Forward Thinking Structures: Meta-Modelling Supply Chain Predictive Analytics for Real-World Disruptions

Ava Reilly Beckman

Trinity Washington University

I have adhered to the university policy regarding academic honesty in completing this assignment

Submitted to Dr. Kelley Wood on behalf of the faculty of the School of Business and Graduate Studies in partial fulfillment of the degree requirements for the Masters of Science in Administration in Federal Program Management

Spring 2015

Acknowledgements

First and foremost, I would like to express my most sincere gratitude to Dr. Kelley Wood and to Trinity Washington University for the opportunity to write this capstone. Dr. Wood's guidance, patience, and enthusiasm was pivotal in my work; and in helping me develop a theoretical model, and find the broader context of this work as a management tool, not simply a data modelling exercise.

I also wish to deeply thank my colleagues at Booz Allen Hamilton and at the Partnership for Supply Chain Management (PFSCM) for the professional opportunity to work within the fields that made this research possible. Special thanks must be given to Nathan Vasher and Charles Davenport at PFSCM for the remarkable quality of thought they dedicate to predictive analytics within the supply chain management context on a daily basis. All of my work grew on the foundation that they and many other researchers have built.

I would be remiss in not thanking my parents, brother, and friends for their patience and encouragement throughout the completion of this degree. A few sentences on an acknowledgement page are hardly adequate recompense for your constant love and support, but know that I remain ever grateful.

Lastly, I should also thank the traffic patterns on Interstate 95 and 395 for the non-optional thinking time they include in my day. In the absence of this egregious gridlock, this would have been forty pages shorter.

Abstract

Supply chain managers face large amounts of uncertainty at both the operational and strategic levels of supply chain management. A large part of this uncertainty stems from an inability to access real-time data and see broader historical trends when making decisions. Predictive analytics models can offer insight into both of these kinds of data while also generating statistically tested supply chain predictions that mitigate pieces of this uncertainty. Designing a predictive analytics model that addresses the nuance of each specific supply chain and associated data collection, storage, and retrieval mechanisms is a complex process. On order to address this need while maintaining the highest level of generalizability across the field of supply chain management, this research designs a meta-model, a model of the model, to cope with uncertainty. This meta-model offers insight into the predicted performance of the orders and shipments occurring within a supply chain via statistical computational algorithms. Additionally, it is designed to provide the necessary flexibility for managers to respond to actual and perceived disruptions in the supply chain in real-time. The meta-model is structured around seven stages: (1) purpose of the model, (2) categorical drivers, (3) variance testing, (4) data cleaning, (5) process time calculations, (6) prediction generation, and (7) final model usage. In each meta-model stage, key decision points are presented and the determined resolution provides the individual structure of the resultant predictive model.

Keywords: supply chain, supply chain management, predictive analytics, proactive management tools, meta-modelling, supply chain analytics, performance management

Table of Contents

Page

Introduction 6
Statement of the Problem 6
Purpose of the Study 7
Significance of the Study 7
Theory 7
Research Method 8
Delimitations 8
Limitations 9
Summary 9
Literature Review 10
Supply Chain Uncertainty 11
Predictive Analytics 16
Theoretical Framework 18
Summary 21
Research Method 24
Research Questions 25
Data Source 27
Ethical Considerations 31
Research Strategy 32
Summary 34
System Design and Analysis 36
Categorical Drivers 36
Variance Testing 42
Data Cleaning 53
Process Time Calculations 55
Prediction Generation 72
Summary 76
Discussion 79
Research Questions 80

Conclusions 85
Recommendations and Implications 88
Summary 90
References 92

List of Figures

Page

Figure 1. Categorization of supply chain uncertainty research. 13
Figure 2. Two-by-two matrix of prediction and explanation. 17
Figure 3. Supply chain meta-model decision design as a theoretical framework. 19
Figure 4. Theoretical model of the interaction of categorical variables per process step. 21
Figure 5. Stages of research design. 33
Figure 6. Illustrative matrix of completed independent categorical driver variance testing 50
Figure 7. Illustrative matrix of independent categorical driver variance reduction and breadth... 51
Figure 8. Frequency of process times in days for process x. 54
Figure 9. Matrix of AD/UD solutions based on storage mechanism and process design. 58
Figure 10. Illustration of predictions made through simple exponential smoothing 66
Figure 11. Illustration of predictions made through Holt's trend-corrected exponential smoothing 67

Introduction

How does one drive a car while looking only through the rear-view mirror? It is certainty possible, but general consensus would show that it is probably not the safest way to arrive at work. And yet, this perilous position is one in which many supply chain managers, both in private industry and in government, find themselves. Future decisions are made based on past experience, or reflecting on performance after it occurred. This position of being able to react to supply chain performance and events after the fact as opposed to foreseeing and potentially mitigating future problems is problematic due to the importance of supply chains to the global economy. Given that supply chains control critical aspects of our society, from moving food and water from rural farms to urban centers, to the manufacturing and delivery of essential medicines across the world, reactive management structures present concerns to the dynamic environment of supply chain management.

Statement of the Problem

Supply chain management is fraught with uncertainty. Managers often consider the performance of a supply chain retroactively, through looking at performance after it has already occurred. In response to poor performance, root cause analysis must be performance and corrective actions implemented. In the event of high-quality analysis and a culture responsive to corrective action, performance issues can be mitigated in a timely manner. In the absence of these factors, profound issues can linger and tarnish the efficacy of a supply chain. However, even when these factors are present, they do not address the impact of poor performance in real-time. This presents a problem.

Not only are supply chain managers not able to recognize performance problems in advance, but they are often not able to recognize performance problems as they occur. It is only after the presentation of monthly or quarterly performance reports that these circumstances come

to light for managers to take action. This also presents the converse problem that good performance cannot be recognized in real-time, and when it can be recognized after the fact, it is often eclipsed by discussions of poor performance. This hinders the understanding of best practice successes and slows them from spreading to other less mature aspects of the supply chain.

Purpose of the Study

This study examines that deficit of real-time and future insight of supply chain uncertainty. It then responds with a meta-model of predictive analytics and a discussion of the process of implementing a predictive analytics model that is customizable to any supply chain. The design and discussion surrounding this model offer a new framework through which to view supply chain data.

Significance of the Study

The impact of this research is the framework for new supply chain management structures. Through the utilization of this model, supply chain managers would metaphorically be able to drive their car while looking through their front windshield, as one generally recommends. The implementation of a predictive system based on historical data provides for an alerts-based management tool that allows supply chain managers to be aware of and respond to issues in the supply chain as they arise or are forecasted to arise. Through having this knowledge in advance, they would be able to either mitigate the problem and improve performance, or begin to manage the expectations of the client and thereby improve customer satisfaction.

Theory

This theory blends a meta-analysis of supply chain uncertainty as conducted by Simangunsong, Hendry, and Stevenson (2012) as well as several other authors, with the newer field of predictive analytics (Mayer-Schönberger & Cukier, 2013; Waller & Fawcett, 2013). The

purpose of this co-mingling is to propose predictive analytics as a new way to respond to supply chain uncertainty. Within this school of predictive analytic thought, the academic paradigm is reversed. The data collected is simply too big to apply these traditional modes of thought and academic formulism. Rather than use a theoretical framework to develop a hypothesis which is then tested, predictive analytics favors a big data approach that applies statistical testing to data to glean insight.

Research Method

This mixed methods study embraces both a qualitative and quantitative approach in turn. It is qualitative in terms of the research questions that structure this study and in the phased data collection and analysis. However, the answers to these questions and the phases of analysis are designed to be quantitative. This mixing of method echoes the significance of this theory in creating new, forward-thinking management structures.

Predictive analytics allows managers to make data-information decisions in real-time. Decision making itself is a qualitative process, while the phrase "data-driven" implies quantitative rigor. This method of answering qualitative questions with quantitative rigor is both the end-goal and the initial, driving force behind the research methodology developed for this meta-modelling.

Delimitations

This theory is capable of creating a framework for organizations to implement a predictive analytics system in their supply chain. This meta-model and associated discussion construct and test a model from the base components of building a data set to the manipulation of this model to ascribe it enough flexibility to process atypical world events. A model with this historical, data-driven core, combined with a flexible qualitative understanding of a supply chain and of world events allows managers to see a clearer picture of supply chain performance in real-

time as well as looking into the future. This theory could additionally support other predictive quantification efforts within the field of supply chain management.

Limitations

If supply chains do not possess a centralized source of data to conduct analysis or lack reliable technologies to share the insight gleaned from this model, this theory will not be able to offer a direct path to creating a forward-thinking management structures. A level of quantitative rigor and business intelligence technology are required to implement this kind of algorithm across a large data set. By necessity, these assumptions require a level of maturity from a supply chain that may or may not yet be present in organizations that seek to implement a forward thinking management structure. In those circumstances, this theory might be used to understand the requirements of future technology acquisitions. Additionally, it might be used to model an aspirational state of technology and of supply chain management.

Summary

This theory is designed to highlight power of predictive analytics in response to supply chain uncertainty. It possesses the ultimate goal of creating forward-thinking management structures which allow those at the helm of supply chains around the world to make decisions on real-time and predicted performance with a level of confidence provided by a statistically rigorous and sufficiently flexible predictive model. To this end, it creates a model of what such a model would look like, called a meta-model. This research will be mixed methods, aiming to answer qualitative questions with quantitative rigor.

Literature Review

When a country sends military personnel to a foreign country, there is a supply chain to deliver their provisions and supplies across oceans and through hostile territory. When a government offers much-needed food and medical supplies in the wake of a disaster to another country, there is a supply chain to coordinate the delivery of those relief supplies into places where the roads may no longer be passable and the warehouses no longer standing. When one takes a trip to the neighborhood convenience store to buy a jar of apple sauce, there is a supply chain that starts at the seed bank and travels through the orchard that grows the apples, to the factory that processes the apples into apple sauce, to the truck that moves it to the grocery store shelf from which it travels its last mile home to a pantry, fridge, or lunch box. Every physical thing that can be purchased arrives in the hands of its new owner through supply chains either great or small.

In this global economy, supply chains have the ability to become exceedingly complex very quickly. A batik fabric produced in a Hindi-speaking factory India is hand-dyed from natural dies imported from South America, with patterns drawn in wax brought in from West Africa, on cotton bought from China. It is wrapped on bolts sourced from cardboard made at a recycling plant over 80 miles away in Punjab, and packed in shipping crates and sent via freight to the United States. From there, it is put on a truck and stored in a warehouse in Utah until it is used to fulfill a bulk order to a fabric shop where it will then be sold by the cut to individuals.

This hypothetical supply chain crosses five countries, multiple continents, and speaks at least six languages. An issue in West Africa might mean the wax delivery will not arrive and that the fabric cannot be dyed. Does one elect to dye solids to sell later at discount instead of batiks in an effort to maintain normal factory production as well as to not overstock the very small storage warehouse with loose cotton (a fire hazard) since there is little market for bare

fabrics for this company locally? Does one find an alternative wax procurement source in China? Does one simply wait for the West African shipment to arrive and use this forced downtime to perform maintenance and safety checks on the factory machines? That depends; how long will it take and is it an option to hold back the impending cotton shipment? There is a lot of room for uncertainty in a system with so many moving parts. Because this uncertainty exists in the backbone of private industry and of the public sector, it is surrounded by a high level of concern.

Supply Chain Uncertainty

Supply chain uncertainty is a problem for all managers (Hult, Craighead, & Ketchen, 2010). Davis (1993) identified the three principle causes of supply chain uncertainty: (1) demand, (2) manufacturing process, and (3) supply uncertainty. Davis became one of the first to argue that this uncertainty was critical to study given that the disruption of the complex network of a supply chain can have huge impacts on the economy, on public health such as with the annual distribution of the influenza vaccine, and on personal safety such as with the quality control of critical automotive parts. Bhatnagar and Sohal (2005) attribute this uncertainty to the increasing complexity of supply-chain networks in a global economy, which offer an increased potential for quality concerns and delivery delays. However, an increased concern in the contemporary geopolitical landscape, when supply chains cross international borders, is the impact of political event, security risks, and epidemiological factors also are critical features of supply chain uncertainty. Since Davis' assertion, a wealth of literature on the subject of supply chain uncertainty and risk has developed, spanning industries and focal points to generate a broad understanding of the multiple contexts of uncertainty that supply chain managers face.

Uncertainty versus risk. Across this body of literature there are many common terms; however, some terms face a contextual definition as opposed to a universal one, such as 'supply

chain uncertainty'. An understanding of 'supply chain uncertainty' in this context, comes from an analysis of its academic relationship to the term 'supply chain risk.' Given the interchangeability of these terms in some research (Peck, 2006; Ritchey & Brindley, 2007), a definition by contrast becomes imperative. Authors like Courtney, Kirkland, and Viguerie (1997) insist on a clear distinction between 'risk' and 'uncertainty,' while other authors blur the lines between the terms to give a general understanding that there is little value to creating a distinction between the two (Juttner, Peck, & Christopher, 2003; Li & Hong, 2007). The difference in question relates to the expected outcome related to each term. Some assert that the term 'risk' is solely used to refer to issues with an expected negative outcome (Wagner & Bode, 2008; Peck, 2006). In contrast, the term 'uncertainty' maps to issues, which can present positive or negative outcomes, or even positive and negative outcomes. In this dichotomy, the risks presented by a volcanic eruption interrupting flight patterns will only lead to problems in the supply chain that impact delivery, but the uncertainties surrounding vendor production lead times for a new product could potentially result in positive or negative outcomes. Given this respective understanding of the terms, 'supply chain uncertainty' takes on a broader context of events, referring to both items that can be traditionally labelled as risks as well as to other scenarios that present uncertainty (Simangunsong, Hendry, & Stevenson, 2012). This research favors the broader use of the term 'supply chain uncertainty' and the associated definition of Van der Vorst and Beulans (2002), as follows:

> Supply chain uncertainty refers to decision making situations in the supply chain in which the decision maker does not know definitely what to decide as [s/]he is indistinct about the objectives; lacks information about (or understanding of) the supply chain or its environment; lacks information processing capacities; is unable to accurately predict the impact of possible control actions on supply chain behavior; or, lacks effective control actions (noncontrollability) (p. 5).

Academic responses to supply chain uncertainty. Across the academic literature, there are two primary academic responses to uncertainty: modeling and management (Simangunsong et al., 2012). The first responds to uncertainty through the perspective of identifying sources of said uncertainty. The second focuses on the management of uncertainty within the supply chain. As this research focuses not on the identification of sources, but rather the management of uncertainty, it will focus its analysis on the second category. Figure 1, adapted from Simangunsong et al. (2012), illustrates the diverging branches of discourse around the subject of supply chain uncertainty management.

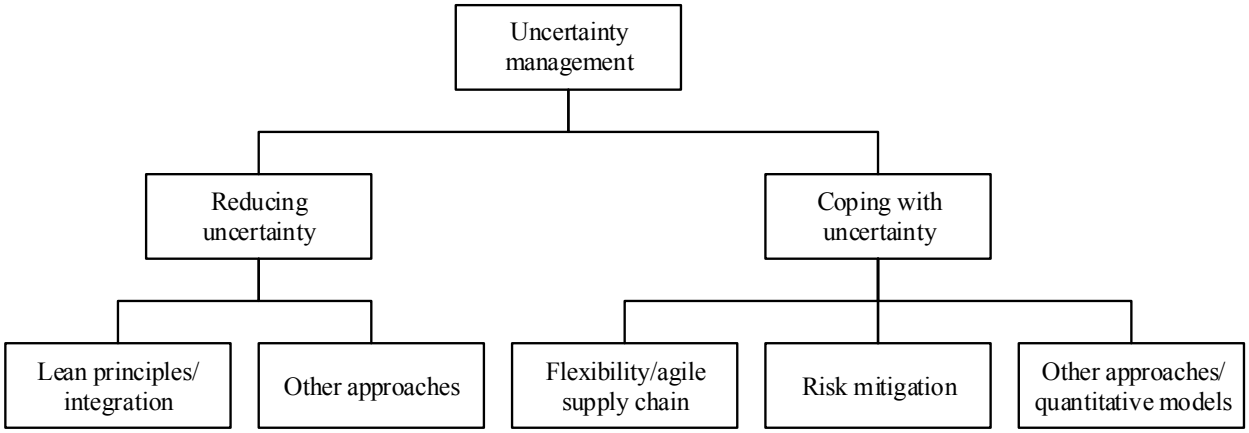

Figure 1. Categorization of supply chain uncertainty research.

Uncertainty management in the supply chain. As shown, this second category diverges into (1) reducing uncertainty and (2) coping with uncertainty. Reducing uncertainty takes an overall approach to the supply chain, by applying lean models and processes to create a seamless supply chain, thereby reducing uncertainty. Geary, Childerhouse, and Towill (2006), key contributors to this field, focus on the reduction of uncertainty related to control and chaos, i.e. when there are difficulties in production planning due to sales orders that are too small when compared to the production-batching system, or errors caused by inaccuracies in reporting from external supply-chain partners. Other approaches include process performance control (Davis,

1993), and supply chain infrastructure redesign (Van der Vorst & Beulans, 2002; Bhatnagar & Sohal, 2005), among others.

The research in this field revolves around the assumption that once uncertainty is identified, it can be mitigated through an overall lean supply chain. However, there are many kinds of uncertainty that effect a supply chain that exist outside of the control of supply chain managers. Quite simply, there are some kinds of uncertainty that simply cannot be mitigated. Instead of focusing on lean or agile principles, or 'leaglity' as this hybridized idea is referred to by some authors (Vinodh & Aravindraj, 2012), other authors are concerned with coping with uncertainty through risk analysis and other systems.

Coping with uncertainty, as a field of supply chain literature, breaks into three subfields: (1) flexibility/agile supply chain, (2) risk mitigation, and (3) other approaches/quantitative measures. Prater, Biehl, and Smith (2001), working in this subfield, respond to the uncertainty related to chain configurations, infrastructure, and facilities, discussing such things as how the availability of dependable communication impacts process times and reduces flexibility. Risk mitigation remains a very popular topic of discussion across the supply chain field (Miller, 1992; Christopher & Peck, 2004; Savic, 2008), each part of the tradition of building a risk matrix, or a risk model to be used as a lens through which to evaluate one's own supply chain for mitigation.

The final category, 'other approaches/quantitative measures,' diverges the most widely. First, there is the study of *postponement*, championed in the academic literature by B. Yang (Yang, Yang, & Wijngaard, 2007; Yang & Yang, 2010; Yang, Qin, & Zhou, 2013). This field studies the impact of company's coping with uncertainty by completing activity no sooner than absolutely necessary. For example, the uncertainty of market demand for a product should be met with a marked hesitation in production until the demand has solidified. Another field, *lead time management* (Prater et al., 2001) looks at the uncertainty that can be handled through the

quoting of longer lead times on customer orders than the manufacturing time. Advanced analytical techniques, such as the work of Davis (1993) and Christopher and Peck (2004), among others, apply rigorous statistical analysis and modeling to quantitatively assess supply chain uncertainty. Piedro and et al (2009) conducted a review of available quantitative literature which will not be replicated herein. *Buffer stock discourse* discusses coping with uncertainty through the presence of an amount of stock that can meet a certain level of immediate demand in the face of uncertainty (Davis, 1993; Van der Vorst & Beulans, 2002; Wong & Arlbjorn, 2008). However, one should note that, buffer stock can at times be diametrically opposed to the concept of a lean operating model, which assumes a zero inventory. Lastly, there are also *advanced simulation models*; however these constitute a scope too broad to consider here, but can be found within the following references: Koh and Sade (2002), Gupta and Maranas (2003), and Kwon, Im, and Lee (2007).

R**esponding to uncertainty with data.** Lacking a metaphorical crystal ball, supply chain managers must nonetheless respond to the levels of uncertainty present in the supply chain. Modelling provides a way to structure a supply chain and its various inputs. Through the model, a variety of inputs can be selected and manipulated and the resulting outputs of the system can be quantified. The key to building a model is data; and fortuitously, supply chains collect a large amount of data. Supply chain managers, therefore, are in a unique position to handle uncertainty through the application of data models. A primary means of using data to mitigate uncertainty comes in the field of predictive analytics. Through a predictive model, managers can construct an understanding of future performance based on historical data, putting them in a position to steer performance rather than reflect upon it.

Predictive Analytics

At its heart, predictive analytics is a business' way of taking data it already collects, and using it to generate insight. It is not a form of deductive reasoning; it does not start with a hypothesis which one must go on to prove or reject. Instead, predictive analytics is an inductive process, one in which a set of specific observances are used to generate a board observation.

In the hypothetical supply chain context, a data set with 150,000 lines exists in which each line represents a line item within a shipment. Within this single line, there are a series of categorical variables that identify key attributes of that line item, and the dates that said item passed each milestone in the supply chain, from the triggering procurement step to final delivery at destination. If predictive analytics is a deductive process, one would struggle to at first identify a hypothesis in such a sheer mass of data.

If that person has extensive experience with the supply chain, they might recall a personal or anecdotal observance, such as shipments travelling across the ocean by ship in the colder months are often delayed. They could then test that hypothesis by selecting all of the ocean-shipped orders, identifying the time it took to navigate the freight and logistics portion of the supply chain, and analyze the process time performance across periods of time. At length, an answer would be found, and the researcher would have an answer to the veracity of that single hypothesis. In the meantime, a hundred thousand possible insights would remain unknowable.

In the era of big data, answering questions about this supply chain one hypothesis at a time is both cost-prohibitive and less useful for managers who want to use their data to begin making decisions. This is why predictive analytics takes an inductive approach. Eschewing the traditional ways of looking at the data, such as the time-honored performance metric, this process of analysis runs tests against the entire data set, neither looking to prove a specific hypothesis nor to define a causal relationship. Instead, it begins with a vague question: what is actually going

on here? This seeming non-rigor is by design: insisting upon a theoretical model before analyzing data as large at the data sets used for these kinds of analysis ensures that businesses will not be able to utilize their data in a timely manner to make management decisions. While the development of a theoretical model certainly has use in long-term root cause analysis, or in publishing technical reports on the functioning of a supply chain, predictive analysis offers flexibility, agility, and an approach tailored to the amount of data it needs to generate meaningful information (Waller & Fawcett, 2013). As Mayer-Schönberger and Cukier wisely note:

> Society will need to shed some of its obsession for causality in exchange for simple correlations: not knowing why but only what. This overturns centuries of established practices and challenges our most basic understanding of how to make decisions and comprehend reality (p. 7).

Waller & Fawcett (2013), in response, developed a two-by-two matrix comparing predictions and explanations, reproduced below in Figure 2.

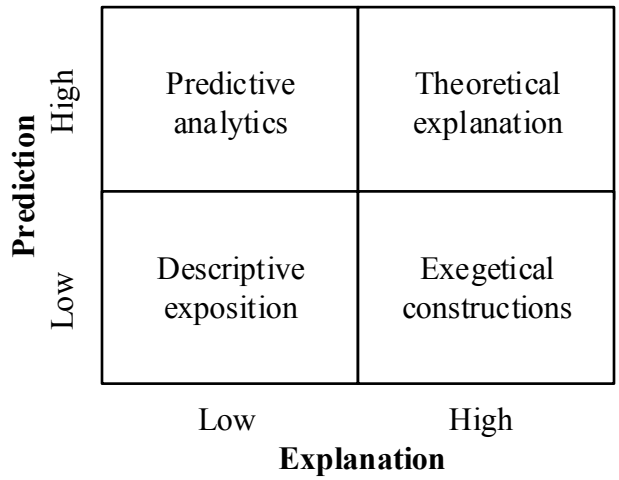

Figure 2. Two-by-two matrix of prediction and explanation.

Predictive analytics, according to this model, realizes that it may sometimes be more important to understand exactly what is going on than to understand each of the nuances associated with the underlying causation. The upper right-hand quadrant of theoretical

explanation represents the goal of most journal articles. It is essential to the advancement of an academic discipline. However, given that new concepts and phenomena in this millennium emerge at a pace faster than society is able to generate theories to support them, society must endeavor to use all of the analytical tools available within the other three quadrants. This enables a high-value understanding of flexible and even temporary systems in real-time without the immediate overhead of theoretical explanation.

Thus, a conundrum evolves in which emerging analytic tools offer insight that is able to exceed existing explanation. This means that some research at the cutting edge of technological advancement, while making a very real contribution to solving real-world problems, cannot be theoretically grounded and may be dismissed by academia. Since the goal of research is to solve real-world problems, even when that research is atheoretical, Waller and Fawcett (2013) argue that despite its nature, predictive analytics should be pursued as a course of research.

Theoretical Framework

Given the nature of predictive analytics as an entity necessarily separate from a theoretical framework, one does not begin this research with a model that it then sets out to validate but rather ends up with a data model based on statistical evidence. While this backwards practice may struggle to offer a traditional academic structure, there are some parallels that can be made in the absence of a model in these initializing stages of theorizing.

There are two primary kinds of data required as input for this model: (1) categorical variables, and (2) predicted process times. Categorical variables are assigned to an entity in the supply chain to categorize it. A multitude of possible variables exist that are individual to the supply chain at hand. The selection of categorical variables is discussed in depth in the following chapter. Predicted process times are created through the utilization of an averaging

mechanism that includes all the individual process times for a collection of lines as created by their categorical grouping.

The interaction of these data is determined through the selections made along each step of designing the model. Decision points occur at each of the seven steps described below in Figure 3. Supply chain meta-model decision design as a theoretical framework. Each will be discussed at length in the System Design and Analysis chapter.

Figure 3. Supply chain meta-model decision design as a theoretical framework.

Independent and dependent variables. Given the definitions above, categorical variables represent the independent and moderating variables in this analysis, and the predicted process times represent the dependent variable. While the difference between a moderating and an independent variable can be demonstrated by significant interaction effects at a later stage in the analysis, one must acknowledge that the magnitude of variables at play in a supply chain, nor

does any of the academic research conducted thus far offer any immediate and generalizable insight in the difference. Essentially, there are no papers to cite that lists independent and moderating variables for supply chain performance by process step.

At this juncture, one must consider again the words of Mayer-Schönberger and Cukier (2013) cited above. The researcher stresses the quote in a supplication for patience. A model of the interaction of independent and moderating categorical variables in relation to dependent process times will be quantified. In fact, there will be a unique model for each process step in the supply chain. However, a defining, single framework to structure the supply chain cannot be supplied *before* analysis is conducted. There is not enough research to support a hypothesis of this complexity. As such, only a generalized model is suggested below in Figure 4. As this research seeks to generate a meta-model of supply chain predictive analytics, the precise drivers in each box of this framework for the meta-modelling of a single process step are irrelevant as this meta-model is based off imputed data. However, the analysis conducted to generate this model demonstrates how to utilize this documentation in modelling real-world supply chains.

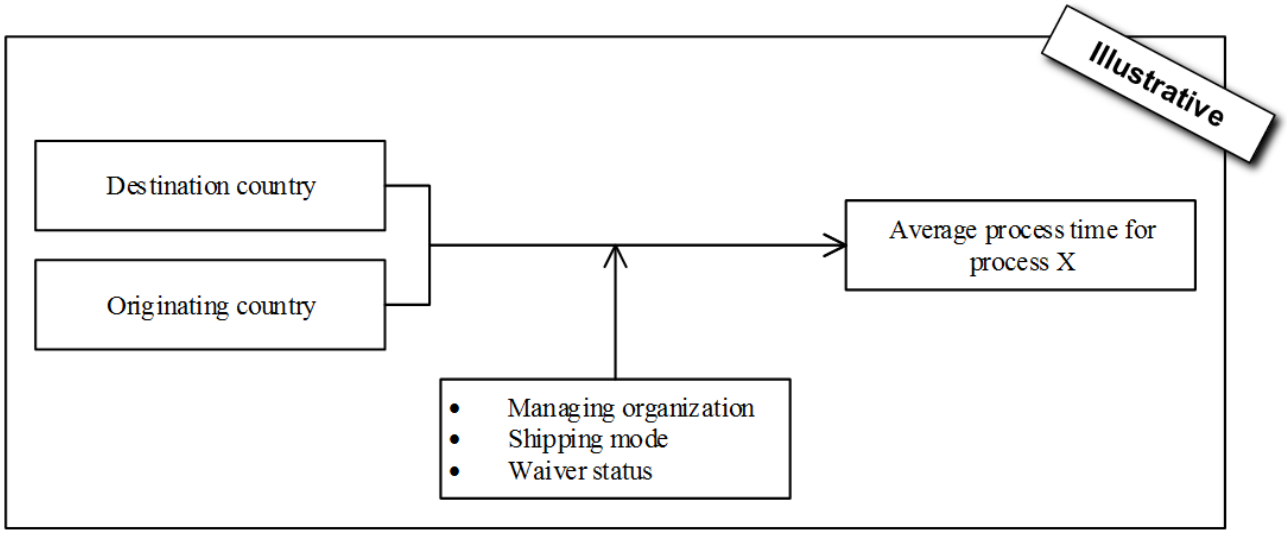

Figure 4. Theoretical model of the interaction of categorical variables per process step.

Summary

Supply chains are a critical piece of an industrialized and global society. They impact nearly every facet of private industry as well as public initiatives. Given their criticality, there is very little tolerance on the part of supply chain managers for the levels of uncertainty and risk that naturally exist within the supply chain. Risk, in this research, differs from uncertainty in the nuance that risk is applied to events that always expect a negative impact on the supply chain, i.e. a health epidemic, like the Ebola epidemic in eastern Africa, whereas uncertainty is associated with events which could have either a positive or negative impact on the supply chain, i.e. the impending award of a grant or other funding source.

Favoring 'uncertainty' as the term of choice in the research, as it is the broader of the two terms, a review of existing literature notes that studies revolving around uncertainty fall into two primary categories: (1) identifying sources of uncertainty, and (2) managing uncertainty. Within the second category, there are two subcategories: (1) reducing uncertainty, usually through the implementation of lean supply chain management principles, and (2), coping with uncertainty, which is the broader of the two fields and includes flexibility, risk mitigation, and quantitative models of uncertainty.

Following along the path of quantitative modeling of uncertainty, one enters into the realm of predictive analytics. Predictive analytics is an atheoretical field of analysis. Its inductive ability to use technology to churn through massive amounts of data in real-time allows data scientists to generate insight into a supply chain without the generation of a theoretical model to explain that insight prior to beginning analysis. Despite operating in an academically non-traditional pattern of creating answers before asking questions, predictive analytics is an in-demand capability in industry and government, and academia in the past two years has begun to acknowledge that and build slowly upon those strengths.

This capstone exists at the intersection between supply chain management uncertainty and predictive analytics. It seeks to utilize the large existing body of theoretical knowledge, documented best practices, and descriptive analysis on the subject of supply chain management. It does so while looking through the lens of predictive analytics in order to generate new insight in the form of an adaptive supply chain-wide predictive delivery model. In the absence of extant research identifying the key drivers of supply chain process performance from a quantitative perspective, it offers a meta-model built by the remainder of the research through phases of statistical analyses. Therein lays the value of this field of discourse.

This research does not seek to describe the relevant drivers of supply chain performance for a single data set. Instead, if offers a meta-model to structure the modelling of any supply chain data set. The application of rigorous predictive analytics across the field of supply chain management allows managers to look into the future to gauge performance. It ensures that managers are aware of impending issues before it becomes too late to solve them, allowing them to either take action to find solutions proactively, or begin to manage client expectations around the timeliness of deliveries. This added insight is invaluable.

Research Method

This study takes a pragmatic approach to science, rather than a positivist/post-positivist/social constructivist paradigm. This theory aims to utilize the methods and technologies which appears best suited to the research problem, without losing the purpose and impact of the work in philosophical debates about the best approach. Utilizing a framework of pragmatic research, therefore grant a freedom to mix methods, techniques and procedures across quantitative or qualitative research as necessary. The researcher recognizes that every method employed in the study will have its limitations, and that the different approaches can be complementary.

To illustrate this mixture, this theory structures its research questions qualitatively, to appreciate the necessity of responding to a different series of questions while designing this meta-model. This allows individuals looking to apply this methodology to their own supply chain to work through their own development by asking and answering these same questions in terms of their unique circumstances. Also, this qualitative structure provides for data to be collected and analyzed in stages, rather than collected all at once as is typically done in quantitative analysis. The value of this iterative analysis is in the building, error-testing, and manipulation of this meta-model. However, the bulk of the analysis does respond to these qualitative questions with quantitative analysis.

This integration of quantitative answers to qualitative questions creates a natural bridge between the theoretical design of a meta-model for supply chain analysis and the statistical rigor necessary to generate and test a predictive analytics model. This integration echoes the purpose of this tool in an industry or governmental supply chain setting. Managers can use this tool for data-driven decision-making. This action is inherently a mixed methods process as it uses quantitative findings to make qualitative decisions, and this parallel is continued herein.

Research Questions

The research questions that follow provide the guiding insight into this inquiry surrounding the building and testing of a predictive analytics model in supply chain management. The central research question is essentially broad: how can we use supply chain data to create forward a forward-thinking management structure? This overarching question bifurcates into two larger themes. First, how does one utilize supply chain historical data? Second, how does one take advantage of the insight that it offers? This dual nature of the question provides structure to the nature of this theory, while breaking into associated sub-questions as follows.

It is of note that this theory responds to qualitative questions with quantitative models and data in a form of exploration. Further research might then posit hypotheses for quantitative testing and analysis. As discussed in the Theoretical Framework section of the previous chapter, hypotheses cannot be immediately derived from such a necessarily complex collection of data. This task becomes even more difficult when there is no previous research to identify the interactions between independent and moderating variables across each process step of a supply chain. Taking this into account alongside Mayer-Schönberger and Cukier (2013) position, the research questions will be discussed in terms of the questions that the meta-model seeks to answer, rather than hypotheses it seeks to prove or disprove.

Research question one (RQ1). How might historical supply chain data be used to predict future supply chain performance?

Sub-question one a (RQ1a). What categorical variables affect supply chain performance across different process steps?

Sub-question one b (RQ1b). How does the model deal with incomplete data?

Sub-question one c (RQ1c). How might process times be calculated depending on data availability and supply chain patterns?

Proposition one. After mapping the supply chain to identify key milestone dates and determining the breadth of categorical drivers to test, historical data can be used to create an additive predictive model. By testing the relevance of categorical data to the various process times between milestones, one gains greater insight into which factors in the supply chain drive performance. After creating the model, one can apply a variety of quantitative tests to the model's outputs in order to test the accuracy of its predictions. However, the availability and structure of data will determine which tests can be performed.

Research question two (RQ2): How might a predictive model be used to manage atypical supply chain events?

Sub-question two a (RQ2a). How does the model account for the unexpected, inconsistent, and/or unknown variables that effect shipments?

Sub-question two b (RQ2b). What other logical contingencies can be applied to predictions to account for supply chain realities?

Sub-question one c (RQ2c). What other logical contingencies can be applied to predictions to account for real-time intervention?

Proposition two. In applying the model to making real-time management decisions, it becomes necessary at times to disrupt the additive model through contingencies. These quantitative disruptions mirror the disruptions that supply chains face in real life. First, for single shipments or orders that are stalled at some point in the process, a *today's date contingency* applies, which updates the projected delivery date to account for the passage of time occurring due to the disruption.

However, in dealing with circumstances that affect more than one shipment or order—such as environmental, political, geopolitical, epidemiological situations—delay codes are other observations are used instead to understand the larger effects these kinds of events have on a supply chain as a whole.

Data Source

The researcher generated the data source of 20,000 lines for this modeling research based on a system of repeating logics. These logics were developed to impute a large data set that maintains some level of statistical relationships for the purposes of generating relevant patterns in subsequent analysis. Despite that implicit patterning, the logics also allow a sufficiently random assignation of process time to ensure that this data, insomuch as possible, mirrors the entropy of an actual supply chain. This data set exists at a shipment level. Each of the 20,000 lines represents a single shipment or PO that has yet to reach the shipment stage. Therefore, each shipment identification number creates a unique primary identifier for that historical line of data. Orders that have not yet reached the shipment stage will have a unique PO number in the data.

Justification for the use of an imputed data source. While it may seem counterintuitive to frame research on creating new supply chain management structures within a data set not established in a specific instance of a supply chain, the decision to do so is actually foundational to the structure and applicability of this theory. The motivation to utilize a constructed data set for this theory is three-fold: (1) generalizability, (2) practicality, and (3) confidentiality.

Generalizability. Given that the purpose of this theory is to design a model of a model of supply chain predictive analytics, it is an intentional decision to use a data source not based in a real-world supply chain. As such, a theoretical data set allows for this model's application to a broader range of supply chain than would be possible were it designed solely for the delivery of

health commodities, or for a supply chain operated from end-to-end by a military organization. The intentional vagary of a logics-build data set provides for a high level of generalizability while situating this theory firmly within the grounds of meta-modelling as opposed to the actual modelling of a specific real-world supply chain.

Practicality. Supply chain management is a data-heavy field, from its procurement documentation to shipment confirmation. In mature supply chains, much of these data are stored in an *enterprise resource planning* (ERP) system. In less mature or multiply-owned supply chains, the data may not be so centrally stored. The diversification of data sources allows for decreased data quality, in terms of data formatting and referential integrity. Additionally, in supply chains with a potential for circuitous date records, such as may occur when certain milestones are revisited during the supply chain process, the potential for non-representative negative process times exists.

Given this potential, the use of an actual supply chain data set requires extensive work to validate its integrity before modelling can occur. While this theory will discuss the necessity of and processes by which to conduct an integrative *data quality assessment* (DQA) as part of creating a model, it will not actively conduct DQAs on the theoretical data. Because of its logics-based origin, the data set of record for this theory is fully cleaned and referentially sound. It includes neither outliers nor negative process times and was built with the referential integrity to guarantee a one-to-many relationship between identifying numbers, and a strict one-to-one relationship between these identifying numbers and categorical drivers. This is a practical decision to focus this discussion neither on the consideration of outlying data points, nor the sheer importance of a logics-enforced data management system, but rather on the details of model generation as they apply across all manners of supply chains.

Confidentiality. Due to the strategic and critical nature of supply chain information (i.e. vendor selection, pricing, client promised dates) and the political sensitivity surrounding supply chain performance, most organizations and governments consider a master data set that includes the data required to generate this system to be confidential information. As such, the publication of these data and their conclusions as academic research—even in such a way as to scrub identifying information about the supply chain—is considered a risk that many are unwilling to take. Due to the nature of the required information, locating the amount of historical data and associated legal permissions to use it is untenable. However, as this theory focuses on the predictive meta-modelling of the supply chain, not on the testing of the model in a specific instance of supply chain management, the utilization of imputed data is in no way prohibitive to the validity of the conclusions drawn.

Structure of the data set. The structure of this data set mirrors the requirements of any data set one would need to build a predictive analytics model. Given this parallel, this subsection defines the requirements of an underlying data set to build this model. There are four kinds of data that this model requires: (1) identifying information, (2) categorical information, (3) supply chain milestone dates, and (4) assigned promised dates.

Identifying information. Identifying information marks which order or shipment each lines in the data set represents. The content or amount of this information will change depending on the procurement and shipping mechanisms in place in the supply chain. Many supply chains segment the different mechanisms of supply by procurement, production, and shipping, allocating a different identification number for each part of the process.

Illustratively, a price request (PR) is made by the client and assigned a number. This PR then becomes a price quote (PQ) that is returned to the client for the goods in question. After some manner of negotiation between parties, this PQ is turned into a purchase order (PO), which

is sent to the appropriate production entity. After production is complete, the order will be forwarded to a shipping entity, who will assign it an advanced shipping notice (ASN) number.

The most important thing to understand about how a shipment or order moves through the supply chain is the relationship between these numbers. POs and ASNs can exist in a one-to-one relationship (one PO becomes one ASN every time), a one-to-many relationship (one PO is broken into multiple ASNs depending on shipping availability), a many-to-one relationship (multiple POs are consolidated into a single ASN, such as what can happen when many small orders are scheduled for delivery at or around the same time), or even theoretically, albeit unlikely, a many-to-many relationship (different POs are broken up across different shipments depending on the availability of product and shipping mechanisms). Any collection of identifying numbers can be used to mark a line of data in the system, provided that an understanding exists of the relationship between these identifiers.

The data set for this model contains two identifying numbers: (1) PO number, and (2) ASN number. These numbers exist in a one-to-many relationship, meaning that each PO number may break into multiple ASNs. However, each ASN will have only a single PO worth of items contained within it.

Categorical information. Categorical information contains many different kinds of content but all serve the same purpose: to describe the shipment. These descriptive fields are used to group shipments into like categories for the purposes of prediction. These groupings operate on the assumption that shipments subject to the same criteria will experience similar performance. For example, one can assume that all shipments going to the same country will perform more similarly on processes that are found to be related to the country of delivery than would shipments going to different countries.

Supply chain milestone dates. These fields constitute the supply chain milestone dates for each shipment. To create an additive supply chain model, one must assume that these dates fall consecutively and sequentially, like dominoes. If there are circuitous processes in the supply chain, it is recommended that the system does not attempt to make predictions or base process times off of these dates. In this data set, there are no missing dates for a shipment. If a date is not present in a line, it is because that line has not reached that milestone and therefore the date must be predicted. In a real supply chain data set, missing dates may exist. The handling of those missing dates will be discussed in the next chapter.

Assigned promise dates. These fields are the dates that the supply chain uses to measure performance. Given that this data set includes vendor production and delivery to client in the supply chain, there are two present in this system: (1) vendor promised fulfillment date, and (2) client promised delivery date. These dates represent fixed stopping points in the system, through which it is possible to understand how the shipment is performing. These fields are critical for using the model to facilitate management decision-making. For example, if the model predicts a shipment to arrive two months after the client promised delivery date, this will trigger actions for management to take to either mitigate that lateness or manage client expectations accordingly.

These dates could also come in the form of target process times. This would allow a comparison between predicted performance of a shipment and the time that management specifies that these processes should take. Target process times are not included in this data set, but the mechanism for measuring performance is essentially the same as measuring against promise dates.

Ethical Considerations

The meta-modelling of supply chain predictive analytics is an area of study exempt from most ethical considerations due to its exclusion of all human subjects and interaction. Since this

meta-modelling analysis is based on an imputed data set built on statistically - significant logics, there is no sensitive information, including supply chain performance data or personally or organizationally identifiable information, contained within. Despite these exclusions, all data was handled in an ethical manner. To insure that all research conducted complies with the highest ethical standards, the researcher completed the National Institute of Health's Institutional Review Board (IRB) training course for Protecting Human Research Participants.

The content and findings of this research seeks not to diminish, devalue, or critique any extant research in the field; nor does it aim to support any negative cause. Its sole aim is to develop a meta-model of supply chain predictive analytics in order to facilitate forward-thinking management structures that allow supply chains to operate in an environment that accounts for real-world delays.

Research Strategy

The research strategy for this occurs in three stages, as shown in Figure 5. These stages must be completed in order to allow the next stage of analysis to be conducted.

Data set design. The first stage involves the coding of a program to generate the necessary data set, and the successful execution of that program. This is a foundational part of this research as it provides the working data which fuels the further stages of this research.

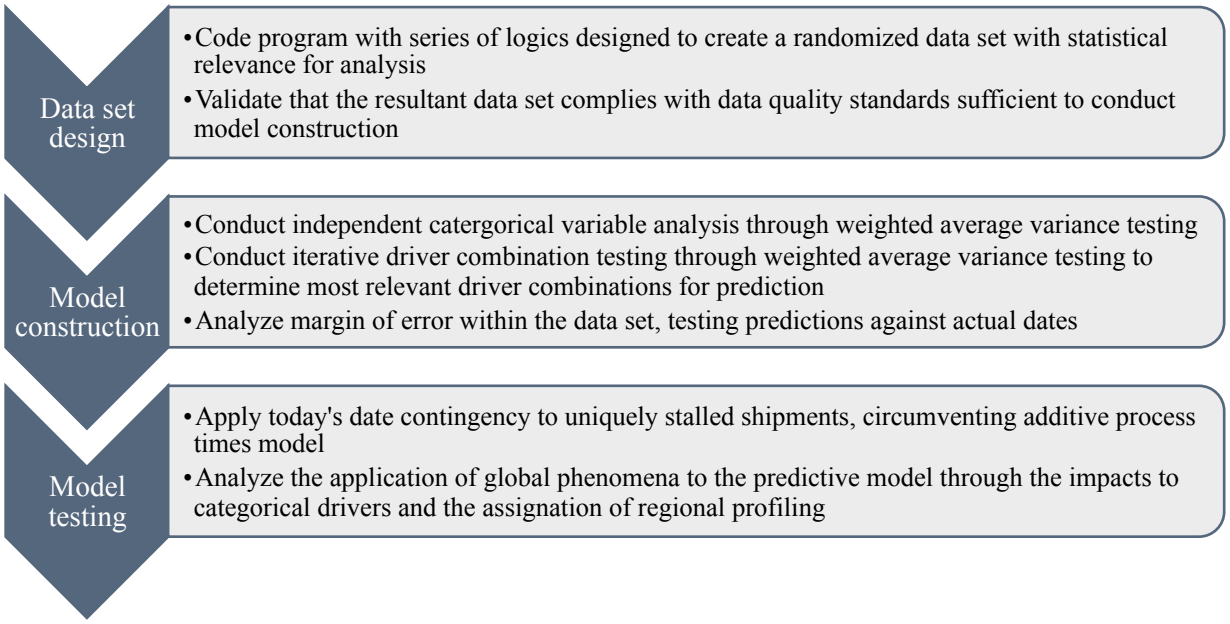

Figure 5. Stages of theory design.

Model construction. This stage of the theory seeks to answer RQ1 and its associated sub-questions. It will use the constructed data set to generate a predictive model. Data undergoes manipulation and statistical assessment to determine the most relevant combination of categorical variables. This combination, known as the categorical drivers for a process, best groups the data on the most likely factors that affect performance per process step. These drivers will then be used to segment the data and generate a table of process times. These process times will feed an additive algorithm that makes date predictions for current shipments based on historical performance. After creation, its performance is assessed through a series of tests designed to compute its precision and modify driver combinations accordingly. At this stage of completion, the theory enters the third and final stage of analysis.

Model testing. This stage of the theory seeks to answer RQ2 and its associated sub-questions. This phase of the theory introduces a series of interruptions into the additive model which mirror real world events in an effort to create an algorithm with enough flexibility to model real-world events that effect supply chains every day. These events could be

epidemiological (e.g. the Ebola outbreak in West Africa), political (e.g. acts of terror or civil unrest), or even natural (e.g. earthquake).

Summary

The need for supply chain managers to anticipate supply chain performance generated a series of research questions revolving around the generation of predictive analytics models. Managers need systems that are flexible enough to handle supply chain uncertainty, while also being statistically rigorous enough to provide a tool for day-to-day management of the supply chain. Answering this need in the form of research takes on three distinct stages.

First, as this is a meta-modelling scenario, a data set must be generated to support this analysis. The decision to generate a data set rather than appropriate one from a mature and actual supply chain was three-fold. This theory and its findings are designed to be highly generalizable. Shying away from an industry-specific supply chain structure allows for a broader discussion on the way to model a model of a supply chain, rather than simply modelling a single supply chain. It also was a decision based in practicality. Using a generated model allowed the research to operate on clean and referentially sound data, rather than manipulating a data set from an actual supply chain. Also, due to the strategic nature of the supply chain, this information is general assumed to be confidential. Out of respect for the importance of the confidentiality of these data and wanted to maintain the highest ethical standards for this research, the researcher concluded that a model data set built off logic would provide the strongest combination of quantitative rigor and ease of explanation for this meta-modelling research.

The data set was built with four kinds of data included: (1) identifying information, (2) categorical information, (3) supply chain milestone dates, and (4) assigned promised dates. These values were imputed by a series of logics and random assignation. This data set is a

shipment-level data set, meaning that each line in the data set represents a single shipment in the system, each associated with one order in a many-to-one ratio. This means that each shipment line is identifiable by its unique ASN. When an order has not yet reached the shipment stage, it is identifiable by its unique order number.

After the data set has been generated, the theory advances to the second stage: Model construction. This stage is designed to answer RQ1 and sub-questions. In this stage, the predictive meta-model is built, including discussions on a variety of data situations encountered when creating this model. First, supervised segmentation is used to determine which categorical information is most relevant to computing process times. Next, categorical identifiers are combined and tested iteratively to determine which combination of drivers provides the best predictive validity. Once drivers are identified for each process step, process times are computed to create a predicted process time for each driver for each process time. These predicted process times are then added to actual dates in the data set to generate predicted milestone completion dates. Once this computation is complete, the accuracy of the system can be predicted, which permits the research to move the third and final stage.

Lastly, the data enters the model testing stage. This stage answers RQ2 and sub-questions by testing the model against real-world disruptions to the supply chain. There are two mechanism by which the model accounts for disruptions to the supply chain: (1) today's date contingency, and (2) categorical reevaluation based on historical data of disruptive events and/or country profiling.

Through these stages of analysis, the historical data generates a predictive model that can be repurposed across any industry. With this system and its build-in flexibility in place, managers will be able to not only forecast supply-chain performance, but also create an alerts-based management system responsive to the dynamic environment in which supply chains exist.

System Design and Analysis

A model, in its most basic form, is a simplified representation of reality that is created with a specific purpose in mind (Provost & Fawcett, 2013). This is done in such a way as to preserve the relevant information in a more simplified form to help the model better serve its purpose. This simplification of reality occurs through the making of assumptions about what falls into the scope of importance for that specific purpose. An example of this kind of simplification is the picking of selected categorical variables. While those selected may not represent every nuance of what impacts a supply chain's performance, they capture enough to provide a very good idea about the interactions at hand. Assumptions can also be based on the constraints of data availability or traceability, such as with the integration of delay codes discussed further on.

A predictive model, such as the kinds that this metamodeling work seeks to create, establishes a complex formula for estimating a heretofore unknown value that is of interest. This value is called the *target variable*. The formula to do this could take the form of a mathematical equation, such as the process time calculations that follow, or a logical statement, like the business rules required to make assumptions and impute values for the model. The model we seek to build in this instance is a hybrid of both. The metamodel breaks this process of model design into two parts. First, there is model induction, the process of creating the model. This includes the selection of categorical drivers and variance testing. Next is the process of deduction, of using the model to make predictions. This includes the sections of data cleaning, process time calculations, prediction generation, and real-time intervention.

Categorical Drivers

When determining whether a piece of categorical information is relevant, there are four criteria for inclusion (1) performance impact, (2) uniform presence, (3) multi-line inclusion, and

(4) historical repeatability. Not all categorical information collected on a shipment is a good categorical driver for performance. While there is an intuition to test all available kinds of data, given that statistical testing requires resources to perform, it is cost-prohibitive for a business to test categorical drivers that do not automatically meet these four criteria.

Performance impact. First, categorical drivers must have a demonstrated impact on performance. This can be determined primarily through a logical understanding of the supply chain. Let us assume that a supply chain collects data on the type of plane used to deliver air shipments. While that is categorical information that can be used to segment the data, assuming that all of the planes used are up-to-code, modern aircraft piloted by licensed pilots, the performance of those different models of aircraft has little to no demonstrated impact on the time it takes to fly a shipment from point A to point B. Therefore, segmenting data based on the model of plane for delivery is low-value. However, the charter company operating the plane may be a good categorical driver, and the fact that a shipment is moving by plane as opposed to by sea is very likely a strong categorical indicator of the length of time it takes for the shipment to arrive.

Uniform presence. Second, categorical drivers must be uniformly present in the data. In order to use this information to segment the supply chain, it needs to be present for all or nearly all lines in the data at some specified point in the supply chain. While in the PO stages, not all categorical driver information may be present for the ASN stages of the shipment. However, there should be some assigned point in the supply chain where all categorical drivers used for computation are present in the data. If a piece of information is only put into the system on occasion, or only for a small percentage of orders, it is not going to provide as much statistical accuracy when segmenting the data. In fact, it might introduce noise into the model by grouping unlike lines of data based on the fact that they contain blanks in those fields. By contrast, data

that is present for every line at the same time in the supply chain will offer much more stability as a categorical driver.

Multi-line inclusion. The purpose of a categorical driver is to group lines of data that one can expect to have similar performance. Without this parameter, we will oversegment the supply chain and lose the ability to make valid predictions. A more specific discussion of the correct level of multi-line inclusion will take place when discussing supervised segmentation.

Historical repeatability. A categorical driver must repeat across the data set, meaning it must be present in historical lines as well as in active lines where predictions are being made. This is not to say that new values in categorical drivers cannot emerge. New values can be introduced in the event that the supply chain is operating under different conditions. Predictions made on these new values should be understood to be less precise until there is sufficient historical data supporting these categories. Disregarding the occasional new value, the bulk of values in a categorical driver should be reflective of the shipments that are currently in the supply chain. Illustratively, historical data built off of shipments travelling within North America is going to offer less predictive validity when trying to make predictions for shipments going to South America. It is for this reason that identifying numbers—such as the PO number which is omnipresent, includes multiple lines of data, and can demonstrate performance impacts when tested statistically—do not represent good categorical drivers since they are not repeated across periods in the data set.

Regional profiling and derived categorical drivers. The origin and destination locations for a shipment are generally always relevant to the length of time a supply chain needs to make a delivery. Things going up the road will arrive in a different timeframe than things crossing national borders. Geographic locations provide an opportunity to derive categorical variables in addition to those assigned to the shipment by buyers or managers. Similar

geographic locations can be grouped together to form an overarching regional category. There are three primary reasons one would implement a regional categorical driver.

First, regional drivers can help make predictions on new values. While a supply chain might never have shipped to Laos, performance of the supply chain in Vietnam of Laos might offer insight into Laotian performance.

Second, regional groupings can offer insight at the strategic level for high-level managers to make decisions. For example, a Caribbean country pays very high freight costs for ship products in from Asia because more Caribbean manufacturers charge more for the individual product. Regional variables permit analysis to occur which can answer the question of whether it is more cost-effective to pay a local premium and save of freight costs or to continue importing goods from Asia.

Third, regional categorical drivers can help avoid oversegmentation of the data. When grouping data by categorical drivers, a higher level driver will break the data down into less groups then its lower level counterpart. On processes where a geographical indicator plays some role but may not be the primary drive of performance, it could be useful to instead use a regional driver than a specific location driver. In this way, geography is still accounted for, but the system is also able to maximize the number of lines of data included in each group to make more accurate predictions.

There are two principal ways to derive regional drivers: (1) by geography, and (2) by statistical profiling. Each will be discussed in turn.

Regional profiling by geography. Geographic regions are relatively straightforward to designate. They could be as broad as continents (i.e. Asia, Europe, et al.) or more specific to the regions within continents (i.e. Southeast Asia, Central Europe, et al.). In a national supply chain,

they could break down to the local governing area, such as a state or county, presuming there are multiple geographic locations that would fit into that subcategory.

This kind of regional driver would likely be most useful during when grouping the origin country. Place of origin may show sharp statistical relevancy to a given shipping process. This makes sense given that where something starts will make a difference as to how long it will take to get to a destination. For example, a shipment travelling 100 miles will take a different amount of time than a shipment travelling 1,000 miles. However this origin country is likely to show some relevancy in other process times as well. It will likely lack relevancy in the earliest stages of an order because at this juncture it is unlikely that origin country is designated, since origin country would be a reflection of the manufacturing site where the product is created which may or may not yet be confirmed. If it shows up in the manufacturing process, it is likely a reflection of the vendor or manufacturer less so than the country itself. Given the potential prevalence of the relevancy of this field, it might be more useful to look at origin region as opposed to origin country when looking at non-shipping or export process steps. This is one way to help avoid oversegmenting the data.

Regional profiling by statistical profiling. Using statistical methodology to create regional profiles is a more sophisticated method of deriving a regional categorical variable. It is most useful for large, international supply chains working across several geographic regions. For smaller supply chains, this approach can offer similar insight, but the analysis required might become cost-prohibitive to the amount of information that it subsequently offers. For smaller scale supply chains, regional categorical drivers based on geographic region may in fact be the most useful derived category.

This method can be especially useful when trying to discern patterns of performance across countries, and for predicting new spaces within the supply chain based on historical data.

A regional profile is determined by assessing the likeness between various countries based on similar characteristics. The modifier 'regional' is assigned to this profile because the geographic region has some marked relevancy to the process, although it is not the only facet of the country to be analyzed when determining the profile. To illustrate the reasoning behind this kind of driver, we can look at a hypothetical example. Zimbabwe and South Africa are adjacent and potentially might both grouped into a South African region including: Lesotho, Namibia, and Botswana. This assumes that shipping performance in those regions looks similar based on geography.

However, South Africa as country has access to seaports that Zimbabwe lacks as a non-coastal country. Anything shipped by ocean to Zimbabwe would require additional time on a truck to reach its destination than would something travelling through South Africa. Additionally, the waiver requirements to import a product may be different between the two countries, as is perhaps the road quality, the security of the shipments moving through the country, and the accessibility of the various destinations. Given these sharp difference between countries within the same region, a geographic region may introduce more variance than similarity when predictive performance. If this same supply chain were to move into Swaziland and used a derived categorical driver based on geographic region, the system would predict its performance as an average of Zimbabwe, South Africa, Lesotho, Namibia, and Botswana, weighted by the volume of shipments going into each country. This may or may not be accurate. What would be more accurate is to predict future performance in Swaziland based on a country or set of countries that mirrors its geographic region, import/export laws, and infrastructure. This is where the regional categorical driver by statistical profiling would provide the most utility.

There are a variety of ways these profiles can be determined. In an effort to not derail the study of the larger system with the nuances of this very small facet, only a few options will be presented. Determining these profiles could occur as simply as deciding upon a series of categories across which to measure all countries, including waiver requirements, infrastructure, accessibility, etc. When determining the likeness of countries, these categories could be used to find the nearest proxy for a new country. A more sophisticated, statistical methodology involves formal segmentation processes, such as k-means segmentation and clustering. However, we will not explore those options in this research at this time.

Variance Testing

The analysis of variance is a valuable tool for exploring data to explain observations. It give us the ability to formalize intuitive judgements or hypotheses for the purposes of analyzing experimental data and developing models. The true strength of this method for the purposes of predictive analysis as compared to similar statistical tools like correlation is that not all data must be numeric in this context. Additionally, it provides confidence in explanatory relationships. The most commonly used statistical tests for analyzing variance come from the ANOVA family. However, these are not necessarily the correct tests to make in this circumstance.

Assumptions of ANOVA tests. An ANOVA test makes three basic assumptions. First, the response variable occurs on a normal distribution. Second, there is homogeneity of population variances within the sample. Third, each sample is an independent random sample. Please note that in an effort to focus on the theory of model building, the equations for running the various statistical analyses listed below have not been included. However, their original publications have been cited, and the mechanisms to run these tests are available in most statistical software packages, like SPSS or STATA, or in open source software, like R.

Normality of distribution. This can be testing using the Shapiro-Wilk test, or other similar tests of normality, such as the Anderson–Darling, Kolmogorov–Smirnov, Lilliefors, and tests. However, the Shapiro-Wilk test was empirically demonstrated to have the best power, or sensitivity, for correctly rejecting the null hypothesis (Razali & Wah, 2011). This test carries the null hypothesis that a population is normally distributed. Therefore, if the p-value is less than the chosen alpha level, we can reject the null hypothesis. This means that there is evidence that the data testing are not part of a normally distributed population. The converse is also true. If the p-value is greater than the chosen alpha level, the null hypothesis cannot be rejected, and the data comes from a normally distributed population. (Shapiro & Wilk, 1965).

Supply chain process times data, in the authors experience, does not come from a normal distribution and therefore rejects the null hypothesis of this test. This requires us to resort to one of two options: (1) to transform the data using various algorithms in order to create a normal distribution, or (2) to utilize the non-parametric Kruskal-Wallis H Test, a test which does not require the assumption of normality (Kruskal &Wallis, 1952).

Homogeneity of variance. The occurrence of equal variances across samples is referred to as the homogeneity of variances (Snedecor & Cochran, 1989). To determine whether or not homogeneity of variances exists in the data set, Barlett's homogeneity of variance test must be run (Bartlett, 1937). This test is used to determine if k samples are from populations with equal variances. Again, supply chain process times data that is segmented into k samples by a given categorical variable does not satisfy this assumption.

Fortunately, two tests become applicable when the assumption of homogeneity of variances has been violated: (1) Welch test, also known as the unequal variances t-test (Welch, 1947), or (2) the Brown and Forsythe test (Brown & Forsythe, 1974). The Brown and Forsythe test is an alternative to the Bartlett test that is less sensitive to departures from normality. Also,

in similar fashion to the option provided in the event of the violation of the first assumption, a Kruskal-Wallis H Test could be run instead.

Independent random sampling. This assumption is frequently cited as the most important assumption to fail. Often, there is little that can be done that offers a good solution. In this supply chain context, independent cases cannot be verified. Thus, neither ANOVA nor the Kruskal-Wallis H Test are perfect options. Instead, we must turn to data mining techniques in order to find a solution.

Supervised segmentation. The goal of categorical variance analysis is to use historical data, or training data as it's called, to make a predictive model. This process is known as induction, since it involves using all available observations from the training data to draw conclusions. Induction differs from deduction, which is the consequent practice of using the model. After the completion of this stage to build the model, it can then be used to make predictions. To achieve this goal, the training data will be put through a process of supervised segmentation.

Segmentation is the process by which one determines how to segment a population, in this case shipment data. Into groups that differ from each other with respect to a specified quantity of interest. In this circumstance, the quantity of interest is the amount of time it takes for shipments or orders to undergo a particular process within a supply chain. The differentiating factors are what we know as categorical variables.

Adding the word 'supervised' to segmentation is a term which categorizes the kind of segmentation used on these data as part of supervised learning. *Supervised learning* is a broad term used to encompass the varied techniques in use, like induction algorithms, to identify relationships between independent attributes (i.e. categorical variables) and their designated dependent attributes (i.e. process times). Therefore, *supervised segmentation* is the process by

which we will induct this predictive model through a systematic assessment of how to segment the population (Provost & Fawcett, 2013).

This process will happen uniquely for each step in the supply chain. For example, if an order goes through six stages for procurement and eight stages for shipping, then supervised segmentation would occur independently for each of those 14 steps because different categorical drivers would theoretically impact each process step in a different way.

Information gain and entropy. Supervised segmentation in this context relies on the principle of using information about shipments to reduce the level of uncertainty about how long it will take that shipment to arrive. For example, knowing that a shipment is travelling from Ireland to Haiti will not tell us for certain how long that shipment will take to arrive, but it does reduce in some small measure the amount of uncertainty surrounding that shipment's arrival. By picking better pieces of information, the more uncertainty around process times is reduced. This is the ultimate goal of the process: to reduce uncertainty.

Given this, we aim to segment data via categorical variable into groups that are *pure*, or *homogenous*. This means that every member of this group yields the same target variable. For an illustrative example, all shipments that travel by chartered aircraft take three days to travel from point A to point B. It is exceedingly rare to find a wholly pure group because attributes rarely split a group perfectly. For the most part, one ends up reducing the impurity of the group as much as possible without achieving a homogenous group.

Given this difficulty and the additional complications that not all attributes are binary (i.e. yes or no) and that some attributes take on numeric values, a purity measure is used to resolve these complications. A *purity measure* is a formula used for the evaluations of an attribute's ability to segment data with respect to a chosen target variable (Provost & Fawcett, 2013).

Most commonly, a purity measure is calculated through the concept of information gain, based on entropy, a type of purity measure (Shannon, 1948). Entropy, as a measure of disorder, can be applied to a set such as the segments created through grouping data by a categorical variable. In assessing how the values of the target variable of the members of that group, a measure of disorder indicates how impure that segment of data is given the specified properties of interest (Provost & Fawcett, 2013).

Information gain measures how much a specific attribute decreases entropy over the whole segmentation that it creates. As a parent set is split into child sets by a given attribute, information gain is a function of both the parent set and of the children resulting from the partitioning of the parent set based on that attribute. The relative purity of each of the children in this instance decides how much information the attribute is able to provide. One should note that the entropy of each child will be weighted by the proportion of instances belonging to that child. Illustratively, there exists an attribute that can create pure children for six out of seven possible values in that attribute. However, the seventh value, which contains the most lines of data, is highly impure (i.e. it was the miscellaneous value into which many lines of other data were grouped). This attribute could be measured as providing less information gain than an attribute which provides a more average amount of entropy reduction across all values.

Through this process of comparing attributes between parents and children, a classification tree or probability estimation tree can be induced. These trees, types of models, can then be used to generate business rules in order to make predictions. The attentive reader should note that this methodology is not readily applicable to the situation at hand. A tree model will not enable its user to deduct a process time. In this model, we are instead looking at a regression problem where the target variable is numeric.

Regressions with numeric target variables. The fundamental concept is similar: the reduction of the impurity of the child. However, entropy-based information gain is not the correct measure, because it is based on the distribution of properties within the segmentation instead of the purity of numeric values within the children. *Variance* is a natural measure of impurity amongst numeric values. The square root of variance is the familiar standard deviation. If the child set contains entirely the same value for the target variable, then that set is pure and would have a variance of zero. If the numeric target values within the child vary wildly, then the set would have a very high variance.

Variance and floating point arithmetic. It should be noted that the selection of software and or variance equations should not be random. With floating point arithmetic which is ubiquitous in computer systems (Goldberg, 1991), the equation for variance that is equivalent to the second cumulant of the probability distribution for random variable *x* (shown below) should not be used.

$$Var(x) = E[x^2] - (E[x])^2$$

In this computing environment, this equation suffers from catastrophic cancellation, or loss of significance. *Catastrophic cancellation* occurs when the performance of a mathematical operation on two numbers increases relative error more significantly than it increases absolute error. This can occur, for example, when subtracting two nearly equal numbers. As numerically stable alternatives exist for calculating variance, those should be used in this scenario.

Weighted average variance testing. Therefore, in the spirit of entropy-based information gain, we use *weighted average variance testing* to determine which categorical variables are most relevant to each process step. This process groups the parent data set into child subsets and analyzes the variance reduction of the target variable within that subset. Having

discussed the theory behind this selection in the previous section, we will now turn to the practical application of it to a data set.

As a preliminary note, a discussion of natural weighting within the data set occurs in the following 'Data Cleaning' subsection of this text. The use of multiple data sets to de-weight orders that multiply into many shipments that applies within that context also applies herein. Similar precautions should be taken to clean the data for average weighted variance testing. Assuming a clean and un-weighted data set in prepared for analysis; the following subsections will guide the process of determining categorical variables with which to induct the predictive model.

(1) Compute process times for each prediction date. For each date that is to be predicted by the model, a process time must be computed. Process times are the difference between that date (d) and the date preceding it (d_{-1}): [$d-d_{-1}$]. This should be computed for every line of data in the data set that is targeted to be a part of categorical variable analysis.

(2) Clean process time data. This step is potentially optional dependent on the data quality of the parent data set. The primary business rule that should not be violated is that a process time must be greater than or equal to zero: [$d-d_{-1}\geq 0$]. Additionally, one should ensure that null date values are not being interpreted as a zero and therefore returning numbers in excess of 35,000. This can occur when using software like MS Excel, which is highly ill-advised for this analysis, because of the way in which it stores date values numerically: [$d-d_{-1}<35000$]. Outlier removal can also occur on a process step by process step basis via the selected method of the individual choice. A brief discussion of outlier removal occurs in the following 'Data Cleaning' subsection and will not be repeated here.

(3) Calculate overall variance of process times by process steps. For each process step, calculate the variance across each line of data. Please refer to the 'Variance and floating point

arithmetic' subsection of this chapter if debating on technologies and variance equations. This variance calculation represents the variance of the parent set and is the base number to which all further comparisons for variance testing will be analyzed.

(4) Test independent categorical variable variance. After grouping data by subgroups based on a selected categorical variable (x_n), compute the variance of each subgroup (v_1, v_2, $v_3 \ldots v_n$). In order to weight the final number, multiply that variance by the number of instances (i_1, i_2, $i_3 \ldots v_n$), or lines of data, falling into that subgroup: $[(v_1 \cdot i_1)]$. This product for each subgroup of the data should then be summed.

Next, divide the sum by the count of subgroups (n) the parent data set. This calculation now represents the average weighted variance of that categorical variable. In order to understand change in variance (ΔV), the magnitude by which that attribute allows us to reduce uncertainty around the data by using that categorical variable, we must compare it to the overall variance of the parent set calculated in Step 3. This can be easily understood as a percent change between the numbers. This equation involves subtracting the average weighted variance of that categorical variable from the overall variance (V), and dividing that difference by the original overall variance. The simple equation for percent change from overall variance to the average weighted variance of a subgroup can be found below. In the output from this equation, a negative number indicates a percent decrease in variance which indicates that the categorical variable has value in a predictive model. A positive number indicates a percent increase in variance which indicates that the categorical variable does not have value in a predictive model, as it does not reduce the amount of uncertainty around the numeric target variable.

$$\Delta \mathrm{V} = \left(\frac{\left(\frac{\left(\sum_{k=m}^{n} v_k \times i_k = (v_m \times i_m) + \cdots (v_n \times i_n) \right)}{n} \right) - \mathrm{V}}{\mathrm{V}} \right) \times 100$$

This test must be run across each categorical variable for each process step. If one were to test ten categorical variables for a supply chain aiming to create a model that predicts 18 key milestone dates, this test would need to be run 180 times in order to complete the independent categorical variable analysis. Results can be compiled into a matrix with process names for each row and categorical variables for each column, as shown below.

	x_1	x_2	x_3	x_4	x_5	$x_{6...}$
Process A	-7%	-5%	-32%	-16%	-10%	-5%
Process B	-2%	-2%	-1%	0%	-2%	-2%
Process C…	-8%	-7%	-3%	-1%	-8%	-7%

Figure 6. Illustrative matrix of completed independent categorical driver variance testing

(5) Test combined categorical driver variance iteratively. For the purposes of this theory, a combination of categorical variables is called the categorical driver of a process. Drivers should made by combining the categorical variables with the highest reduction of variance as computed in the table above. Variables with the highest relevance from the independent analysis should be tested in combination. This testing should occur iteratively. It is important to note that reductions in variance do not combine additively, multiplicatively, or exponentially when combining variables. This means that all possible combinations that could have relevance must be tested independently. For example, Process A driver tests might include the following: (1) x_3-x_4, (2) x_3-x_4-x_5, (3) x_3-x_4-x_5-x_1. As a general rule, the larger reduction in variance that a single variable provides, the less variables one will need to combine while testing in order to select a driver that provides an acceptable reduction of variance while also avoiding the risk of overfitting the model.

The combination of variables in this stage is then used to sort the parent data into subsets. The subsets are tested using the same weighted average variance testing described in Step 4. Once all of the potential combinations have been tested, this step is complete.

(6) Calculate breadth of categorical drivers across data set. Next, the breadth of the children sets must be calculated. *Breadth*, in this context, refers to the number of subsets within the data containing two or more lines of data. As previously mentioned, variance in a set will be zero when all lines in that subset have the same target variable. The smaller the number of lines of data that fall into this, the easier it is to see a very high reduction in variance.

This calculation is used to help avoid overfitting the model by selecting too many categorical variables and including too few lines of data in each child set. This is calculated independently of the parent set, as the parent set has a breadth of 100% since it represents the entire population. To compute the breadth, first we must count the number of lines of data falling into each child set. Through a binary system (i.e. true/false), we then identify whether those subsets contain more than two lines. Please note that another threshold greater than two can be used for this classification. The greater the number selected, the more lines of historical data will be used to computer process times for the purposes of prediction. The breadth is finally computed by dividing the number of subsets that fall within the range by the total number of children sets that driver creates out of the parent set. Therefore, if two hundred out of three hundred child sets meet the criteria, that driver would have a breadth of 67%.

	x_1	x_2	x_3	x_4	x_5	$x_{6...}$
Process A	-7%	-5%	-32%	-16%	-10%	-5%
	71%	93%	92%	100%	42%	100%
Process B	-2%	-2%	-1%	0%	-2%	-2%
	83%	93%	93%	100%	67%	100%
Process C…	-8%	-7%	-3%	-1%	-8%	-7%
	83%	93%	93%	100%	67%	100%

Figure 7. Illustrative matrix of independent categorical driver variance reduction and breadth

This figure illustrates child set breadth (in grey) in combination with the reduction of variance (in white). While this illustrates breadth at an independent testing level, the

methodology is the same when used to analyze categorical drivers instead of categorical variables. From this chart, we can see that x_5 is likely a categorical variable with many possible values which therefore segments the parent data set into many child sets that are comprised of very few lines. An explanation behind the different breadths between Process A and Process B and C is data availability. While it is optimistic to assume that every date will be available and not removed as an outlier when computing, when these dates are missing or outlier removal cleans these data out, null values will remain. These null values mean that there is less data available to compute breadth. This translates to a delta in breadth between different processes.

(7) Analyze results and select the best path. The final step of weighted average variance testing involves manual analysis and driver selection. Up to this point, nearly all calculations can be done automatically by statistical analyses tools. At this point, one should look at the results of all the categorical driver tests. First, a logical validation should occur. One should be certain that the variables selected for that stage are variables that actually exist in the data set at that point in time. For example, in the very early stages of the process, the shipment origin may not yet be present since the manufacturer has not yet been selected. When looking at the historical line of data once it has been delivered and all categorical variables are present, shipment origin might end up creating a reduction in variance for that early process. However, in those circumstances that variable should not be included in the driver. Its inclusion will create a model in which lines of data that require predictions will not be linked to any historical performance. This grossly increases the amount of variance in the data and makes inaccurate predictions.

Once the drivers are selected, we have completed the induction part of building the model based on training data. The completion of this part permits a system to begin being designed around this model in order to enter the deduction stage. Using the model now we have

determined how categorical variables impact process times is discussed in the following subsections.

Data Cleaning

Data cleaning and validation is a necessary precursor for any piece of analysis. It ensures that data contains the proper referential logics, does not contain duplicates, removes inappropriate values which would create errors during analysis, and contains the correct fields in the correct format as necessary for analysis. This text assumes that these basic validations to ensure that the data set is prepared for analysis have occurred in advance of the building of this system. In an effort to avoid a discussion of elementary data quality analysis, this discussion of data cleaning for this system is solely threefold, including discourse on (1) weighting in the data set, (2) outlier removal, and (3) the assignation of imputed logics.

Weighting in the data set. Depending on the construction of the data set, it is possible for weighting of some variety to occur when making calculations. If an order breaks down into multiple shipments, the performance for that order will be counted multiple times when computing average process times. To avoid this, it is important to ensure that order process times are computed of an order-level data set, while shipment process times are computed off a process level data set.

Outlier removal. Outlier removal is an art more than a science. It is a quantitative way to determine what portion of collected data is likely invalid and introduces unnecessary noise into the conclusions drawn from that data. Most researchers utilize standard deviation to identify and remove outliers from their data. This methodology calculates the mean and standard deviation of a group of numbers and assumes that anything falling outside of two standard deviations of the mean is an outlier.

This approach does not work quite as well in this context, as two standard deviations below the mean would be a negative number which is not possible when measuring process times. Given that the distribution of process times for a given process step are not normalized into a standard bell curve (as discussed extensively in the 'Variance Testing' subsection of this chapter), the standard deviation methodology cannot assume that the removal of anything outside of two standard deviations is an outlier. The figure below illustrates the frequency of process times across a random process, process x, in days.

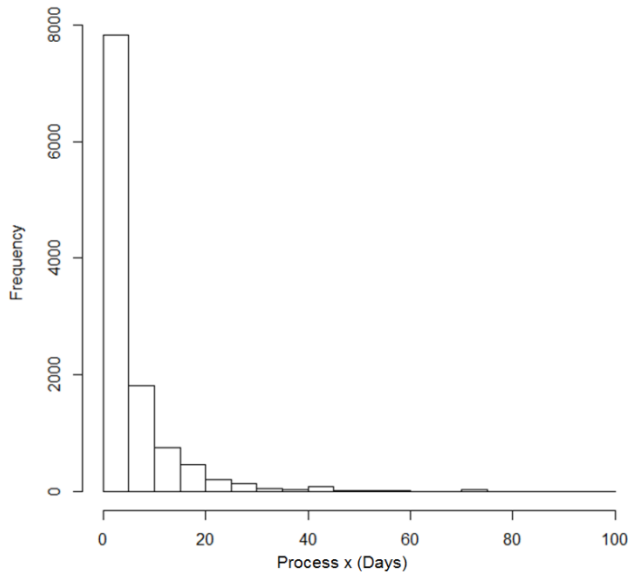

Figure 8. Frequency of process times in days for process x.

The large cluster of lines at one number of days and the inability to have negative process times skews this curve. Rather than discounting lines of data based on standard deviation, which does not fit so neatly to a non-normalized curve, another method must be found. One can decide to not remove excessively long process times. This could be a good option if the point of data entry has strong logics ensuring that false dates cannot be entered in and therefore that excessively large process times cannot be calculated.

If the system cannot control these false date entries or one does not want to compute process times skewed by a few lines of data representing shipments that went astray, a maximum

threshold can be designated on a process step by process step basis. This would ensure the inclusion of lines deemed relevant and the exclusion of abnormally long process times. The method by which the removal of outliers occurs is a decision to be made on a case by case basis based on data quality, prevalence, and management rationale.

Imputed logics. As mentioned in Step 7 of the 'Variance Testing' subsection, categorical variables should only be included if they are present at that state in the process. At the earliest stages when certain categorical variables have not yet been decided, it is necessary to impute those categorical variables in order to permit the system to make predictions up to the delivery date. These logics simply guess what a categorical variable will be before it is completed. These can be as simply as selecting the most common option for that variable, or as complex as generating a tree model to determine the non-numeric value based a variety of other factors. This data mining technique will not be covered in this research.

These imputed logics should never override data that is present in the data set, nor should they enter the final output of this data. Their sole purpose is to place data with incomplete categorical variables into a subgroup of data of the most likely category. Once actual data are entered for those missing variables, the line should bounce back into the subgroup where it is meant to be based on real data.

Process Time Calculations

This subsection, building on the foundation of those previous, uses cleaned data and the categorical drivers decided previously in order to create predicted process times built on historical data. This section trifurcates into distinct but related subjects: (1) the use and impact of delay codes in process times, (2) the mathematical theory of generating predicted process times, and (3) the creation and use of process time matrices.

Delay codes. As a preliminary disclaimer, this section can be wholly excluded from the model in the event that delay codes are not a systematic part of the supply chain or management elects to address them specifically in the model. *Delay codes*, in this circumstance, are codes that a buyer or manager interacting with an order or shipment can use to move a promise date. For a definition of *promise date* and the other kinds of data in a predictive analytics data set, please turn to the 'Data Source' section of the 'Research Methods' chapter. For the purposes of this metamodel, delay codes fall into two categories: (1) acceptable delay, and (2) unacceptable delay. An *acceptable delay* (AD) code is one that can be used to move back a promise date for a reason deemed acceptable, such as the destination warehouse is full and a delay in delivery was requested until more space is made. An *unacceptable delay* (UD) code is one that will not move back a promise date because it is considered an unacceptable type of delay, such as a shipment being held in customs due to incomplete paperwork.

For the purposes of computation, AD/UD Codes are numerical values representing the number of days of delay included in the process. The method by which they are given quantitative relevance to the model depends on two features: (1) the code reporting and storage mechanism, and (2) the code entry process by shipment managers. This is by no means an exhaustive discussion of the ways in which AD/UD codes are used, stored, accounted for, and/or managed. This theory simplifies the myriads of possibilities available to supply chains through their ERPs into two options generated by the requirements necessitated for the performance of various types of calculation, as shown in Figure 9.

AD/UD code reporting and storage mechanism. This feature focuses on the way in which the supply chain's ERP data architecture stores AD/UD codes. There are two broad options that we will consider here: (1) code captured in bulk, or (2) codes linked to a specific process step.

When AD/UD codes are stored in *bulk*, the assignment of the code to a shipment occurs at the shipment level. Meaning: if three AD/UD codes were added to a shipment, each for two days, those six days of delay codes would be linked to that shipment line in the data. Each individual code and reason can be identified, but there is no visibility into which process in the shipment incurred that specific code.

When AD/UD codes are *linked* to a specific process step, the assignation of the code to a shipment occurs at a date level. Meaning: if an AD/UD code for four days is added to a shipment, its addition is associated to the specific milestone day that it impedes. Each individual code and reason can be identified, and there is direct visibility into each process time impacted by that code.

AD/UD code entry process by shipment managers. This feature focuses on the way in which the processes surrounding AD/UD code entry work within a supply chain. The two broad options considered here relate into the perceived purpose of AD/UD codes to the supply chain and is dependent both on management direction and on the practical implementation of the process: (1) performance management focus, or (2) operational management focus.

AD/UD codes that are perceived to have solely a *performance management focus* are those that are used solely for the purpose of moving a promise date in order to mitigate the impact that supply chain delays have on performance metrics. Often, this means that codes are entered all at once during the final stages of the supply chain, on or before the performance is officially tabulated for that shipment or that period of performance.

AD/UD codes that are perceived to have an *operational management focus* also serve a performance management purpose, but the timing of their entry differs. AD/UD codes are entered progressively and iteratively throughout the process as a way to communicate the order

or shipments place in the supply chain to internal and external stakeholders. Codes are entered at or around the time the delays that they signify are occurring.

The way in which a supply chain satisfies these requirements and therefore falls into either category decides which response they can utilize for including AD/UD codes in their predictive model. Figure 9 and the following sub-sections explain in more detail the quantitative mechanisms optionally available.

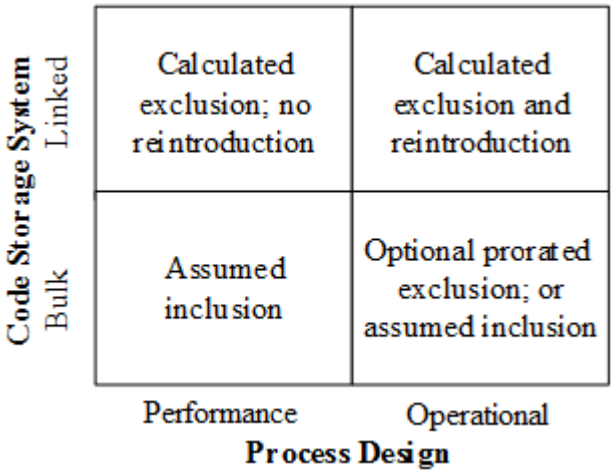

Figure 9. Matrix of AD/UD solutions based on storage mechanism and process design.

The key to a successful model is its ability to simplify the world around it in such a way as to focus on the truly important facets. This real time intervention is critical for the model to work in area of high uncertainty, such as within supply chains. Delay codes provide some of this intervention. For catastrophic events—such as acts of terror, viral epidemics, and natural disasters— to be calculable in the model, they must be quantifiable. Meaning: to understand how an act of terror impacts a supply chain, perhaps by shutting down roads and ports or by theft and diversion we need to be able to understand how many days of delay it adds to the process and add that information into the model via AD/UD codes. Similarly, if viral epidemics, such as Ebola in West Africa in 2014, shut down international travel in afflicted nations and limit in-country staff's ability to work, that quantification in days can enter the system through AD/UD

codes. By absorbing and utilizing this information, the model is able to update predictions accordingly (presuming the code system is sophisticated enough for reintroduction as described below), and give managers a high-level understanding of how these events impact performance across multiple shipments. An order made by a manufacturer recovering from an earthquake that destroyed part of the factory will likely face delays. This delay means the order won't be prepared in time for when the freighter scheduled to carry it departs. The understanding of the implications of that delay well in advance of the order not making it onto the freighter allow for alternate provisions to be made. While the buyer of order point of contact might be well aware of this delay and is working to mitigate it, in large supply scales without a centralized predictive and performance data set such as the model described here, it will be hard to get an overarching sense of the scope that such events have on the supply chain as a whole. With that criticality in mind, we can move into a discussion of the various ways a model can be set up to handle AD/UD codes.

Calculated exclusion and reintroduction. For this calculation, when process times are computed, the number of days assigned by AD/UD codes is additionally removed from the process. Therefore if the initiating date and the concluding date of a process differ by 32 days, and 12 days of AD codes are entered, that process should be recorded as 20 days. An argument can be made that UD codes should not be removed from the process time since it is indicative of the process times outside of supply chain control that are faced by shipments on that lane.

Since an operational management focus is required for this process, AD/UD codes are entered into the system as they are brought to light, they can be included in the predictions. This will likely only impact the first prediction being made, but it will help to communicate the expected completion of that milestone as accurately as possible given available data. This is an important feature in order to make the model responsive to actual events occurring within a

supply chain. In this way, both historical and real-time data are merged to provide the most accurate predictions possible.

Calculated exclusion; no reintroduction. The calculation here mirrors the above. The number of days assigned by AD/UD codes is removed from the process. The same UD code argument for exclusion exists. Since this methodology assumes that AD/UD codes are entered towards the end of the shipment and cannot be counted upon to appear in the data proactively as the shipment moves through the supply chain, this calculation ends here, and no AD/UD days are reintroduced during the prediction-making part of the model.

Prorated exclusion; no reintroduction. In the event that AD/UD codes are stored in bulk and the decision is made to exclude them from predicted process time calculations, proration offers a solution. This method is most successful when used in conjunction with a binary target threshold designation. A *binary target threshold designation* is a field that determines whether a process time falls within a designated target. This requires that target process times are set for each process used to make a prediction as is necessary for the target process time contingency described in the 'Process Time Matrices' subsection. This binary designation would compare the process time for a line against that target, designating whether it is at or below the target threshold ("0"), or above target ("1"). Therefore, if the target for process A is 3 days, a line in which the process took two days would return a "0", while a five day process would return a "1". Next, sum all the "1" processes that occur on that line. This is the proration denominator. After computing the prorated factor, apply it to each threshold-exceeding process time via subtraction.

To illustrate, let us assume that the given supply chain predicts 20 dates, meaning there are 20 predicted process times to calculate. For a given shipment, delay codes providing for 21 days of delay are assigned. Please note that the designation between AD or UD is non-differentiating. It is determined that seven process steps exceed their designated target. The

quotient of 21 and seven is three. Therefore, three days should be subtracted from each of the seven process steps that exceed threshold.

In the event that the quotient is a non-integer, there are two ways to negotiate the decimal. First, one can simply round; a simple solution to a simple problem. Conversely, if quantitative designation of maximum exceedance of threshold is added to the qualitative binary target threshold designation, the remainder after division could be additionally attributed to the process that maximally exceeds its target. The quantitative field might best calculate based on the percentage above target, rather than the count of days above target if one prefers to disallow the respective length of target times to weight the calculation. As an illustration within the same supply chain described above, a buyer designates 22 days' worth of AD/UD codes. In the qualitative field, six process steps are found to exceed target. In the quantitative field, process H is found to be in maximal exceedance of its target. Six divides evenly into 22 three times with a remainder of four. Therefore, three days are removed from each of the above-target process times, and the remainder of four days is also removed from process H (removing a total of seven days from process H). Similar to the previous sub-section, AD/UD days will not be reintroduced when making predictions as they do not enter the data set proactively.

Assumed inclusion. This is the easiest solution to implement: do nothing. If codes are stored in bulk, there is no true way to assign them to a specific process time short of pro-rating (discussed above). In this circumstance, on can assume that any delays that would require codes are simply part of historical process times, and therefore any process time computed is inclusive of the expect delays that said order or shipment will incur on a historical basis. In this circumstance, it assumes that shipments with like categorical variables will face like delays. It does not account for the changes in process that would eliminate delays.

For example, all orders sent to a specific vendor take five days for confirmation to occur due to some internal process specific to that vendor alone. Let us assume that this is assigned a four day acceptable delay code on top of the one day that it usually takes to confirm an order. The vendor changes its processes to eliminate this costly internal process. Depending on the magnitude of historical data behind that process time, the model will continue to predict five days for that process to occur for a potentially substantial amount of time before the one day process to filter into predictions. The model has no way to account for that process change other than to use a relatively short time frame of historical data to induct the model.

Once the decision has been made as to how to calculate process times and which categorical variables link to each process step as discussed in the previous discourse, a decision must be made about how to compute predicted process times for each child group of the parent data set.

Mathematical procedures for predicted process time calculation. The end purpose of variance testing and data cleaning is to find the best possible way to group the data in subsets in which lines are expected to form similarly. In this way, when new lines are created for which the system must generate predictions, one can align that line with the subset it fits most closely. Historical performance from other members of that group can be used to predict the performance of this newest member. The question remains: how do we compute the predicted process time of those lines of data within a child group? There are many options, depending on the nature of the data, the supply chain, and the managers involved. Please note that this text will focus primarily on the decision points surrounding the use of that calculation, rather than the intricacies of performing every step of that calculation.

Averaging. The simplest way to achieve this is of course with an average. At this point, there should be no weighting in the data with which one can be concerned, so one can simply

sum all the available process time and divide by the number of process times. This methodology is most successful when shipments and orders are expected to perform roughly the same across time and throughout all seasons. This option is also best for data that is highly segmented, where there are not enough process times to generate enough data points for an accurate regression. The only note of caution worth adding when using this calculation is that it is best to use a more narrow set of training data, since a blunt average assumes that all historical values are of equal importance. Too much historical performance on a supply chain that has been steadily improving or worsening will factor too much historical performance that may no longer be relevant into the process times.

Exponential smoothing techniques. Exponential smoothing techniques can be used to base a future predicted process time off of past data where the most recent observations in the parent data set are given more weight than the older observations. This allows us to factor in performance trends and seasonality when predicting process times in the future. This weighting is realized through smoothing constants. *Smoothing constants* determine the level at which previous observations influence a prediction.

Larger constants create faster changes in a fitted line, while small weights create slower changes in the fitted line. Accordingly, the larger the weight used, the closer that the smoothed values will follow the data resulting in a less smooth line. Equally, the smaller the weight; the smoothed values will create a smoother line resulting in slightly greater divergence from the data. As such, small weights are usually recommended for series that contain a high level of noise around the signal pattern, while large weights are usually recommended for a series with a small noise level around the pattern. Smoothing constants are used in each of the following procedures. It is up to the implementing management team looking at the data at hand to determine where to set smoothing constants for creating the best predictions.

A strong empirical way to select a smoothing constant (α) exists. For each possible value of α (generally between 0 and 1, by steps no larger than 0.2), a set of predicted process times is generated using the desired smoothing procedure. Next, these sets are compared to the actual observations within the time series. The value of α that returns the smallest sum of squared errors is selected. Evolutionary algorithms are easily used and highly recommended by the author to optimize α. This optimization for one-step error should take place for every one of the methods selected below for every predicted process step. As it applies uniformly to each calculation, it will not be repeated below.

Also, when considering exponential smoothing techniques in supply chain predictive analytics, one must pay attention to the dates used to parse each line into the correct time period. While forecasting procedures which use similar techniques only have to select one date to use, in this model a different date must be used for each predicted process time. That date will always be the initializing date of the process. Returning to the process calculation equation used in variance testing ($d\text{-}d_{1.}$), the parsing date to determine the time frame for each line of data is always going to be the d_1 of that process.

Managers will also need to decide the span of time frames to use in analysis. Does it make more sense to make predictions based on weekly time periods, monthly, quarterly? This depends on how sophisticated a methodology is in use, how much change one can expect to see across the period, and how segmented the data are. Heavily segmented data will need to use a longer time period in order to ensure there are enough lines of data for each time period. Less segmented data can use more segmented time periods in order to spread out the data into small enough chunks to closely follow trends and seasonality.

Simple exponential smoothing. Simple exponential smoothing (SES) only uses one smoothing constant. It assumes that the data, once broken out into its time series has two

components: (1) a level, and (2) some amount of error around that level. A *level* is the straight mean of the data points, assuming that all historical values are of equal importance. The predicted process time is simple to compute. For a given time period, the process time is the level plus the error around that level at that given time period.

But as levels shift over time, equal weighting across historical data makes less sense. A more calculated way to do this exists. First, we find the initial estimate of the level: $level_0$. $Level_0$ is the mean of the process time for those lines of data in that time period in that subgroup. We then use $level_0$ to assume $level_1$, that same subgroup in the next consecutive time period. We find the difference between $level_1$ and the actual mean process time of $level_1$. This error is factored into the equation as follows:

$$level_{current\ period} = level_{previous\ period} + \alpha \times (mean_{curent\ period} - level_{previous\ period})$$

α is the smoothing constant discussed at the beginning of this subsection. Once this is computed for all time periods across the training data, the final estimate of the level is what would be used as the predicted process time for future time periods. This can be seen in Figure 10, where the red line shows how the calculation generates a flat line for all future periods. At the last level, the latest process time observations will count more than those previous which have had their error adjustments multiplied by α as many times as the count of time periods. This final level estimate is going to be the best one available to predict the future. . This prediction may be very accurate in the short term, but will likely lose accuracy as the model tries to make predictions for time periods further into the future (Foreman, 2014).

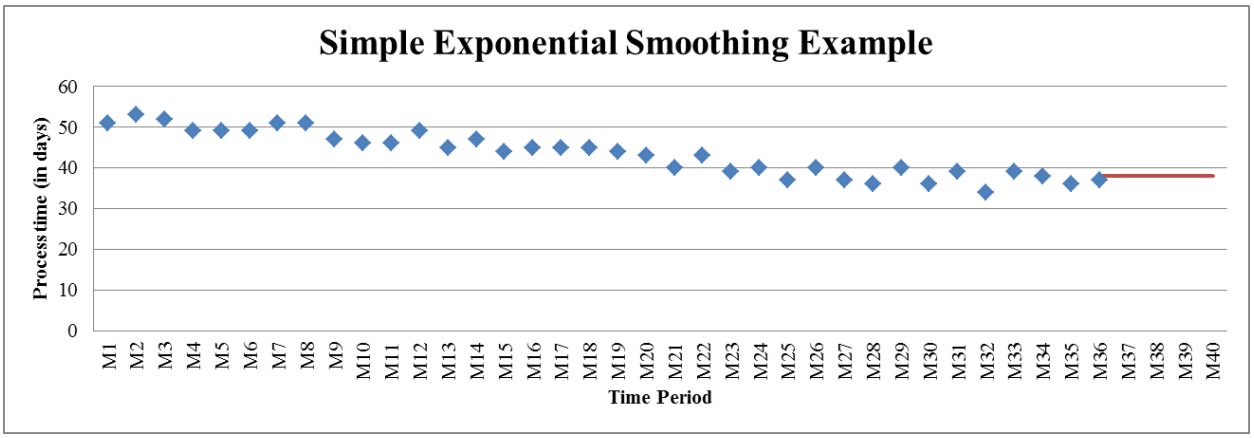

Figure 10. Illustration of predictions made through simple exponential smoothing

Holt's trend-corrected exponential smoothing. Holt's trend-corrected SES expands on SES as described above to create predictions from the data with a linear trend. Before deciding to use a methodology that accounts for trend, it must be demonstrated that one exists. To determine whether a trend exists, a linear regression must be fit to the data. Next, a *t test* must be performed on the slope of the trendline to determine statistical significance. If the slope of that line is nonzero and has a p value less than 0.05, one can assume a level of confidence around the fact that the data has a trend. Next the slope of the regression line must be determined. A positive slope indicates that process times are going up and things are requiring a longer lead time. A negative slope indicates that orders and shipments are moving through milestones in more streamlined ways.

In this calculation, the demand at a given time period is going to be equal to the level plus the trend at that given time in addition to the random error that we expect to see. Accordingly, the equation for calculation is:

$$level_{current\ period} = level_{previous\ period} + trend_{previous\ period} + \alpha \times (mean_{curent\ period} - (level_{previous\ period} + trend_{previous\ period}))$$

The trend equation is very similar, but adds another smoothing factor: γ. γ multiplied by the amount of error incorporated into the same level update is used to adjust the level based on shifting trend estimation. Trend is calculated as follows:

$$trend_{current\ period} = trend_{previous\ period} + \gamma + \alpha \times (mean_{curent\ period} - (level_{previous\ period} + trend_{previous\ period}))$$

This creates a predicted process time that accounts for trends in supply chain performance, as shown in Figure 11.

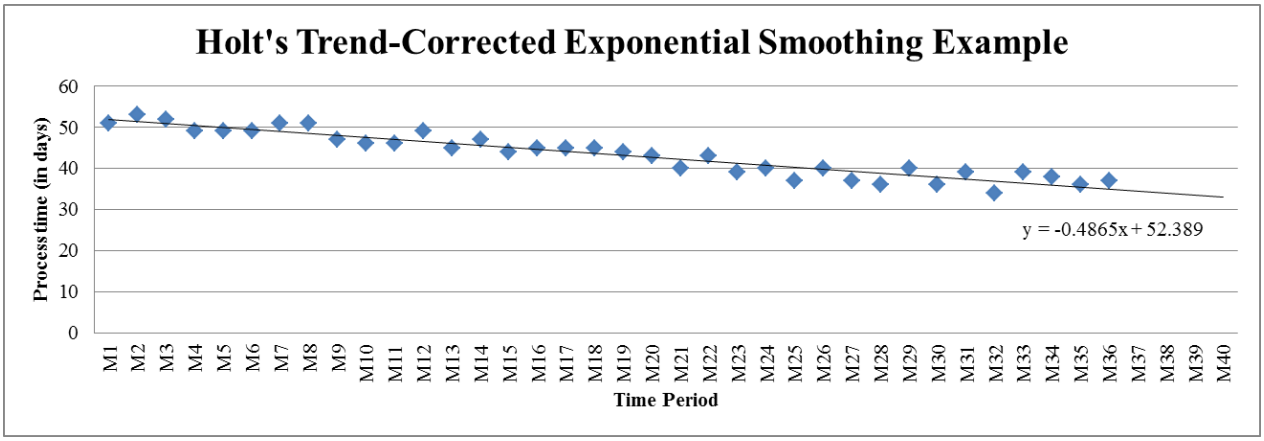

Figure 11. Illustration of predictions made through Holt's trend-corrected exponential smoothing

Multiplicative Holt-Winters exponential smoothing. After optimizing α in the previous kind of calculation, the one-step errors remaining should be random. However, if a pattern is hidden in those errors at a repeated interval, this may be an indication of a seasonal trend. To confirm this, we need to look at autocorrelations. *Autocorrelations*, also known as serial correlations, are the cross-correlation of a signal with itself. It indicates the similarity between observations as a function of the time periods extant between them. This mathematical tool aids in finding repeating patterns, such as the presence of a periodic signal obscured by noise. In this context, it can be used to find patterns in supply chain performance. For a detailed explanation

of methodology, the author recommends Foreman (2014). If the resultant correlogram created by this analysis indicates a seasonal cycle, than this version of exponential smoothing is appropriate for creating predicted process times. Seasonality, it should be noted, does not necessarily have to fall on a 12 month scale.

Seasonal adjustments, unlike the adjustments we have seen previously, are multiplicative, not additive, hence the title of this section. The premise behind this kind of exponential smoothing is that demand at a given point in time is equal to the level plus the trend at a given period multiplied by the seasonal adjustment for a given period and random error.

$$level_{current\ period} = level_{previous\ period} + trend_{previous\ period} + \alpha \times \frac{(mean_{curent\ period} - (level_{previous\ period} + trend_{previous\ period}) \times seasonality_{last\ relevant\ period})}{seasonality_{last\ relevant\ period}}$$

As with double exponential smoothing, the trend is updated in relation to the level in the same way (Foreman, 2014).

$$trend_{current\ period} = trend_{previous\ period} + \gamma + \alpha \times (mean_{curent\ period} - (level_{previous\ period} + trend_{previous\ period}))$$

Similarly, the seasonal factor must be updated. It is very similar to the trend update equation, but it instead adjusts the last relevant seasonal factor using δ, another smoothing factor, that was ignored up to this point (Foreman, 2014).

$$seasonality_{current\ period}$$

$$= seasonality_{last\ relevant\ \ period}$$

$$+ \delta \times (1 - \alpha)$$

$$\times \frac{(mean_{curent\ period} - (level_{previous\ period} + trend_{previous\ period}) \times seasonality_{last\ relevant\ period})}{seasonality_{last\ relevant\ period}}$$

After optimization, autocorrelations must again be checked. If there are no correlations above the critical value (p = 0.33), the model is capturing the structure of process times fluctuation well (Foreman, 2014).

As a final note, whatever method is selected must be done for all child subsets available for each process that the system will predict. More importantly, this needs to be able to run each time that they process times matrices are updated, so it cannot be a manual process. If a supply chain does not have the mathematical tools in place to manage this kind of analyses in a fully automated way with relatively quick speed across very large data sets, than this methodology will collapse under the weight of its own complexity. While a model that accounts for every nuance is most desirable for making accurate predictions, a model that is too complex for its environment will never successfully integrate into a larger management structure.

Process times matrices. Once all of these predictive process times have been calculated, they must be assembled into a matrix. This matrix serves three purposes. First, it gathers process times in a single place which provides a way for the model to make predictions without rerunning every calculation made previously. Second, this matrix provides a single place towards which buyers and managers can turn when quoting lead times for shipments during the order process. Thirdly, it provides a place in the model to account for certain contingencies.

A *contingency*, in this context, is a logical rule that is used to override the extant logic of the system. These contingencies are used to mitigate lines of data where predictions cannot be

made due to lack of available historical data. This occurs when a new categorical variable is entered into the system.

Regional process time contingencies. Regional contingencies occur when the driver combination includes a geographic variable, like place of origin and/or shipment destination. They are used in the event that the driver subgroup does not provide enough information to compute an average process time. In this event, the regional variable replaces the more specific geographical variable, which creates a larger pool of data from which to computer. The process time is then assigned from the regional subgrouping which is computed previously alongside the non-regional drivers and only used when this contingency is triggered.

This is a simple concept when there is only one geographic variable to replace. For a given process, the drive is created from three variables: product type (t), shipping method (m), and destination country (d), If the contingency is triggered by a lack of data, the following would occur to provide a bigger group of data for calculation:

$$t\ m\ d_{geographic} \rightarrow t\ m\ d_{regional}$$

However, when there are two (or more) variables to replace. It is best to approach regional variables iteratively. For example, process K is predicted by a categorical driver with two geographic variables: (1) origin or pick-up country (p), and (2) destination country (d). During independent variance testing origin country was shown to reduce variance by 8%, and destination country reduced it by 17%. This means that destination country is likely a slightly better predictor of performance than origin country. In that event, the variance reduction offered by this variable should attempt to be retained as long as possible while entering into a contingency prediction for the other geographic driver(s). In the event that this does not provide a prediction, the variables can be reversed. Finally, both can be tested as regional drivers to

provide a solution. In the illustration below, ‘g’ denotes a geographic driver, while ‘r’ denotes a regional one.

$$p_g d_g \rightarrow p_r d_g \rightarrow p_g d_r \rightarrow p_r d_r$$

In the unique scenarios where more than two geographic variables are used in a driver, the concept is similar. In the following example, origin and destination country are joined by a third geographic variable, x. Variance reduction showed that x reduces variance more than p, but that d is the primary reducer of variance in this driver. Rather than fall back through every possible combination of geographic and regional variable, this model will simplify the combinations based on the assumption that the geographic variable for d should be retained as long as possible.

$$p_g x_g d_g \rightarrow p_r x_g d_g \rightarrow p_g x_r d_g \rightarrow p_r x_r d_g \rightarrow p_r x_r d_r$$

One should note that other contingencies that for non-geographic variables can be written and used in the same manner. For example, perhaps a supply chain uses a relatively large range of vendors, meaning that new vendors are frequently entering the data and requiring predictions. Reflecting on the selection of categorical driver discussion, we know that such an unstable field quite likely will provide difficulty as a categorical variable since it will likely oversegment the data. However, if this piece of information is still deemed relevant to the model, the contingency above can be used. Perhaps the process in question relates to manufacturing time. It can be assumed that while the vendor in question plays a definitive role in determining how long an item takes to manufacture, that the kind of product being manufactured plays a similar, if less precise role in determining manufacturing process times. In this instance, the vendor (v) would take the place of the geographic variable, while the product type (t) would become the regional contingency. If this process also included origin country (p), the contingency would look as follows:

$$p\ v \rightarrow p\ t$$

In the event that the process used both origin country (p) and product type (t) as variables, translating vendor to product type would become duplicative. Instead, the vendor variable can be allowed to drop off entirely to create more generalized predicted process times when new vendors enter the system.

$$p\ v\ t \rightarrow p\ t$$

Target process time contingency. The target process time contingency is the last and least preferred way to derive a predicted process time. In the event that all other possible contingencies have been exhausted, a predicted process time can simply revert to a designated target process time. These process times are also used in the event of prorated AD/UD code exclusion and in generating performance metrics, such as whether the shipping process occurred within the target lead time.

Prediction Generation

Prediction generation, after all of the math it takes to get us to this point, is an elementarily simple process. There are two basic concepts to keep in mind about prediction generations. First, predictions are created additively, by adding the predicted process time in days to the previous date. For example, if the previous date occurred on May 3 and the process is predicted to take 4 days, the prediction for the next milestone date would be May 7. Second, predictions fall like dominoes, one after the other. This means that most predicted dates will be built off another predicted date, which is built off another predicted date, etc. This is important to note because prediction mechanisms in the implementing software must be designed to make predictions based on other predictions while utilizing different predicted process times. This need to have unique drivers for each process time but be able to use them additively creates a requirement for the process times matrix described above.

Types of blank dates. There are three primary reasons that a date might be blank within the source data set. First, the date may simply have not been entered or was entered incorrectly. Second, the order or shipment might have been cancelled or short-closed, meaning all future dates are blank. Third, the order or shipment is in process, but has not yet reached the milestone date. It is only in the third case that a prediction must be made. A set of logical rules is easily derived in order to pull out making unnecessary predictions on the first two kinds of blank dates in the data set, so that only legitimate predictions are being made.

Prediction contingencies. Like with process time calculations, it is possible to add contingencies in the prediction process. Whereas with the process time calculations, contingencies are used to attempt to make accurate predictions in the face of lacking data, prediction contingencies are used to make predictions more accurately model real-world events.

Promise date contingencies. Promise date contingencies are a simple logic used to update a prediction to a specified promise date in the event that the prediction falls short of the promise date. For example, a vendor promises an order to be prepared for shipment on September 2. Given the historical performance of that vendor, the model predicts the order to be prepared for shipment on August 12. In this event, the logic uses September 2 instead of August 12 as the predicted date for that process milestone. This is done based on the assumption that the promise date is contractually binding and determined within the context of that individual order. Historic performance can be used to make accurate predictions, but it cannot take into account such facts as a charter plane was hired for the second week in September, so it is highly unlikely that the order would be fulfilled a month early. This is also a beneficial contingency when vendors are given orders well in advance of their manufacturing lead time, as it correctly excludes the period of time in which the order sits waiting for manufacturing to begin. Conversely, if the predicted date falls beyond the fulfillment date it is allowed to remain as such.

This provides an alert to managers to let them know that this vendor may exceed the fulfillment date given their past performance.

Another way this contingency could be implemented is to look at the delta between promise date and delivery date. In this circumstance, each subgroup is measured not on the amount of time that it takes for them to undergo the manufacturing process (which may differ greatly depending on a variety of factors), but rather on the number of days early or late that they tend to be with regards to the vendor promise date. With this contingency in place, once a vendor promise date is entered into the system, a vendor's historical performance against that date would populate the predictions of the rest of the process. Before the entry of the promise date (which may not be available in the earliest stages of an order), the predicted process time would be used to predict that date.

Today's date contingency. This contingency exists to handle orders and shipments that exceed their allocated process times. The purpose of this contingency is to ensure that every prediction the model makes is a future date. This contingency states that when a prediction is made that falls into the past because a shipment has stalled, that the first predicted milestone on that line is updated to today's date. Since predictions fall like dominoes, this ensures that every prediction following it will be updated to reflect that delay. As an example, a shipment was predicted to arrive at port on January 17. It is now January 18, and the shipment has not arrived. Instead of the prediction remaining static at January 17, it is updated to January 18. If the shipment still does not arrive today, it will be updated tomorrow to January 19, moving out the predicted final delivery date accordingly. This will continue to occur until the shipment moves into the next process. In the event that there is a lag between milestone occurrence and the data being entered into the system and flowing into this data set, a buffer can be built into this contingency to avoid updating milestones that quite likely happened, but are not yet recorded in

the data set. In this circumstance, assuming a buffer of five days, the today's date contingency would not begin taking effect until January 22.

Weekend contingency. This contingency exists to avoid predicting dates on days of the week where they are not likely to occur. For example, order stages are not likely to be completed on a Saturday or Sunday. Shipment stages may however happen on a Saturday, but not a Sunday depending on the entity responsible for delivery. Optionally, holidays can also be included. Once these business rules are determined according to the realities of the supply chain, a system logic can be written to move that predicted date to the nearest business day. One should note that the system is only concerned with not predicting on a weekend day, there is no discounting of weekend days in the initial process calculations or the predicted process time, since the same amount of weekend days can be assumed to occur in both time periods.

Delay code contingency. The final contingency that will be discussed involves the reintroduction of delay codes into the prediction. This can only occur according to the parameters described above regarding the calculated exclusion and reintroduction of AD/UD codes. This contingency is engaged when an AD/UD codes is entered into the system for a date that is currently being predicted. In this circumstance, the prediction would be moved forward to include those AD/UD days through addition. For example, a shipment is currently predicted to arrive at its destination on November 1. However, the shipment was rerouted due to flooding across critical piece of infrastructure. The buyer proactively added an AD code of three days to account for the truck's rerouting. The prediction would be updated to November 4. At times, the utility of this contingency can be lost under the today's date contingency, particularly is the order is in a relatively short process. Where this contingency becomes most useful is in managing processes with longer process times, where delays are incurred in advance of the triggering of the today's date contingency.

Summary

This rather substantial chapter walked through the key phases of model induction and deduction with the intent of presenting a variety of algorithmic options that are designed to be scalable across different supply chain circumstances. Five key sections are addressed within: (1) categorical drivers, (2) variance testing, (3) data cleaning, (4) process times calculations, and (5) prediction generation.

Categorical drivers are used to segment historical data for the purposes of determining the factors that driver process times. Categorical variables should be selected based on four qualities: (1) performance impact, (2) uniform presence, (3) multi-line inclusion, and (4) historical repeatability. The assignation of regional categorical variables to all variables that are geographic data by nature allows for higher level groups that will enable fall back contingencies when predicting process times. Regional variables can be assigned either by geography or through country profiling, which can happen qualitatively through assigned factors, or quantitatively though statistical analysis.

Once categorical variables are selected, variance testing can occur. Variance testing is used to ascertain the relative importance of each variable to the variance of the process times in the subgroups that said variable creates. While ANOVA tests and other non-parametric statistical analyses are not acceptable for these data because of the failed assumption of independent random samples, average weighted variance analysis can be used. This testing computes the variance of child subgroups as weighted by the number of lines that fall into that subgroup. First, independent testing is done. Pursuant to those results, iterative combined testing is used to determine how to best combine variables into drivers that reduce the amount of variance within their process step and subgroup without oversegmenting the data.

Next the further stages of the model require some preliminary data cleaning. First, one must ensure that there is not weighting in the data set that would skew results based on duplicated performance. Next, outliers must be removed. Lastly, system logics must be used to impute categorical variables if they are absent in the data set.

After the data are cleaned, process times can be calculated. There are many ways for process time calculation to occur. A mean of historical data within a subgroup is the simplest solution. However, exponential smoothing provides for more elegant solutions. Three were presented herein. First, simple exponential smoothing and its single smoothing constant were discussed. Next, Holt's trend-corrected exponential smoothing permitted us to account for an upward or downward trend in the data via a second smoothing constant. Lastly, multiplicative Holt-Winter's exponential smoothing allowed for both the trend correction described about in addition to the aspect of seasonality through a third smoothing constant. This concept of contingencies is introduced as a way to use regional drivers and target process times to supplement historical data in making predictions. One should be careful to note that while models can be generated to accurately account for a variety of situations, no model should be used that is too complex for the software or management environment in which it exists. Even the best models, when implemented in the wrong place, can become extinct.

After process times are calculated, they must be gathered into a process times matrix. This matrix helps with the creation of predictions in the next stage, as well as in generating an overall lead time matrix for buyers and managers to use when quoting lead times. Once a matrix is in place, predictions can be made. Predictions operate on a simple additive model (previous date + time) and are designed to fall like dominoes: one after the other. This theory discusses four prediction generation contingencies: (1) the promise date contingency, (2) the today's date

contingency, (3) the weekend contingency, and (4) the delay code contingency. Each of these contingencies serves to create a model that is flexible enough to handle supply chain uncertainty.

Through the integration of AD/UD codes and prediction contingencies, this model is able to change course on historical data and make predictions in real-time with a level of flexibility. This flexibility makes a model built around these broad concepts that is ideal for responding to the uncertainties that exist in supply chain management. Of course, now that we have a model built of training data and calibrated to the idiosyncratic needs of its supply chain, the larger challenge comes: creating the process and management structures necessary to make the most us of this predictive tool.

Discussion

People begin their search for answers most often with the purpose of confirming a truth that they are predisposed to accept. People, in other words, tend to find the things that they are looking for. An anecdote becomes a theory. A limited observation becomes an overarching assumption. Experience, in this way, becomes expertise. This is not a bad thing. This expertise cannot be considered nullified by virtue of its observational origin. However, it also cannot be considered exhaustive as it lacks the understanding wrought by detailed analysis. Experience grows organically and iteratively based on the subjects in which the holder of said experience chooses to observe. This is an accurate level of expertise to select a new household appliance, but less so for decision-making in a larger and more nuanced system.

Predictive analytics offer temperance to this system of thought. Instead of observing and reacting, one can glean insight from data in an effort to focus those reactions towards data-validated observations, rather than the more limited scope offered by the focus of a single manager. At times there exists in business a dichotomy between what managers hold to be the source of a problem and what evidentiary support proves to be true. For example, a manager might note that shipments are more likely to be late in the winter months of the year and blame delays on disrupted transit due to weather. However, data indicates that the delays are incurred in the customs clearance processes for the shipment due to customs officials taking time off for the holidays. Holding the preliminary observation as true without further analysis and data-driven evidence means that corrective actions taken to improve circumstances are actually focused around and incomplete understanding of the problem.

Predictive analytics offers some nullification to this dichotomy. It provides managers with real-time insight into the going-on of complex systems, permitting their expertise to grow through a data-driven context in addition to an experiential one. To do so, managers must have

at their hands a strong predictive model for their supply chain: one whose findings they trust and upon which they are willing to rely to make their experiential observation. With this in mind, the most important part of generating a predictive analytics model for a supply chain is not in fact the math or the data. Despite the focus of the preceding chapters, the most important facet of a system such as this is the buy-in of operational management. In many cases, it is important that models be understood and easily explained. This is both useful for the team that is building them, but also for communicating results to stakeholders and managers not knowledgeable about the field of data modelling. Given this truth, it is generally best to create a model that works well and can be easily explained, even if it is not the most accurate model that one can produce from a given data set (Provost & Fawcett, 2013).

Research Questions

Despite the quantitative data science of the previous chapters, the research questions that designed this study are qualitative. There are no null or alternate hypotheses to reject.

Research question one (RQ1). How might historical supply chain data be used to predict future supply chain performance?

Sub-question one a (RQ1a). What categorical variables affect supply chain performance across different process steps?

Sub-question one b (RQ1b). How does the model deal with incomplete data?

Sub-question one c (RQ1c). How might process times be calculated depending on data availability and supply chain patterns?

Proposition one. A set of training data built from historical supply chain deliveries is used to induct a model. This training data contains, at minimum, a set of categorical variables and numeric target variables (in days) for each date that the system must predict. In order to understand which categorical variables impact the performance of orders and shipments at each

stage in the process, weighted average variance testing of categorical variables is done. These variables are then combined to create categorical drivers which segment the data in order to make process time predictions. Incomplete and outlying data are handled through data cleaning, with removes outliers and data quality errors while imputing categorical variables based on business rules. These imputed categorical variables are not used to change the source data, but rather to sort the line of data into its most likely subgroup in order to facilitate predictions being made for that line before all of the categorical variables have been assigned.

The first step to determining how to calculate predictive process times is to determine how to handle delay codes. This theory offers four possible solutions based on the binary options for two parameters. Depending on the way that delay codes are stored in the system and the timing with which delay codes are assigned to a shipment, different options present themselves: (1) calculated exclusion and reintroduction, (2) calculated exclusion and no reintroduction, (3) prorated exclusion and reintroduction, and (4) assumed exclusion.

There are a variety of options discussed as for how to calculate predicted process times. The simplest solution is a blunt average of historical data. This option is best where there is a relatively short time span of data, or data becomes oversegmented when split into time periods for a more elegant quantitative solution. This simply solution contains no exponential smoothing and assigns equal value to observance in the historical data, despite its age.

Borrowing techniques from forecasting, predicted process times can also be calculated through exponential smoothing. There are three kinds of exponential smoothing discussed in this metamodel. This first is simple exponential smoothing (SES). SES, like a blunt average, assigns the same value to all future predictions, regardless of time frame. However, SES weights more recent observations heavier than past observations. SES uses a single smoothing constant, which must be determined though optimization.

If the data appears to have a trend that is confirm by linear regression and t-testing, then Holt's trend-corrected exponential smoothing is in order. Through the use of two smoothing constants, standard error is built into the prediction as well as the general trend of performance. Therefore, future predictions will be expected to carry on similar trends, and the number of days predicted will vary across time periods.

After analyzing the data by trends, if a cyclical pattern appears via testing for autocorrelations, multiplicative Holt's-Winters exponential smoothing is in order. This technique uses three smoothing constants and is able to account for both trending and seasonality when making predictions. Like trend-corrected exponential smoothing, these equations weight more recent performance heavier than past performance. However, future performances will not be a smooth, sloped line. This smoothing will make predicted process times based on a regression that spikes or drops cyclically as according to seasonal demand.

After selecting a methodology by which to compute predicted process times, process times must be computed for every process step for every subcategory. If there are twenty milestone dates to predict and the categorical drivers, on average, break a parent data set into 85 subgroups for analysis, 1,700 process times must be computed. If regional drivers are being used, those process times must also be computed. If multiple regional contingencies are in effect, each possible combination of regional and geographic drivers that was decided upon must be run as well. This could result in several thousand calculations for this system to make.

Next, a process time matrix is compiled. If the first two techniques are utilized (averaging or SES) to compute process times, this matrix will be two dimensional comparing subgroups to process steps. If the other two (Holt's and Holt-Winter's) are utilized, then the matrix should be considered in a three-dimensional light, comparing subgroups to process steps to time periods of prediction. The process time matrix allows for contingencies to be used in

order to make predictions when sufficient data are lacking. This metamodel discusses a regional contingency and a target contingency. In the event that drivers have geographic variables in their drivers and cannot make a predicted process time, that variable is substituted with the corresponding regional variable and that broader regrouping of the data yields a predicted process time. If there are multiple geographic variables in a single driver, a logic in how to replace geographic variables is in place based on retaining the variable most contributive to reduction in variance for as long as possible. In the event that regional drivers are not able to return a prediction, then a target process time can be used. By the end of the chain of available contingencies, there should be no blank value in the process time matrix. Through these calculations and logics, the model is created. Once lines of data are entered, given the model built from these steps, supply chain milestone dates can now be predicted.

Research question two (RQ2). How might a predictive model be used to manage atypical supply chain events?

Sub-question two a (RQ2a). How does the model account for the unexpected, inconsistent, and/or unknown variables that effect shipments?

Sub-question two b (RQ2b). What other logical contingencies can be applied to predictions to account for supply chain realities?

Sub-question one c (RQ2c). What other logical contingencies can be applied to predictions to account for real-time intervention?

Proposition two. A model that can make prediction based on historical data is useful. However, a model that can make predictions on historical data with input from new information as it becomes available is far more valuable as an operational tool. There are several features that this metamodel considers as to how to achieve that real-time intervention.

Stalled shipments are handled through a today's date contingency. This contingency updates the predicted date to today's date in the event that it surpasses its original predicted date based on the last present date input into the system. This mechanism will always ensure that predicted dates fall into the future, while allowing the user to see how the delay in the first predicted date will impact the expected delivery of a shipment. This contingency can be implemented with a grace period to account for any habitual delay in entering data into the system.

Certain supply chain realities can be quantified with logical rules and therefore be added into the model. The weekend contingency provides one example. If a predicted date is scheduled to fall on a non-delivery day (like Sundays and perhaps Saturdays) or a holiday, the predicted day is moved to the next working day. The promise date contingency represents another supply chain reality that the model can accommodate. While a manufacturer's lead time can be computed from historical data, it does not always allow us to make accurate predictions, given that not all relevant data are present in the system. For example the model may not have the ability to understand that the buyer specifically sent the order early to the manufacturer, but requested they hold it for 100 days before beginning manufacturing. Their lead time may be the same, but the predicted date would fall 100 days early because it doesn't account for that wait time. This is mitigated by a logic that utilizes the vendor promise date for an order in the event that the predicted vendor fulfilment date precedes the promise date.

Lastly, real-time intervention also comes in the form of delay code contingencies. Delay code contingencies can be reintroduced when making predictions. By using delay codes that are linked to specific process steps and entered at the time they occur in the real world, the system is able to make predictions off the most up-to-date information that a buy or manager has.

Conclusions

This theory is designed to describe the multitude of ways a supply chain predictive analytics model could be built. Other options not contained herein surely exist that can be added in supplement. Accordingly, the theory steers very clearly away from the drawing of specific conclusions regarding the superiority of one decision made over another. Each option is right in its own context, and implementing a model based on the assumption that the most complicated solution is going to yield the most precision is incorrect.

Therefore, in an effort to conclude but to not limit one's potential implementation by biasing mathematical procedures by personal opinion, the conclusions drawn here will not take the form of what decisions should be made from this theory. Rather, we will discuss what kinds of conclusions one must make when attempting to implement the system described within. Thus, what follows are the questions a manager should ask themselves before undertaking a project such as this.

(1) How do I intend to use this system? This is an important foundational question. If the system is intended to tell clients when a shipment is arriving, there will be less tolerance for it being wrong as opposed to if the dates are used only for internal purposes. If the model is being used primarily to understand lead times across the supply chain, then more attention should be paid to accurate and descriptive process calculations, and less to prediction mechanisms and contingencies. If the model is going to be used primarily to tell buyers when something is at risk of late delivery, then perhaps a more conservative prediction mechanism is at order. Understanding the purpose of why this system is going to be implemented is quite like the most important question to answer in deciding how to build the system, which parts are the most important, and what will be done with it once it is ready to go.

(2) What are my technological limitations? This analysis requires a stable ERP and statistical software to run the basic analysis. If the ERP cannot integrate this kind of model, it must be housed outside of the system, which requires space to store it and people to maintain it. This kind of model and resultant workflow will look different than one built directly into the ERP. This is an important thing to know at an early stage.

In order to build a functioning model to begin making predictions, knowledge of data architecture, database design, and computational algorithms is also required. A strong understanding of the software resources supporting a supply chain can help manager's know what their capabilities are as it comes to building and using data models in their existing environment.

Open source software can be used to mitigate the cost of multiple software licenses (i.e. R instead of SPSS for statistical analysis; MySQL for databases). However, the tradeoff with open source software is that it is not an out-of-the-box piece of software. This means that this is generally a sharper learning curve as this software is designed by developers for other developers. Fortunately, vast and growing knowledge bases exist online to support the use of these systems. Essentially, it becomes a question of spending money on the software license or spending money on the qualified personnel to write open source code.

(3) What are my staff limitations? While this theory was designed to broadly present the options and structures available, and the predicates and choices surrounding those options; it is by no means a complete textbook on the subject. Knowing that a Kruskal-Wallis H Test could be run presuming there is independent random sampling is very different than knowing how to run a Kruskal-Wallis H Test, or knowing how to pull and manipulate the necessary data to run this test correctly and with the right tools. There is a level of quantitative, technological, and statistical expertise that is assumed in the carrying out of the broad concepts portrayed in this

text. Without personnel resource(s) available with sufficient understanding of the supply chain and its relevant data systems, this model will become very time-consuming to implement.

(4) Am I collecting the right data to induct this system? This is an indication of whether or not the data collection tools within a supply chain are sophisticated enough to even consider a predictive analytics model, or if this kind of data analysis is primarily aspirational. This can be answered by reviewing the 'Research Methods' section on data source requirements.

(5) How will I measure the accuracy and precision of this system? There are many ways margin of error can be measured in a system like this. This theory steers away from detailing options because of the dependency those options have on the many possible ways a model can be built as described herein. Margin of error calculations can be set up to measure how well the system can predict the training data based on its current algorithms. This can be a good indication of its accuracy when predicting lines outside that group of data. Calculations can also be run to determine that accuracy of the delivery date based on the first predicted delivery date when that line entered the data set as opposed to the last prediction made before the shipment was delivered. This kind of measurement would indicated how much more accurate the system grows as more information is added to it throughout the milestones. If one process step is markedly unsuccessful at predicting dates, then that is an indication that its categorical driver should be revisited. If one process step seems to be particularly inaccurate, all process steps after it will likely be inaccurate due to the domino effect. Perhaps, in that event, it makes more sense to look at the accuracy of predicted process times as opposed to the accuracy of predicted dates. These are just a few suggestions, given that another entire study could be devoted to the best ways to quantitate a predictive analytics model's success.

(6) How can I implement managerial processes to make this system most effective? The power of this model is not the individual calculations it makes or the dates it predicts. Its

ability lies in the way this information can be used. At this point, it is good to envision how a model like this will fit into performance management processes. A data set like this would nicely facilitate an alerts-based management system which would enable managers and buyers to act on problematic shipments while there is still time to find a solution. The process to do this is simple, but the business rules surrounding its accomplishment are not. Under what conditions are alerts generated? When should alerts be sent? To whom are they sent? All of these kinds of questions fall into this bucket and should be considered proactively.

(7) How do I intend to visualize this data? This question is very intentionally saved for last because it is an important this to consider but there are more important things to consider first. For data to be widely consumed by a non-technical audience, it must be able to be visualized. When interfaced with GIS systems, these data can be used to make map-based dashboards. When performance metrics are included, these data can be used to create traditional looking dashboards with up-to-date and future performance metrics. While one should not build their dashboards before they have fully designed their data set, this is an important thing to at least keep in mind in the initial stages both to help communicate the vision of this system to others, but also to ensure the system is enabled to move strategically towards this goal.

Recommendations and Implications

The theory defines a model of how predictive analytics models can be built for supply chains. Blending qualitative inquiry with quantitative solutions, a metamodel emerges that contains scalable solutions depending on the level of maturity and the requirements of a given supply chain. The value of models of this kind is in their ability to use extant data to generate insight. Essentially, they create a way of leveraging information to make information. This can be invaluable for supply chain managers as a way to mitigate some of the uncertainty inherent to supply chain management.

While this metamodel offers a wealth of options for constructing a model, one of the most critical premises is that a model should never be built without the business infrastructure to sustain it. This is to say that any complexity undertaken by this model is one that the supply chain's ERP or other technology is able to implement in a sustainable way with regards to staff level of effort. For example, while multiplicative Holt-Winter's exponential regression is a sophisticated structure for tracking fluctuating performance and trends across time, if an ERP cannot handle the sophistication of running that calculation potentially thousands of times to make a process time matrix, then it is non-optional in the business context in which it lives. Keeping this in mind, there is no guarantee that the more sophisticated operations that the model can implement will guarantee levels of predictive accuracy more significant than the simpler versions. Given that predictive analytics is a way by which we are able to teach machines to make the best possible educated guess about the future based on historical data; we must never lose sight of the fact that it is, at the end of the day, just guessing.

While a statistical model of supply chain performance offers remarkably more precision and a greater understanding of the ways in which supply chain performance occurs in real-time, there is some likelihood that it will not always be accurate. Further, this model is designed to exist in a feedback loop. One should have the ability to understand the mechanics of the model through the answering of certain questions:

(1) What percentage of lines are predicting off of target process times?

(2) In which process steps are predictions the most different from actual dates?

(3) What is the margin of error of this system?

(4) How are predictions changing over time?

By ensuring the model has the visibility to answer some or all of these questions and the great many other that could be used to analyze system performance, we are able to find ways to

refine the model. This is not a system that is inducted once and then left to run on its own devices. A successful implementation of this capability is inducted iteratively and refined throughout its usage. Categorical variables make change. Supply chain conditions may necessitate new contingencies. Business rules will evolve. Targets will require updating. In this way, a data model is like an organic creature. As it is used, it will divulge ways in which it can be used better. Incorporating that feedback is the most critical path for success in using data to generate insight.

Summary

It is exceedingly difficult to drive a car by looking solely through the real-view mirror. The picture of where an organization has been is not always the most useful information to have when trying to determine where to go next. And yet, many supply chains handle performance management in a reactive way. They respond to issues after they are identified through performance management processes with corrective actions and mitigation plans. While these plans are successful, what if a model could be built that would enable manager to see and understand tends as they are occurring?

Rather than looking at key performance indicators for the quarter and noticing that shipments managed by a particular office in India are being delivered late, a predictive model would be able to tell managers proactively that process times for certain actions managed by this office were creeping above target at an unacceptable level. In this way, the issues were identified early enough to put a corrective action in place in advance of the late delivery. A system like this could lessen the rate at which clients are dealing with supply chain uncertainty by notifying buyers and managers of it early enough to allow time for fixing the problem.

A predictive model can be designed by moving through the key decision points highlighted in this meta-model. First comes the ‘purpose of the model’ stage, one must map

their supply chain (i.e. understand what data is available and which dates will be predicted), determine how the model will be used, and what technology is available to support it. This lays the groundwork for understanding the unique ways the system must be implemented for a given supply chain. Second, categorical drivers must be determined. Third, the data set must undergo variance testing in order to assess which drivers contribute to the performance of an order or shipment in the supply chain. Fourth, the criteria for outliers must be determined and data cleaning logics must be assessed and written as part of data cleaning in preparation for the next steps. Fifth, the data must be segmented and undergo calculations in order to create a process time matrix that provides the raw material for generating predictions. At this stage, regional and target contingencies must also become available to allow the system to create predicted process times in the event that data is missing. Sixth, the system must make predictions on all of the active orders and shipments in the system, given a series of contingencies that provide for real-time intervention in the model. Seventh, the model must be put into use as dictated by its environment and purpose, and the decisions made up to this point. By moving through these stages, this meta-model is translated into a model unique to an individual supply chain as proscribed for its intended purpose. In this way, managers can create a system that allows them to turn their data into insight.

References

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Statistical Society*, Series A 160, 268–282.

Bhatnagar, R., & Sohal, A. S. (2005). Supply chain competitiveness: measuring the impact of location factors, uncertainty and manufacturing practices. *Technovation*, *25*(5), 443–456. Retrieved from www.journals.elsevier.com/technovation/

Brown, M. B., & Forsythe, A. B. (1974). Robust tests for equality of variances. *Journal of the American Statistical Association* 69: 364–367. Retrieved from: http://dx.doi:10.1080/01621459.1974.10482955

Christopher, M., & Peck, H. (2004). Building the resilient supply chain. *The International Journal of Logistics Management*, *15*(2), 1-14.

Courtney, H., Kirkland, J., & Viguerie, P. (1997). Strategy under uncertainty. *Harvard Business Review*, *75*(6), 66-79. Retrieved from http://www.ncbi.nlm.nih.gov/

Davis, T. (1993). Effective supply chain management. *Sloan Management Review*, *34*(4), 35–46.

Foreman, J. W. (2014). *Data smart*, Indianapolis, IN: John Wiley & Sons

Geary, S., Childerhouse, P., & Towill, D. R. (2006). On bullwhip in supply chains – historical review, present practice and expected future impact. *International Journal of Production Economics*, *101*(1), 2–18. Retrieved from www.elsevier.com/locate/ijpe

Goldberg, D. (1991). *What every computer scientist should know about floating-point arithmetic*, New York, NY: Association for Computing Machinery, Inc.

Hult, G. M., Craighead, C. W., & Ketchen, D. J. (2010). Risk uncertainty and supply chain decisions: a real options perspective. *Decision Sciences*, *41*(3), 435-458. Retrieved from: http://dx.doi.org/10.1111/j.1540-5915.2010.00276.x

Juttner, U., Peck, H., & Christopher, M. (2003). Supply chain risk management: outlining an agenda for future research. *International Journal of Logistics: Research and Applications*, *6*(4), 197-210. Retrieved from: http://dx.doi.org/10.1080/1367556031000162701 6

Kruskal, W., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association* 47(260): 583–621. Retrieved from: http://dx.doi:10.1080/01621459.1952.10483441

Kwon, O., Im, G. P., & Lee, K. C. (2007). MACE-SCM: a multi-agent and case-based reasoning collaboration mechanism for supply chain management under supply and demand uncertainties. *Expert Systems with Applications*, *33*(3), 690–705.

Li, J., & Hong, S.-J. (2007). Towards a new model of supply chain risk management: the cross-functional process mapping approach. *International Journal of Electronic Customer Relationship Management*, *1*(1), 91-107. Retrieved from: http://dx.doi.org/10.1504/JECRM 2007.014428

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. New York, NY: Eamon Dolan/Houghton Mifflin Harcourt.

Miller, K. D. (1992). A framework for integrated risk management in international business. *Journal of International Business Studies*, *23*(2), 311-331.

Peck, H. (2006). Reconciling supply chain vulnerability, risk and supply chain management. *International Journal of Logistics: Research and Applications*, *9*(2), 127–142. Retrieved from: http://dx.doi.org/10.1080/13675560600673578

Prater, E., Biehl, M., & Smith, M. A. (2001). International supply chain agility: tradeoffs between flexibility and uncertainty. *International Journal of Operations and Production Management*, *21*(5-6), 823–839.

Provost, F., & Fawcett T. (2013). *Data science for business* (1st ed.). Sebastopol, CA: O'Reilly Media.

Razali, N., & Wah, Y. B. (2011). Power comparisons of Shapiro–Wilk, Kolmogorov–Smirnov, Lilliefors and Anderson–Darling tests. *Journal of Statistical Modeling and Analytics* 2 (1): 21–33.

Ritchey, B., & Brindley, C. (2007). Supply chain risk management and performance: a guiding framework for future development. *International Journal of Operations and Production Management*, *27*(3), 303–323.

Savic, A. (2008). Managing IT-related operational risks. *Ekonomski anali*, *53*(176), 88-109.

Shannon, C. E. (1948). *A mathematical theory of communication.* Champaign, IL: University of Illinois Press.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika,* 52(3–4): 591–611. Retrieved from: http://dx.doi.org/10.1093/biomet/52.3-4.591

Simangunsong, E., Hendry, L. C., & Stevenson, M. (2012, August 15). Supply-chain uncertainty: a review and theoretical foundation for future research. *International Journal of Production Research*, *50*(16), 4493–4523. Retrieved from: http://dx.doi.org/10.1080/00207543.2011.613864

Snedecor, G. W., & Cochran, W. G. (1989). *Statistical methods* (8th ed.). Ames, IA: Iowa State University Press.

Van der Vorst, J. G., & Beulans, A. J. (2002, March). Identifying sources of uncertainty to generate supply chain redesign strategies. *International Journal of Physical Distribution and Logistics Management*, *32*(6), 409–430. Retrieved from: http://dx.doi.org/10.1108/09600030210437951

Vinodh, S., & Aravindraj, S. (2013). Evaluation of legality in supply chains using fuzzy logic approach. *International Journal of Production Research*, *51*(4), 1186–1195. Retrieved from: http://dx.doi.org/10.1080/00207543.2012.693960

Wagner, S. M., & Bode, C. (2008). An empirical examination of supply chain performance along several dimensions of risk. *Journal of Business Logistics*, *29*(1), 307–325. Retrieved from http://enterrasolutions.com/

Waller, M. A., & Fawcett, S. E. (2013). Click here for a data scientist: Big data, predictive analytics, and theory development in the era of a maker movement supply chain. *Journal of Business Logistics*, *34*(4), 249–252. Retrieved from: http://dx.doi.org/10.1111/jbl.12024

Welch, B. L. (1947). The generalization of 'Student's' problem when several different population variances are involved. *Biometrika* 34(1–2): 28–35. Retrieved from: http://dx.doi:10.1093/biomet/34.1-2.28.MR 19277

Wong, C. Y., & Arlbjorn, J. S. (2008). Managing uncertainty in a supply chain reengineering project towards agility. *Journal of Agile Systems and Management*, *3*(3), 282–305.

Yang, B., & Yang, Y. (2010, April). Postponement in supply chain risk management: a complexity perspective. *International Journal of Production Research*, *48*(7), 1901–1912. Retrieved from: http://dx.doi.org/10.1080/00207540902791850

Yang, B., Yang, Y., & Wijngaard, J. (2007, February 15). Postponement: an inter-organizational perspective. *International Journal of Production Research*, *45*(4), 971-988. Retrieved from: http://dx.doi.org/10.1080/00207540600698886

Yang, P., Qin, J., & Zhou, W. (2013, February). Multi-commodity flow and multi-period equilibrium model of supply chain network with postponement strategy. *Journal of Networks*, *8*(2), 389-397.