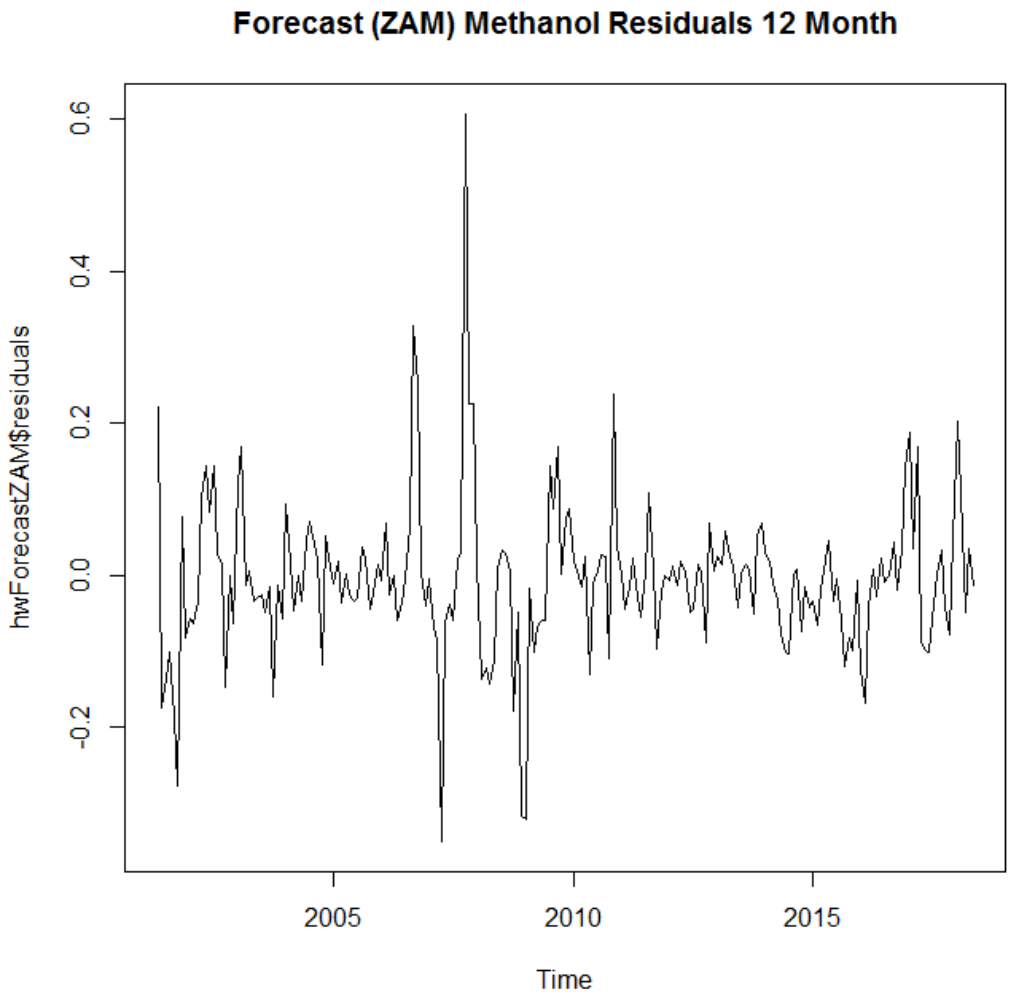



*Figure 38. 12 Month Forecast Residuals for Methanol Price (Model=ZAM)*

**Decompose** Methanol Data set Time Series Object (*confirm seasonality*)

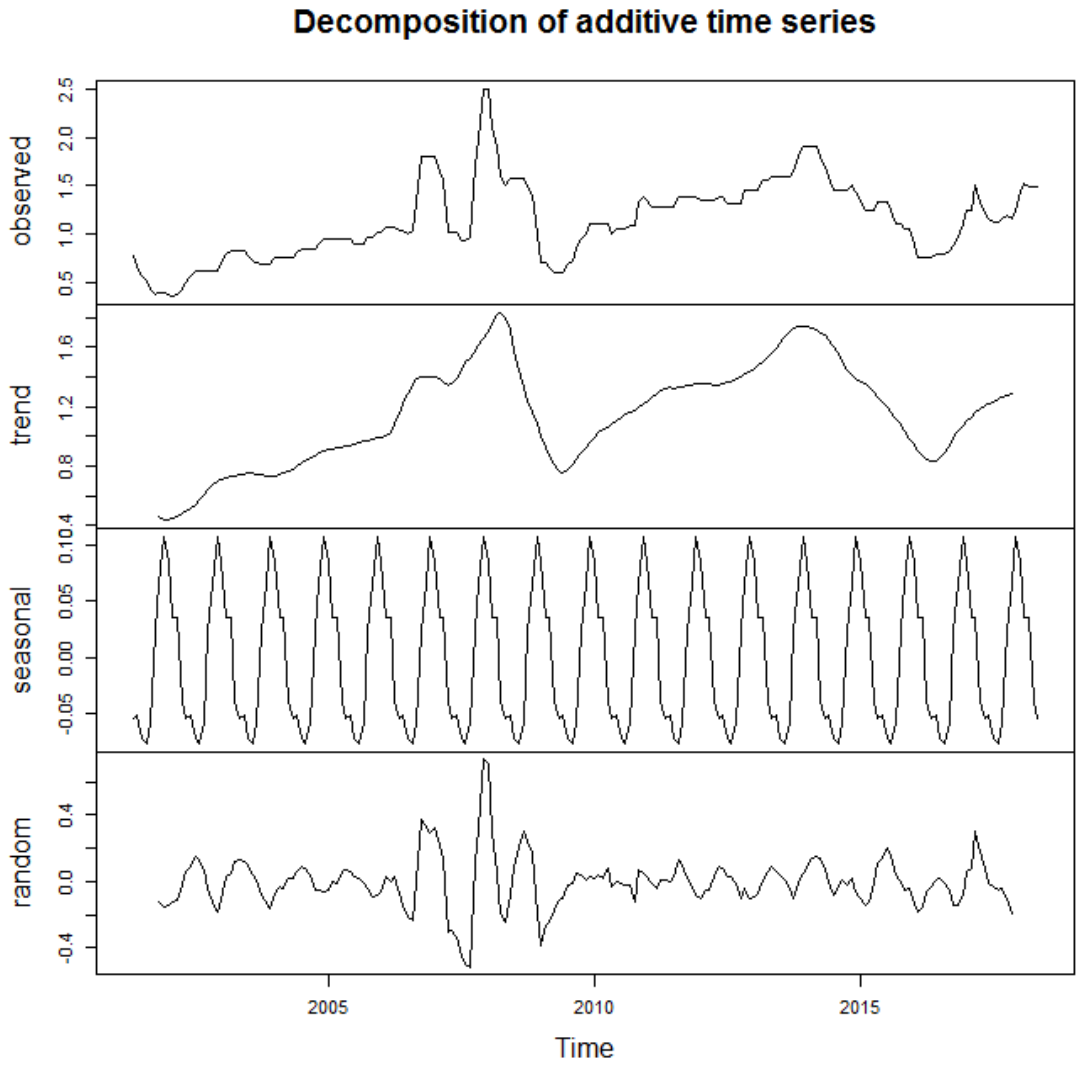



*Figure 39. Decomposition for Methanol Price Time Series*

Plot **ACF** for **Holt-Winters** Forecast Model

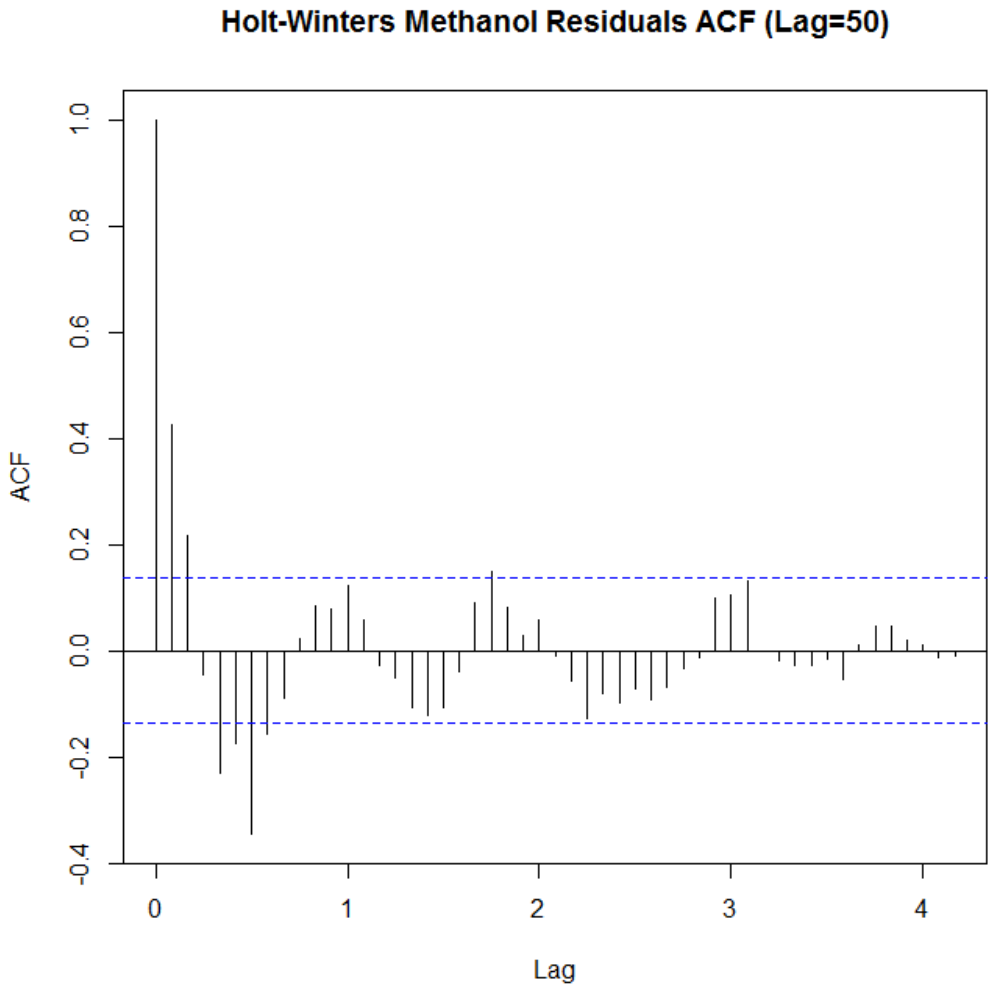



*Figure 40. ACF for Methanol Data set (Holt-Winters)*

Plot **ACF** for **MAN** Forecast Model

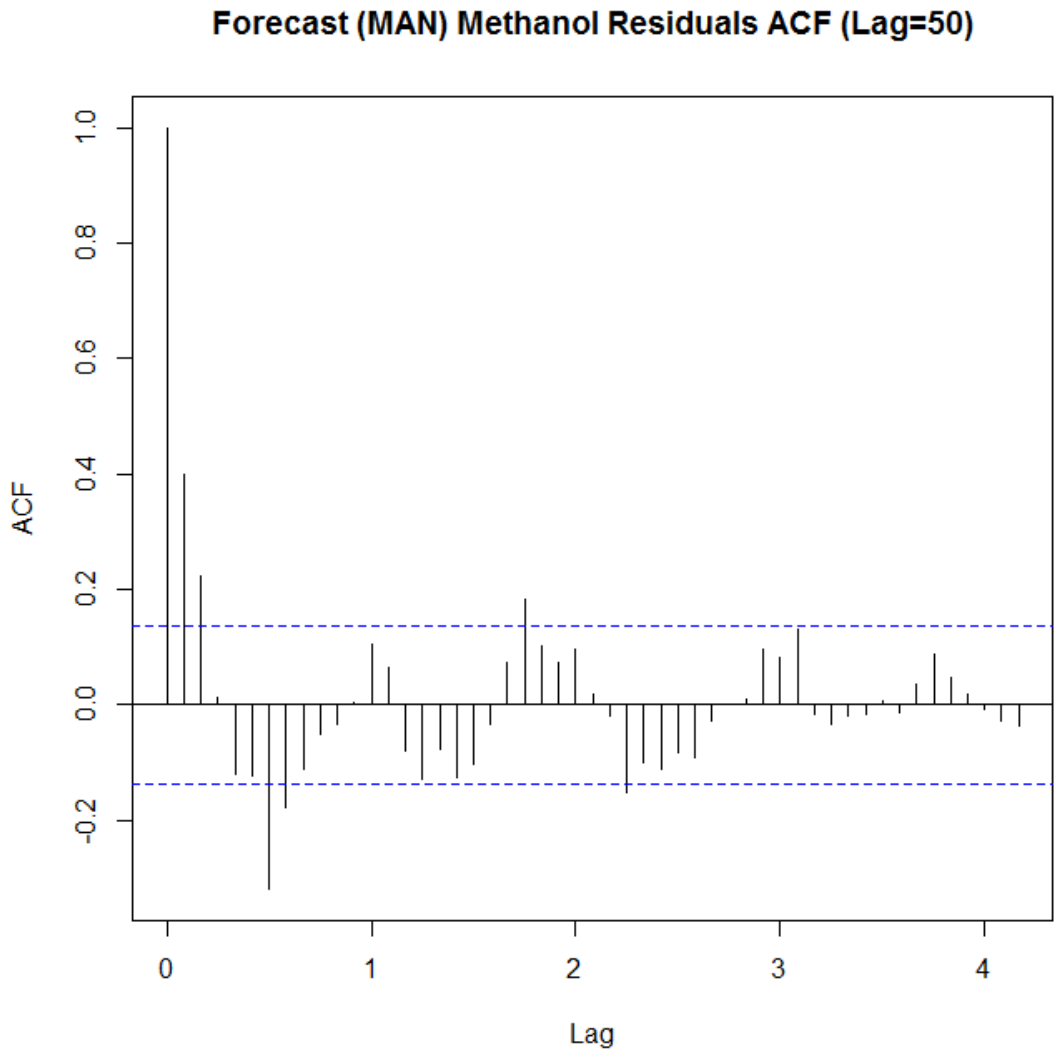



*Figure 41. ACF for Methanol Data set (Model=MAN)*

Plot **ACF** for **ZAM** Model Forecast Model

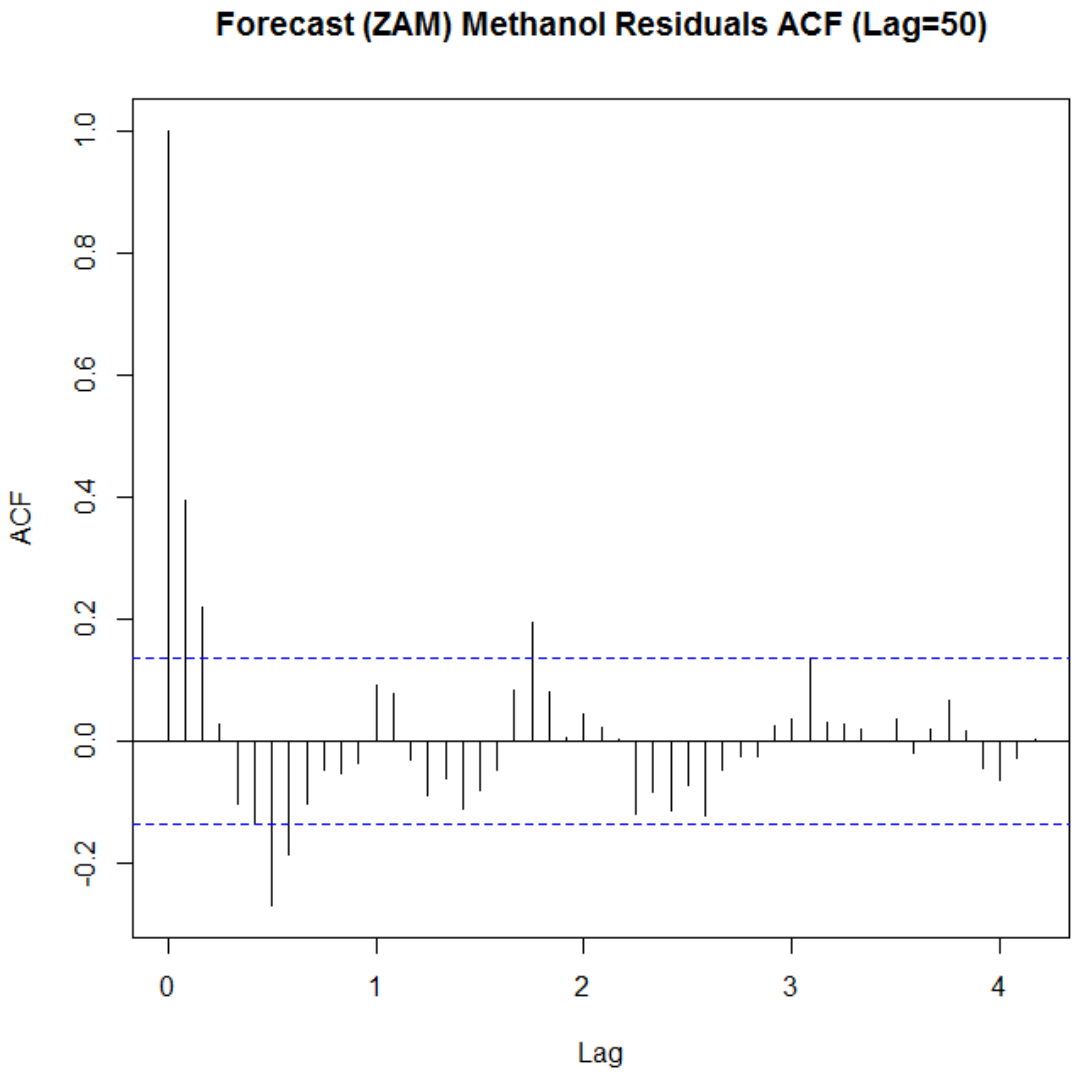



*Figure 42. ACF for Methanol Data set (Model=ZAM)*

Generate **ForecastErrors Histogram** on previous **Holt-Winters** Forecast output data

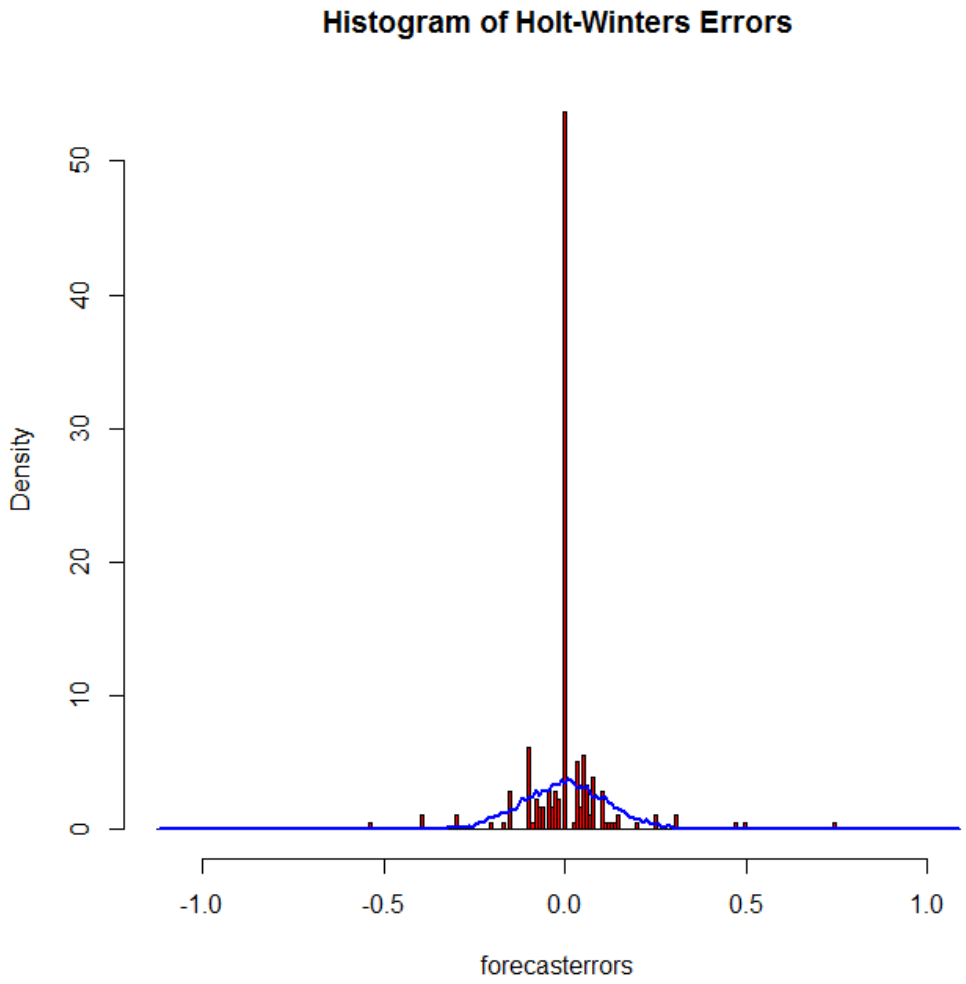



*Figure 43. Histogram of Errors: Methanol Data (Holt-Winters)*

Generate **ForecastErrors Histogram** on previous **MAN** Model Forecast output data

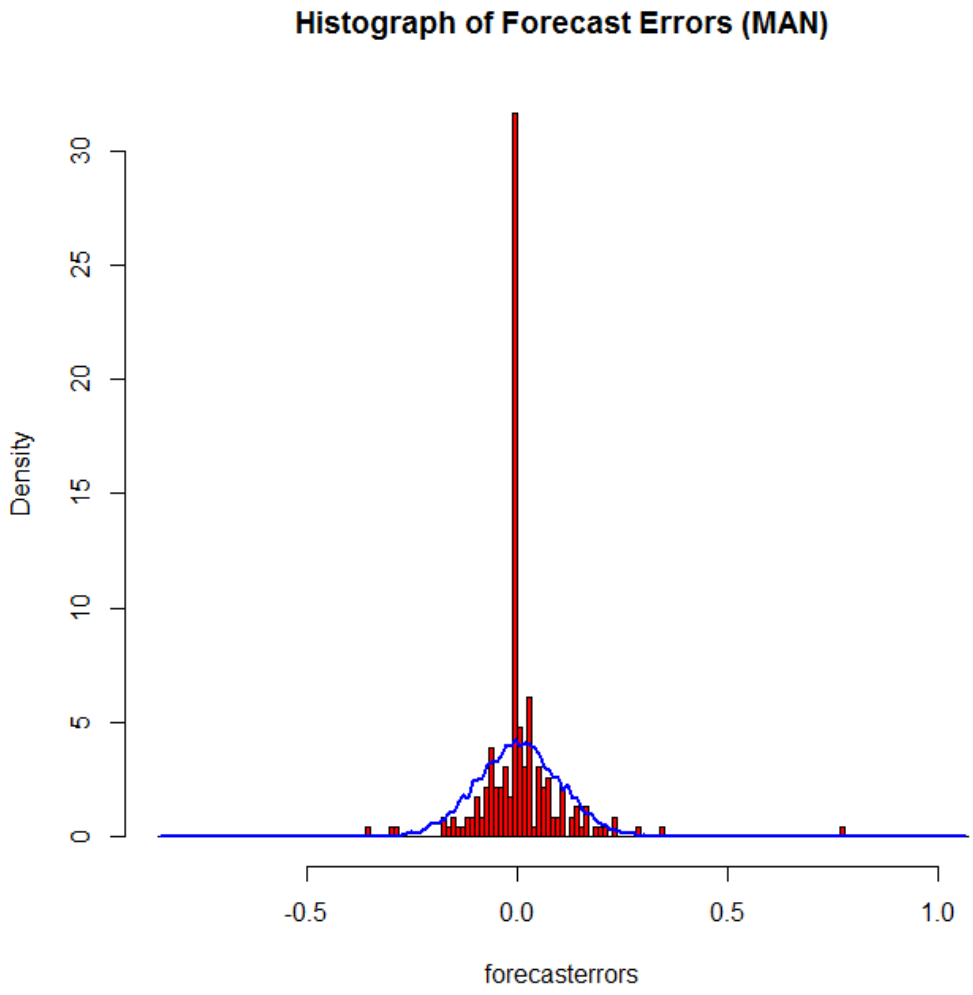



*Figure 44. Histogram of Errors: Methanol Data (Model=MAN)*

Generate **ForecastErrors Histogram** on previous **ZAM** Model Forecast output data

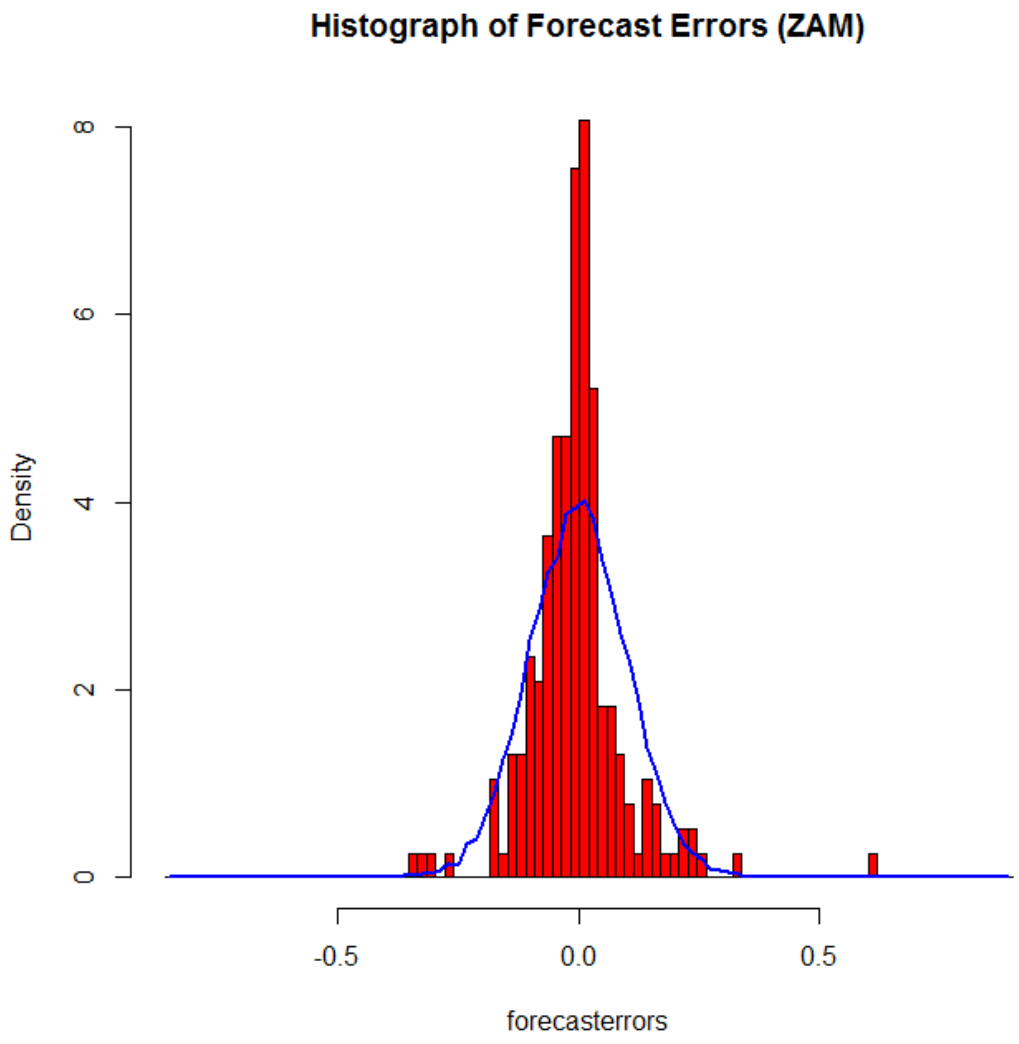



*Figure 45. Histogram of Errors: Methanol Data (Model=ZAM)*

Plot **Forecast for 2013-2014** Comparison

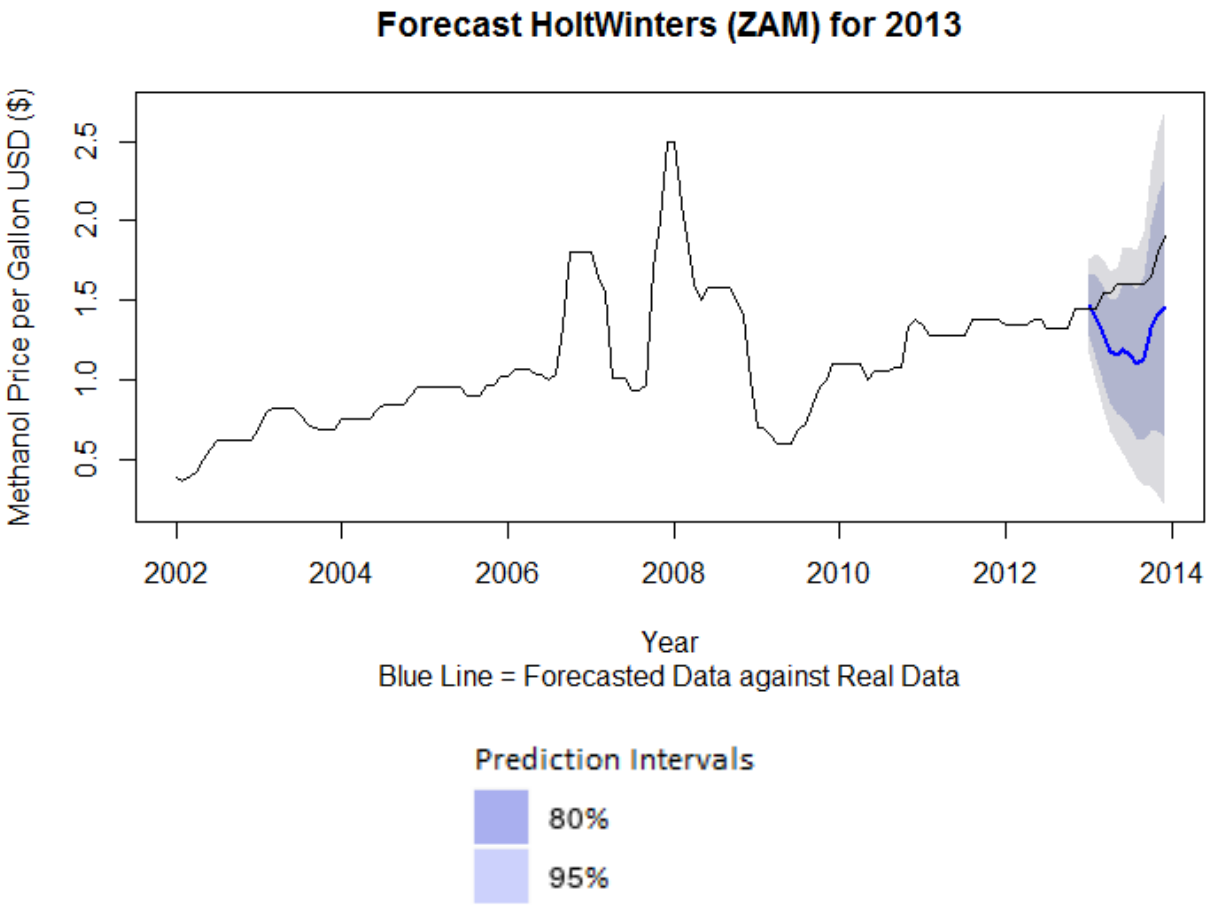



*Figure 46. Methanol Forecast 2013-2014 Comparison*

Plot **Forecast for 2013-2014+** Comparison against Real Data

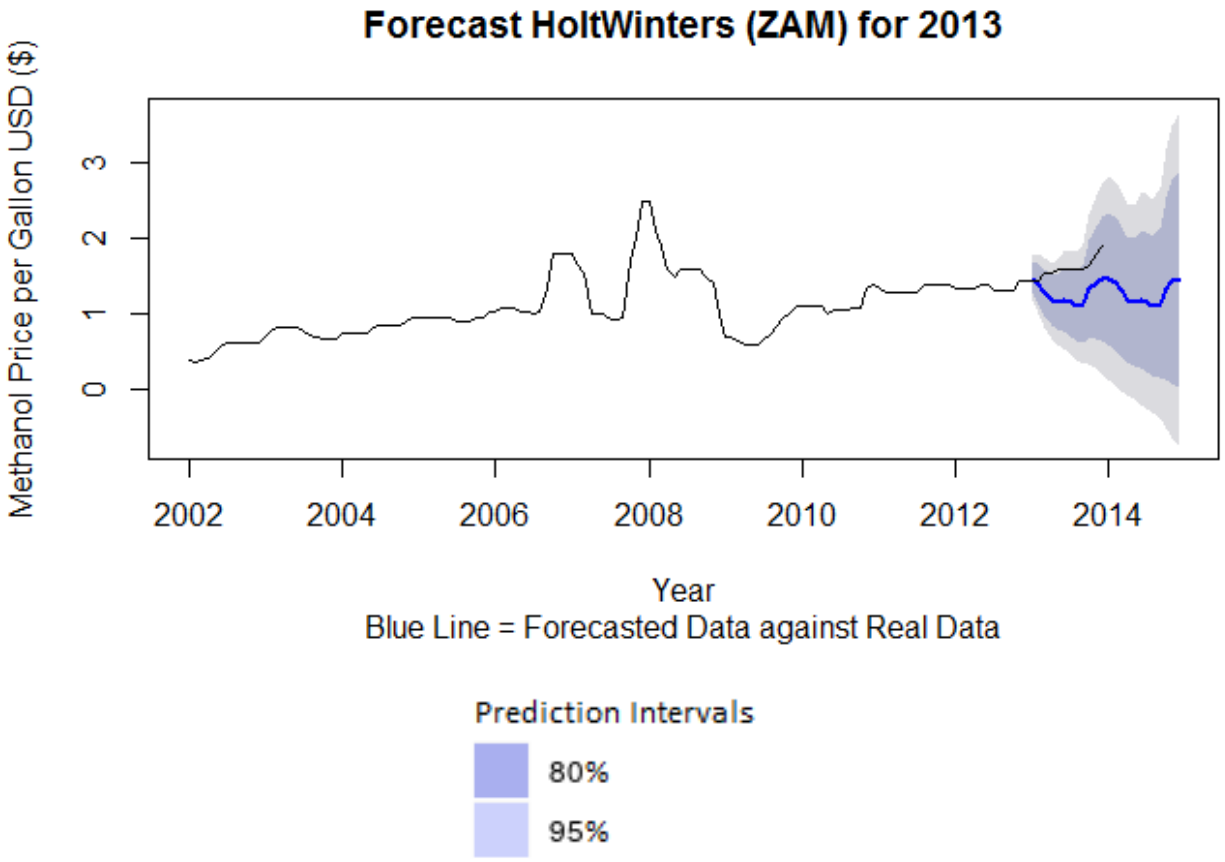



*Figure 47. Figure 40. Methanol Forecast 2013-2015 Comparison*

Plot **Forecast for 2013-2018** Comparison against Real Data

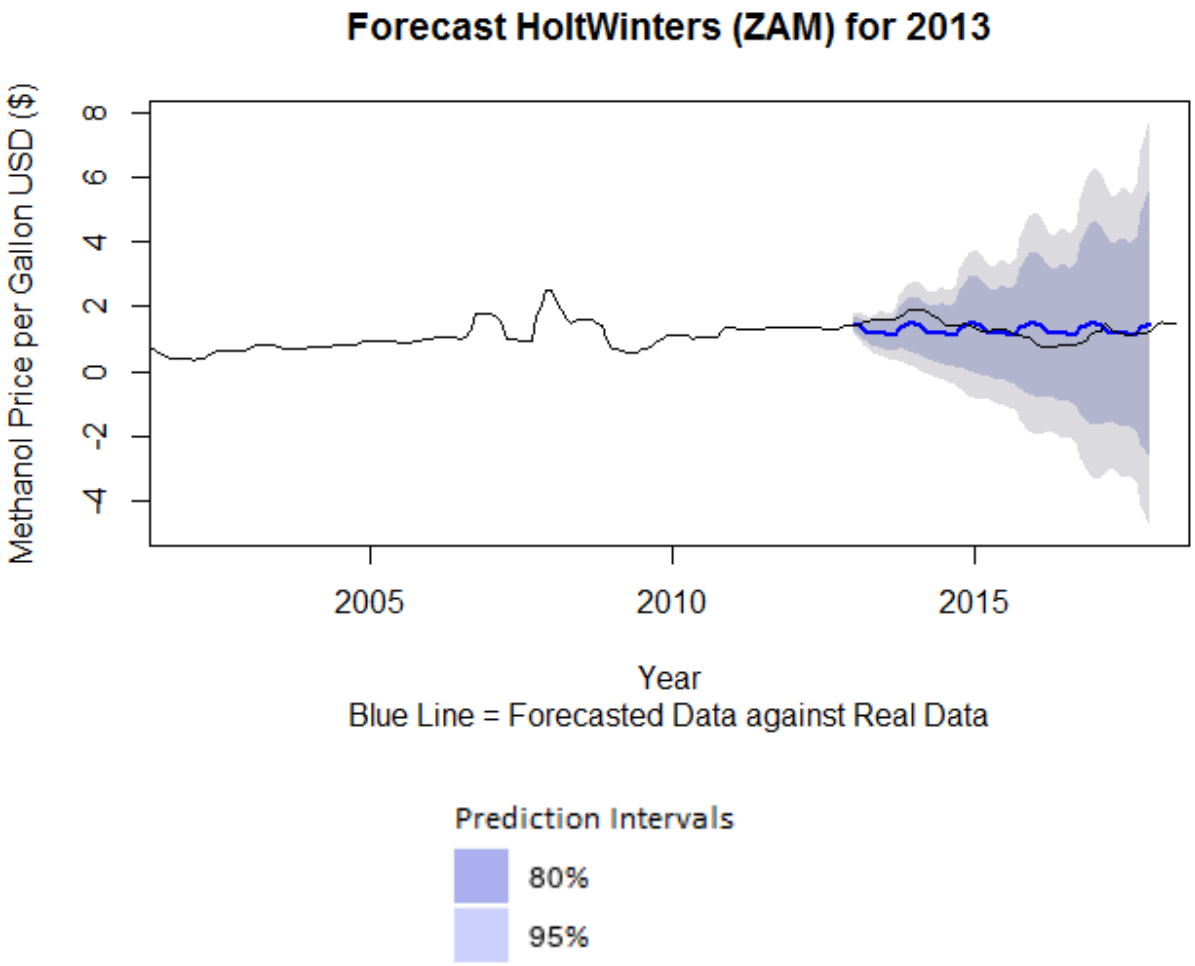



*Figure 48. Five Year Methanol Data Forecast: 2013-2018 for Comparison*

Plot **Forecast for 2013-2018** Comparison against Real Data (Resized)

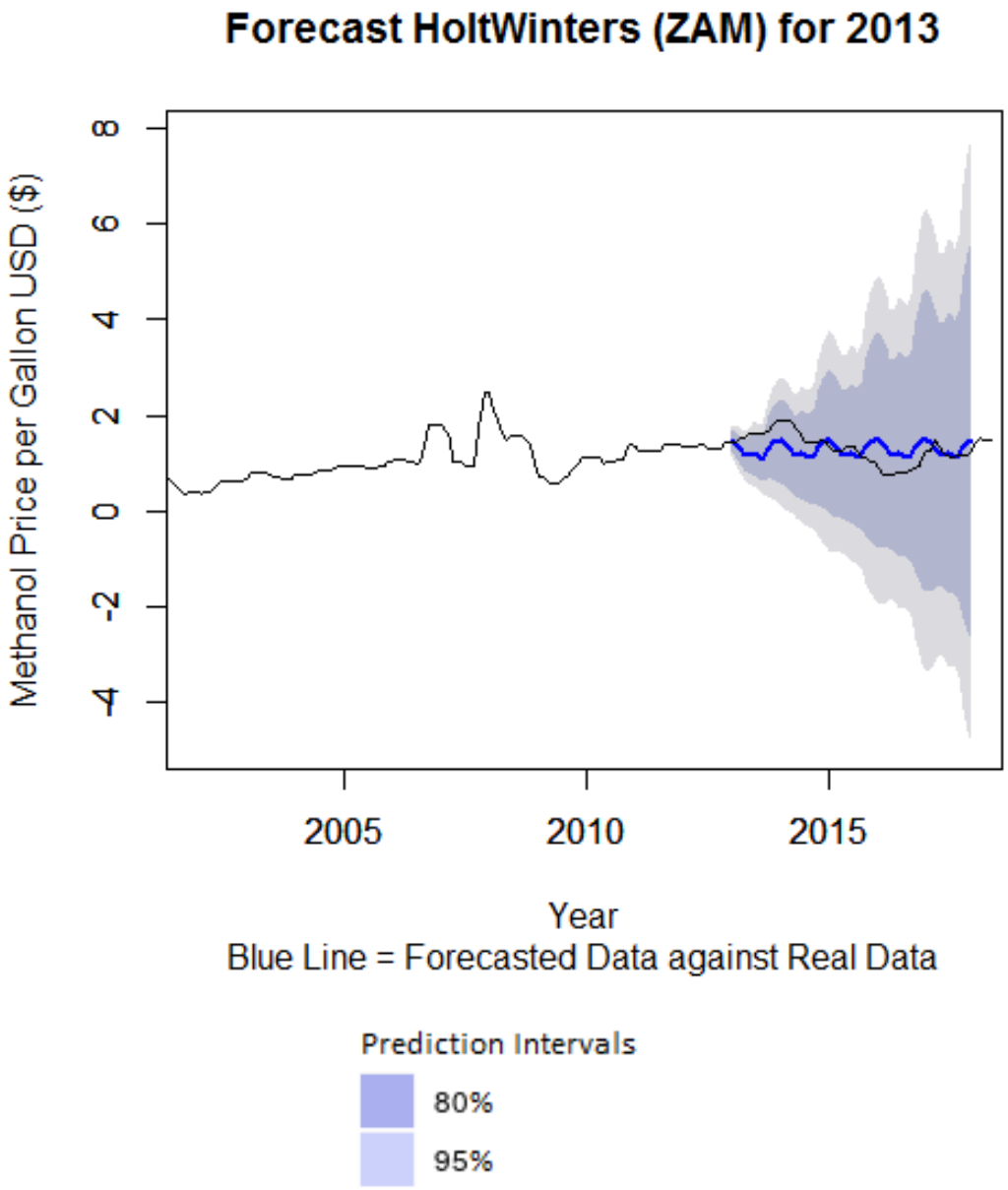



*Figure 49. Five Year Methanol Data Forecast: 2013-2018 for Comparison - Resized*

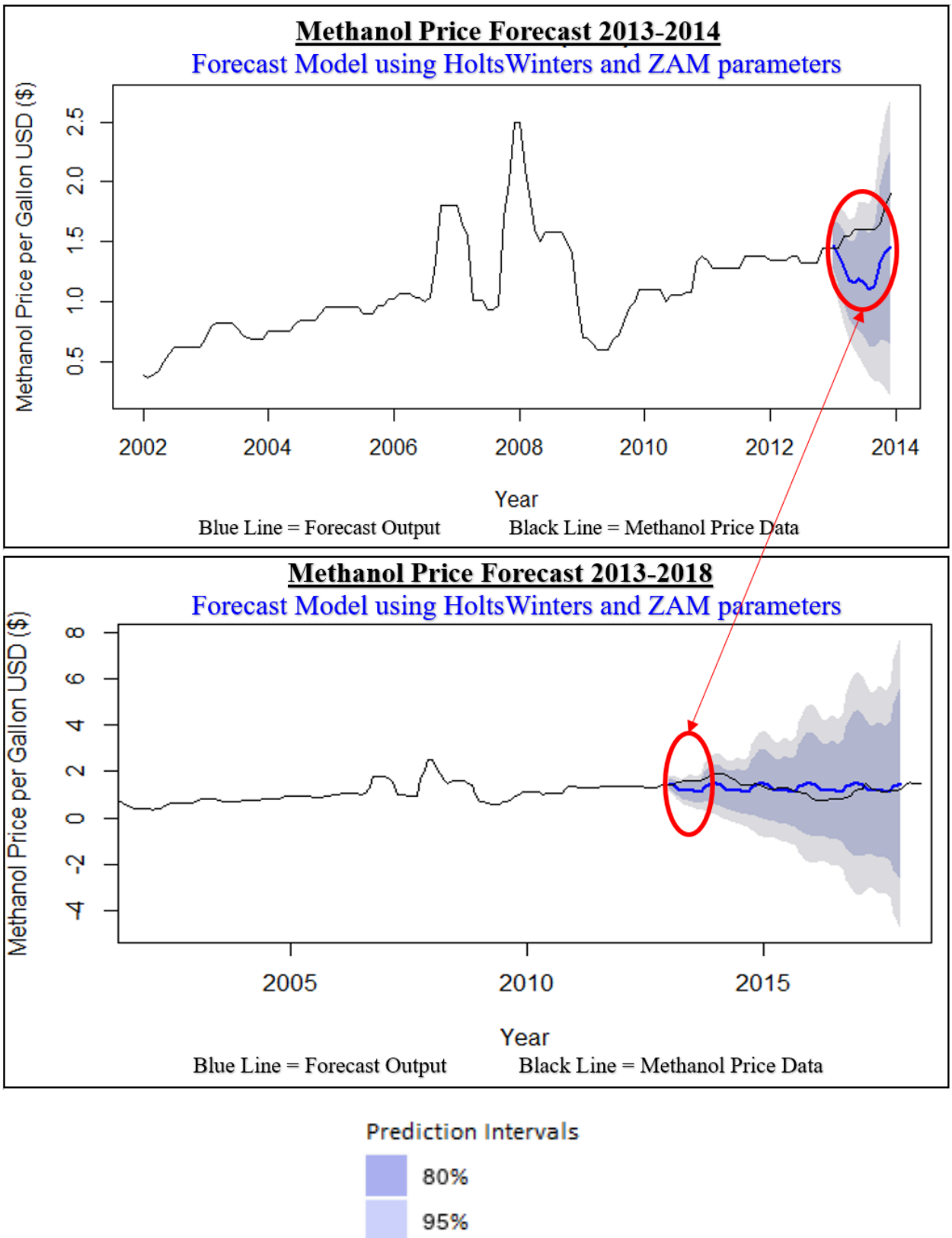



*Figure 50. Why Validation is Necessary: Comparison of Methanol Price Forecasting*

## VI. Simulation Takeaways

After assessing the results, the biggest takeaway that shows evidence of the need for validation is the comparison of graphs displayed in Figure 50. In Figure 50, the second (or bottom) forecast titled, "Methanol Price Forecast 2013-2018" shows the forecast output in blue of next 20 quarters (points) starting from 2013 (5 years) to 2018. Actual methanol price data is represented in black. Qualitative or Exploratory Data Analysis shows that this forecast is a forecast with seasonality that fits the real data realistically. Note the Y-axis units and potential for framing due to the stretched-looking graph which should be carefully assessed. In addition, the price cannot be below 0 in this real-world data set. The data scientist must also account for ethics regarding visual flaws in outputs and avoid manipulation.

The same forecast model used for both. The red oval represents the same data and output to show the potential for visual traps. The only difference is that Figure 50's second (or bottom) forecast length parameter is 20 quarters instead of 4, and the graph physical/visual size remains similar. This example demonstrates a visual trap that can stop a data scientist from using the more accurate forecasting technique and choose a less accurate one due to not validating. Due to visual stretching, not comparing different forecast lengths, and validating the labeling of the Y-axis there are many flaws, traps, and potentials for failure and confusion that a data scientist needs to overcome or mitigate. This simulation demonstrates the need for validation and for awareness of these traps.

# Chapter 6: Neural Network Approaches to Improve Forecasting

***Executive Summary:*** *This chapter will review new approaches to increasing accuracy and precision in forecasting models by implementing new technology such as machine learning, artificial intelligence and neural networks applied to understand state-of-the-art forecasting. New current and innovative approaches are becoming more common to improve accuracy of forecast models by implementing neural network technology into forecasting modeling, predictive analytics, and data science. However, this increases the complexity of validation and by reviewing different types of neural networks this Chapter will also attempt to show these levels of complexity and associated needs for additional computational science validation.*

## I. Introduction to Forecasting with Neural Networks

Artificial neural networks are widely used in many disciplines for their recognized capability of improving accuracy of forecasting models. This is due to the mapping of the input and output vectors "without pre-assuming any fixed relationship between them" and their major components: the network topology, algorithm tuning, the division of subsets of data, and reduction in dimensionality (Gonzalez-Carrasco & Garcia-Crespo, 2012). It is recognized in the literature that using neural networks to solve a real problem is not easy and there must be investigation into finding the most appropriate neural network framework for the goal of their research due to the complexities in the configurations and architectures. Therefore, a recommended first step is to discover the best classifier for the problem (Gonzalez-Carrasco & Garcia-Crespo, 2012). Some of the popular classifiers to note in the research from 2012 are: multi-layer perceptron (MLP), support vector machines (SVM), and RNNs (Gonzalez-Carrasco & Garcia-Crespo, 2012). However, discovering a strong classifier does not necessarily have to be the first step and one can change classifiers or train a neural network to create a learning model that can be applied to other classifiers, inherently complicating the validation.

Researchers such as Gonzalez-Carrasco and Garcia-Crespo propose artificial neural network frameworks to attempt to optimize different neural networks in order to discover the most appropriate classifier for the data and use a genetic algorithm to do so since it is an adaptive method (Gonzalez-Carrasco & Garcia-Crespo, 2012). This is a potential way to validate an artificial neural network framework, if a decision-maker or data scientist can use various tests such as analysis of variance (ANOVA) to find the one that results in the highest accuracy of forecasting outputs. This step can be very time-intensive and there could still be better

frameworks than the ones tested, however, this is a potential way to include validation of the classifier.

There are examples of short-term energy and power grid load forecasting using time series data in many publications. One of the more complex, or ensemble models, for conducting a forecast is proposed by Zhang & Wang (2018). In order to more accurately predict electrical load in the energy industry due to the complexities of economics, climate conditions, population, and policy, more advanced models must be developed rather than the limitations proposed by, "traditional statistical methods, such as multiple regression methods, exponential smoothing, [auto regressive moving average models], and seasonal ARIMA" (Zhang & Wang, 2018). Their research proposes a new 'decomposition-ensemble model' using seasonal patterns of short-term electrical loads. Figure 50 shows the forecast framework they used, which can also be used as a guideline for stages that would require validation. In their research, they show the validation and use of this framework by using a data subset for comparison in the electrical load markets of New South Wales and Singapore (Zhang & Wang, 2018). If this method was used with other data or disciplines such as in finance, validation should be done at each stage, as recommended in the proposed validation framework previously shown in Figure 8 in this dissertation to ensure proper execution. This reaffirms that using different industry-specific forecasting models, ensembles, hybrid-approaches, and frameworks can work, but may require an experienced data scientist to look at each step qualitatively, or by using exploratory data analysis to confirm that the methods are appropriate.

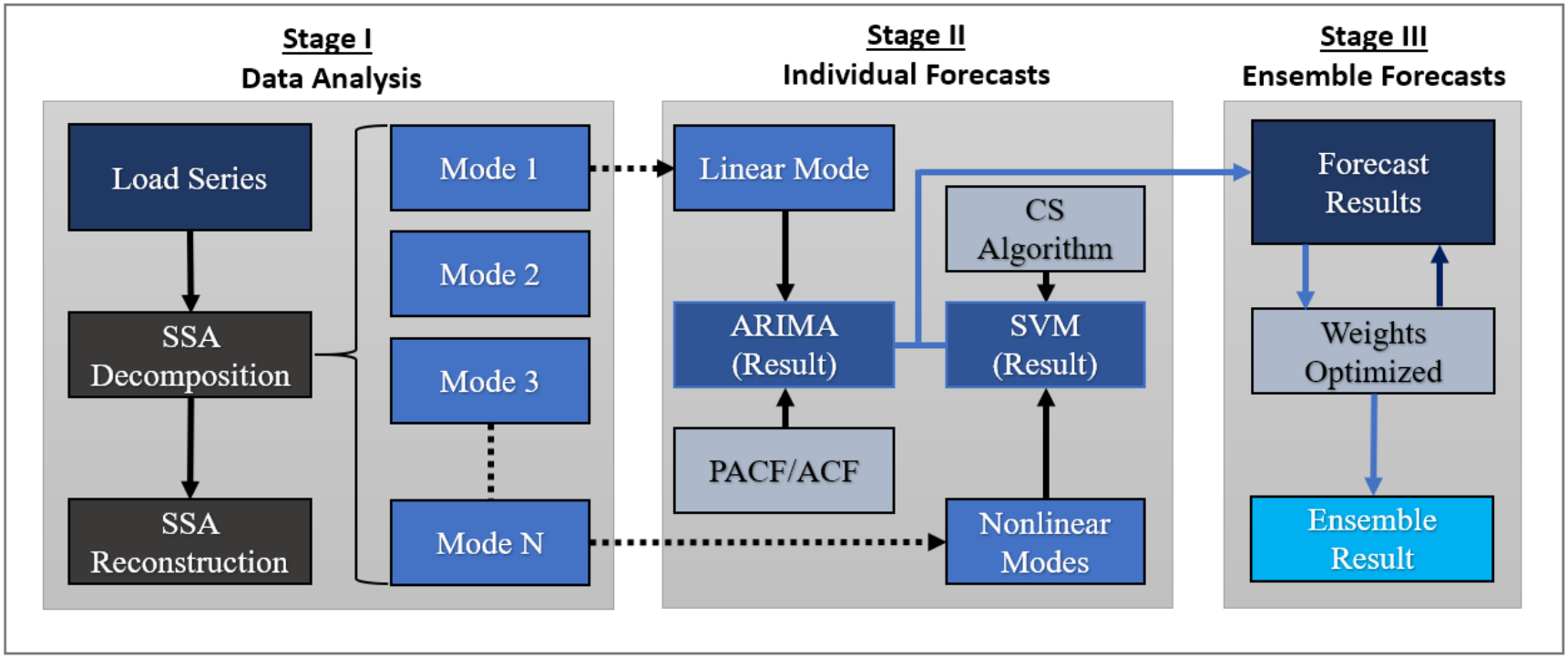



*Figure 51. Short Term Load Forecasting Forecast Section of Flowchart (Re-Colorized)*

*(Zhang & Wang, 2018)*

A more in-depth look at their framework in Figure 50 shows that this is a hybrid model that uses non-linear and linear data in a forecasting model. They used hourly time series data which provided successful results and employed a singular spectrum analysis (SSA) to decompose and recompose the initial data set to discover hidden characteristics. This is a good example of integrating validation within the model and their results are positive using this proposed decomposition-ensemble method. However, it should be noted that there are significant changes from the raw time series that may require further validation. The researchers did recognize this flaw or trap, and state, "[a]fter the individual forecasting, to cancel out the diverse errors generated by single predictors, the weighting-based approach based on the CS algorithm is adopted to aggregate the individual forecast outcomes." This method allowed them to avoid errors that may have been created by single predictors (Zhang & Wang, 2018). These are the types of traps, flaws, and biases that can occur in basic forecasts and are only more likely to occur in complicated ones. A data scientist needs to be cognizant and always looking out for these statistical or data science traps, especially when implementing a neural network approach to forecasting.

The next sections of this chapter will briefly discuss forecasting methods associated with various types of commonly used neural networks and give a brief overview and understanding of why certain neural networks are used and how they are applied to forecasting.

**II. Convolutional Neural Network (CNN)**

To give a background on Convolutional Neural Networks (CNNs) and how they are applied to forecasting models today it is important to understand the interdisciplinary uses that they are mostly used and developed for. In 1998 Yann LeCun developed LeNet which used backpropagation to allow a computer to understand and interpret numbers that were handwritten, even if they overlapped or crossed into each other by the human who wrote them. His early functioning neural network was used to read financial checks – a use in the financial and banking industry that has nothing to do with forecasting financial data which many neural networks are used for today. CNNs are also used heavily in imagery analysis – from scientific to satellite imagery since the imagery can be fed to the network and a trained classifier can be implemented to decipher objects or activity. Convolutional neural networks essentially "scan" an image by a resolution size, for example a 50px x 50px square section of an image and looks to detect an object or activity. It is then fed through what are called convolutional layers that care about the 50px x 50px cells next to them to try to identify additional similar objects. They also can use pooling layers for additional parameters such as colors (Lecun, Bottou, Bengio, & Haffner, 1998). Moreover, Bayesian approaches to neural networks are also used today to strengthen accuracy, object, and activity detection. An end-to-end pipeline, as described in the literature for a convolutional neural network can be seen in Figure 52. This specific pipeline flow chart is for the intended goal of detecting objects in imagery, and each convolution can piece back together the original raw data lineage.

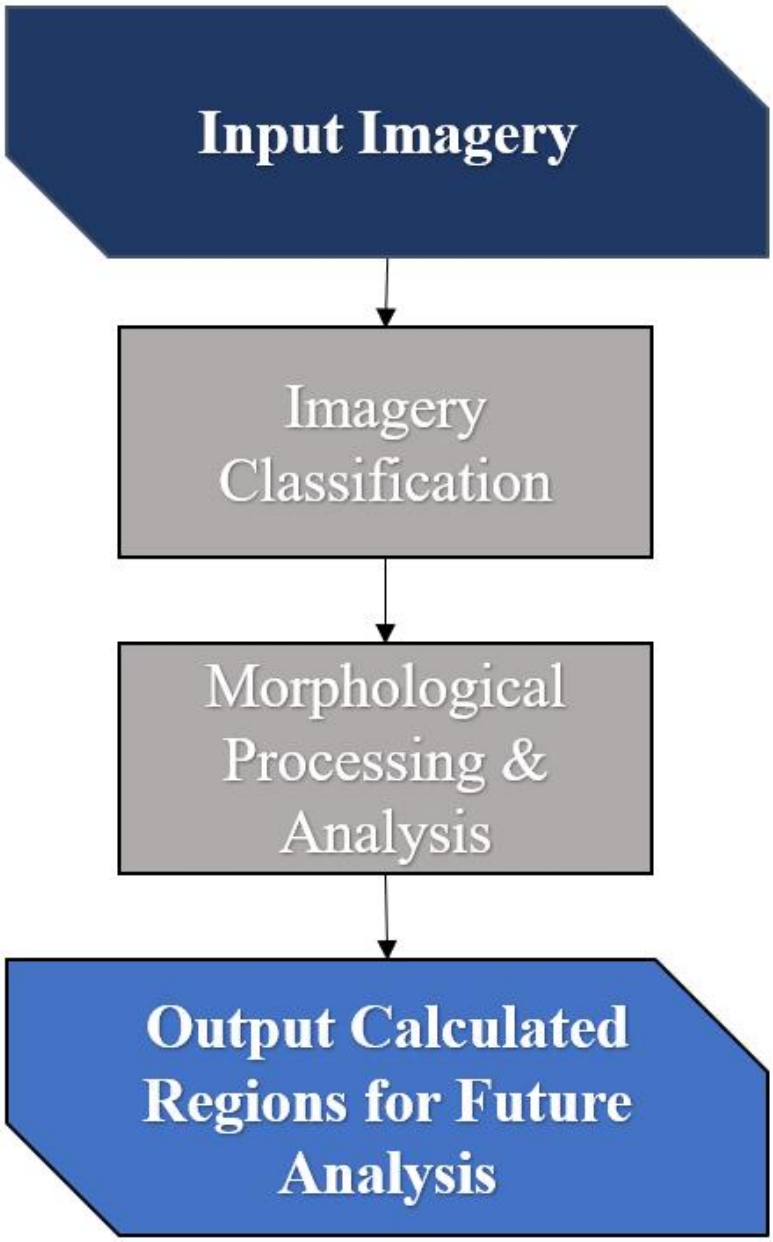



*Figure 52. End-to-end Pipeline Flowchart*

*(Cisek, Dale, Pepper, Mahajan, & Yoo, 2016)*

An example of the neural network architecture used for a simple scan of counting cars in a parking lot is shown in Figure 53 that uses the end-to-end pipeline flowchart described in Figure 52. Financial firms, hedge funds, and venture capital firms are just some of the types of companies that are interested in using neural networks on satellite imagery to incorporate data on the number of cars in stores' parking lots into their forecasting models to forecast future sales (Cisek, Dale, Pepper, Mahajan, & Yoo, 2016). When using an architecture such as this, the values of the patch sizes may need to be validated at different granularities and computer processing time and costs may need to be considered, however, this is a good example of what a convolutional neural network looks like for demonstrative purposes.

| Type | Patch Size/Stride | Output Size |
|---|---|---|
| Data | --- | 3 x 60 x 60 |
| Convolution | 3 x 3 / 1 | 64 x 58 x 58 |
| Max Pool | 2 x 2 / 2 | 64 x 29 x 29 |
| Convolution | 3x3 / 1 | 64 x 27 x 27 |
| Max Pool | 2 x 2 / 2 | 64 x 14 x 14 |
| Convolution | 3 x 3 / 1 | 64 x 12 x 12 |
| Max Pool | 3 x 3 / 2 | 64 x 6 x 6 |
| Inner Product | --- | 500 |
| Softmax | --- | 500 |

*Figure 53. Example of Convolutional Neural Network Architecture*

*(Cisek, Dale, Pepper, Mahajan, & Yoo, 2016)*

Recently, given the superior performance of neural networks over Support Vector Machines (SVMs) at other computer vision tasks, imagery and computer scientists tend to lean towards using a convolutional neural network for many imagery classification problems, since it has a specialized architecture for handling images (Cisek, Dale, Pepper, Mahajan, & Yoo, 2016). This also touches upon transfer learning, which allows knowledge gained through the training of a neural network, known as the source domain, to improve the training process of a network on a different but similar target domain (Cisek & Mahajan, 2017). When implemented, transfer learning can reduce the amount of data points necessary and overall data required for training a model and can be less computationally intensive (Cisek & Mahajan, 2017). There are many limitations that may require a lot of data preparation or tweaking which increases the necessity for validation.

Understanding CNNs is valuable when using neural networks, however, for forecasting, recurrent neural networks seem to be more appropriate. There are many recent examples of using convolutional neural networks in the literature for forecasting applications such as "smart-grid short-term load forecasting using an ensemble approach combining a convolutional neural network with K-means clustering (Dong, Qian, & Huang, 2017) or for particulate matter (pollutants) such as published in a paper discussing deep convolutional neural networks combined with Long Short-Term Memory (LSTM) for smart city forecasting (Huang & Kuo, 2018). To continue the discussion of convolutional neural networks used for imagery and applied to forecasting, another example is discussed next. This convolutional neural network focuses on time series classification where researchers were able to combine it with a relative position matrix by transforming the time series data into imagery data. The architecture for this CNN converts raw time series data into an image map, or 'image representation' to reduce dimensionality (Chen & Shi, 2019). Therefore, it

is important to understand the history regarding how these computer science derived technologies such as convolutional neural networks are being used for time series forecasting. Figure 54 shows the architecture of this neural network framework which retains the time series data using a transform equation.

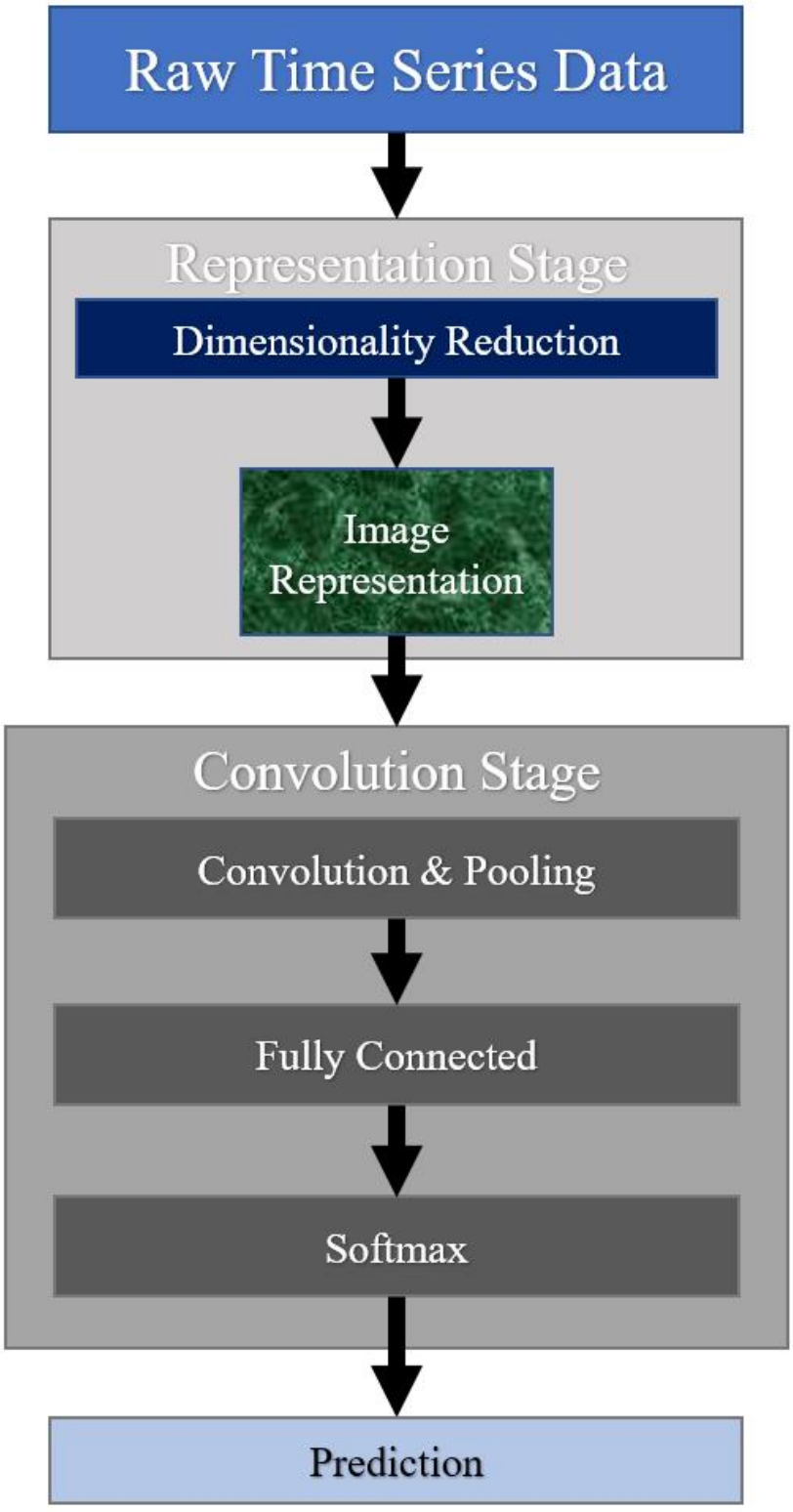



*Figure 54. Architecture of Relative Position Matrix Convolutional Neural Network (Chen & Shi, 2019)*

The time series data is retained through the equation:

$$M = \begin{bmatrix} x_1 - x_1 & x_2 - x_1 & \cdots & x_m - x_1 \\ x_1 - x_2 & x_2 - x_2 & \cdots & x_m - x_2 \\ \vdots & \ddots & \ddots & \vdots \\ x_1 - x_m & x_2 - x_m & \cdots & x_m - x_m \end{bmatrix}$$

*Equation 7. Transform Equation to prepare data before final matrix (Chen & Shi, 2019)*

By using a transform equation, "every two time stamps of the time series are connected by *M* to obtain the information of the whole series by taking a certain time stamp as a reference point" (Chen & Shi, 2019). The authors also state that this can also be considered a 'data augmentation' method. To validate a method such as this, the researchers used other 2-dimensional convolutional neural networks (2D-CNNs) at the representation stage to see which performed better. They also noted that certain methods included exponential smoothing and that their future plans are to test this on more complex time series data (Chen & Shi, 2019). It should be noted that this method has a high amount of extract, transform, and load, which can make validation difficult. Converting time series data to an image to reduce dimensionality can have significant effects on the outputs as discussed in earlier chapters and shown in the examples from the venture capital and methanol industries. The decision-maker or data scientist should check the image representation data using common statistical and imagery tests before moving from the 'representation stage' to the 'convolutional stage' to verify outputs are correctly calculated.

In a more recent example of using convolutional neural networks from 2017, a Long Short-Term Memory (LSTM) approach was used with a spatially supervised recurrent convolutional

neural network for applications in imagery and full-motion video exploitation, or "visual object tracking." Their approach uses a convolutional recurrent neural network (CRNN) using deep learning to create a history of features and learn different objects. This was based on box regression methods and implements the LSTM to apply these methods to video exploitation (Ning, et al., 2017). It is important to understand these applications and neural network architectures as they are now being used in forecasting models, as ensemble forecasts, and for data preparation. All new computer-science derived approaches require additional and possibly more technical and mathematical validation than the previous chapters and simulations described in this dissertation. To reiterate this, another example of a CRNN can be seen in finance to discover share price trend prediction where a CNN is used with LSTM as shown in Figure X. This shows how features are extracted out of a CNN and then ingested into an LSTM to handle, "issues of gradient disappearance and expansion of the time series data" (Yu, Chu, Chan, & Wang, 2019).

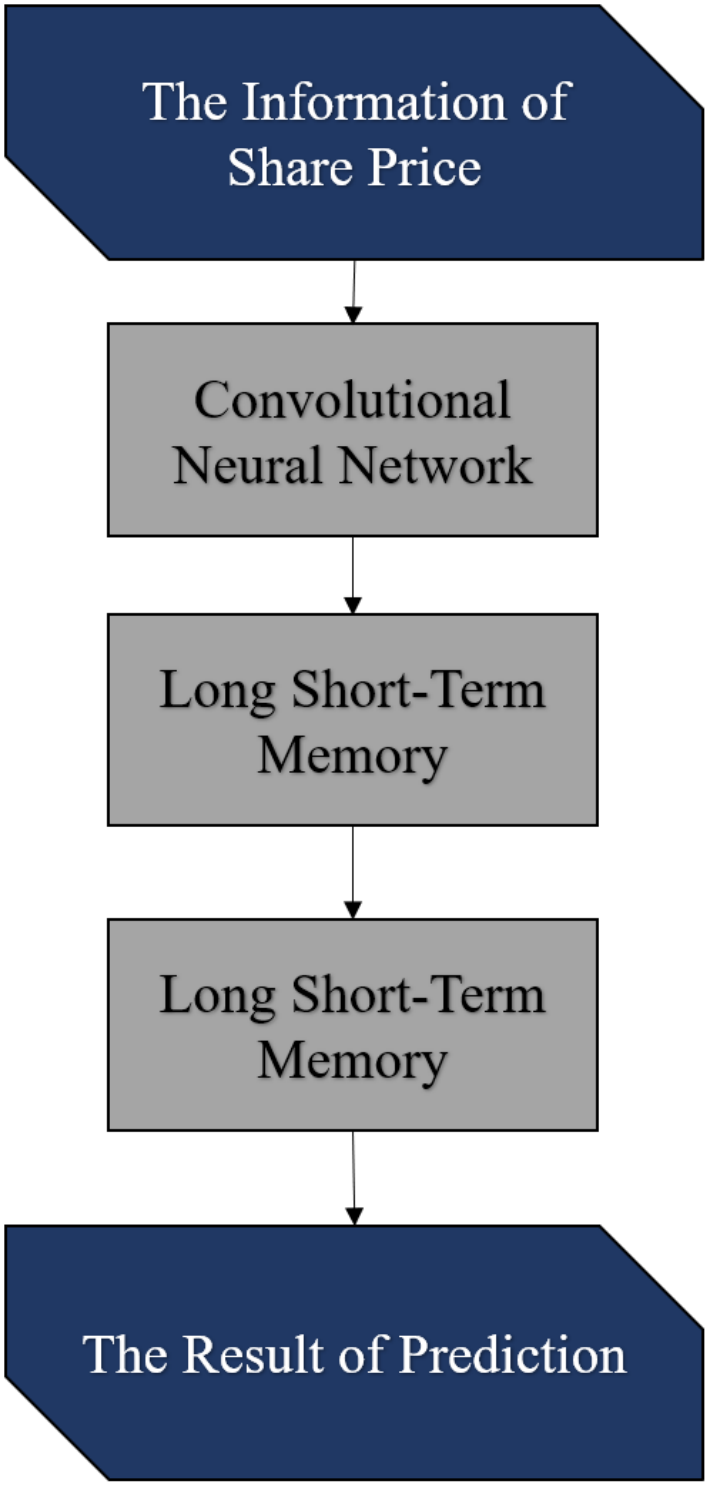



*Figure 55. Share Price Trend Prediction using CRNN with LSTM Structure*

*(Yu, Chu, Chan, & Wang, 2019)*

Many researchers use LSTM in their ensemble or hybrid forecast models to assist in validation or verification such as (Mehdiyev, Lahann, Emrich, & Enke, 2017), (Sang & Massimo, 2019), (Yu, Shao-Wei, & Chan, 2019). However, LSTM is also being improved constantly in the literature, as seen in Sang (2019), where a neural network using a TensorFlow LSTM is used to improve trading technical analysis (Sang & Massimo, 2019). They justify their approach by stating that this is the best way to fit non-linear algorithms due to the information propagation weights being adjusted through the neuron layers. In their work they are focused on using an LSTM ANN model to better improve technical analysis, which as defined by them is referring "to the analysis methodology for forecasting the direction of prices through the study of patterns in past market data, primarily price and volume," (Sang & Massimo, 2019) by using common financial statistical methods such as Simple Moving Average (SMA), Relative Strength Index (RSI), and Moving Average Convergence Divergence (MACD) (Sang & Massimo, 2019).

**III. Recurrent Neural Network (RNN)**

Since some approaches using neural networks combined with forecasting models can use recurrent neural networks, a brief background on what they are and how they are used is necessary for a data scientist or decision-maker implementing them in a forecasting model. An example from the early 1990's is where Japanese researchers at Fujitsu Laboratories used a recurrent neural network to develop a posture recognition system for Japanese sign language with strong results – 98% recognition rate (Murakami & Taguchi, 1991). They used a recurrent neural network because they were able to use time-series data and allow a hidden layer to feedback upon itself. This allowed the neural network to, "trace [ … ] processing at the previous time slice" due to the context layer (units) being stored in 'memory' (Murakami & Taguchi, 1991). This is why many recurrent neural networks are also called 'Auto Associative' which is due to their ability to handle time series and store information in memory (Poznyak, Oria, & Poznyak, 2019). A recurrent neural network framework is shown in Figure X. This is similar to a feed forward neural network as previously seen in Figure 2, except it adds the copied context layer which feeds back into the hidden layer to do trace processing to understand time series data using back propagation (Murakami & Taguchi, 1991). It should also be noted that by 1994, researchers identified limitations that caused recurrent neural networks to not be widely used for the control of nonlinear dynamical systems. This was due to the potential ineffectiveness of training algorithms and proposed algorithms such as the decoupled extended Kalman filter (Puskorius & Feldkamp, 1994).

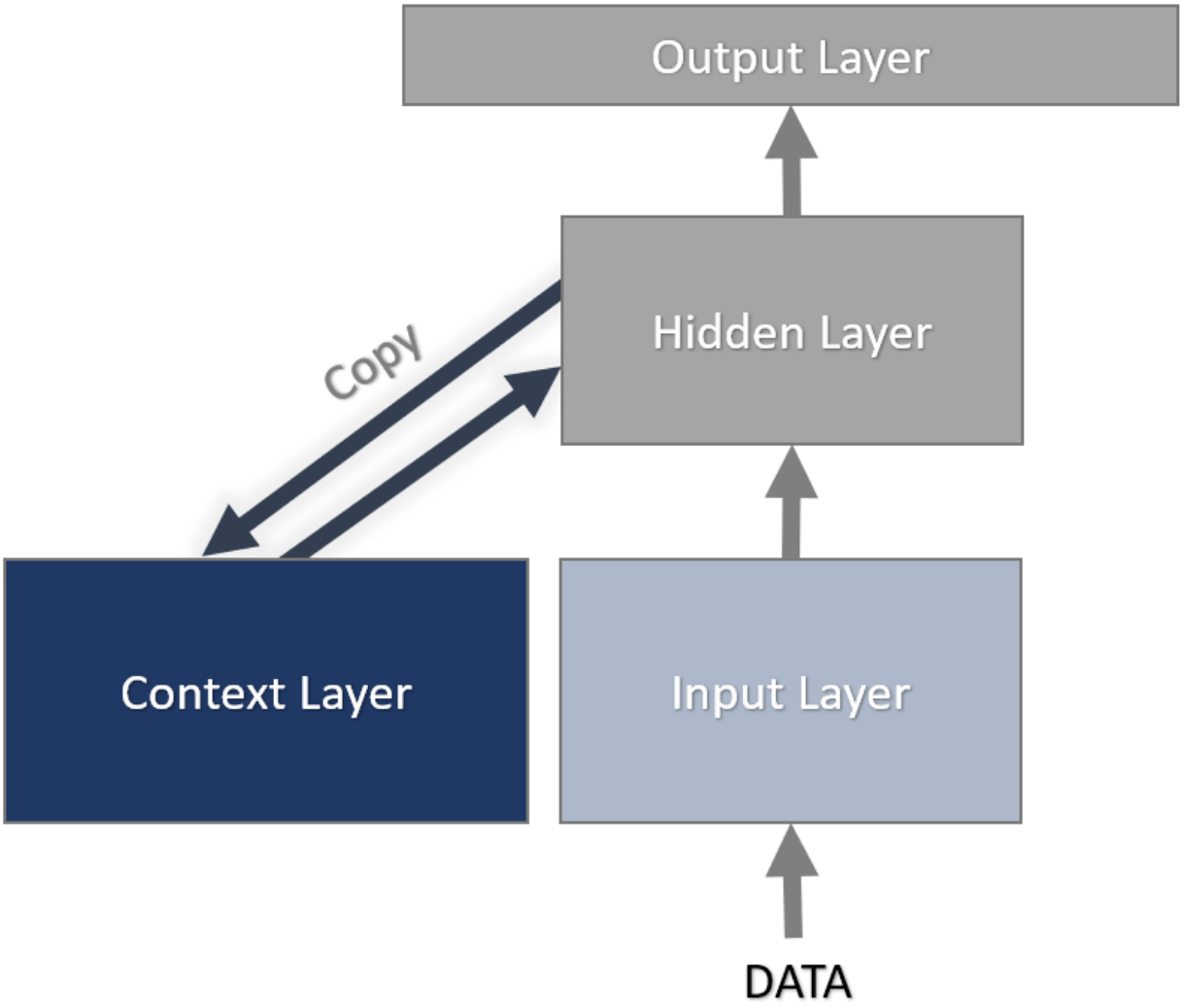



*Figure 56. Recurrent Neural Network in 1991 (Upconverted)*

*(Murakami & Taguchi, 1991)*

In 2011, researchers combined wind speed forecasting models with time-series data to compare multivariate and univariate accuracies of ARIMA models with recurrent neural networks. They discovered the recurrent neural networks outperformed the ARIMA models, and the multivariate ARIMA outperformed the univariate ARIMA (Cao, Ewing, & Thompson, 2012). Their recurrent neural network model is shown in Figure 57. Note that this differs from the 1991 model used by Murakami & Taguchi where the context layer interacted with the hidden layer. This network feeds the output layer into the context layer and back into the hidden layer. Moreover, there are various types of recurrent neural networks, such as Hopfield networks, Elman recursive networks, Jordan networks, and others. The recurrent neural network in Figure 57 follows a Jordan network architecture in regards to the fact that it is feeding the output back into the context layer (Poznyak, Oria, & Poznyak, 2019).

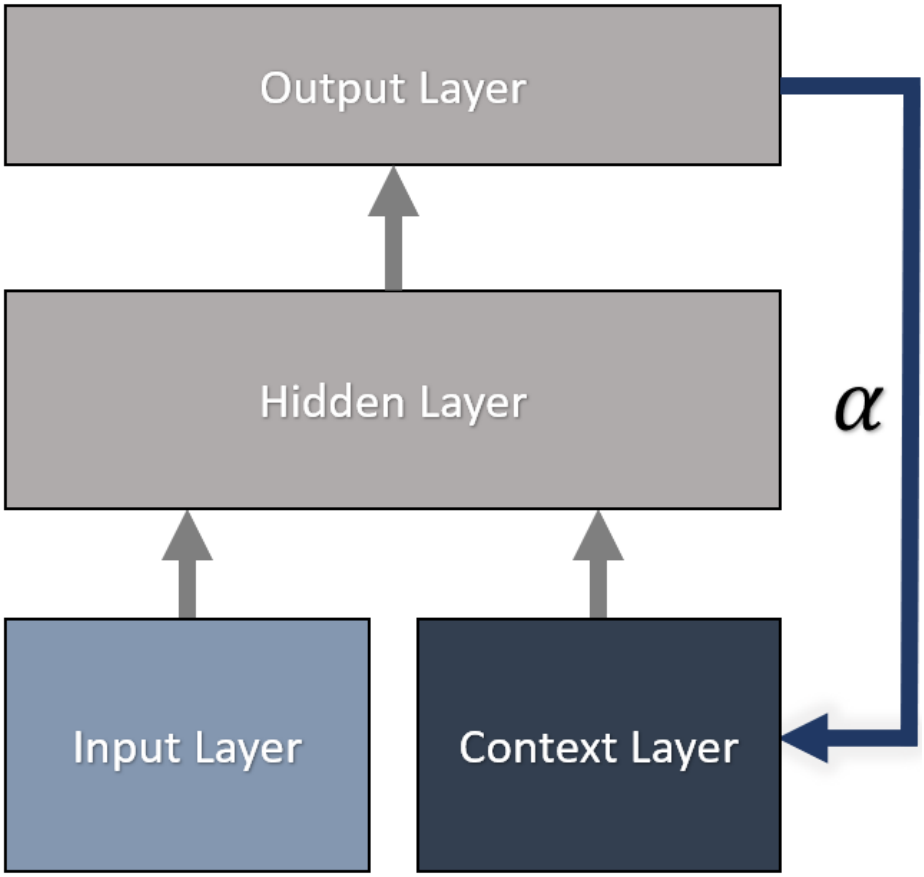



*Figure 57. Recurrent Neural Network in 2012 (Colorized)*

*(Cao, Ewing, & Thompson, 2012)*

A current example is using neural networks in the energy industry to increase the accuracy of power-grid failure forecasting. This is implemented by using a neural network with a binary classification forecasting model (Haseltine & Eman, 2017). These models include variables such as historical and forecasted weather, load on the power grid, and power reserve margins for training data. This application of new technology applied to forecast modeling for the energy sector is due to the previous forecast failures of predicting blackouts such as "The Southwest Blackout of 2012, the Florida Blackout(s) of 2008, and the major Northeast Blackout of 2003" (Haseltine & Eman, 2017). However, using neural networks in forecasting is not new – the neural network technology is now mature enough to apply. Previous forecasting models using neural networks for energy industry forecasting in the early 1990's recognized flaws and that further research was needed to investigate "whether a neural network result is system dependent, in that the model may vary from utility to utility, and whether it is case dependent where the accuracy may vary from one time period to another, are crucial for a utility company in order to evaluate a new technique for load forecasting" (Lu, Wu, & Vermuri, 1993). In addition, the early studies done on applying neural networks into forecasting models recognized flaws, traps and biases, such as ensuring data preparedness by cleaning the data by removing errors and pre-processing (Azoff, 1994).

## IV. Generative Adversarial Network (GAN)

After understanding some of the architectures of neural networks, Generative Adversarial Networks can be implemented to allow competition between neural networks to increase accuracy of many applications, one of which, is forecast modeling. GANs were proposed in a publication in 2011 by Ian Goodfellow of University of Montreal by providing a framework for "estimating generative models via an adversarial process, in which we simultaneously train two models: a generative model $'G'$ that captures the data distribution, and a discriminative model $'D'$ that estimates the probability that a sample came from the training data rather than $'G'$." The point of using two models simultaneously is by training $'G'$ one can maximize the probability of $'D'$ making a mistake (Goodfellow, Pouget-Abadie, Mirza, & Xu, 2014). This method of using adversarial neural networks allowed for training to be done without Markov chains, and other existing training methods for creating samples while allowing backpropagation training for the system (Goodfellow, Pouget-Abadie, Mirza, & Xu, 2014). This led to researchers in 2019 publishing works with GANs such as: "Stock Market Prediction Based on Generative Adversarial Network"; "Improving trading technical analysis with TensorFlow Long Short-Term Memory (LSTM) Neural Network"; "Modeling financial time-series with generative adversarial networks." Adversarial neural networks used to model financial time series are currently being researched in many publications because of the complex problem of learning and training a network on data while maintaining statistical properties. Moreover, as stated by researchers such as Takashi Shuntaro from the University of Tokyo, "The GAN model produces a time-series such as the linear unpredictability, the heavy-tailed price return distribution, volatility clustering, leverage effects, the course-fine volatility correlation, and the gain/loss asymmetry" (Shuntaro, Yu, & Kumiko, 2019). Using the GAN approach to generate a financial time-series, Shuntaro is

able to overcome some of the problems with financial time-series data due to the complexity of financial data, for example, he states that "[m]odeling stylized facts, however, requires the model to capture the global structures of time-series, which may not be acquired through training for the prediction of each value. GANs architecture trains the generator $'G'$ using the signals of the discriminator $'D'$, which only concerns the properties of the whole time-series" (Shuntaro, Yu, & Kumiko, 2019). Other current researchers have built on using GAN architectures to conduct stock market predictions where they introduce new models using GANs with multi-layer perceptrons (MLP) combined with popular methods such as Long Short-Term Memory (LSTM) models to predict the closing price of a stock (Kang, Zhong, Dong, & Wang, 2019). In their research they used S&P 500 as a data set and were able to outperform other machine learning methods and base their research on the fact that ARIMA performs well on linear data such as stationary time-series but not work as well on non-linear data. They also compared an artificial neural network (ANN) and a convolutional neural network (CNN) using the same stock price data. In Figure 58 the generator for their GAN is shown and in Figure 59, the discriminator is shown. However, these are parts of a more complex forecasting model and framework due to the use of adversarial neural networks.

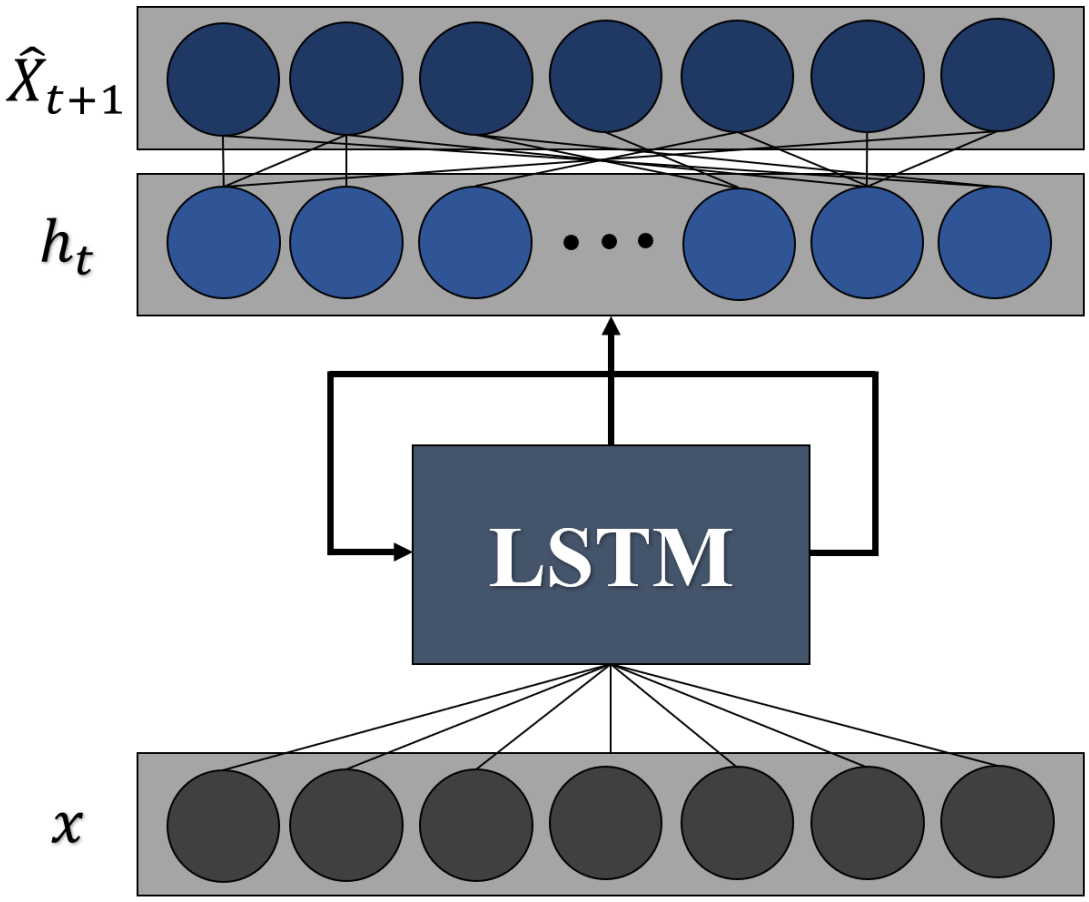



*Figure 58. Generator Designed with LSTM (Reformatted)*

*(Kang, Zhong, Dong, & Wang, 2019)*

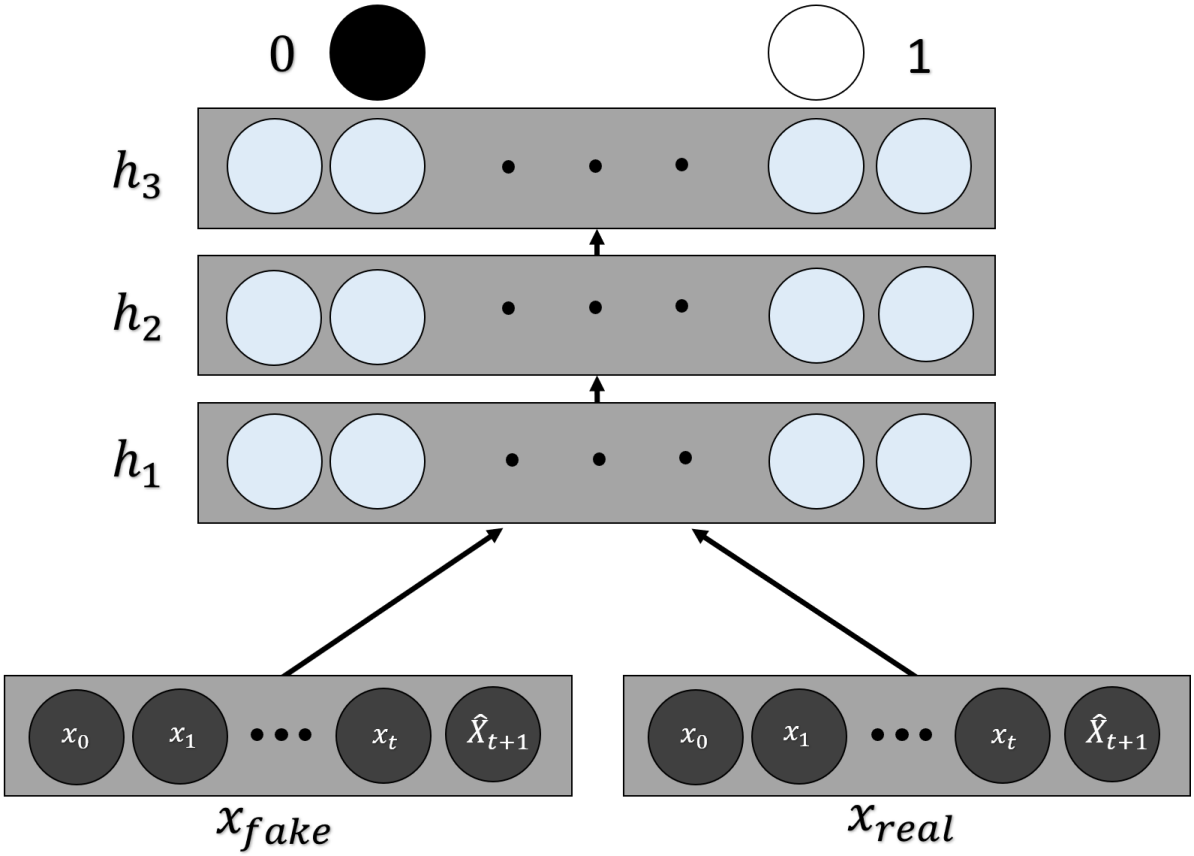



*Figure 59. Discriminator designed using an MLP (Reformatted)*

*(Kang, Zhong, Dong, & Wang, 2019)*

To attempt to mitigate and error-check, they attempt to ‘fine tune’ or calibrate the data by checking if the discriminator is correct against the generator, therefore, creating an adversarial network. Using a GAN architecture, they concluded that they were able to outperform Support Vector Regression (SVR), ANN/CNN, and LSTM which is what they used as baselines for comparison in their research.

**V. Hybrid, Combination, and Ensemble Neural Network Forecasting Models**

Since the 1970's a popular forecasting method is using autoregressive integrated moving average (ARIMA) to conduct forecasting and more recently, creating a hybrid model with an artificial neural network, researchers were able to take advantage of the linear and nonlinear modeling aspects of each to improve forecasting accuracy (Zhang G. P., 2003). However, researchers such as G. Peter Zhang recognized there are flaws in this model since there may be linear correlation structures left over in the residuals, and therefore certain patterns will be left undetected. Therefore, it is important to discover errors, run an autocorrelation function (ACF) on the outputs, and use exploratory data analysis techniques on each data set as conducted in the simulations in Chapter 4 and Chapter 5. Zhang also notes that there is "currently no general diagnostic statistics for nonlinear autocorrelation relationships. Therefore, even if a model has passed diagnostic checking, the model may still not be adequate in that nonlinear relationships have not been appropriately modeled" (Zhang G. P., 2003). This is why the hybrid approach used implements an artificial neural network (ANN) to attempt to increase accuracy by discovering the nonlinear relationships with $n$ input modes:

$$e_t = f(e_{t-1}, e_{t-2}, \dots, e_{t-n}) + \varepsilon_t$$

*Equation 8. ANN Residuals*

*(Zhang G. P., 2003)*

Zhang also mentions that "the correct model identification is critical, and the combined forecast of *Lt* (linear component), $N_t$(nonlinear component), and $e_t$ (the residual at a time *t* from the linear model) to produce $y_t$ (the combined forecast) using the formulas in Equation 8 and Equation 9. By taking account for the residuals, accuracy seems to be increased in most examples from the literature since there is less of a loss of data. However, proper validation should be conducted to ensure that the data is not transformed so much that it loses the added value from this method. Therefore, there is inherent risk increased in complicated hybrid, ensemble, and combination approaches.

$$\hat{y}_t = \hat{L}_t + \hat{N}_t$$

*Equation 9. Combined Forecast using ARIMA and ANN (Linear and Nonlinear)*

(Zhang G. P., 2003)

Since Recent hybrid approaches can become extremely complex and difficult to validate, it is worth continuing to explore examples of complex approaches. In 2019 research on improving the accuracy of time-series forecasts using an autoregressive integrated moving average and artificial neural network hybrid method (ARIMA-ANN) was published comparing various forecasts and including "empirical mode decomposition" to avoid making the mistakes of other methods (Büyüksahin & Ertekin, 2019). This research is focused on handling large amounts of time series data with the knowledge that combining linear and nonlinear forecasting models has shown accuracy increase in previous research. Furthermore, the researchers were aware of the restrictions, and in order to mitigate them, they applied an additional method called the "Empirical Mode Decomposition (EMD) Technique." This approach uses an ARIMA-ANN

method to decompose the data and combine "linear and nonlinear models throughout the hybridization process" to increase forecasting accuracy and performance, and afterwards, combine an EMD technique to increase forecasting accuracy (Büyüksahin & Ertekin, 2019). According to the results, their methods with EMD had less error than other forecast models and stated that by adding EMD, they can solve volatility problems with ARIMA. The researchers conclude by stating, "As a future work, further time series analysis methods can be applied to each EMD component as a pre-processing step to choose the most proper method to apply," and that this can be used in a multi-step-ahead forecast model, not just a one-step-ahead forecasting model (Büyüksahin & Ertekin, 2019). However, this complex pre-processing and advanced hybrid model should be validated carefully before using with other data sets. This is denoted in the Proposed Forecast Model Validation Framework in Chapter 3 as optional since this adds a new set of validation issues. Incorporating neural networks may increase accuracy but can increase the risk and complexity of proper execution which is why it is important to validate carefully and choose the correct type of model.

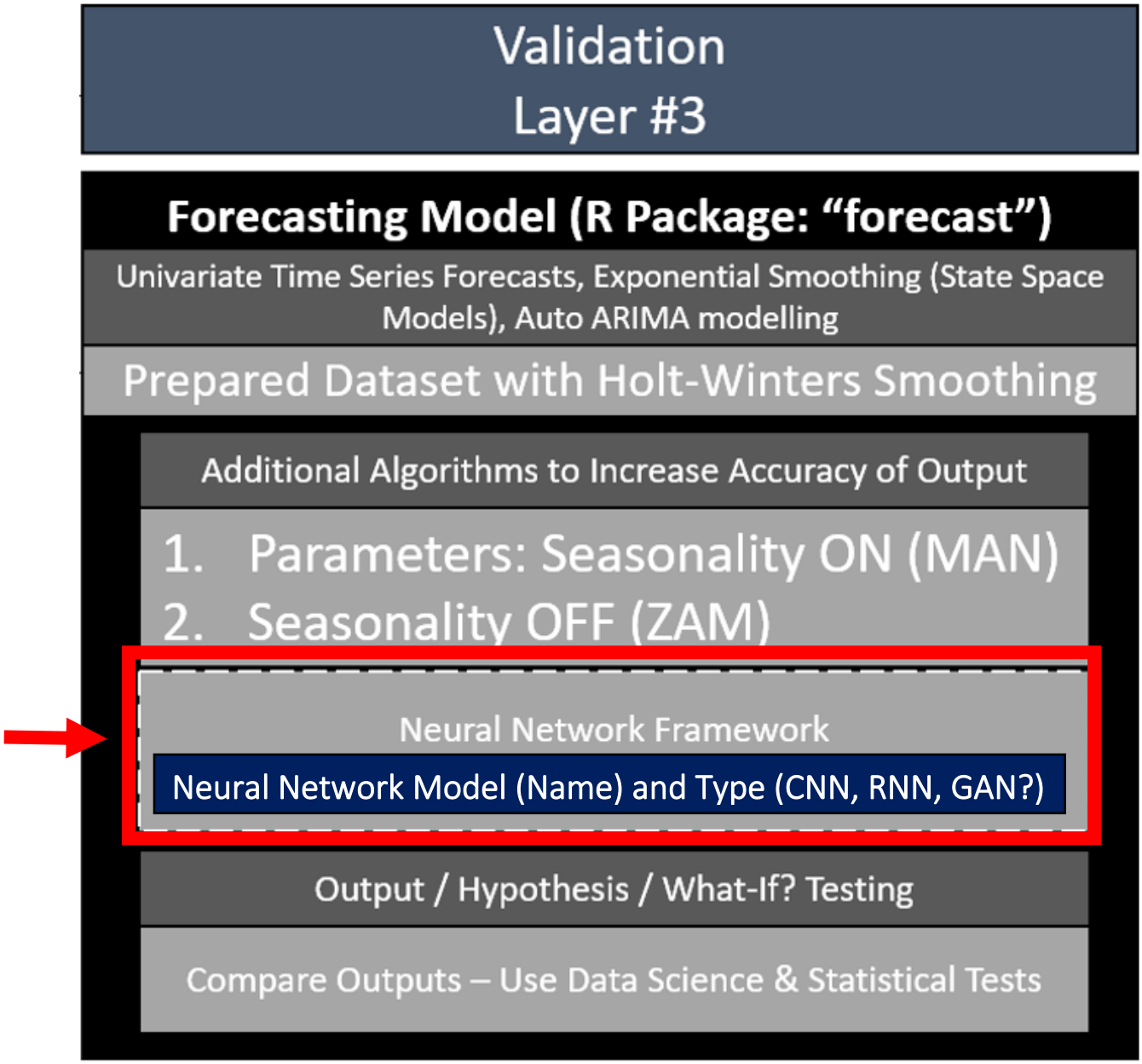



*Figure 60. Example of Proposed Forecast Model Validation Framework - Validation Step #3*

Alternatively, there are approaches which claim to outperform ARIMA, such as the multi-layer perceptron neural network and Elman Recurrent Neural Networks in terms of root mean square error (RMSE) and lower variance (Lasheras, Francisco, & Sanchez, 2015). The wind speed forecasting example mentioned previously in Section III of this chapter which also concluded that the neural network (in this case, a recurrent neural network) outperformed univariate and multivariate ARIMA models (Cao, Ewing, & Thompson, 2012).

Current hybrid or 'ensemble' approaches to neural networks have generally resulted in positive performance in recent literature in works. Smith & Jin discuss how ensemble methods have had better performance than single models but the selection and actual combination is a complicated endeavor and focus their research on multi-objective generation of time series RNN (Smith & Jin, 2014). Other examples are ones such as using deep learning in the process industry for time series data (Mehdiyev, Lahann, Emrich, & Enke, 2017), and discovering the optimal combination for neural network based time series forecasting (Wang, Wang, & Liu, 2018). These publications propose various frameworks of using neural networks to increase performance of basic forecasting models and like most published works, discuss some limitations or traps, however, it seems that qualitative approaches and educated judgments may some of the more realistic ways to discover the best model of the data set.

In regards to the recent literature on the topic of hybrid, ensemble, and combination forecasting with neural networks, in 2014, researchers began realizing that ensembles, or hybrid approaches to forecasting, were "shown to provide better performance than single models," but the problem is creating and choosing the correct combinations (Smith & Jin, 2014). The research discusses methods to mitigate the risk and states, "when constructing an ensemble it is important that each individual model is both accurate and diverse" (Smith & Jin, 2014). This research uses

a Pareto set of solutions in order to assist with decision-making between models shown in Figure 61. The data scientist or decision-maker should also be aware of this as another validation point and complexity that is added into an already complicated method. Similarly this is also noted by Medhiyev in 2017, stating, “The necessity to overcome the shortcomings related to the learning processes of the traditional neural networks motivated the researchers to innovate diverse algorithmic approaches, such improving the optimizers, exploring novel approaches for parallelization, applying locally connected networks, etc., resulting in groundbreaking studies in the last decade” (Mehdiyev, Lahann, Emrich, & Enke, 2017). The recognition of this new area of forecasting science and understanding of its complexities propose areas for future work. One of the areas that currently needs and will need improvement is how to truly validate these large complicated approaches. These methods require additional validation as seen in an example in the manufacturing and process industry where they monitored German steel producers and used Principal Component Analysis (PCA) to reduce the dimensionality of the data streams with an LSTM autoencoder model (Mehdiyev, Lahann, Emrich, & Enke, 2017). The autoencoder was used to “encode the high dimensional input data to the hidden layers using the relevant activation functions and then try to reconstruct the original inputs through the decoder layer as accurate as possible” (Mehdiyev, Lahann, Emrich, & Enke, 2017). Using these encoders in a neural network and making the decision of the architecture all require additional validation.

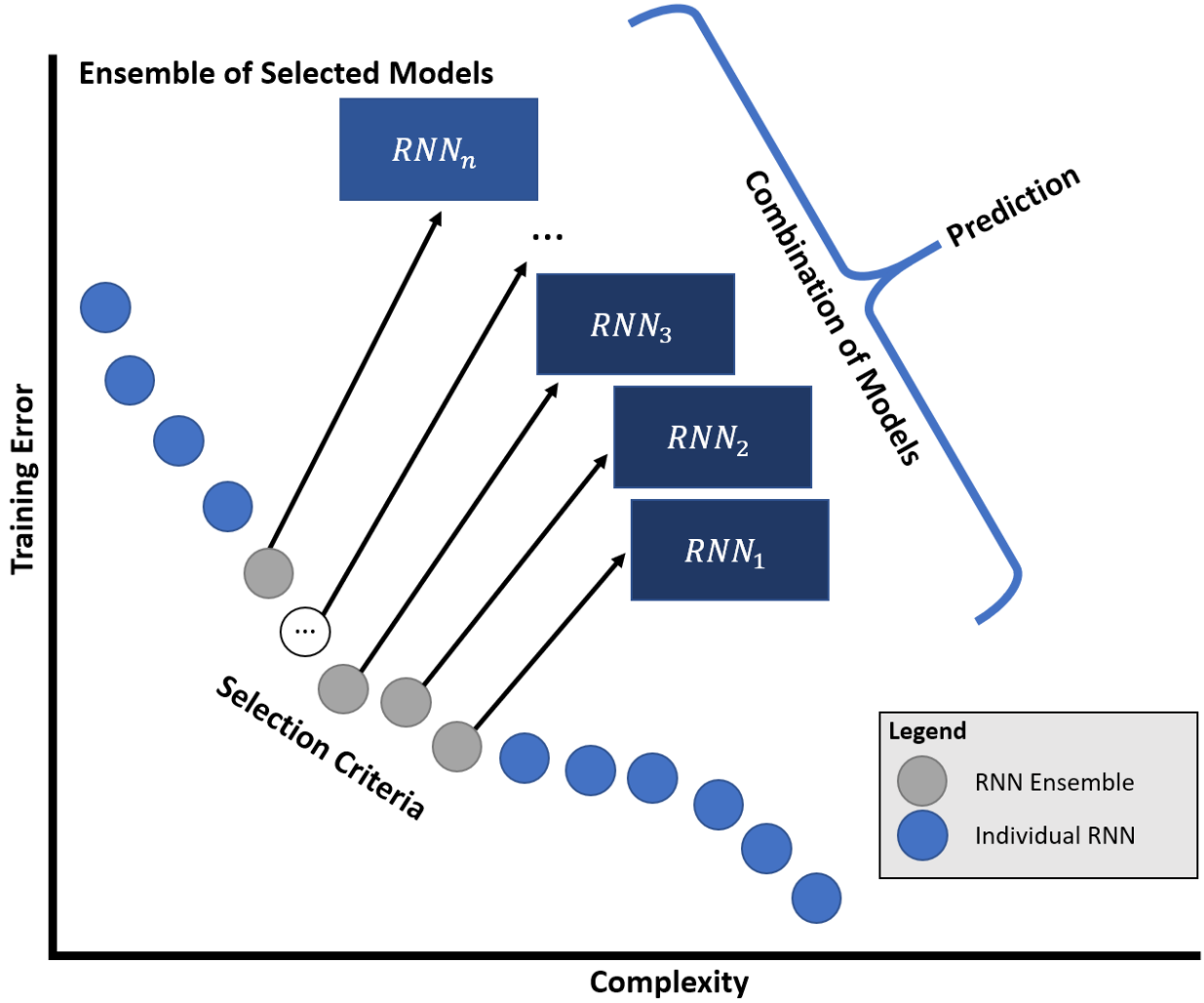



*Figure 61. Pareto Set of Solutions (updated)*

(Smith & Jin, 2014)

Another approach was proposed in 2018 using an 'optimal combination method' (Figure 61) where the researchers concluded that the successes were based on "how well the component models are selected and combination weights are determined," (Wang, Wang, & Liu, 2018) which inherently implies a thorough validation process within the neural network. They recognized that having a strong forecasting model for the data set that results in acceptable results on its own is the best ways to improve performance overall and recommend that artificial neural networks are the best approach for combination modeling. This is because ANNs have an advantage of being "self-adaptive and data driven [ … ] that can suggest an appropriate data-generating process for both linear and nonlinear time series" (Wang, Wang, & Liu, 2018). Their framework is shown in Figure 62.

In these examples, combination, ensembles, and hybrid approaches are all recognized as difficult due to order of magnitude, weighting, decision-making and many other problems. However, if validated correctly, error-tested, and thinking critically with validation at each step, one can confirm these issues are mitigated, and can produce improved accuracy and better forecasting results. To reiterate, "By reducing each individual's generalization error and increasing their ambiguity, the overall generalization error of the ensemble will reduce. However, by increasing the ambiguity of an individual predictor there is an increase in the individual's error. Diverse ensemble members can be either implicitly or explicitly created" (Smith & Jin, 2014). This quote from Smith points out the potential for failure or a flaw, which is what this dissertation is focused on and shows why implementing a neural network with a forecasting model can be extremely complex – and complex to validate.

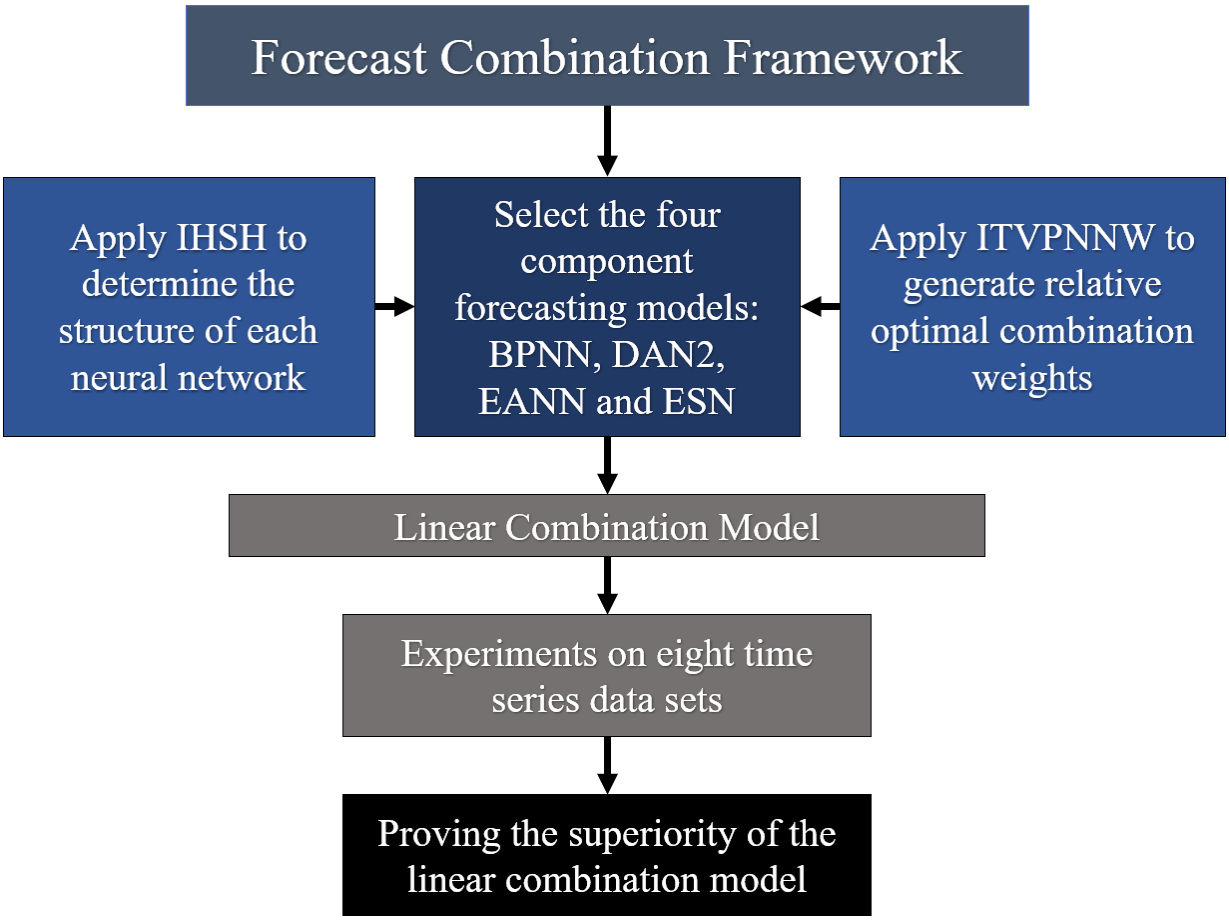



*Figure 62. Flowchart for forecast combination (re-colorized)*

(Wang, Wang, & Liu, 2018)

# Chapter 7: Conclusion – Continuous Validation

***Executive Summary:*** *This chapter will review the conclusions from earlier chapters which focused on basic forecasting models, forecasting frameworks, ensemble and hybrid forecasting models, and how to validate each. It will give recommendations, best practices, and lessons learned from using forecasting in the strategic high-risk innovation driven industry level in venture capital, the methanol time-series pricing subset, and how to implement new approaches using neural networks and deep learning architecture. In addition, heuristic traps and flaws, subjective and unethical data smoothing, and how to better improve forecasts using the proposed validation framework and other validation and statistical testing methods. By reviewing these predictive analytic techniques focusing on forecasting and applying new innovative technologies such as neural network approaches to forecast modeling, this chapter will inform the reader of methods to increase the quality of their forecasting and quality of their data science methodology. This is appropriate for R&D investors, government budget authorities, program managers, science and technology policy makers, venture capitalists, decision-makers, and technologists.*

## I. Takeaways

After reviewing the literature throughout this dissertation from the earlier chapters regarding forecast modeling, validation, and frameworks, the simulations of Chapters 4 and Chapter 5 prove that there are many issues that can arise from leaving parameters default and using forecasting models 'out-of-the-box'. This dissertation provides valid evidence that simple and complex forecasting approaches can cause forecasting failures, misleading results, and non-real-world answers. It is up to the data scientists and decision-makers within organizations as well as organizational leadership to enact a business process that requires the use of a strategic level forecasting framework. The recommendation is to use a framework such as the Proposed Forecasting Validation Framework and tailor it to meet the requirements of the intended line of business, corporation, product line, or academic topic. Force-fitting a framework from other industries, products, or academic practices may work but can also fail. Due to the amount of literature and impacts of forecasting in nearly every aspect of business, academia, and society which can impact policy making, it is imperative to create a quality assurance or validation process to double-check and try to mitigate the risk of non-real-world results in a forecast. Another takeaway from this research is the recommended use of an experienced data scientist or team of data scientists who are familiar with the current literature and experts in all aspects of the model, especially if it brings in computational science algorithms or neural networks, as they can complicate the validation process. Since most small businesses or individual researchers may not have this luxury, again it is recommended to defer to creating a validation process that does not just include the model but also the data set and data preparation phases, along with the potential failures and output types shown in Figure 63. Next, the conclusions from the work presented in this thesis are highlighted.

The results of the simulations demonstrate that there is a strong need for validation throughout the entire forecasting process since there are many stages and variables at each stage that can impact the final output. The proposed forecasting framework is suggested as a strategic guideline for a multi-disciplinary approach to checking, testing, and ensuring that the forecast output is accurately executed. Furthermore, it is meant to provide a high-level understanding and acknowledgement of the intricacies at each stage in forecasting that are commonly overlooked. The framework can serve as a template for a decision-maker or data scientist to get in the habit of checking their data and the model at each step. Often, data is ingested into a model and the output is used at face value, however, the results of the simulations show how different the outputs can be, and how the more accurate real-world outputs was different than the default outputs. Through the simulations and minor adjustments there is a significant different in outputs, which are all mathematically sound, but may be wrong or less accurate than others if used in the real-world.

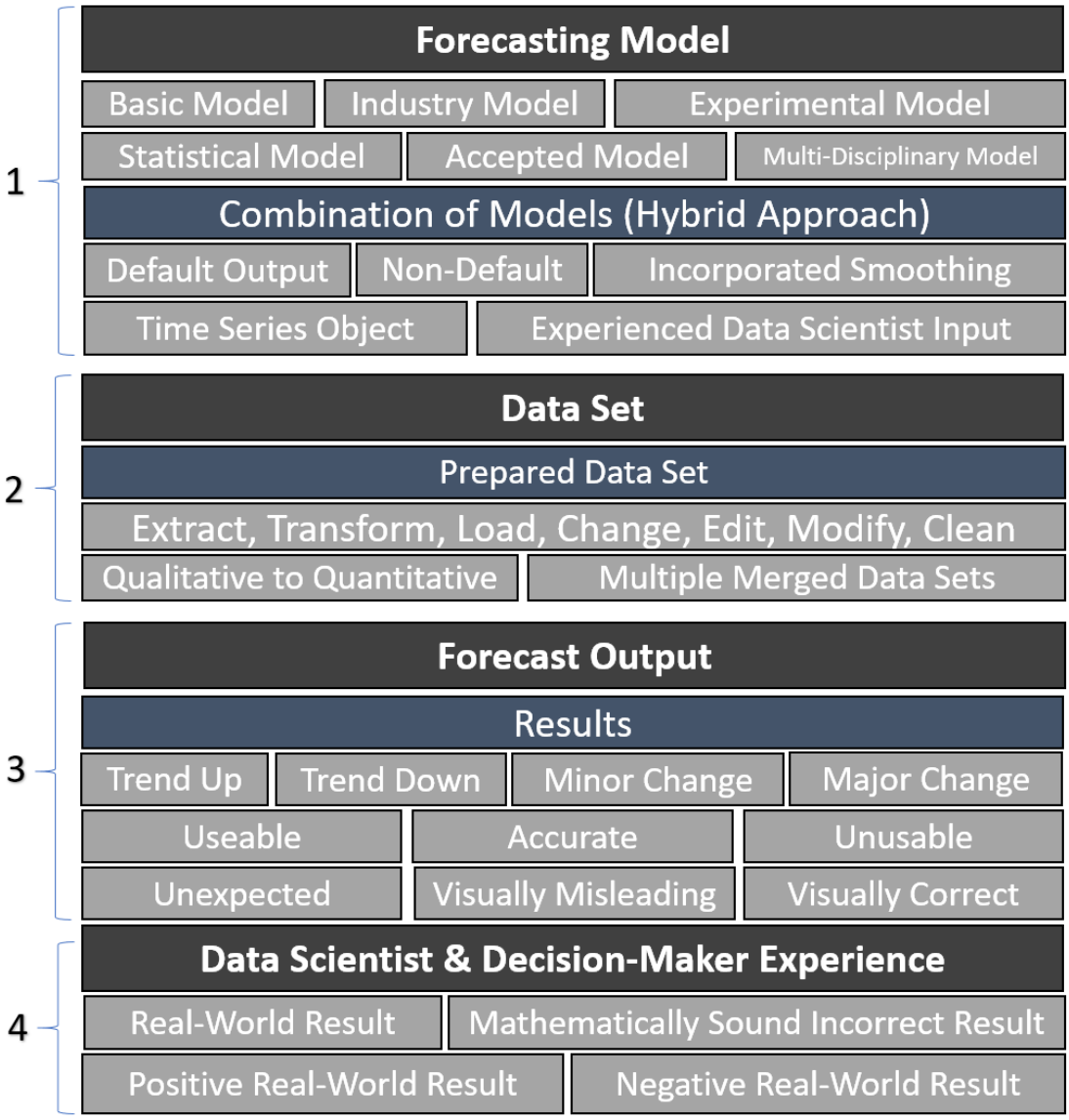



*Figure 63. Forecast Validation Elements*

To reaffirm the takeaways and recommendations of the simulations, Figure 63, Forecast Validation Elements, attempts to codify many of the potential issues that can arise and justify why validation is necessary based on other validation frameworks in various interdisciplinary topics. In Figure 63, Section #1 is assuming optimal model selection has been completed, the data scientist now has a 'forecasting model' chosen. This model can be one of many different types such as: a basic model using something as simple as the NIST forecast formula, an industry model that has been used for a long time in a certain industry or for forecasting a certain product or supply chain, an experimental model from academia or developed in-house, a statistical model, an accepted model that could work well with the data set and forecasting output goals, or a multi-disciplinary model that generally meets the data requirements are just a few examples. Moreover, the chosen model could easily be a combination or hybrid approach which the data scientist must be aware of. There could be innate algorithms built-in or forgotten about such as exponential smoothing or data transformation after ingesting the data into these hybrid models that could also disrupt the data set. Therefore, the data lineage is important to the data scientist or decision-maker. These transformations and ETL algorithms can change the outcome or hide and ignore residuals. This is also the area where the data scientist needs to understand the potential parameters and to ensure they are properly using the model. Starting off a forecasting with default parameters may assist the data scientist in getting a benchmark before editing the variables or parameters initially, but they should also realize if the data needs to be converted in a specific way, such as placed in a time series object, which may require going back to the data preparation stage if the data does not meet specific requirements or has missing data points. This is a crucial stage where the data scientist must use their experience and make strong judgment

calls as well as explore parameter options that may be unique due to the data set being used, which leads to Section #2.

In Section #2, the data set is prepared for use within the model, which can be a difficult stage due to decision-making regarding data transformation. Many forecasts use time series objects, which may require specific data sets to be modified. For example, a forecast that uses monthly data may require the data scientist to take an average of daily data points to produce a new monthly data set, or there could be certain data points missing or removed due to other events (government shutdown, emergency closures, market corrections, abrupt policy changes) that can impact the time series data set. Moreover, according to the literature, one of the more popular and difficult to validate approaches is converting qualitative to quantitative data with new algorithms such as natural language processing (NLP), sentiment analysis, and other social media and big data harvested and mined information. There are many other algorithms in place at the lower level which require validation to attempt to determine how good, realistic, or beneficial the data set is and if it will have a positive mathematical impact on the forecasting outputs, or cause additional error, misleading outcomes, or noise, which leads to Section #3.

In Section #3 and Section #4 the forecast output and results must be interpreted by the data scientist and compared, assessed, analyzed, and tested for real-world value. This could be as simple as testing different parameter outcomes as seen in the simulations in Chapter 4 and Chapter 5, or become complicated as seen in the example in Chapter 5 where the more accurate forecast was misleading due to the amount of forecasted data points. This is where the data scientist must look for visual traps, biases, and flaws and repeat the forecast to ensure there is nothing they did not expect to occur. If not, this could lead to a mathematically sound, but incorrect result that is not realized. The outputs also could have positive real-world results or

negative real-world results. This brings in an ethical judgment call for the data scientist and decision-maker to not just use the more attractive or positive output for their cause, pitch, stockholders, program managers, or other stakeholders they are conducting the forecast for. These items are listed in Section #3 where there may be very minor changes between forecasting results to major changes where interpretation and experience may have to decide if an upward trend or downward trend seems more realistic. Furthermore, it is possible that the outputs could be useless or inconclusive, which could signal the data scientist to return to the model selection stage, which would be the whole point of conducting validation in the first place.

## II. Summary & Future Work

In summary, the conclusions of this dissertation attempt to show that forecast modeling is very complex and the model or approach, can be impacted heavily by the data set and proper validation. When used appropriately many models may seem extremely accurate as proposed in the literature or in a certain industry when they in fact may be misleading or have inherent inaccuracies. Some of these inaccuracies may not be in the execution or math but are inaccurate regarding real-world results. This is why qualitative experience of an informed decision-maker or data scientist is invaluable, and why a framework to self-check and verify the model and data at each step of transformation as the data flows through the model should be conducted. This can be implemented as a business process or best practices in a large enterprise or for an individual researcher.

Neural network and machine-learning approaches attempt to mitigate these issues as best they can, however, these technologies can add multitudes of complexity since they use advanced computer science, biological, and mathematical sciences to create a deep learning model. In addition, it is difficult to understand the exact way the neural network learned and created a model based on the initial training. Many self-learning or self-improving models can change drastically over time which can cause significant differences in outputs of the same data set on a more developed model. This dissertation hopes to recognize some of these issues to create a better way for data scientists to conduct forecast modeling by use of the proposed forecast validation framework. Since newer approaches and more advanced data sets are being used in the future, this is an on-going and continuous validation process.

Future work on creating a more in-depth validation framework that includes validation processes for neural networks would be the obvious next step to this research. The limitations

posed by the Proposed Forecast Validation Framework are that although artificial neural networks are acknowledged, it needs more development to handle the relationship between the data set moving through the various steps and how it can change within the neural network. As it stands, it is up to the data scientist, decision-maker, or organization to create extra validation stages within the framework to ensure that the neural network or deep learning model is cohesively working with the forecasting model to increase accuracy and output a real-world result. Moreover, as the availability and size of data increases, each with its own individual algorithms, transformations, and ingestion methods, these models will require additional validation. Another future research investigation would be to simulate other data sets from other venture capital industries as well as gas price or crude oil price by the barrel. Following a stringent framework edited to fit the market, topic, or subject, the data scientist and decision-maker should be able to improve the quality and validations for their forecast results.

# Bibliography


1. Adhikari, R. (2015). A neural network based linear ensemble framework for time series forecasting. *Neurocomputing, 157*, 231-242. doi:https://doi.org/10.1016/j.neucom.2015.01.012
2. Adomavicius, G., Bockstedt, J. C., Gupta, A., & Kauffman, R. J. (2008, December). Making sense of technology trends in the information technology landscape: a design science approach. *MIS Quarterly, 32*(4), 779-809.
3. Ailon, N., Jaiswal, R., & Monteleoni, C. (2009). Streaming k-means approximation. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. Williams, & A. Culotta (Ed.), *NIPS'09 Proceedings of the 22nd International Conference on Neural Information Processing Systems* (pp. 10-18). Vancouver, British Columbia, Canada: Curran Associates Inc.
4. Aman, S. G. (2014). *Empirical comparison of prediction methods for electricity consumption forecasting.* University of Southern California. IEEE.
5. Ames, E., & Reiter, S. (1961, September). Distributions of correlation coefficients in economic time series. *Journal of the American Statistical Association, 56*(295), 637-656. Retrieved from https://www.jstor.org/stable/2282085
6. Armstrong, S. J. (2001). Evaluating forecasting methods. In J. S. Armstrong, & U. o. Pennsylvania (Ed.), *A Handbook for Researchers and Practitioners* (pp. 443-472). Springer. Retrieved from http://repository.upenn.edu/marketing_papers/146
7. Asur, S., & Huberman, B. A. (2010, November 1). Predicting the future with social media. *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. doi:10.1109/WI-IAT.2010.63
8. Atiya, A., El-Shoura, S., Shaheen, S., & El-Sherif, M. (1999, March). A comparison between neural-network forecasting techniques-case study: river flow forecasting. *IEEE Transactions on Neural Network, 10*(2), 402-409.
9. Azoff, E. M. (1994). *Neural Network Time Series Forecasting of Financial Markets.* New York, NY: John Wiley & Sons, Inc.
10. Bastardi, A., & Shafir, E. (1998). On the pursuit and misuse of useless information. *Journal of Personality and Social Psychology, 75*(1), 19-32. doi:0022-351
11. Bates, J. M., & Granger, W. J. (1969, December). The combination of forecasts. *Operational Research Society, 20*(4), 451-468. doi:10.2307/3008764
12. Bergmeir, C., J.Hyndman, R., & Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis, 120*, 70-83. doi:https://doi.org/10.1016/j.csda.2017.11.003

13. Bildosola, I., Gonzalez, P., & Moral, P. (2017). An approach for modelling and forecasting research. *Scientometrics, 112*, 557–572.

14. Brettel, M., Mauer, R., Engelen, A., & Küpper, D. (2011). Corporate effectuation: Entrepreneurial action and its impact on R&D project performance. *Journal of Business Venturing, 27*(2), 167-184. doi:https://doi.org/10.1016/j.jbusvent.2011.01.001

15. Brown, R. G. (1959). *Statistical forecasting for inventory control.* New York, New York, USA: McGraw-Hill. Retrieved from https://catalog.hathitrust.org/Record/001125212

16. Büyüksahin, Ü. Ç., & Ertekin, S. (2019, October 7). Improving forecasting accuracy of time series data using a new ARIMA-ANN hybrid method and empirical mode decomposition. *Neurocomputing, 361*, 151-163. doi:https://doi.org/10.1016/j.neucom.2019.05.099

17. Cao, Q., Ewing, B. T., & Thompson, M. A. (2012, August 16). Forecasting wind speed with recurrent neural networks. *European Journal of Operational Research, 221*(1), 148-154. doi:https://doi.org/10.1016/j.ejor.2012.02.042

18. Ceron, A., Curini, L., Iacus, S. M., & Porro, G. (2014, March 1). Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New Media & Society, 16*(2), 340-358. doi:https://doi.org/10.1177/1461444813480466

19. Chambers, J. C., Mullick, S. K., & Smith, D. D. (1971, November). How to choose the right forecasting technique. *Research Management, 49*, 45-71.

20. Chen, W., & Shi, K. (2019). A deep learning framework for time series classification using Relative Position Matrix and Convolutional Neural Network. *Neurocomputing, 359*, 384-394. doi:https://doi.org/10.1016/j.neucom.2019.06.032

21. Christensen, C. (1997). *The innovator's dilemma: when new technologies cause great firms to fail.* (H. B. Press, Ed.) Boston, MA: McGraw-Hill. Retrieved from https://www.hbs.edu/faculty/Pages/item.aspx?num=46

22. Cisek, D., & Mahajan, M. (2017, October 26). A transfer learning approach to parking lot classification in aerial imagery. *2017 New York Scientific Data Summit (NYSDS)*, 1-5. doi:10.1109/NYSDS.2017.8085049

23. Cisek, D., Dale, J., Pepper, S., Mahajan, M., & Yoo, S. (2016, November ). Parking lot delineation and object detection using a localized convolutional neural network. *2016 New York Scientific Data Summit (NYSDS)*, 1-5. doi:10.1109/NYSDS.2016.7747821

24. Coghlan, A. (2015). *A little book of R for time series* (Vol. Release 0.2). Cambridge, UK: Wellcome Trust Sanger Institute. Retrieved from http://www.calvin.edu/~stob/courses/m344/S15/a-little-book-of-r-for-time-series.pdf

25. Crone, S. F. (2010). *Forecasting time series forcasting competition*. Retrieved from Forecasting Competition for Artificial Neural Networks & Computational Intelligence: http://www.neural-forecasting-competition.com/NN5/index.htm

26. DARPA. (2018, June). *The Heilmeier Catechism*. Retrieved from DARPA: https://www.darpa.mil/work-with-us/heilmeier-catechism

27. de Jongh, P. J. (2017). A proposed best practice model validation framework for banks. *South African Journal of Economic and Management Sciences, 20*(1). doi:10.4102/sajems.v20i1.1490

28. de Treville, S., Petty, J. S., & Stefan, W. (2014). Economies of extremes: lessons from venture-capital decision making. *Journal of Operations Management, 32*(6), 387-398.

29. Derbyshire, J., & Giovannetti, E. (2017). Understanding the failure to understand new product development failures: mitigating the uncertainty associated with innovating new products by combining scenario planning and forecasting. *Technology Forecasting & Social Change, 125*, 334-344.

30. Dohnal, M., & Doubravsky, K. (2016). Equationless and equation-based trend models of prohibitively complex technological and related forecasts. *Technological Forecasting & Social Change, 111*, 297-304.

31. Dong, X., Qian, L., & Huang, L. (2017, February). Short-term load forecasting in smart grid: A combined CNN and K-means clustering approach. *2017 IEEE International Conference on Big Data and Smart Computing (BigComp)*, 119-125. doi:10.1109/BIGCOMP.2017.7881726

32. Eberhart, R., & Dobbins, R. (1990, September). Early neural network development history: the age of camelot. *IEEE Engineering in Medicine and Biology Magazine, 9*(3), 15-18. doi:10.1109/51.59207

33. Eliaz, K., & Schotter, A. (2010, February). Paying for confidence: an experimental study of the demand for non-instrumental information. *Games and Economic Behavior, 70*(2), 304-324.

34. El-Sharkawi, M. A., Marks II, R. J., & Weerasooriya, S. (1991). Neural networks and their application to power engineering. *Control and Dynamic Systems, 41*(1), 359-461. doi:https://doi.org/10.1016/B978-0-12-012741-2.50012-9

35. Franses, P. H. (1994). A method to select between gompertz and logistic trend curves. *Technology Forecasting and Social Change, 46*, 45-49.

36. Franses, P. H., & Haldrup, N. (1994, October). The effects of additive outliers on tests for unit roots and cointegration. *Journal of Business & Economic Statistics, 12*(4), 471-478. doi:10.2307/1392215

37. Frausto-Solís, J., Chi-Chim, M., & Sheremetov, L. (2015, February). Forecasting oil production time series with a population-based simulated annealing method. (S. S. B.V., Ed.) *Arabian Journal of Science and Engineering*. doi:10.1007/s13369-015-1587-z

38. Gardner, E. S. (1983). Automatic monitoring of forecast errors. *Journal of Forecasting, 2*, 1-21. doi:https://doi.org/10.1002/for.3980020103

39. Gentry, T. W., M. Wiliamowski, B., & Weatherford, L. (2002, September). Comparison of traditional forecasting techniques and neural networks. *Journal of Revenue and Pricing Management, 1*(4), 765-770. Retrieved from https://www.researchgate.net/publication/281317510_Comparison_of_traditional_forecasting_techniques_and_neural_networks

40. Gonzalez-Carrasco, I., & Garcia-Crespo, A. (2012, August 30). Towards a framework for multiple artificial neural network. *Expert Systems, 31*(1), 20-36. doi:10.1111/j.1468-0394.2012.00653.x

41. Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., & Xu, B. (2014). Generative adversarial nets. *NIPS'14 Proceedings of the 27th International Conference on Neural Information Processing Systems. 2*, pp. 2672-2680. Cambridge, MA, USA: MIT Press.

42. Goodwin, P. (2018). The holt-winters approach to exponential smoothing: 50 years old and going strong. *Foresight: The International Journal of Applied Forecasting*(19), 30-33.

43. Granger, C. W. (1969, August). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica, 37*(3), 424-438. doi:https://doi.org/10.2307/1912791

44. Griffith, E. (2013, October 15). *Steve Blank explains why accelerators should mimic "Moneyball*. Retrieved from Pando: https://pando.com/2013/10/15/steve-blank-explains-why-accelerators-should-mimic-moneyball/

45. Haseltine, C., & Eman, E. E.-S. (2017). Prediction of Power Grid Failure. *2017 16th IEEE International Conference on Machine Learning and Applications.* IEEE. doi:10.1109/ICMLA.2017.0-111

46. Hoenen, S., Kolympiris, C., & Schoenmakers, W. (2014). The Diminishing Signaling Value of Patents Between Early Rounds of Venture Capital Financing. *43*.

47. Holt, C. C. (2004). Forecasting seasonals and trends by exponentially weighted moving averages (1975). *International Journal of Forecasting, 20*, 5-10.

48. Huang, C.-J., & Kuo, P.-H. (2018, May 31). A deep CNN-LSTM model for particulate matter (PM2.5) forecasting in smart cities. *Sensors 2018, 18*. doi:10.3390/s18072220

49. Humphrey, G. B., Maier R., H. R., Wu, W., Mount, N. J., & Dandy, G. C. (2017, June). Improved validation framework and R-package for artificial neural network models.

*Environmental Modelling & Software, 92*, 82-106. doi:https://doi.org/10.1016/j.envsoft.2017.01.023

50. Hyndman, R. (2018, June). *Forecasting functions for time series and linear models*, 8.12. (R. Hyndman, Producer, & CRAN) Retrieved from CRAN: https://cran.r-project.org/web/packages/forecast/forecast.pdf

51. Kang, Z., Zhong, G., Dong, J., & Wang, S. (2019). Stock market prediction based on generative adversarial network. *Procedia Computer Science, 147*, 400-406. doi:https://doi.org/10.1016/j.procs.2019.01.256

52. Kyebambe, M. N., Cheng, G., Huang, Y., He, C., & Zhang, Z. (2017). Forecasting emerging technologies: A supervised learning approach through patent analysis. *Technological Forecasting & Social Change, 125*, 236-244.

53. Kyriazi, F., Thomakos, D. D., & B., G. J. (2019). Adaptive learning forecasting, with applications in forecasting agricultural prices. *International Journal of Forecasting, 35*(4), 1356-1369. doi:https://doi.org/10.1016/j.ijforecast.2019.03.031

54. Laha, S., & Roy, K. M. (2018). *Performance Comparison of Various Forecasting Techniques.* (B. i2i, Ed.) Retrieved 2019, from Bridge i2i: Whitepapers: https://bridgei2i.com/performance-comparison-of-various-forecasting-techniques/

55. Lasheras, F. S., Francisco, J. D., & Sanchez, A. S. (2015, September 15). Forecasting the COMEX copper spot price by means of neural networks. *Resources Policy, 45*, 37-43.

56. Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998, November). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278-2324. doi:10.1109/5.726791

57. Lemke, C., & Gabrys, B. (2010, March). Meta-learning for time series forecasting and forecast combination. *Neurocomputing, 73*(10-12), 2006-2016. doi:https://doi.org/10.1016/j.neucom.2009.09.020

58. Li, J., Xu, Z., Yu, L., & Tang, L. (2016). Forecasting oil price trends with sentiment of online news. *Information Technology and Quantitative Management, 91*, 1081 – 1087. doi:https://doi.org/10.1016/j.procs.2016.07.157

59. Li, M., Wong, W., & Leung, S. (2014). Comparison study on univariate forecasting techniques for apparel sales. *International Journal of Clothing Science and Technology*.

60. Lovallo, D., & Kaneman, D. (2003, July). Delusions of success: how optimism undermines executives' decisions. *Harvard Business Review, 81*(7), 56-63. Retrieved from https://hbr.org/2003/07/delusions-of-success-how-optimism-undermines-executives-decisions

61. Lu, C., Wu, H., & Vermuri, S. (1993, February). Neural network based short term load forecasting. *IEEE Transactions on Power Systems, 8*(1), 336-342.

62. Mehdiyev, N., Lahann, J., Emrich, A., & Enke, D. (2017). Time series classification using deep learning for process planning: a case from the process industry. *Procedia Computer Science, 114*, 242-249. doi:https://doi.org/10.1016/j.procs.2017.09.066

63. Methanex Corporation. (2018, June). *Pricing*. Retrieved from Methanex: https://www.methanex.com/our-business/pricing

64. Methanol Institute. (2019, May 2019). *The Methanol Institute*. Retrieved from The Methanol Industry: https://www.methanol.org/the-methanol-industry/

65. Methanol Institute. (2020, 3). *Methanol Price*. Retrieved from Methanol Institute: https://www.methanol.org/methanol-price-supply-demand/

66. Murakami, K., & Taguchi, H. (1991, March). Gesture recognition using recurrent neural networks. *CHI '91: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 237–242. doi:10.1145/108844.108900

67. *National Venture Capital Association*. (2018, June). Retrieved from NVCA: https://nvca.org/

68. Nguyen, H. T., & Le, H. M. (2012). Modified feed-forward neural network structures and combined-function-derivative approximations Incorporating exchange symmetry for potential energy surface fitting. *The Journal of Physical Chemistry, 116*, 4629-4638. doi:dx.doi.org/10.1021/jp3020386

69. Nguyen, T. H., Shirai, K., & Velcin, J. (2015, December 30). Sentiment analysis on social media for stock movement prediction. (D. B. Lin, Ed.) *Expert Systems with Applications, 42*(24), 9603-9611. doi:https://doi.org/10.1016/j.eswa.2015.07.052

70. Ning, G., Zhang, Z., Huang, C., Ren, X., Wang, H., Cai, C., & He, Z. (2017). Spatially supervised recurrent convolutional neural networks for visual object tracking. In IEEE (Ed.), *2017 IEEE International Symposium on Circuits and Systems (ISCAS)* (pp. 1-4). Baltimore, MD: IEEE. doi:10.1109/ISCAS.2017.8050867

71. NIST. (2015, November). *Exploratory Data Analysis*. (NIST, Editor, & NIST, Producer) Retrieved November 2019, from Engineering Statistics Handbook: https://www.itl.nist.gov/div898/handbook/eda/section1/eda121.htm

72. NIST. (2018). *Forecasting with Single Exponential Smoothing*. Retrieved from Engineering Statistics: https://www.itl.nist.gov/div898/handbook/pmc/section4/pmc432.htm

73. Ortiz, M., Ukar, O., Azevedo, F., & Múgica, A. (2016, May). Price forecasting and validation in the Spanish electricity market using. *Electrical Power and Energy System, 77*, 123-127. doi:doi.org/10.1016/j.ijepes.2015.11.004

74. Pitchbook. (2018, April 25). *The 12 most active VC investors in US cleantech*. Retrieved from Pitchbook: News & Analysis: https://pitchbook.com/news/articles/clean-tech

75. Porter, M. E. (1985). *Competitive Advantage: Creating and Sustaining Superior Performance.* New York: Free Press.

76. Poznyak, T. I., Oria, I. C., & Poznyak, A. S. (2019). *Ozonation and biodegradation in environmental engineering: dynamic neural network approach.* Cambridge, MA, USA: Elsevier. doi:https://doi.org/10.1016/C2016-0-03865-2

77. Puskorius, G., & Feldkamp, L. (1994, March). Neurocontrol of nonlinear dynamical systems with Kalman filter trained recurrent networks. *IEEE Transactions on Neural Networks, 5*(2), 279-297.

78. R-core@R-project.org, R.-c. (2019). *acf function | R Documentation*. Retrieved from R Documentation: https://www.rdocumentation.org/packages/stats/versions/3.6.0/topics/acf

79. Rojas, R. (1996). *Neural Networks - A Systematic Introduction.* Berlin: Springer-Verlag. Retrieved from https://www.amazon.com/Neural-Networks-Introduction-Raul-Rojas/dp/3540605053

80. Rosenblatt, F. F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review, 65*(6), 386-408.

81. R-Project.org. (2018). *Time-Series Objects*. Retrieved from RDocumentation: https://www.rdocumentation.org/packages/stats/versions/3.5.1/topics/ts

82. Sang, C., & Massimo, D. P. (2019, March). Improving trading technical analysis with TensorFlow Long Short-Term Memory (LSTM) Neural Network. (KeAi, Ed.) *The Journal of Finance and Data Science, 5*(1), 1-11. doi:https://doi.org/10.1016/j.jfds.2018.10.003

83. Sharma, R. (2014). The ever-emerging markets: why economic forecasts fail. *Foreign Affairs, 93*(1), 52-56.

84. Shuntaro, T., Yu, C., & Kumiko, T.-I. (2019, April). Modeling financial time-series with generative adversarial networks. *Physica A: Statistical Mechanics and its Applications, 527*. doi:https://doi.org/10.1016/j.physa.2019.121261

85. Smith, C., & Jin, Y. (2014). Evolutionary multi-objective generation of recurrent neural network ensembles for time series prediction. *Neurocomputing, 143*, 302-311. doi:https://doi.org/10.1016/j.neucom.2014.05.062

86. Stone, M. (1974). Cross-validatory choice and assessment of statistical prediction (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological), 36*(2), 111-147. Retrieved from https://www.stat.washington.edu/courses/stat527/s14/readings/Stone1974.pdf

87. Sun, W., & Trevor, B. (2018, June). A stacking ensemble learning framework for annual river ice breakup da. *Journal of Hydrology, 561*, 636-650. doi:https://doi.org/10.1016/j.jhydrol.2018.04.008

88. Taleb, N. N. (2007). *The Black Swan: The Impact of the Highly Improbable.* New York, NY, USA: The Random House Publishing Group.

89. Taylor, J. W. (2009). Triple seasonal methods for short-term electricity demand forecasting. *European Journal of Operational Research, 204*(1), 139-152. doi:https://doi.org/10.1016/j.ejor.2009.10.003

90. Tealab, A. (2018, December). Time series forecasting using artificial neural networks methodologies: a systematic review. *Future Computing and Informatics Journal, 3*(2), 334-340. doi:ttps://doi.org/10.1016/j.fcij.2018.10.003

91. Tukey, J. W. (1962). The Future of Data Analysis. *The Annals of Mathematical Statistics, 33*(1), 1-67.

92. Valente, T. W. (1996, January). Social network thresholds in the diffusion of innovations. *Social Networks, 18*(1), 69-89.

93. Wang, L., Wang, Z., & Liu, S. (2018). Optimal forecast combination based on neural networks for time series forecasting. *Applied Soft Computing, 66*, 1-17. doi:https://doi.org/10.1016/j.asoc.2018.02.004

94. Wang, M., Zhaoc, L., Duc, R., Wang, C., Chen, L., & Tiana, L. (2018, March). A novel hybrid method of forecasting crude oil prices using complex network science and artificial intelligence algorithms. *Applied Energy, 220*(15), 480-495.

95. Wang, Y., & Hongguang, L. (2018, August). A novel intelligent modeling framework integrating convolutional neural network with an adaptive time-series window and its application to industrial process operational optimization. *Chemometrics and Intelligent Laboratory Systems, 179*, 64-72. doi:https://doi.org/10.1016/j.chemolab.2018.06.008

96. Wikipedia. (2019, November 3). *Steve Blank*. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Steve_Blank

97. Winters, P. R. (1960, April). Forecasting sales by exponentially weighted moving averages. *Management Science, 6*(3), 324-342.

98. Woike, J. K., Hoffrage, U., & Petty, J. S. (2014). Picking profitable investments: the success of equal weighting in simulated venture capitalist decision making. *68*(8). doi:https://doi.org/10.1016/j.jbusres.2015.03.030

99. Yu, S.-S., Chu, S.-W., Chan, Y.-K., & Wang, C.-M. (2019). Share Price Trend Prediction Using CRNN with LSTM Structure. *Smart Science, 7*(3). doi:10.1080/23080477.2019.1605474

100. Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing, 50*, 159-175.

101. Zhang, G., & Tang, C. (2017). How could firm's internal R & D collaboration bring more innovation? *Technology Forecasting & Social Change*, 399-308.

102. Zhang, G., Hu, M. Y., & Patuwo, E. B. (1999). Artificial neural networks in bankruptcy prediction: general framework and cross-validation analysis. *European Journal of Operational Research, 116*(1), 16-32. doi:https://doi.org/10.1016/S0377-2217(98)00051-4

103. Zhang, X., & Wang, J. (2018, April). A novel decomposition-ensemble model for forecasting short-term load-time series with multiple seasonal patterns. *Applied Soft Computing, 65*, 478-494. doi:https://doi.org/10.1016/j.asoc.2018.01.017

104. Zhao, L.-T., Wang, Y., Guoa, S.-Q., & Zeng, G.-R. (2018). A novel method based on numerical fitting for oil price trend forecasting. *Applied Energy, 220*(15), 154-163. doi:https://doi.org/10.1016/j.apenergy.2018.03.060

# Appendix I: R Code for VC Energy Simulation

```
#Comment: Load Forecast Package
library(forecast)
par(mfrow=c(2,2))

#Comment: Load Main Data Table for ALL VC Data

v = read.table("c:/Manoj/95_2015_TSVCquarterlyrPercent.csv", header=TRUE, sep=",")
y = (v$YearQ / 100) / 1000
e = v$EnergyIndust

TSVCquarterly <- ts(v[,"EnergyIndust"], start=c(1995,1),frequency=4)
TSVCquarterly

#Comment: Start Forecast for 4 Future Data Points
future4Q = forecast(TSVCquarterly, 12)
#Comment: HoltWinters Normalization/Smoothing Filtered Data
HWEnergyVC= HoltWinters(TSVCquarterly, beta=FALSE, gamma=FALSE)
#Comment: Display HoltWinters Filtered Data
# HWMethanol
plot(HWEnergyVC)
dev.new()
#Comment: Plot Forecast for Data Set
plot(future4Q, ylab=" VC Energy Investment in Millions USD ($)", xlab="Year", main =
"Forecast for next 3 years *Default")
dev.new()
#Comment: Plot Default HoltWinters Data
plot(HWEnergyVC)

#Comment: Plot Forecast for Holt-Winters Normalized Data Set
dev.new()
plot(forecast(HWEnergyVC,12), main = "Forecast Holtwinters for the next 3
years*Default",ylab="VC Energy Investment in Millions USD ($)", xlab="Year")

#Comment: Plot ETS/HW (Error: Multi, Trend: Auto, Season: None)
dev.new()
hwForecastMAN <- forecast(ets(TSVCquarterly, model="MAN", damped = FALSE), h=12)
plot(hwForecastMAN, main = "Forecast HoltWinters (MAN) for next 3 Years", sub="Error:
Multi, Trend: Additive, Season: None", ylab=" VC Energy Investment in Millions USD ($)",
xlab="Year")

dev.new()
#Comment: Plot ETS/HW (Error: Auto, Trend: Additive, Season: Multi)
hwForecastZAM <- forecast(ets(TSVCquarterly, model="ZAM", damped = FALSE), h=12)
```

```
plot(hwForecastZAM, main = "Forecast HoltWinters (ZAM) for next 3 Years", sub="Error:
Auto, Trend: Additive, Season: Multi", ylab=" VC Energy Investment in Millions USD ($)",
xlab="Year")

#Comment: Decompose Methanol Data set for ACF and Confirm Seasonality
dev.new()
EnergyDECOMP <- decompose(TSVCquarterly)
plot (EnergyDECOMP)

#Comment: Function for Bar Graph with Forecast Errors
plotForecastErrors <- function(forecasterrors, title)
  {
    # make a histogram of the forecast errors:
    mybinsize <- IQR(forecasterrors)/4
    mysd   <- sd(forecasterrors)
    mymin  <- min(forecasterrors) - mysd*5
    mymax  <- max(forecasterrors) + mysd*3
    # generate normally distributed data with mean 0 and standard deviation mysd
    mynorm <- rnorm(10000, mean=0, sd=mysd)
    mymin2 <- min(mynorm)
    mymax2 <- max(mynorm)
    if (mymin2 < mymin) { mymin <- mymin2 }
    if (mymax2 > mymax) { mymax <- mymax2 }
    # make a red histogram of the forecast errors, with the normally distributed data overlaid:
    mybins <- seq(mymin, mymax, mybinsize)
    hist(forecasterrors, col="red", freq=FALSE, breaks=mybins, main=title)
    # freq=FALSE ensures the area under the histogram = 1
    # generate normally distributed data with mean 0 and standard deviation mysd
    myhist <- hist(mynorm, plot=FALSE, breaks=mybins)
    # plot the normal curve as a blue line on top of the histogram of forecast errors:
    points(myhist$mids, myhist$density, type="l", col="blue", lwd=2)
  }

#Comment: Holt-Winters for ACF (Residuals)
HWEnergy12 <- forecast.HoltWinters(HWEnergyVC,12)
dev.new()
acf(HWEnergy12$residuals, lag.max=50, main="Holt-Winters Default Energy VC Residuals
ACF (Lag=50)")
dev.new()
plot.ts(HWEnergy12$residuals, main="Residuals for Holt-Winters Default Energy VC Residuals
(Quarterly)")

#Comment: Residual Plots - MAN
dev.new()
acf(hwForecastMAN$residuals, lag.max=50, main="Forecast (MAN) VC Energy Residuals
ACF (Lag=50)")
```

```
dev.new()
plot.ts(hwForecastMAN$residuals, main="Forecast (MAN) Energy VC Residuals (Quarterly)")

#Comment: Residual Plots - ZAM
dev.new()
acf(hwForecastZAM$residuals, lag.max=50, main="Forecast (ZAM) VC Energy Residuals ACF
(Lag=50)")
dev.new()
plot.ts(hwForecastZAM$residuals, main="Forecast (ZAM) VC Energy Residuals (Quarterly)")

#Comment: Plot the Histograms
dev.new()
plotForecastErrors(HWEnergy12$residuals, title="Histogram of Holt-Winters Errors")

#Comment: Histogram for MAN
dev.new()
plotForecastErrors(hwForecastMAN$residuals, title="Histogram of Forecast Errors (MAN)")

#Comment: Histogram for ZAM
dev.new()
plotForecastErrors(hwForecastZAM$residuals, title="Histogram of Forecast Errors (ZAM)")
```

# Appendix II: R Code used to Generate Methanol Simulation

```
#Comment: Load Forecast Package
library(forecast)

#Comment: Load Main Data Table for Methanol Data from Methanex
v = read.table("c:/Manoj/methanol-yyyy-mm.csv", header=TRUE, sep=",")
y = v$DATE
e = v$PRICE

methanolPRICE <- ts(v[,"PRICE"], start=c(2001,5),frequency=12)
methanolPRICE

#Comment: Start Forecast for 12 Future Data Points
FCTwelve = forecast(methanolPRICE, 12)

#Comment: HoltWinters Normalization/Smoothing Filtered Data
HWMethanol = HoltWinters(methanolPRICE, beta=FALSE, gamma=FALSE)
#Comment: Display HoltWinters Filtered Data
#HWMethanol

dev.new()
#Comment: Plot Forecast for Data Set
plot(FCTwelve, ylab="Methanol Price per Gallon in USD ($)", xlab="Year", main = "Forecast
for next 12 Months *Default")

dev.new()
#Comment: Plot Default HoltWinters Data
plot(HWMethanol)

#Comment: Plot Forecast for Holt-Winters Normalized Data Set
dev.new()
plot(forecast(HWMethanol,12), main = "Forecast Holtwinters *Default",ylab="Methanol Price
per Gallon in USD ($)", xlab="Year")

#Comment: Plot ETS/HW (Error: Multi, Trend: Auto, Season: None)
dev.new()
hwForecastMAN <- forecast(ets(methanolPRICE, model="MAN", damped = FALSE), h=12)
plot(hwForecastMAN, main = "Forecast HoltWinters (MAN) for next 12 Months", sub="Error:
Multi, Trend: Additive, Season: None", ylab="Methanol Price per Gallon USD ($)",
xlab="Year")

dev.new()
#Comment: Plot ETS/HW (Error: Auto, Trend: Additive, Season: Multi)
hwForecastZAM <- forecast(ets(methanolPRICE, model="ZAM", damped = FALSE), h=12)
```

```
plot(hwForecastZAM, main = "Forecast HoltWinters (ZAM) for next 12 months", sub="Error:
Auto, Trend: Additive, Season: Multi", ylab="Methanol Price per Gallon USD ($)",
xlab="Year")

#Comment: Decompose Methanol Data set for ACF and Confirm Seasonality
dev.new()
methanolDECOMP <- decompose(methanolPRICE)
plot (methanolDECOMP)

#Comment: Function for Bar Graph with Forecast Errors
plotForecastErrors <- function(forecasterrors, title)
  {
    # make a histogram of the forecast errors:
    mybinsize <- IQR(forecasterrors)/4
    mysd   <- sd(forecasterrors)
    mymin  <- min(forecasterrors) - mysd*5
    mymax  <- max(forecasterrors) + mysd*3
    # generate normally distributed data with mean 0 and standard deviation mysd
    mynorm <- rnorm(10000, mean=0, sd=mysd)
    mymin2 <- min(mynorm)
    mymax2 <- max(mynorm)
    if (mymin2 < mymin) { mymin <- mymin2 }
    if (mymax2 > mymax) { mymax <- mymax2 }
    # make a red histogram of the forecast errors, with the normally distributed data overlaid:
    mybins <- seq(mymin, mymax, mybinsize)
    hist(forecasterrors, col="red", freq=FALSE, breaks=mybins, main=title)
    # freq=FALSE ensures the area under the histogram = 1
    # generate normally distributed data with mean 0 and standard deviation mysd
    myhist <- hist(mynorm, plot=FALSE, breaks=mybins)
    # plot the normal curve as a blue line on top of the histogram of forecast errors:
    points(myhist$mids, myhist$density, type="l", col="blue", lwd=2)
  }

#Comment: Holt-Winters for ACF (Residuals)
HW12Methanol <- forecast.HoltWinters(HWMethanol, h=12)
dev.new()
acf(HW12Methanol$residuals, lag.max=50, main="Holt-Winters Methanol Residuals ACF
(Lag=50)")
dev.new()
plot.ts(HW12Methanol$residuals, main="Residuals for Holt-Winters 12 Month")

#Comment: Residual Plots - MAN
dev.new()
acf(hwForecastMAN$residuals, lag.max=50, main="Forecast (MAN) Methanol Residuals ACF
(Lag=50)")
dev.new()
```

```
plot.ts(hwForecastMAN$residuals, main="Forecast (MAN) Methanol Residuals 12 Month")

#Comment: Residual Plots - ZAM
dev.new()
acf(hwForecastZAM$residuals, lag.max=50, main="Forecast (ZAM) Methanol Residuals ACF
(Lag=50)")
dev.new()
plot.ts(hwForecastZAM$residuals, main="Forecast (ZAM) Methanol Residuals 12 Month")

#Comment: Plot the Histograms
dev.new()
plotForecastErrors(HW12Methanol$residuals, title="Histogram of Holt-Winters Errors")

#Comment: Histogram for MAN
dev.new()
plotForecastErrors(hwForecastMAN$residuals, title="Histogram of Forecast Errors (MAN)")

#Comment: Histogram for ZAM
dev.new()
plotForecastErrors(hwForecastZAM$residuals, title="Histogram of Forecast Errors (ZAM)")
```

**Code used for Comparisons (Real Data vs. Forecast Data)**

```
#Comment: Load Forecast Package
library(forecast)
#Comment: Load Main Data Table for Methanol Data from Methanex May 2001 - May 2012
v = read.table("c:/Manoj/methanol-02-12.csv", header=TRUE, sep=",")
methanolPRICE2012 <- ts(v[,"PRICE"], start=c(2002,1),frequency=12)
methanolPRICE2012
#plot to check data
#plot(methanolPRICE2012)

dev.new()
#Comment: Plot ETS/HW (Error: Auto, Trend: Additive, Season: Multi)
hwForecastZAM <- forecast(ets(methanolPRICE2012, model="ZAM", damped = FALSE),
h=12)
plot(hwForecastZAM, main = "Forecast HoltWinters (ZAM) for 2013", sub="Blue Line =
Forecasted Data against Real Data", ylab="Methanol Price per Gallon USD ($)", xlab="Year")

#Comment: Load Main Data Table for Methanol Data from January 2002 December 2013

v2013 = read.table("c:/Manoj/methanol-02-13.csv", header=TRUE, sep=",")
y2013 = v$DATE
e2013 = v$PRICE

methanolPRICE2013 <- ts(v2013[,"PRICE"], start=c(2002,1),frequency=12)
methanolPRICE2013
par(new=TRUE)
lines(methanolPRICE2013)


#---------


#Comment: Load Main Data Table for Methanol Data from Methanex
fullMData = read.table("c:/Manoj/methanol-yyyy-mm.csv", header=TRUE, sep=",")

methanolPRICE <- ts(fullMData [,"PRICE"], start=c(2001,5),frequency=12)
methanolPRICE

dev.new()
#Comment: Plot ETS/HW (Error: Auto, Trend: Additive, Season: Multi)
hwForecastZAM <- forecast(ets(methanolPRICE, model="ZAM", damped = FALSE), h=12)
plot(hwForecastZAM, main = "Forecast HoltWinters (ZAM) for next 12 months", sub="Error:
Auto, Trend: Additive, Season: Multi", ylab="Methanol Price per Gallon USD ($)",
xlab="Year")
```

```
#----------


#Comment: Load Forecast Package
library(forecast)
#Comment: Load Main Data Table for Methanol Data from Methanex May 2001 - May 2018
v = read.table("c:/Manoj/methanol-02-12.csv", header=TRUE, sep=",")
methanolPRICE2012 <- ts(v[,"PRICE"], start=c(2002,1),frequency=12)
methanolPRICE2012
#plot to check data
#plot(methanolPRICE)

dev.new()
#Comment: Plot ETS/HW (Error: Auto, Trend: Additive, Season: Multi)
hwForecastZAM <- forecast(ets(methanolPRICE2012, model="ZAM", damped = FALSE),
h=65)
plot(hwForecastZAM, main = "Forecast HoltWinters (ZAM) from January 2013 to May 2018",
sub="Blue Line = Forecasted Data against Real Data", ylab="Methanol Price per Gallon USD
($)", xlab="Year")

#Comment: Load Main Data Table for Methanol Data from January 2002 December 2013

v2013 = read.table("c:/Manoj/methanol-02-13.csv", header=TRUE, sep=",")
y2013 = v$DATE
e2013 = v$PRICE

methanolPRICE2013 <- ts(v2013[,"PRICE"], start=c(2002,1),frequency=12)
methanolPRICE2013
par(new=TRUE)
lines(methanolPRICE)
```

Code used for **ForecastErrors** Function modified from Avril Coglan's plotForecastErrors code (Coghlan, 2015).

This function is designed to display the Forecast Errors in a histogram that can be used as a tool for analysis. This will allow for discovering if the forecast errors are normally distributed.

```
plotForecastErrors <- function(forecasterrors, title)
 {
    #Comment: make a histogram of the forecast errors for data scientist to interpret:
    mybinsize <- IQR(forecasterrors)/4
    mysd   <- sd(forecasterrors)
    mymin  <- min(forecasterrors) - mysd*5
    mymax  <- max(forecasterrors) + mysd*3
    # Comment: generate normally distributed data with mean 0 and standard deviation mysd
    mynorm <- rnorm(10000, mean=0, sd=mysd)
    mymin2 <- min(mynorm)
    mymax2 <- max(mynorm)
    if (mymin2 < mymin) { mymin <- mymin2 }
    if (mymax2 > mymax) { mymax <- mymax2 }
    #Comment: make a histogram, in red, of the forecast errors, with the normally distributed data
overlaid:
    mybins <- seq(mymin, mymax, mybinsize)
    hist(forecasterrors, col="red", freq=FALSE, breaks=mybins, main=title)
    # freq=FALSE ensures the area under the histogram = 1
```

```
  # generate normally distributed data with mean 0 and standard deviation mysd
  myhist <- hist(mynorm, plot=FALSE, breaks=mybins)
  #Comment:  plot the normal curve as a blue line on top of the histogram of forecast errors:
  points(myhist$mids, myhist$density, type="l", col="blue", lwd=2)
}
```