2017-06-01

The Application of Artificial Neural Networks for Prioritization of Independent Variables of a Discrete Event Simulation Model in a Manufacturing Environment

Rebecca Pires dos Santos
Brigham Young University

The Application of Artificial Neural Networks for Prioritization of Independent Variables of a Discrete Event Simulation Model in a Manufacturing Environment

Rebecca Pires dos Santos

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

Jason M. Weaver, Chair
Yuri Hovanski
Douglas L. Dean

School of Technology

Brigham Young University

ABSTRACT

The Application of Artificial Neural Networks for Prioritization of Independent Variables of a Discrete Event Simulation Model in a Manufacturing Environment

Rebecca Pires dos Santos
School of Technology, BYU
Master of Science

The high complexity existent in businesses has required managers to rely on accurate and up to date information. Over the years, many tools have been created to give support to decision makers, such as discrete event simulation and artificial neural networks. Both tools have been applied to improve business performance; however, most of the time they are used separately.

This research aims to interpret artificial neural network models that are applied to the data generated by a simulation model and determine which inputs have the most impact on the output of a business. This would allow prioritization of the variables for maximized system performance. A connection weight approach will be used to interpret the artificial neural network models.

The research methodology consisted of three main steps: 1) creation of an accurate simulation model, 2) application of artificial neural network models to the output data of the simulation model, and 3) interpretation of the artificial neural network models using the connection weight approach.

In order to test this methodology, a study was performed in the raw material receiving process of a manufacturing facility aiming to determine which variables impact the most the total time a truck stays in the system waiting to unload its materials.

Through the research it was possible to observe that artificial neural network models can be useful in making good prediction about the system they model. Moreover, through the connection weight approach, artificial neural network models were interpreted and helped determine the variables that have the greatest impact on the modeled system.

As future research, it would be interesting to use this methodology with other data mining algorithms and understand which techniques have the greatest capabilities of determining the most meaningful variables of a model. It would also be relevant to use this methodology as a resource to not only prioritize, but optimize a simulation model.

Keywords: discrete event simulation, artificial neural networks, connection weight approach, data mining.

ACKNOWLEDGEMENTS

I would like to thank God for the many blessings, enlightenment and courage that I received in these months; Dixon Duke Cowley for his great generosity in financing my studies; Jorge, Laurici, Camilla and Leonardo for supporting my decisions and for their love; Deirdre Paulsen for her encouragement and desire to help; Professor Charles Harrell for his sensitivity and great advisement during the first year of my masters; Professor Russell K. Anderson for his willingness to help by changing his software to fit my research purposes; Professor Jason M. Weaver for his advisement and direction; Professor Yuri Hovanski for joining my committee half way through and for his contributions; Professor Douglas L. Dean for being willing to be involved in my research, for his great insights and for inspiration even in personal decisions; Ruth Ann Lowe for always being willing to help and for her constant smiles and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

1.1 Challenges Faced by Companies Today

With more competition taking place in the market over the years, businesses have felt the need to reduce costs, improve service performance and satisfy customers. However, the high complexity that exists in manufacturing systems today makes it hard to accomplish those goals and be distinct in a competitive market. According to Çiflikli and Kahya-Özyirmidokuz (2010), even the most experienced engineer faces complex challenges in order to make quality consistent, costs low and lead time short. These complexities make it hard for managers to make accurate decisions about a business that will improve its performance.

Thus, the present research aims to develop a method for determining the most important factors to efficiency improvement of a manufacturing system. This will be done by the ranking of variables of a simulation model through the interpretation of artificial neural network models applied to the output of a discrete event simulation.

1.2 Simulation as a Tool to Help Managers Make Better Decisions

Decisions can be better made if tools are used to support the decision maker. Discrete event simulation is a proven tool for improving the efficiency of a system and helps managers make better decisions. This tool consists in artificially creating a set of conditions for a real situation in order to be able to study or experience it. It gives more confidence that a good

decision will be made, as it is tested beforehand and results are known. It also minimizes risk, saves time and reduces the cost of decisions made on the actual system. It has been applied successfully to diverse areas, from surgery training (Johnston et al., 2016) to investment evaluation (Freiberg & Scholz, 2015).

1.3 Data Mining Applied to Simulation

Simulation creates large amounts of valuable information that is not always taken into consideration. This information could be better used if data mining algorithms were applied in order to find hidden patterns in the data. Data mining algorithms are used to create models that can learn from historical data and make predictions on the behavior of a system.

One well known data mining algorithm is artificial neural networks. This algorithm has shown good results in predicting data. However, it has the downside of not being easily interpreted. However, some research has been done to make artificial neural networks more interpretable (Garson, 1991; Gevrey, Dimopoulos, & Lek, 2003; Olden & Jackson, 2002; Olden, Joy, & Death, 2004; Oña & Garrido, 2014). Understanding and interpreting the algorithm would make it possible for decision makers to speed up the process of improving the simulation model. This is due to the fact that the interpretation of the algorithm will make it possible to know which inputs have the most impact on the output of the system. Thus, this powerful algorithm added to discrete event simulation can be beneficial for decision makers in a manufacturing environment.

1.4 Connection Weight Approach

In order to extract more information from artificial neural network models, researchers have created different approaches to facilitate the interpretation of this algorithm. One approach

is the connection weight approach (Olden & Jackson, 2002). In this method a score is given to each variable that is part of the artificial neural network model. The score represents the importance of each input variable to the output of the model. The bigger the score is, the higher the impact of the variable in the outcome of the model. This approach can be useful in determining which variables will be more impactful on the output of the model being studied.

1.5 Variable Prioritization of a Simulation Model

The definition of an importance score to each variable of a model can be helpful in improving the performance of the business being simulated in a discrete event simulation model. The knowledge of which variables are most important can inform managers regarding where to focus their efforts to achieve the desired improvements.

1.6 Thesis Statement

The purpose of this research is to create artificial neural network models from data generated by a discrete event simulation model. This will provide a way to determine which inputs have the most impact on the output of a business. This would allow prioritization of the variables for improving system performance. A connection weight approach will be used to interpret the artificial neural network models.

1.7 Hypotheses

This study aims to confirm the following hypotheses:

1. The connection weight approach applied to artificial neural networks can be used to rank independent variables of a discrete event simulation according to their importance.

2. Manipulation of the most important variables ranked by the connection weight approach in a simulation model can lead to improvement in performance of a business.

1.8 Delimitations

The study is limited to prioritizing variables of a discrete event simulation model by using the connection weight approach to interpret artificial neural networks algorithms. It is assumed that artificial neural network algorithms can be applied to discrete event simulation data.

Furthermore, it is assumed that the simulation model created represents the manufacturing accurately. Thus the improvements observed in the model analysis will represent the improvements that will be observed in the real system.

The data used in the study is limited to the data generated by a simulation model created through the observation of a real manufacturing environment.

There are other data mining algorithms that could also be applied to discrete event simulation models aiming to prioritize independent variables. Some examples are linear regression, decision trees, random forest and others. However, it is not the purpose of this research to study other algorithms.

Scientists have developed different approaches to interpret artificial neural networks. Some instances are Garson's Algorithm, Partial Derivatives, Input Perturbation, Sensitivity analysis, Forward stepwise addition and others. However this study is limited to the connection weight approach created by Olden and Jackson (2002).

1.9 Definitions

Algorithm – A list of procedures that should be performed in order to solve a mathematical problem.

Artificial Neural Networks – Algorithm that make it possible for computers to learn from a dataset, create a mathematical model and make predictions. It imitates the learning process of the brain and has good prediction capabilities even on nonlinear data.

Big Data – Name given to the large amounts of data stored today due to the development of technology and low cost of data collection and storage.

Connection Weight Approach – Artificial neural networks are not easily interpreted. This method was created to interpret artificial neural network algorithms and to rank variables according to their importance.

Data Mining – Science that focus on developing techniques that can be applied to data analysis of big data.

Discrete Event Simulation – Method created to model the behavior of a system through the sequence of events by the use of a computer model.

Machine Learning – The science field that studies the learning process of machines in order to make it possible for computers to be smart and make decisions by themselves.

Neurons – The main unit of an artificial neural network algorithm. Each neuron is represented by a node and has an input and an output. Neurons are connected to each other throughout the artificial neural network sending information to each other and receiving as well.

Overfitting – A data model that can only explain a small dataset but is not applicable to similar datasets.

2 LITERATURE REVIEW

2.1 Introduction

The production of goods has an important role in the world's economy. According to data from the National Association of Manufacturers for every $1.00 spent in manufacturing $1.81 is added to the economy. This is the highest multiplier effect in any economic sector. Specifically in the United States, the National Association of Manufacturers affirms that manufacturers perform more than three quarters of all private-sector research and development and employ 9% of the American work force. Due to the importance of this sector, much has been done to improve its performance in order to increase growth.

However, this task is neither simple nor easy. Manufacturing systems have become more and more complex over the years, having performance goals such as cost reduction and high flexibility that are usually conflicting. In addition to the high complexity, the high competitiveness existing in the market today leaves companies with little margin for error. These factors and others added together make it hard for managers to make good decisions without the use of tools and methodologies that will guide the decision-making process. Some examples are discrete event simulation and artificial neural networks. These tools aim to help leaders see what they could not see otherwise.

2.2 Discrete Event Simulation

One tool that has been used for many years to support complex decision making processes is discrete event simulation. It is defined by Harrell, Ghosh, and Bowden (2011) as follows: “The imitation of a dynamic system using a computer model in order to evaluate and improve system performance.”.

The expansion in the use of simulation was made possible because of the development of computers. According to Sokolowski and Banks (2010), the wide use of simulation only occurred in the 90s when there was a boom in technology. It is stated by Robinson (2005) that during this time computer prices dropped much and this made it possible for the large use in both work and house environments. Likewise, it is said by Robinson (2005) that the new powerful computers facilitated complex models to be developed in a reasonable time. After the 90s, Sokolowski and Banks (2010) states that the tool that once was mostly used for military training could now be applied to different fields, from disease proliferation to human behavior.

This tool has been created specifically to support decisions about a process such as testing between different layout designs, whether, or not, to purchase new equipment, planning a new facility, scheduling, allocating resources, and others. Although this tool can be applicable to the most diverse decision process, not all problems should be solved with the aid of a simulation model. A definition of which criteria determine the application of discrete event simulation is given by Harrell et al. (2011). The criteria are:

1. The decision made has to be operational: This means that the problem involves a quantitative solution. It is not very applicable to behavioral analysis.

2. The process should be repetitive and well defined: If the process only happens once or if it is hard to be defined, then creating a simulation model will be complicated and not beneficial for the decision-making process.
3. Each step of the process should have variability and be interdependent: If there is no variability in each process, the solution to the problem can be easily determined and there is no need to spend resources in doing a simulation. Also if the processes are independent and changes in one variable will not impact the others, then simulation will not be helpful either.
4. The cost associated with the impact of the decision should be greater than the cost of simulating: If the costs of doing a simulation are greater than the costs of the impact of the decision, there is no point in using the tool, as it will only generate more costs.
5. The cost of testing in the real system should be greater than the cost of simulating: If it is cheaper to test in the real system than to simulate, then it is not worth it spending resources in doing a simulation.

The steps above are very important to understand the application of simulation. When they are taken into consideration, the tool is correctly used and will be valuable. This is because the use of a computer model that accurately represents a real system is very beneficial in a decision-making process. With a simulation model, anything can be tried out before it is implemented. According to Jun, Jacobson, and Swisher (1999) it is a technique that makes it possible for professionals to ask what-if questions. These questions will create scenarios in the simulation model that will represent possible solutions for a problem that the real system is facing. After the different scenarios are tested, their results can be compared and a more accurate

decision based on data can be made. Besides the benefits mentioned, other advantages of having a model are listed by Fishman (2013). They are:

1. It is easier to manipulate a model than the real system.
2. A model shortens the time required to perform an analysis.
3. Studying a model is generally less costly than studying the real system.
4. A simulation model permits the modeler to control more sources of variation than the study of a real system.
5. A model will lead to more understanding about the system studied.

However, a model is only useful if it accurately represents the business. A model is created based on a conceptual understanding of the system being simulated. This understanding is many times called the conceptual model and is created through the observation of the system. According to Banks, Carson II, and Barry (2005) the conceptual model represents the assumptions and hypothesis about the system being modeled. This conceptual model created in the head of the modeler can be accurate or not. In order to make sure it is correct, it is important to validate the model. The validation process explained by Harrell et al. (2011) consists of comparing the conceptual model with the real system and assuring the conceptual model correctly reflects the real system.

Validation is not the only process used to check whether the simulation model is accurate of not. It is also important to perform a verification of the model. According to R G Sargent (2013) verification is the process of ensuring the computer programming and implementation of the conceptual model are correct. The author explains that the computer programming is the software used to simulate the model and the computer implementation is the actual model.

The validation and verification of a model are an essential part of the model creation process and will guarantee its validity. According to Robert G. Sargent (2005) each model is created for a purpose and the validity of the model has to be determined in relation to the purpose it was created for. The author further explains that if the model was created to answer a set of questions, the validation and verification processes need to ensure the model can accurately answer each one of the questions.

One more observation about simulation models is that they are created based on assumptions about the nature of a system. These assumptions can come from previous knowledge and common behavior of a system. As it is very complex and even impractical to create all possible scenarios that can exist in a real business, it is important to list all assumptions taken into consideration while creating the model. This will make it clear to all stakeholders what was and was not taken into consideration.

Another advantage of discrete event simulation is that it can be applied to diverse systems such as manufacturing, healthcare, supply chains, service businesses, and others. Many successful applications can be found in the literature (Cigolini, Pero, Rossi, & Sianesi, 2014; Diaz-Elsayed, Jondral, Greinacher, Dornfeld, & Lanza, 2013; Djanatliev & German, 2013; Thiede, Seow, Andersson, & Johansson, 2013).

Although simulation can have many benefits and applications, there are also downsides to it. Some disadvantages in the use of this tool are mentioned by Sharma (2015) in his study. The author states that using this tool requires special training. Moreover, creating a simulation model can also be time consuming and expensive. Lastly, the author states that as there is randomness in the model it is hard to distinguish between randomness and a real result of the interrelationships of the model.

In many simulation problems there is a need to find an optimal solution. However, another downside of discrete event simulation is that it is able to test different scenarios but it does not make a decision on which scenario is the best. This has to be done by the decision maker. Nevertheless, with many variables that exist in complex systems today, simulating all possible scenarios can be time consuming and even impossible. This has caused researchers in the past to develop methodologies that would make the optimization process possible.

Simulation optimization has been defined by Carson and Maria (1997) as a process that intelligently searches for the best solution without having to go through each possible scenario. There are many different methodologies that have been developed on simulation optimization. Some examples are gradient based search methods, stochastic optimization, response surface methods, sample path optimization, heuristic search methods, and statistical methods (Tekin & Sabuncouglu, 2004). These methods have been used over the years in many different applications in risk management, call centers, queues, inventory control, and others showing good results (Marco Better, Glover, Kochenberger, & Wang, 2008; Fu, Glover, & April, 2005).

Although these tools can be very effective there are some problems observed in their application. When doing an optimization the algorithms will test different scenarios, trying to find the best solution. Each algorithm has a different process of finding the "best" solution, but in all of them it is necessary to run multiple scenarios in order to get to the best one. According to Amaran, Sahinidis, Sharda, and Bury (2016) running complex simulations can be expensive if resources, time and money are taken into account.

In order to get away from using complicated optimization algorithms, research has also focused on different perspectives in improving a simulation model. One example of that is the use of data mining algorithms. These algorithms can support the optimization process of a

simulation model. Research has already been done in this field (M. Better, Glover, & Laguna, 2007; Brady & Yellig, 2005; Ghasemi, Ghasemi, & Ghasemi, 2011).

2.3 Data Mining

Developments in information technology have made it possible for data to be easily stored and retrieved in an inexpensive manner. These large amounts of data stored contain important information about a process that is not always extracted and consequently never learned. In order to learn from data companies have found the need of analyzing it so that knowledge can be generated from it. It is affirmed by Seng and Chen (2010) that data and information are different things and that in order to support decision making processes data has to be converted into information and knowledge. Discovering knowledge that is hidden in the data can give competitive advantage to a company. Nonetheless, the larger the dataset the more complicated and time consuming the data analysis can be.

These large datasets created over the past years from the development of information technology are called Big Data. This new term to define data was created because the methodologies used to analyze data in the past have changed because of the new characteristics of Big Data. According to Shmueli, Patel, and Bruce (2016) there are four main characteristics that make data analysis of big data unique and more complex. They are:

1. Variety: Referring to the types of data that are generated. Big data comes from a large variety of sources, each source having a different data type. This makes data analysis more complex.
2. Velocity: Referring to the speed at which data is created. In our digital world data is generated faster than before.

3. Volume: Referring to the size of the data. Big data as the name already suggests is composed of large amounts of data. This characteristic also makes the data processing time slower.
4. Veracity: Referring to the fact that data has been created from many diverse processes that do not go through any kinds of controls or quality checks.

Data mining was created to support the analysis of big data. According to Shmueli et al. (2016) the main fact that drove the growth of data mining is the growth of data. In order to measure the growth in the application of data mining techniques a research done by Liao, Chu, and Hsiao (2012) calculated how many words related to this topic were cited in the literature. The authors observed that from 2000 to 2005 the words related to data mining were cited 48 times. From 2006 to 2011 there was an increase of 292% in citations and the number of words cited was 140 times. This suggests a big growth in the application of data mining techniques that has been happening in the last years and continues to happen.

Data mining is defined by Olafsson, Li, and Wu (2008) as any automated or semi-automated process for extracting knowledge and patterns, unknown but potentially helpful, from large datasets. There are two main objectives in using data mining, according to Anderson (2012): prediction and description.

The purpose of prediction algorithms is to create a model that can make predictions for an outcome variable using input variables. This model is created from a historical dataset that has information on input and output variables. When the model is built, it is possible to use new data on the input variables to make a prediction about an unknown output. The output variable being predicted can be categorical or numeric. If the output being predicted is a categorical variable,

then the algorithm is called a classification algorithm. When the output being predicted is numeric, the algorithm is called a regression algorithm.

Description algorithms have the purpose of finding similarities in the data. There are three types of description algorithms, according to Anderson (2012): cluster, association rules and sequence analyses. Cluster analysis divides the data in groups that have similar characteristics. This can be helpful in a dataset that is diverse. The division of the data in clusters will make it easier to apply other algorithms to each group that need to be described.

Association analysis finds situations that are usually related to another one and occur together. This is very useful in marketing research to understand which items are usually bought together. Sequence analysis tries to find association between different items over time. They are similar to association rules, but instead of looking for items that are bought together, they will look for items that will be bought after the first one was.

The two main objectives in using data mining techniques also define two main processes used in the analysis. They are supervised learning and unsupervised learning. According to Shmueli et al. (2016) supervised learning algorithms are applied to problems where the outcome variable is known and there is data about it. In a supervised learning the algorithm will learn from historical data and will create a model that can predict the outcome of new inputs. This is what prediction algorithms do.

Unsupervised learning algorithms, according to Shmueli et al. (2016), are applied to problems where there is no data information on the outcome variable to learn from. Thus, the model is not used to make predictions. The model is used to find patterns in the data that cannot

be easily observed. Examples of unsupervised learning are the description algorithms, such as cluster, association and sequence analyses.

In order to do a thorough data mining analysis it is important to follow several steps. A list of nine steps that should be taken in order to approach a data mining problem correctly was created by Shmueli et al. (2016) . They are:

1. Determine the purpose of the project: It is important to determine if the project will be done only once or if it is an ongoing process.
2. Define which dataset will be used in the analysis: This step involves sampling from a large database if that is available.
3. Clean and preprocess the data: It is important to determine what will be done with missing values and outliers. Also the analyst has to make sure the data is consistent in units of measurement, time periods and others.
4. Reduce the data, if necessary and split it into training, validation and test datasets: In this moment, it is necessary to understand the importance of each variable in the model. In this step, it is possible that new variables have to be created or eliminated.
5. Determine the task that will be done (prediction or description): In this step, the purpose of the data mining project will be translated to a specific statistic question where the task done in the project will be defined.
6. Choose which technique will be used (regression, artificial neural networks or others): In this step, the analyst will specify which algorithm will be used in the analysis.
7. Apply the algorithms to the data: In this process, usually more than one algorithm is tested or even different variants of the same algorithm. In an iterative process the analyst will look for different possibilities that can yield best results.

8. Interpret the results: In this step, the analyst will choose the best algorithm that will be implemented and also apply the algorithm to the test dataset to observe how it will perform.
9. Deploy the model: This is the final step and it involves incorporating the model with the operational system and using real data to make decisions.

Data mining algorithms have been largely used in business applications doing consumer behavior analysis, consumer relationship management and support for decision making (Hsieh & Chu, 2009; Ngai, Xiu, & Chau, 2009; Seng & Chen, 2010). It has also already been applied specifically to manufacturing environments (Çiflikli & Kahya-Özyirmidokuz, 2010; Harding, Shahbaz, Srinivas, & Kusiak, 2005; Öztürk, Kayalıgil, & Özdemirel, 2006).

There are several data mining algorithms that have been developed over the years. While some are easy to understand, and can be easily implemented, others can be complex and require good computing performance. Artificial neural network is one technique that has shown good prediction capabilities even with nonlinear data.

2.4 Artificial Neural Networks

With the development of computing capabilities there was a need for computers to learn from data so they could be smart and make decisions more like humans. This necessity created a field in the science called machine learning. In this field scientists have studied how computers can learn and then change behaviors. This field is concerned with making computers modify their actions so they will be accurate in reflecting the right ones, according to Marsland (2015). The author uses the example of a computer that can accurately play scrabble against a human and

still win to show how machines can learn and make decisions without the influence of a human being.

There are many techniques that have been created with the purpose of learning from observations and then using that knowledge to judge between different possibilities and pick the right one. One of these techniques is artificial neural networks.

Artificial neural network algorithms were created with the purpose of imitating the learning process of our brains. Scientists studied the learning process of the brain and observed that the same process could be applied to other areas of science. Our brain is composed of neurons, which are cells of the nervous system. The main responsibility of a neuron is to conduct pulses while under specific conditions, according to Silva, Spatti, Flauzino, Liboni, and dos Reis Alves (2016). A representation of a neuron and a description of its parts is shown in Figure 2-1 taken from Silva et al. (2016).

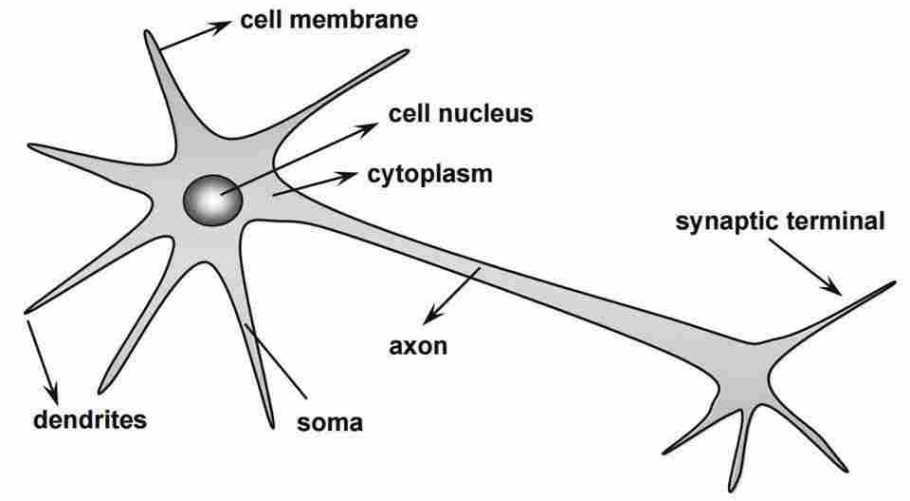

Figure 2-1: Neuron

In order for the brain and body to function properly, it is important that neurons communicate with each other, sending and receiving information. Neurons conduct impulses to one another through a process called synaptic transmission. This process occurs in the synapses, which, according to Silva et al. (2016), are the connections between neurons that make it possible for impulses to be transferred from one neuron to the other. The synaptic transmission occurs when a neuron is activated. This only happens when specific conditions are met. A synaptic transmission process is shown in Figure 2-2 taken from Silva et al. (2016).

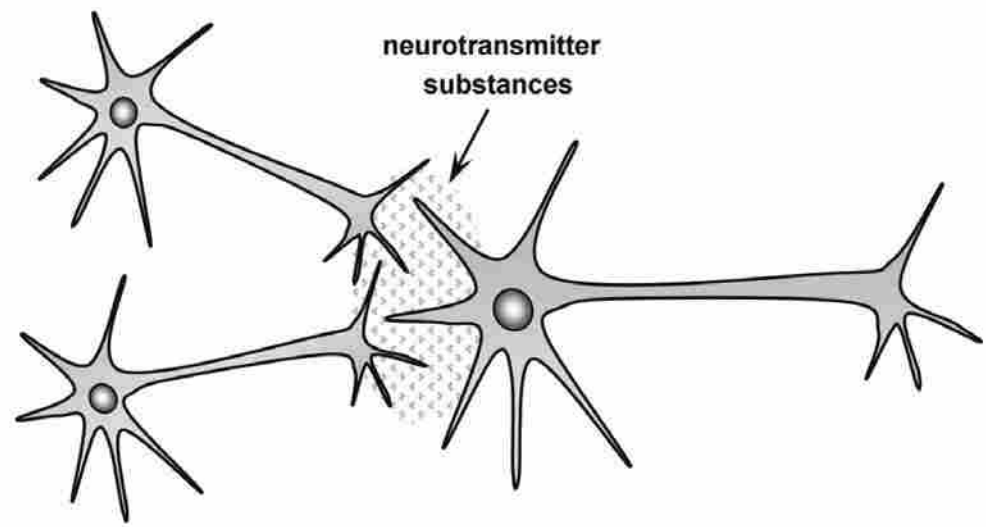

Figure 2-2: Synapses

Researchers observed that in order to imitate the functioning process of the brain it was necessary to describe the way a neuron works first. In order to replicate a neuron, an artificial neuron was created by McCulloch and Pitts (1943). This artificial neuron is a mathematical model that explains the way neurons work. His model is shown in Figure 2-3.

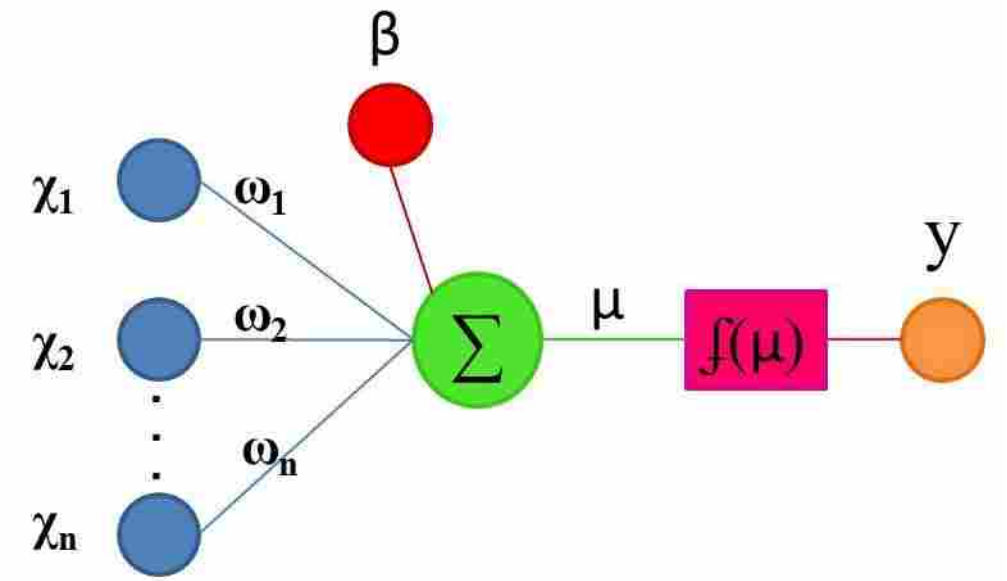

Figure 2-3: Artificial Neuron

Each artificial neuron is composed of seven elements, according to Silva et al. (2016). These elements and their definitions are:

1. Input signals ($\chi_1, \chi_2, \ldots, \chi_n$): Correspond to the values of the independent variables that will be used in the model.
2. Synaptic weights ($\omega_1, \omega_2, \ldots, \omega_n$): These are values attributed to the input variables representing their relevance to the neuron functionality. High weights indicate higher relevance of the input variable in activating the neuron.
3. Linear aggregator ($\sum$): Calculates the weighted sum of input values according to their synaptic weight. The equation used to calculate this is shown in Equation (2-1).

$$\sum_{i=1}^{n} \omega_i \cdot \chi_i \tag{2-1}$$

4. Bias (β): Variable used to adjust the linear aggregator so the neuron will send an impulse correctly.

5. Activation Potential (μ): The result of the difference between the bias and the linear aggregator. If μ > β then the neuron is activated, otherwise the neuron is not activated. The equation used to calculate μ is shown in Equation (2-2).

$$\mu = \sum_{i=1}^{n} \omega_i \cdot \chi_i - \beta \tag{2-2}$$

6. Activation Function (ʃ): This function limits the value of the output to a range that is required.

7. Output signal (y): The result created by the neuron given the input signals. The equation used to calculate y is shown in Equation (2-3).

$$y = ʃ(\mu) \tag{2-3}$$

Different functions can be used as the activation function ʃ applied to μ. The most common ones are linear, exponential and logistic functions. The output of the neuron using a logistic function is shown in Equation (2-4).

$$y = ʃ(\mu) = ʃ\left(\sum_{i=1}^{n} \omega_i \cdot \chi_i - \beta\right) = \frac{1}{1 + e^{(\sum_{i=1}^{n} \omega_i \cdot \chi_i - \beta)}} \tag{2-4}$$

The artificial neuron model created by McCulloch and Pitts (1943) was very useful, however, it had to be adjusted in order to be applied to real world. According to Silva et al. (2016) the brain is composed of 100 billion (10^{11}) neurons. It only works well because neurons are able to communicate. Thus, in order to correctly represent the brain, the mathematical model has to not only describe individual neurons, but rather the communication between them.

This was done when the artificial neuron was transformed into an artificial neural network, where neurons are connected and there is communication happening between them. According to Silva et al. (2016) each artificial neural network has an architecture or a way of being arranged composed of three layers. These layers are:

1. Input layer: This layer represents the layer that receives the input signals from the environment. These input signals are normalized in this layer in order for the results to be more accurate. There is only one input layer.
2. Hidden layers: There can be more than one hidden layer. They are constituted of neurons that will process the data.
3. Output layer: This is the final layer of the neural network where the final results will be processed. There are neurons in this layer as well and there can only be one output layer.

The information in the artificial neural network is passed from one layer to the next, starting at the input layer, and then going to the hidden layers and finally passing through the output layer where the result is defined. Each layer receives a value as input and generates a value as output. The input value of a layer is the output value of the previous one. A picture of an artificial neural network with a single hidden layer is shown in Figure 2-4.

As stated before, artificial neural networks have the capability of learning from observation. According to Silva et al. (2016), this process is done as weights and bias are adjusted until the artificial neural network is able to generalize the results generated by the outputs. When the learning process is complete, the artificial neural network can be used with new inputs to make new predictions or describe existing patterns in the data.

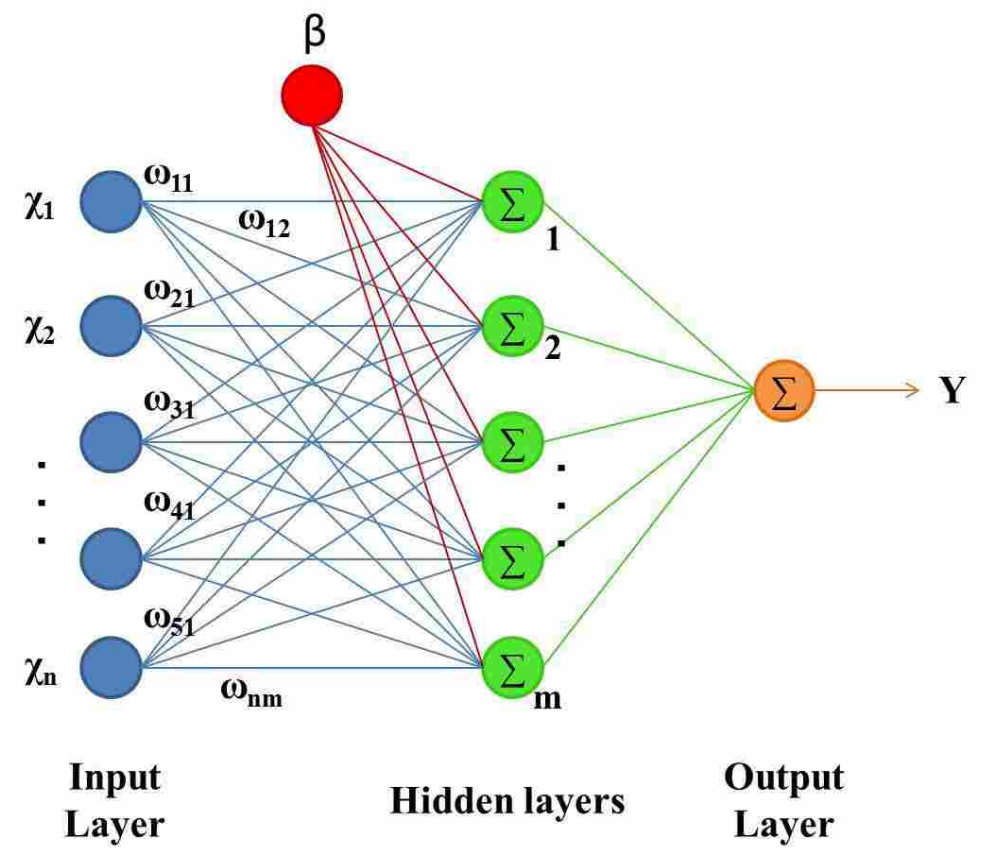

Figure 2-4: Neural Network

An artificial neural networks algorithm can do supervised learning or unsupervised learning. The supervised learning is called by Yegnanarayana (2009) as learning with a teacher because in this learning process the data used to train the artificial neural network has information on the actual output that is being predicted. Thus, it is possible to compare the outputs created by the model with the actual outputs and calculate how well the model predicts.

The unsupervised learning is used when there is no information on the actual output, but there is a need to understand patterns that exist in the data. According to Silva et al. (2016) these patterns are similarities present in the data and the job of artificial neural networks is to find them and organize the data into clusters. In this research only supervised learning will be discussed and applied to the case study.

The great advantage of artificial neural network is that the algorithms are able to learn from linear and non-linear data and this technique has been applied to many different fields (Amato et al., 2013; Dil et al., 2016; Hemmat Esfe, Saedodin, Sina, Afrand, & Rostami, 2015; Joshi, Rana, & Misra, 2010; Zaji & Bonakdari, 2015). However, there is a downside to these algorithms. Artificial neural networks can be quite complex and difficult to interpret in some instances. Therefore, they are sometimes taken as a black box; that is their results are used without attempting to comprehend all components and dynamics of the network. Moreover, if it was possible to interpret the model, more information about the data could be extracted.

According to Olden and Jackson (2002) the main reason why artificial neural networks are known as the "black box" is because of the difficulty in understanding the contributions of input variables to the final outcome of the network. This hinders the possibility of understanding the inter-relationships that may exist between variables and consequently the capacity of generating insights from the model.

In order to increase the information that can be obtained from this algorithm, researchers have been looking for ways to illuminate the "black box", or in other words, be able to better interpret the algorithm and understand the contributions of input variables on the output of the model. Some methods created with this purpose are Garson's algorithm (Garson, 1991), connection weight approach (Olden & Jackson, 2002), partial derivatives (I. Dimopoulos, Chronopoulos, Chronopoulou-Sereli, & Lek, 1999; Y. Dimopoulos, Bourret, & Lek, 1995), input perturbation (Scardi & Harding Jr, 1999), sensitivity analysis (Lek, Belaud, Baran, Dimopoulos, & Delacoste, 1996; Lek, Delacoste, et al., 1996), and others (Gevrey et al., 2003; Olden & Jackson, 2002; Olden et al., 2004). Each method is based on different methodologies, but they all

have the same objective of interpreting artificial neural networks. The present research will focus on the connection weight approach created by Olden and Jackson (2002).

2.5 Connection Weight Approach

In the Connection weight approach, created by Olden and Jackson (2002), the contribution of the input variables on the output of the model is based on the synaptic weights of the artificial neural network. The creators of the approach explain that the contribution of input variables on the output of the model depends on the direction and magnitude of the synaptic weights. Larger synaptic weights indicate variables that have higher importance compared to those with smaller synaptic weights. Moreover, positive synaptic weights indicate an increase in the output value, whereas negative synaptic weights indicate a decrease in the output of the model.

Understanding how synaptic weights impact on the output of the model, the author created a mathematical procedure that calculates a score for each input variable. This score represents the importance of the independent variable to the output of the model. The steps created by Olden and Jackson (2002) are the following:

1. Create several artificial neural network models and select the one with the best results.
2. Record and calculate the following from the artificial neural network:
    a. The contribution (C_{ij} with i representing each neuron from $i = 1, \dots, m$, where m is the total number of neurons and j representing each input from $j = 1, \dots, n$, where n is the total number of input signals) of each input to the output through each hidden neuron. This is calculated as the product

between the input-hidden (ω_{ij}) and hidden-output (ω_{oi}) synaptic weights for each neuron and input. The formula used to calculate this is shown in Equation (2-5).

$$C_{ij} = \omega_{ij}.\omega_{oi} \tag{2-5}$$

b. Overall input contribution (also called importance score) which is the sum of the total input contribution (S_j) of each input to the output through each hidden neuron. The formula is described in Equation (2-6).

$$S_j = \sum_{i=1}^{m} C_{ij} \tag{2-6}$$

3. Repeat the process a considerable amount of times (the creators of the approach tested it 999 times in their study).

An example of an artificial neural network is shown in Figure 2-5. The connection weights approach steps are demonstrated in Table 2-1 and Table 2-2.

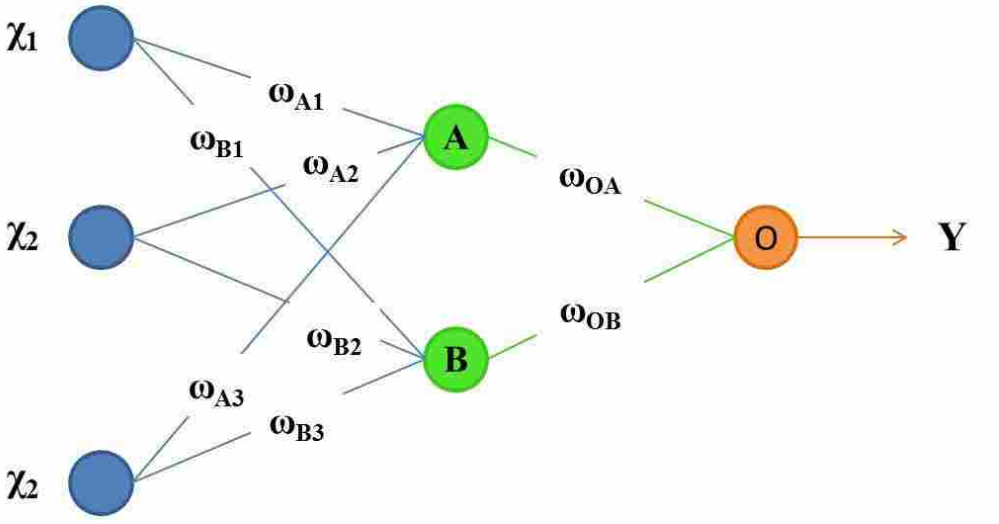

Figure 2-5: Connection Weight Approach

Table 2-1: Synaptic Weights

	Hidden A	Hidden B
Input 1	ω_{A1}	ω_{B1}
Input 2	ω_{A2}	ω_{B2}
Input 3	ω_{A3}	ω_{B3}
Output	ω_{oA}	ω_{oB}

Table 2-2: Input Contributions

	Hidden A	Hidden B	Importance
Input 1	$C_{A1} = \omega_{A1}.\omega_{oA}$	$C_{B1} = \omega_{B1}.\omega_{oB}$	$S_1 = C_{A1} + C_{B1}$
Input 2	$C_{A2} = \omega_{A2}.\omega_{oA}$	$C_{B2} = \omega_{B2}.\omega_{oB}$	$S_2 = C_{A2} + C_{B2}$
Input 3	$C_{A3} = \omega_{A3}.\omega_{oA}$	$C_{B3} = \omega_{B3}.\omega_{oB}$	$S_3 = C_{A3} + C_{B3}$

Studies have shown that this approach is valid and can be used to create a score that represents the importance of independent variables to the output of the model (Olden et al., 2004).

2.6 Data Mining Applied to the Output of Discrete Event Simulation

Data mining algorithms have been successfully applied to many different fields of study. In recent years managers observed opportunities in applying these algorithms to manufacturing environment (Gröger, Niedermann, & Mitschang, 2012). It was noticed that the application of data mining to a manufacturing environment can improve the decision making process and can also give competitive advantage to the company that decides to apply its principles (Kusiak & Smith, 2007; Shao, Shin, & Jain, 2014).

Data mining is also a good fit when coupled with discrete event simulation, as one tool is able to support the other. This happens because in order for data mining algorithms to create accurate models describing the behavior of a system it is necessary the use of large amounts of data. However, in a manufacturing environment it is not always possible to find these large

datasets necessary for the creation of an accurate model. This way data mining can benefit from the use of simulation, as the latter creates large amounts of data about a system. On the other hand, simulation can also benefit from the application of data mining algorithms, as the latter can be used as a support tool that will lead to improvement of a simulation model. Research has already been done coupling the two sciences and has shown good results (M. Better et al., 2007; Ghasemi et al., 2011; Painter, Erraguntla, Gary L. Hogg, & Beachkofski, 2006).

The technique studied in this research, artificial neural networks, has already been coupled with discrete event simulation in manufacturing with the purpose of speeding the process of creating simulation models and the process of making decisions (Fonseca, Navaresse, & Moynihan, 2003; Panayiotou, Cassandras, & Wei-Bo, 2000). However, artificial neural networks algorithms have not yet been used as a tool that will prioritize independent variables and be a guide to the improvement of a simulation model.

The prioritization of independent variables is possible through the interpretation of artificial neural networks. The connection weight approach discussed previously can be used to create an importance score for each independent variable. This score makes it possible for managers to understand which independent variables will make the most impact on the outcome of the model. This knowledge can be used in the simulation model as a guide to where changes should be made or which scenarios should be tested in order to improve the performance of the system being studied. This approach can be helpful in a manufacturing environment as it will support managers with evidence that will lead to better decisions.

3 EXPERIMENTAL PROCEDURE

3.1 Methodology Overview

In order to test the hypotheses of this research, an experiment was performed. The experiment was done in three steps: 1) creation of a simulation model, 2) application of artificial neural network models to the output data of the simulation model, and 3) interpretation of the artificial neural network models using the connection weight approach. Each step will be described in the following sections.

3.2 Problem Description

This study was based on a real problem faced by a manufacturer located in the northeast of Brazil. The company wants to be more efficient in the raw material receiving process. Currently they face fluctuations in the arrival of raw materials. As a consequence there are moments when there are long lines of trucks waiting to unload while at other times there are none. Sometimes there is more raw material than the manufacturer has capacity to receive, while other times there are shortages of materials.

The manufacturer wants to know how to be more efficient in the receiving process to better deal with fluctuations in the arrival of raw materials. Although there are opportunities in improving scheduling of arrivals, the study will focus on dealing with fluctuations and being more efficient internally. The details of the process studied are explained in Section 3.3.

3.3 Simulation Model

The simulation model of the system studied was created using ProModel® software.

3.3.1 Data Collection

There was data collected on all the trucks that arrived at the company, their materials and load quantities, supplier information, arrival and departure dates and times. However, there was not any data collected on the actual processes, i.e. times spent in each operation. This information was collected through observation by the workers of each department involved in the receiving process. All the information used to create the simulation model was based on data collected from January 8, 2016 to Jun 29, 2016.

3.3.2 Locations

In order to represent the business, eleven different locations were created. Each location will be explained below:

1. Arrival Line: Represents the first location of the model. It is the waiting line to enter the manufacturing facility. It is modeled with an infinite capacity; trucks can come to this location any day, any time.
2. Unload Line: This location represents the line where trucks wait for either sample collection or to unload their material. It is modeled with an infinite capacity; trucks can stay in this location twenty-four hours a day, seven days a week.
3. Entrance: This is the first process location, where the truck information is collected and checked. If documents and invoices are correct, the unload process is started.

This location has a capacity of one, as it can only work on one truck at a time. This location is open twenty-four hours a day, seven days a week.

4. Scale: This location is where the truck is weighed before it is unloaded. This location can process one truck at a time. This location is always open, but as it is related to the mill hopper location, it is mostly used when the mill hopper is working as well.
5. Mill Hopper: This is the most important location for the unloading process. It is where a sample of the truck material is collected and also where the material is unloaded. This is the busiest location and it determines the pace of the system. The processes that take the longest are performed here. This location has a capacity of one truck at a time and is only open from 6:00 A.M. to 11:00 P.M. every day. From Monday through Friday there are always two operators working there. On the weekends there is only one, which makes all processes slower.
6. Analysis Line: This is the sample waiting line. It is where samples collected from trucks wait until they are analyzed. The capacity of this location is infinite.
7. Laboratory: This location is where all the analysis of truck samples is performed. At this location only one analysis can be done at a time, thus the capacity is one. The operation hours are the same as the Mill Hopper. However, the number of operators working is the same every day.
8. Group A Silo: This location represents all silos where Group A is stored. The total capacity of these silos is 1764 pounds. The statistical distribution used to represent the consumption of it created by Statistically Fit® software is $N(9.02, 3.36)$ (where $N(\mu, \sigma)$ represents a normal distribution with mean μ and standard deviation σ).

9. Group B Silo: This location represents the silos where Group B is stored. The capacity of Group B silos is 948 pounds. The consumption rate of these silos varies and a statistical distribution that represents it was created by Statistically Fit® software. The distribution is $1.05 + L(3.96, 2.02)$ (where $L(\mu, \sigma)$ represent a lognormal distribution with mean μ and standard deviation σ).
10. Group C Silo: This is the representation of the silo where Group C is stored. The total capacity of this silo is 143 pounds. The consumption statistical distribution also created by Statistically Fit® software is $-0.689 + L(2.36, 0.976)$ (where $L(\mu, \sigma)$ represent a lognormal distribution with mean μ and standard deviation σ).
11. Group D Silo: This is a representation of the silo where Group D is stored. The total capacity of this silo is 130 pounds and its consumption statistical distribution is created by Statistically Fit® software is $5.41e^{-0.002} + L(0.675, 0.614)$ (where $L(\mu, \sigma)$ represent a lognormal distribution with mean μ and standard deviation σ).

3.3.3 Entities

There are various raw material types that are received in the manufacturing. They change according to the season of the year, market prices and availability. Simulating all the different raw materials would be very complicated and unnecessary, as they have similar behaviors. Taking into consideration the different processing times and arrival rates it was possible to split the raw materials into four groups: Group A, Group B, Group C and Group D. In the simulation model an entity was created to represent each group.

3.3.4 Arrivals

In order to describe the arrivals rate of each different entity, statistical distributions were used. These distributions were found through the use of Statistically Fit® software. The distributions are listed in Table 3-1 (where $L(\mu,\sigma)$ represents a lognormal distribution with mean μ and standard deviation σ, $E(\mu)$ represents exponential distribution with mean μ and $T(a, b, c)$ represents a triangular distribution with minimum value a, mode b and maximum value c).

Table 3-1: Arrival Distributions

Group A	Group B	Group C	Group D
$L(21, 319)$	$E(22)$	$E(39)$	$T(12, 90, 239)$

3.3.5 Processing

An important observation about the manufacturer's receiving process is that company policy controls what raw material may be accepted into the manufacturing facility through laboratory tests. Thus, every truck has to have a sample analyzed before its content is unloaded. This process can be short or long depending on the number of trucks in line. Suppliers are aware of the company's policy and accept it. However, a fee is applied to the manufacturer for each day the truck has to wait until it is able to unload its material.

The company's raw material receiving process description is as follows. First the truck arrives at the entrance location where paperwork is done. Then the truck waits for its turn to have its sample collected at the mill hopper location. After collected, the sample goes to the laboratory where it will be analyzed and the truck waits the analysis result. When the analysis is finished, if the raw material is accepted, the truck will wait for its turn to unload its material at the mill

hopper location. After unloading the truck is free to go. If the material is rejected the truck is not allowed to unload. A visual description of the process is shown in Figure 3-1.

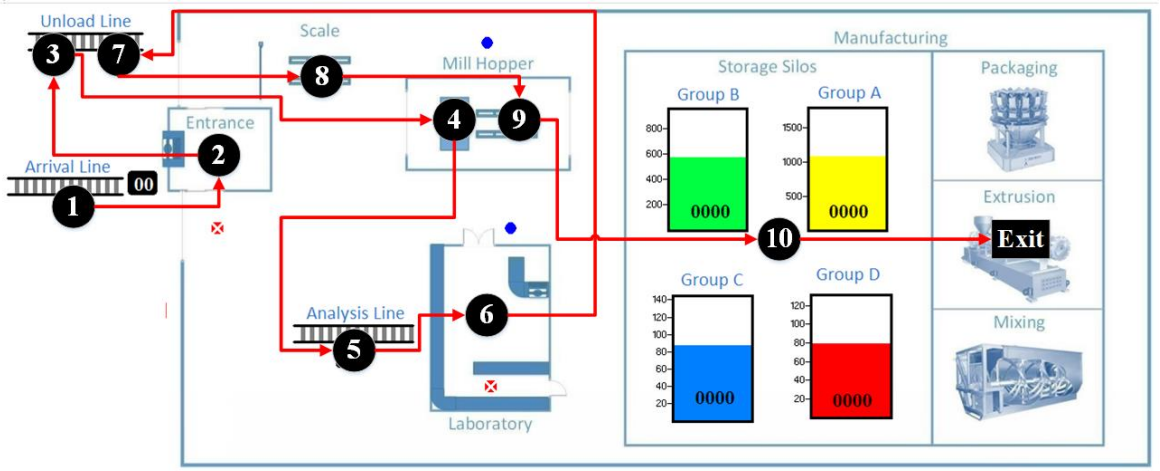

Figure 3-1: Process Description

There are some considerations about this process. The place where the truck has its sample collected is the same place where it unloads. The order in which they process each truck is on a first come first serve basis. However, if there is a shortage of one specific raw material, it will have priority over the other trucks and it will be processed first. Sometimes the truck has a load that is bigger than the free capacity of the manufacturing storage. When this happens, the truck waits until there is enough space for its load to be completely unloaded.

The analysis done in the laboratory depends on the raw material. There are four different analysis types. All raw materials that are part of the Group B are analyzed using one method. The raw materials from Group C and Group D have the same analysis procedure that is different from the method used to analyze Group B. Part of Group A has one analysis specification and the other part is analyzed in a different manner.

The sample collection time varied depending on the raw material and the truck size. Groups B, C and D take the same time to have a sample collected. This is because these materials tend to not have high rejection rates and their collection procedure is standardized. Group A, on the other hand, has stricter rules on sample collection. These materials have a history of having higher rejection rates, and a nonconformance present in the material can be very harmful for the final product made. Thus, the sample collected for Group A is bigger and the process takes longer compared to other groups.

The unload process time also varies depending on the raw material that is received. Some raw materials flow very well through the system and are easily and quickly unloaded. Some others do not flow well and tend to block the system. Consequently, they take longer to unload.

Every truck goes through the same processing stages; however, the processing times are different for each entity. The processing times for each entity are described in Appendix A.

3.3.6 Assumptions of the Model

Every simulation model has assumptions about the business it models, as explained in the literature review. The assumptions for the current model are the following:

1. Samples are immediately analyzed when they arrive at the laboratory.
2. Trucks go immediately to the mill hopper when it is available.
3. Stations never stop during meals and other breaks.
4. There is always an operator or analyst in their work post.
5. The only process stops taken into consideration are those due to raw material shortage. Cleaning stops are not included in the model.
6. The manufacturing process never stops because of holidays.

7. A material will only be considered a priority if the quantity of the raw material stored is less than 10% of the total silo capacity.
8. All machines and equipment never break.
9. During weekends when there are fewer operators at the mill hopper the sample collection and unloading process take 15% more time than on weekdays.
10. When one silo is empty the manufacturing process stops. This means that when there is a shortage of one raw material, there is no consumption of any material. The unloading process continues, however, there is not raw material consumption.

3.3.7 Verification and Validation

As mentioned in the literature review, an important step of the model creation is testing its validity. This is done through model validation and verification. The verification has the purpose of assuring the computer model does what it was set up to do. In this research the modeler did this by checking the output data for reasonableness, observing the animation, reviewing the code and using trace facilities in the software. This was a continuous process that was done as the model was being created until it was completely finished.

The conceptual model was created through the observation of the real system, interviews with operators and managers and learning from existing data. After the model was created the validation process was performed through the comparison of the simulation model data with the real data available. This comparison is shown in the results chapter.

3.3.8 Simulation Specifications

The simulation was performed in one replication of four hundred weeks and in ten replications of forty weeks in order to see if the data generated would cause any impact on the prediction results of the artificial neural network model. The results from the simulations performed are shown in the results chapter.

3.4 Artificial Neural Network Models

3.4.1 Data Preparation

The artificial neural network models were created using VisMiner® software based on the output data generated by the simulation model. In order to decide which variables should be included in the model an analysis of all factors that could impact the total time a truck stayed in the system were taken into consideration. These factors were calculated as variables and are the following:

1. IsGroupA: This represents a dummy variable. It is a binary variable and its value can be one if the material is Group A or zero if it is not.
2. IsGroupB: This is another dummy variable to represent the raw material received. The variable is binary and its value can be either one if the raw material is Group B or zero if it is not.
3. IsGroupC: This is the last dummy variable used to represent the raw materials entered into the manufacturing.
4. IncludesWeekend: This variable is a binary variable that indicates if the truck is waiting to unload during the weekend. If its value is one it indicates that the truck

waited during the weekend, if it is zero the truck unload during a week day. This variable can give important information to the model as unloading processes take longer during the weekend.

5. IncludesNight: This is a binary variable that indicates whether or not the truck stayed overnight. One indicates the truck stayed overnight and zero indicates it did not. This variable can give good information as the mill hopper and the laboratory are closed during the night.
6. Shortage: This is also a binary variable with a value of one when there is a shortage during the time the truck was in the system and zero if no shortage happened.
7. UnloadQuantity: This variable represents the weight of the material in the truck. In the simulation these numbers are presented in thousand pounds
8. WaitedToUnload: This variable is a binary variable that indicates whether or not the quantity loaded in the truck exceeds the silo free capacity at the time the truck arrives. This causes the truck to wait until there is available capacity for it to unload.
9. WasPriority: As mentioned before, when there is a shortage of a material the FIFO rule is broken and the truck is processed in front of other materials. This is a binary variable, it is equal to one when the truck arrives and is given priority status. This value is equal to zero when there is no priority.
10. TimeEntrance: This is the total time it takes for the paper work to be done at the entrance.
11. TimeAnalysis: Time taken at the lab to analyze a sample of the material in the truck.
12. TimeCollection: Time taken at the Mill Hopper to collect a sample of the truck material.

13. TimeUnload: Time taken at the Mill Hopper to unload a truck.
14. TrucksInLine: This variable represents the number of trucks waiting to unload their material at the moment a specific truck arrives.
15. TotalTime: This is the response variable. It measures the total time the truck stayed in the system.

These variables were generated by the simulation model using arrays facilities present in the ProModel® software that collected the data for each truck and automatically exported to an excel file. Thus, there was no need to organize the data, create new columns or make any sort of calculation.

After the data was gathered through the simulation model analysis were made in order to understand which variables were most significant. This was done first by creating a correlation matrix to understand which variables were correlated. Then a trial an error approach was used to determine which combination of variables would create the best prediction results in the artificial neural network models. The results are shown in the results chapter.

3.4.2 Artificial Neural Network Models Creation

After the dataset was prepared it was possible to apply data mining algorithms to the data. The software used in this research was VisMiner®. The research is focused on using only artificial neural network algorithms. However, in order to check which algorithms would have the best prediction capabilities, several algorithms were tested, such as linear regression, nearest neighbors, decision trees, random forest and gradient boost. The results of the application of each algorithm are shown in Section 4.

The software VisMiner® creates artificial neural network models with one hidden layer that are trained by the "backpropagation" algorithm. The models were built using a tool present in the software that permits an interactive creation of the model. This makes it possible for the modeler to train the model until a good prediction result is found. Each analysis shown in the results chapter is based on fifty models that were built manually and their information was exported to an excel file where the analysis was performed.

3.5 Connection Weight Approach

After the creation of the artificial neural network models, it was possible to interpret them using the connection weight approach. This was done in an excel spreadsheet. First, all the information generated by the model was exported to an excel file and then the calculations were performed. In order to make it easier to import the artificial neural network models and perform calculations, macros were created in excel. These macros would automatically make the necessary mathematical operations.

3.5.1 Absolute Value for Overall Input Contribution

Overall input contributions or importance scores are calculated by the formulas given in Section 2.5 of the literature review. It is possible to observe that the formulas do not consider the absolute value of these input contributions. In the articles studied these scores were always positive. However, in this research there were positive and negative values for the contributions, making it hard to make comparisons between them.

It is understood that negative contributions will decrease the output of the model while positive contributions will increase. In the present experiment, negative contributions will lead to

a decrease in the total time a truck stays in the system, while a positive score will lead to an increase in the total time a truck stays in the system. However, the sign of the number does not impact the importance of the variable in determining the output of the model. A variable with high absolute importance score will make a high impact on the output of the model no matter whether its score is positive or negative. What determines the importance is not whether the number is positive or negative, but rather its absolute value.

Thus, in this research all importance scores were calculated following the formulas found in the literature review (Olden & Jackson, 2002). After that their absolute values were calculated. All data discussed in this research is based on the final absolute value. It is important to observe that the results shown in this research will not indicate whether a variable will increase or decrease the output value, but rather, how important the variable is to the dependent variable that will be predicted.

3.5.2 Normalized Scores Instead of Ordinal Rank

In the connection weight approach after the importance scores are calculated, they are given an ordinal number as their rank. Through this research it was possible to observe that some variables have importance scores that are very similar, thus making it hard to differentiate between both. Consequently, when ranks are ordinal they determine that one variable is more meaningful than the other but they do not specify by how much. For some variables the difference is very small, even insignificant.

In order to solve this problem the present research used a normalized rank instead of an ordinal rank. This was done by normalizing the importance score and using this normalized number as their rank. This made it possible to not only understand which variables are the most

meaningful variables to the system while comparing them, but also to know by how much they are better than others. This is very useful when variables have similar scores. All values were normalized using equation (3-1).

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \tag{3-1}$$

3.5.3 Ranking Instability

As observed in the literature review the importance scores have to be calculated several times in order to account for the instability present in artificial neural networks. This instability happens because artificial neural networks begin by selecting random initial weights; then iterate towards a solution. Thus, each time a model is created, the random weights can differ. Since the machine learning process begins with different random weights, there is a lot of variation from one model to the other. As mentioned, fifty artificial neural network models were created and their information was used to rank the independent variables fifty times. This produced a more stable ranking of the most meaningful variables than would result had only one model been created.

3.6 Improvement of the Simulation Model

The normalized scores generated by the connection weight approach gave insight on the parts of the process that had the greatest impact on the output of the simulation model, which was the total time a truck would stay in the system.

The knowledge of the most meaningful variables to the business can be used to improve the efficiency of the real system. In this study case, the knowledge of these variables can guide

the processing of making improvements to the simulation model. This can be done by creating new scenarios that manipulate specifically the most important variables so that the total time a truck stays in the system is reduced. This will simplify the complexity of testing different scenarios without a clue of the variables that impact the system the most.

In order to actually see how the different variables impacted the system, the dataset generated by the simulation model was analyzed. The variables that were considered the most meaningful were listed and different scenarios compared the total time a truck stayed in the system. The findings are shown in the results chapter.

4 RESULTS

4.1 Simulation

4.1.1 Verification and Validation

As mentioned in the methodology, the validation of the model was performed by comparing the model results with the real system results. The main purpose of the simulation model was to understand and predict the time trucks stay in the system. Thus, the validation was focused on that.

Before any statistics was done, a logarithmic transformation was performed in the data. This helped eliminate possible outliers and have a better visualization of the data.

In order to compare real and simulated data, a hypothesis test was performed. It focused on testing whether or not the simulation data followed the same distribution of the real data. Thus, the two hypotheses were:

$H_0: Samples\ follow\ the\ same\ distribution$

$H_1: Samples\ follow\ different\ distributions$

The test used was the Kolmogorov Smirnov test and it was performed using the *ks.test* function in R. The $p - value$ found in the test was $3.108e^{-0.06}$, thus rejecting the null hypothesis.

The test results above indicate that the real data and the simulation data do not follow the same distribution. However, this does not mean that the simulation is not a good representation of the system being modeled. A Kolmogorov Smirnov test is very sensitive to any differences that might exist between two samples. Although it is known the samples do not follow identical distributions, it is important to know how far these distributions are from each other. This was done by comparing the boxplots from both distributions. The boxplots are shown in Figure 4-1. Also, a comparison of the statistical parameters of both distributions is made in Table 4-1.

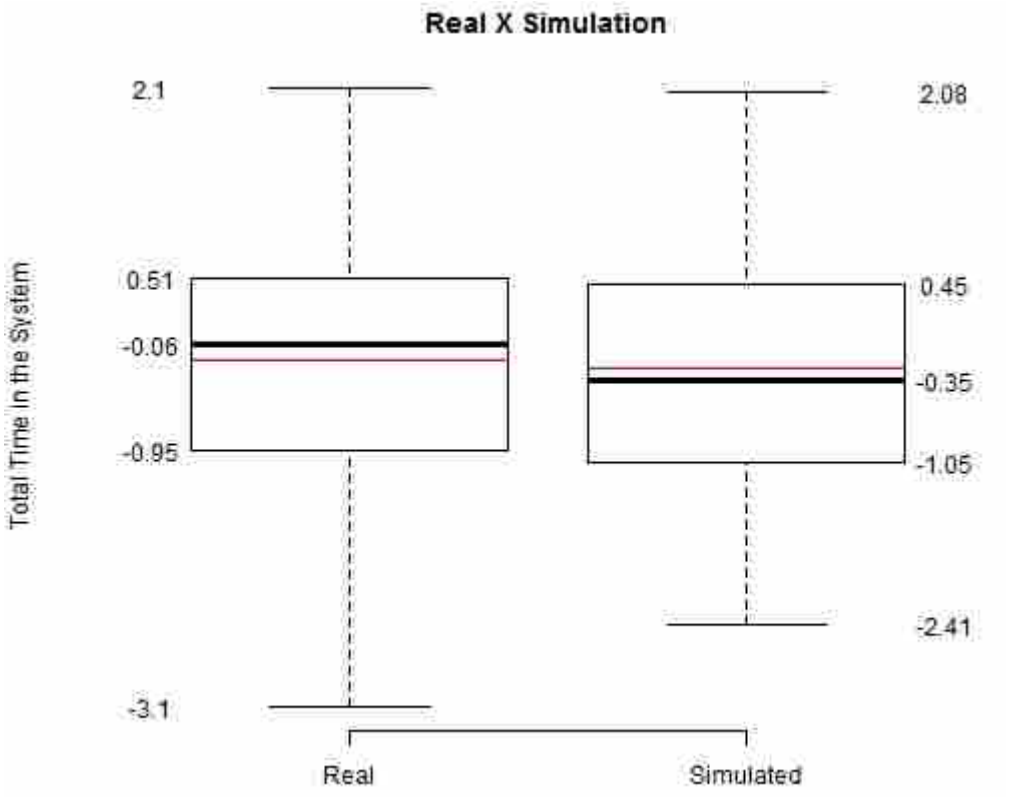

Figure 4-1: Validation of the Simulation Model

Table 4-1: Statistic Parameters

	Real	Simulated	Difference	Difference without logarithmic transformation
Mean	-0.193	-0.261	0.068	0.023
Standard Deviation	0.985	0.990	0.005	-0.134
Median	-0.057	-0.349	0.292	0.239

The boxplots shown in Figure 4-1 representing the real and the simulated data are very similar. This can be confirmed by analyzing the differences listed in Table 4-1. When comparing different data samples the main problems that can come up are either problems with precision or bias. As the standard deviations of both samples are very similar, it means that there are no precision problems. However, it is possible to see that there is some bias as the averages are different in the samples. The simulation tends to under predict the total time the truck is in the system. However, this difference was not considered significant to the study. Thus, the simulation is considered reliable in predicting the real system.

4.1.2 Simulation Specifications

Simulations with two different specifications were performed in order to test whether the number of replications would impact on the prediction capability of the artificial neural network models. The first simulation was done as one replication that was four hundred weeks long. The second was done as ten replications of forty weeks. The simulation data from both datasets was then used to create artificial neural network models. Each model was created with a dataset of 2700 records for training and 1800 records for validation. This made sure that the sample size would not affect the results. Moreover, five artificial neural networks were created for each simulation dataset in order to reduce the impact of randomness existing in the models. The following variables were used in the models: IsGroupA, IsGroupB, IsGroupC, IncludesWeekend,

IncludesNight, Shortage, UnloadQuantity, WaitedToUnload, WasPriority, TimeEntrance, TimeCollection, TimeAnalysis, TimeUnload, TrucksInLine, TotalTime. The results of both simulations are shown in Table 4-2.

Table 4-2: Comparison of Different Simulation Specifications

Artificial neural network models	Validation R^2	
	1 replication of 400 weeks	10 replications of 40 weeks
Model 1	0.774	0.765
Model 2	0.776	0.770
Model 3	0.771	0.768
Model 4	0.768	0.762
Model 5	0.773	0.770
Average	0.7724	0.7670

The prediction capabilities of the artificial neural network models did not change much based on the number of replications and length of the simulation. Thus, the author decided to use one replication of four hundred weeks throughout the research for simplicity.

4.2 Artificial Neural Networks

4.2.1 Dataset Size

Different dataset sizes were used to create artificial neural network models in order to see if there would be a difference in the prediction results of these models. As mentioned in the previous section, the models were created using the dataset produced by the simulation with one replication and a four hundred hour length. The results are listed in Table 4-3.

Table 4-3: Impact of Dataset Size on Model Prediction

Artificial neural network models	Validation R^2	
	1000 records	4000 records
Model 1	0.765	0.797
Model 2	0.770	0.794
Model 3	0.783	0.794
Model 4	0.775	0.797
Model 5	0.765	0.798
Average	0.7716	0.7960

Through the results it is possible to see some improvement in the model prediction results with a larger dataset. However, the difference is not very significant. This shows that the simulation is stable and is producing consistent results. In the research a larger dataset will be used.

4.2.2 Testing Other Data Mining Algorithms

Although the purpose of the research is to focus on artificial neural network models, it is interesting to compare the prediction results observed by applying different data mining techniques to the data. This comparison is shown in Table 4-4. The dataset used to create the models is the simulation output data of one replication of four hundred weeks. The dataset consists of 2961 records for training and 1973 records for validation. The variables included in the models are: IsGroupA, IsGroupB, IsGroupC, IncludesWeekend, IncludesNight, Shortage, UnloadQuantity, WaitedToUnload, WasPriority, TimeEntrance, TimeCollection, TimeAnalysis, TimeUnload, TrucksInLine, TotalTime. It is possible to observe that for the current dataset produced by the simulation model, the artificial neural network algorithm had the best prediction performance, having the highest R^2 and the lowest RMSE in the validation dataset.

Table 4-4: Data Mining Algorithms Prediction Results

	Training Dataset			Validation Dataset		
	MAPE	RMSE	R^2	MAPE	RMSE	R^2
Linear Regression	84.5	1174	0.700	88.0	1158	0.640
Random Forest	37.6	871	0.835	43.6	974	0.745
KNN (Equal weights)	36.8	960	0.799	40	1011	0.725
KNN (weights relative to distance)	39.6	1134	0.720	39.4	1108	0.670
Gradient Boosting	42.6	892	0.827	48.1	1010	0.726
Regression Trees	29.2	805	0.859	39.3	1113	0.667
Artificial Neural Network	35.6	869	0.835	40.1	930	0.770

4.2.3 Which Variables to Include in the Model

Artificial neural networks can be very complex and the greater the number of variables in the model, the higher is the complexity. Also, if variables are included that do not contain useful predictive information, it can confuse the machine learning process. Thus, the purpose of making a variable selection is to reduce the input variables to the ones that have useful predictive information.

In order to understand which variables are the most significant for prediction purposes in the artificial neural network models, a correlation matrix was created. The correlation matrix is shown in Figure 4-2, and was created using VisMiner® software. The correlation coefficients in the picture represent how a variable can be predicted by the other variable to which it is correlated. A positive correlation coefficient indicates that as one variable increases the other increases as well. A negative correlation coefficient indicates that as one variable increases the other decreases. In the picture, correlations coefficients are represented in different colors. Positive correlations are represented in blue shades and negative correlations are represented in red shades. The darker the colors are, the higher the correlation.

Figure 4-2: Correlation Matrix

According to [28] adding highly correlated variables or variables that are not related to the outcome studied in the model can lead to overfitting. When these variables are present, it can be likely that they might not be necessary in the model and also that their information is duplicated. Knowing that, it is important to do a correlation analysis and decide which variables should stay in the model.

Through the correlation matrix it is possible to observe high correlations between the following variables: IsGroupA and IsGroupB and IsGroupA and UnloadQuantity. These high correlations can mean that a variable is well described by the other variable to which it is correlated to and thus indicating that it may not be needed in the model.

In order to decide whether or not a variable was contributed predictive information to the artificial neural network models, I used an iterative trial-and-error approach, where a series of different combinations of variables were tested and their results compared. First I included all

variables that came as an output of the simulation model. Then, each variable was removed in turn to see if the predictive quality of the model increased or decreased, as measured by the R^2 of the resulting model on the validation dataset. Depending on whether R^2 increased or decreased, the variables were taken out of or left in the model. Table 4-5 shows the results of the models that contain different input variables.

As mentioned, the first model tested all variables that were created. The second model tested variables that had high correlations with other variables. As the correlation matrix showed a high correlation between UnloadQuantity and IsGroupA, a model was created without the UnloadQuantity variable. The results show that removing the UnloadQuantity variable did not worsen the results. Thus, it does not seem to be impactful on the model prediction. As a result, it was removed from subsequent models.

Model 3 was then created without TimeEntrance, as this operation has a short time that does not seem to impact the total time a truck stays in the system. As expected, the results confirmed that this variable is not impactful on the final prediction of the model, so TimeEntrance was also removed from subsequent models.

In Model 4, the variable TrucksInLine was tested. When the variable was removed from the model, the prediction results got worse, showing that this variable has important information and should be in the model.

Model 5 removed the variable TimeCollection, which produced results very similar to the results from Model 3. This demonstrates that TimeCollection does not seem to impact the prediction capabilities of the artificial neural network models.

Table 4-5: Variable Selection

	Models														
Variables	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
IsGroupA	X	X	X	X	X	X	X	X	X	X	X	X		X	X
IsGroupB	X	X	X	X	X	X	X	X	X	X	X	X	X		X
IsGroupC	X	X	X	X	X	X	X	X	X	X	X	X	X	X	
IncludesWeekend	X	X	X	X	X		X	X	X	X	X	X	X	X	X
IncludesNight	X	X	X	X	X	X		X	X	X	X	X	X	X	X
Shortage	X	X	X	X	X	X	X		X	X	X	X	X	X	X
UnloadQuantity	X														
WaitedToUnload	X	X	X	X	X	X	X	X	X	X		X	X	X	X
WasPriority	X	X	X	X	X	X	X	X	X	X	X		X	X	X
TimeEntrance	X	X													
TimeCollection	X	X	X	X											
TimeAnalysis	X	X	X	X	X	X	X	X							
TimeUnload	X	X	X	X	X	X	X	X	X						
TrucksInLine	X	X	X		X	X	X	X	X	X	X	X	X	X	X
Validation R^2 (average of 5 models)	0.804	0.803	0.808	0.717	0.815	0.786	0.792	0.723	0.814	0.819	0.794	0.805	0.758	0.786	0.810

In Model 6, the variable IncludesWeekend was removed. The prediction results of the model got worse, indicating that this variable has meaningful information about the business. The same happened with IncludesNight (Model 7) and Shortage (Model 8), as both were also considered meaningful to the study.

Then Models 9 and 10 tested variables TimeAnalysis and TimeUnload, respectively. There was not a meaningful impact on the model results; thus, both were removed from the subsequent models.

The next models (Models 11-15) tested the following variables: WaitedToUnload, WasPriority, IsGroupA, IsGroupB and IsGroupC. Each of these variables had useful predictive information. Thus, Model 10 was chosen. It included the following variables: IsGroupA, IsGroupB, IsGroupC, IncludesWeekend, IncludesNight, Shortage, WaitedToUnload, WasPriority, and TrucksInLine.

Through this process of variable selection and testing, I discovered that five variables were not meaningful in determining the outcome of the system. Thus, these variables could be dropped from the model.

Moreover, the chosen model can be very useful for prediction purposes. If the manufacturer decides to predict how long it will take for a truck to unload its materials as soon as the truck arrives in the system, this model can be applied. When the truck arrives there is no information on variables such as TimeAnalysis, TimeUnload, TimeEntrance and TimeCollection. These variables will only be known after the truck leaves the system. However, variables such as IsGroupA, IsGroupB, IsGroupC, IncludesWeekend, IncludesNight, TrucksInLine, WasPriority, WaitedToUnload, can be known right at the time the truck arrives in

the system. Using the model can make it possible for the manufacturer to make predictions right away about how long the process will take.

4.3 Connection Weight Approach

This section describes the process used to determine the importance of the input variables. Three steps were performed to define the importance scores and to rank the relative importance of the variables. In each test fifty artificial neural network models were created and their resulting weights were recorded. In the first test, the artificial neural network models included all variables listed in Section 3.4.1. In the second test, the models included only those variables that were considered meaningful according to the study done on variable selection. Tests were performed to determine the most accurate rank of the variables. Results of these testing showed that some variables tended to always be important, while others fluctuated much more in terms of importance and ended up with scores that are very similar to those of other variables.

The third test excluded the variable TrucksInLine because it usually dominated the models as the most important input variable. Removal of the TrucksInLine variable made it possible to get a clearer picture of the value of the remaining input variables. The tests meant to understand whether the quantity of variables impacted the ranking results. Specifically, it was meant to observe whether or not there would be a higher distinction between remaining variables after TrucksInLine was removed.

4.3.1 Ranking When All Input Variables Are Included

The results found after applying the connection weight approach to all variables are shown in Figure 4-5 and Figure 4-4. The graph shown in Figure 4-3 represents the average importance scores for each variable of all fifty artificial neural network models created. While the graph shown in Figure 4-4 represents each importance score obtained and how spread out they are. The red lines in each boxplot shown in Figure 4-4 represent the average.

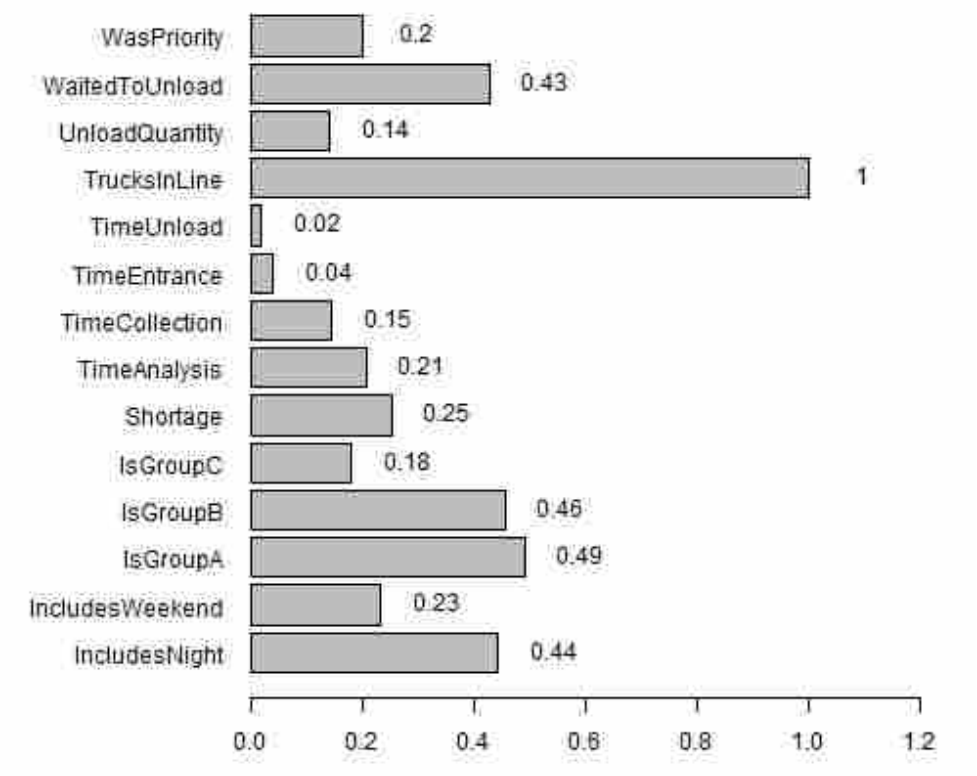

Figure 4-3: Average Importance Scores Including All Variables

The variable TrucksInLine had the highest importance score in all models. Thus, its score is represented by one, which is the highest score possible. The next most important variables are IsGroupA, IsGroupB, IncludesNight and WaitedToUnload. However, they have very similar

average scores, making it hard to define which variables are actually the most meaningful to the model.

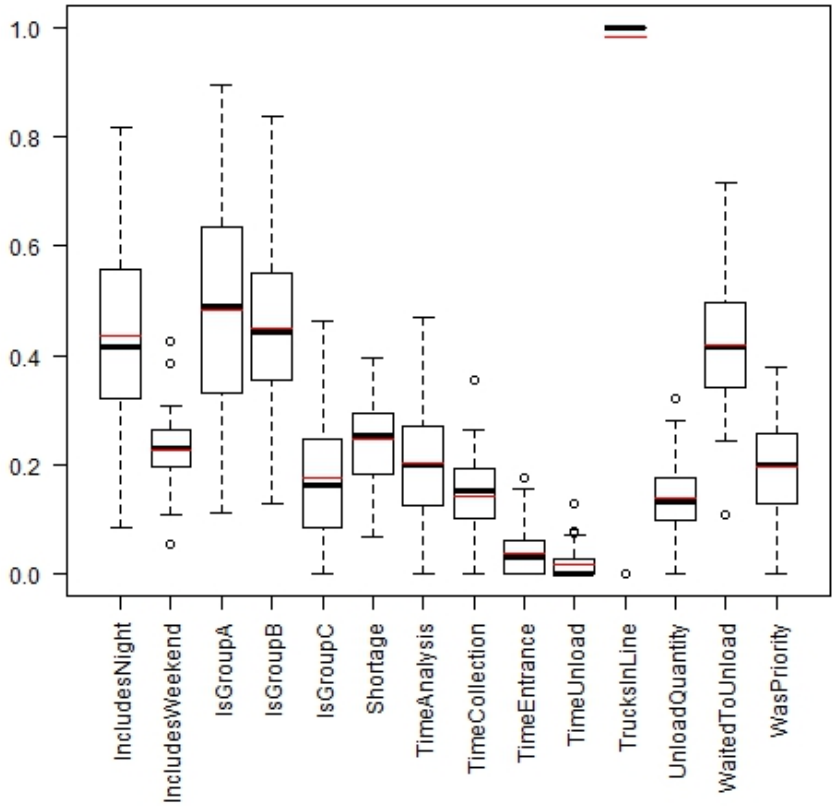

Figure 4-4: Importance Scores from All Models Including All Variables

As shown in Figure 4-3, those variables that were considered less meaningful to the artificial neural network models in the variable selection process had the smallest importance scores in the connection weight approach.

4.3.2 Ranking When Only Meaningful Variables Are Included

In the second test, only the variables that contributed predictive information to the model were included. These variables were listed in Section 4.2.3. The new results are shown in Figure 4-5 and Figure 4-6. Again, Figure 4-5 shows the average of the overall input contribution, while

Figure 4-6 shows all values and how they are spread out. The thin red lines in each boxplot shown in Figure 4-6 represent the average.

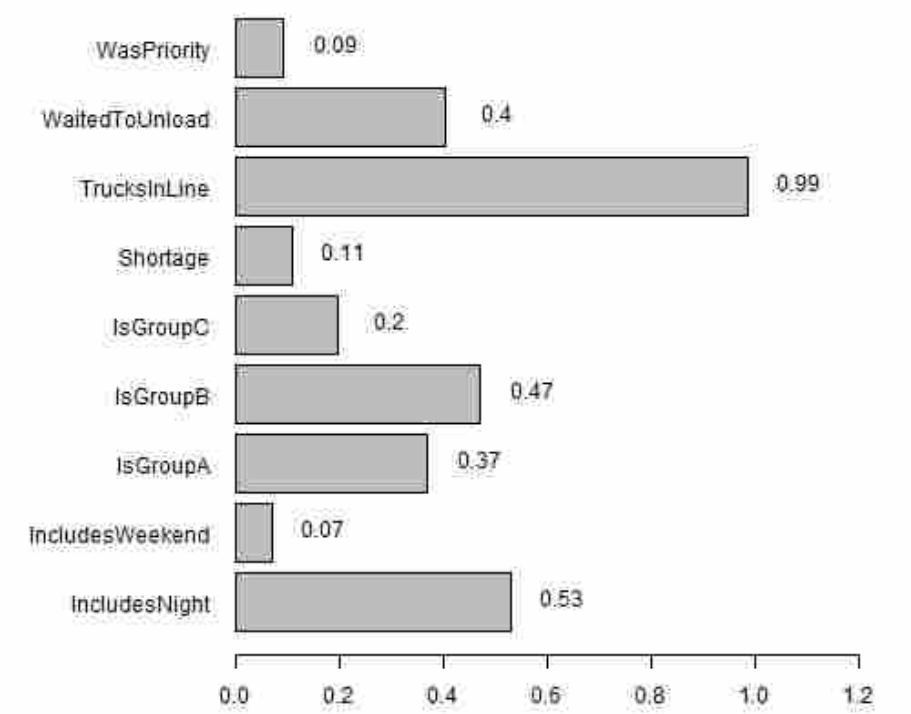

Figure 4-5: Average Importance Scores Including Meaningful Variables

Through the graphs above it is possible to observe that the variable TrucksInLine is still the most important variable in predicting the outcome of the system, followed by IncludesNight, IsGroupB, WaitedToUnload and IsGroupA, in this order. The new scores are in a different order from the previous rank. Thus, removal of variables that did not contribute predictive information to the model made the relative importance of the remaining variables clearer.

Variables IncludesNight, IsGroupB, WaitedToUnload and IsGroupA still have similar scores in the second test. However, the model is more sensitive to existing differences. This is shown as the scores differences are higher than in the previous test.

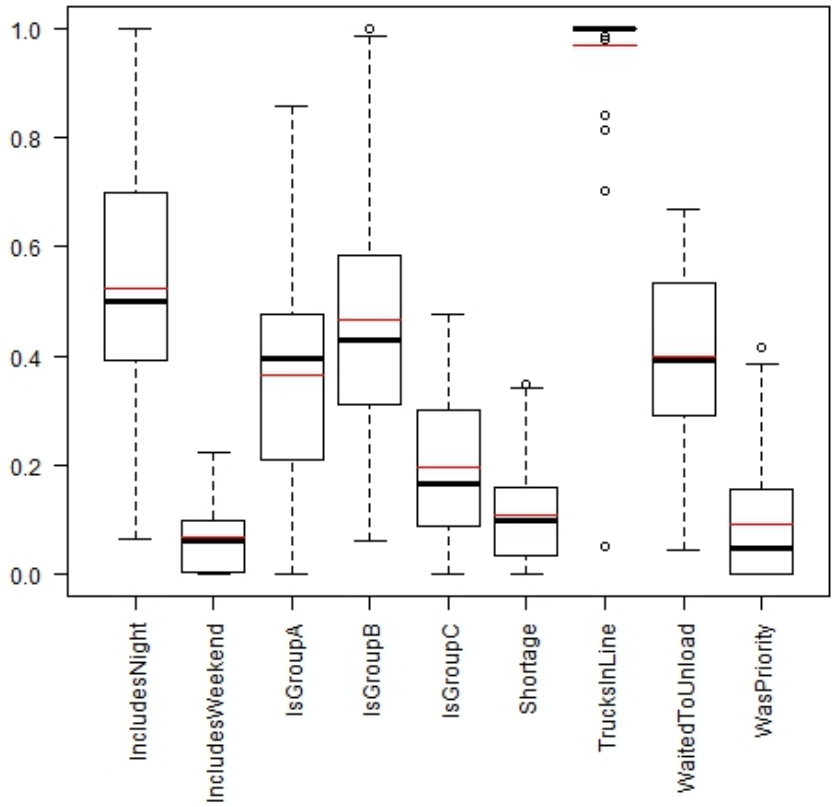

Figure 4-6: Importance Scores from All Models Including Meaningful Variables

The comparison between both scores from the first and second test is shown in Table 4-6 and in Figure 4-7. It is possible to see that there was not much discrepancy in the average importance scores for the variables TrucksInLine, IsGroupB, IsGroupC and WaitedToUnload. On the other hand, there were some differences in the scores of the other variables, the highest difference being 0.16.

Table 4-6: Mean and Standard Deviation Comparison

	All Variables		Meaningful Variables		Mean Differences	Standard Deviation Differences
	Mean	Standard Deviation	Mean	Standard Deviation		
TrucksInLine	1.00	0.00	0.99	0.05	0.01	0.05
IsGroupA	0.49	0.18	0.37	0.21	0.12	0.03
IsGroupB	0.46	0.15	0.47	0.24	0.02	0.09
IncludesNight	0.44	0.16	0.53	0.22	0.09	0.06
WaitedToUnload	0.43	0.11	0.40	0.15	0.02	0.04
Shortage	0.25	0.07	0.11	0.09	0.14	0.02
IncludesWeekend	0.23	0.06	0.07	0.07	0.16	0.01
WasPriority	0.20	0.09	0.09	0.12	0.11	0.03
IsGroupC	0.18	0.12	0.20	0.15	0.02	0.03

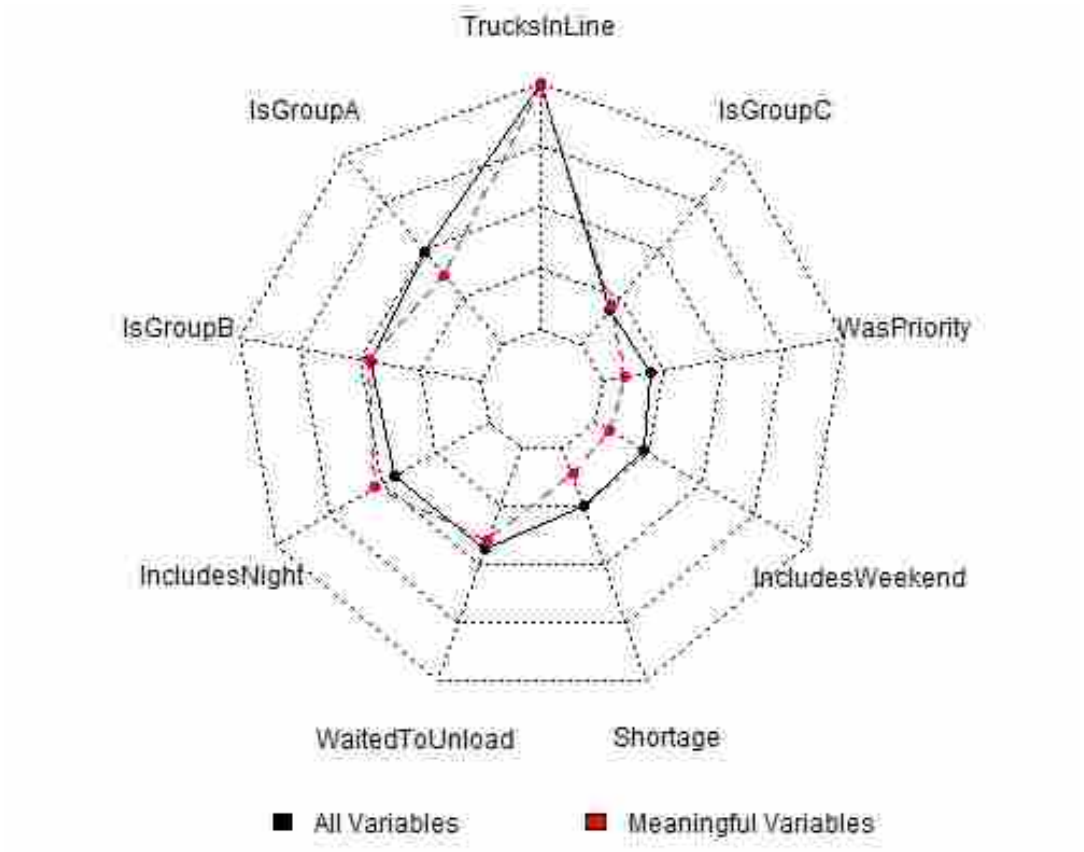

Figure 4-7: Mean Comparison

4.3.3 Ranking Excluding the Most Meaningful Variables

The first and second test did not give many insights on how to rank the variables that have very similar scores. In order to understand better how to make a distinction between variables IncludesNight, IsGroupB, WaitedToUnload and IsGroupA a third test was performed. In this test variable TrucksInLine was excluded from the artificial neural network models. As TrucksInLine was the most influential input variable, it was possible that this dominated the models such that other variables could not differenciate themselves from the others that had similar scores. The results from the third test are shown in Figure 4-8 and Figure 4-9.

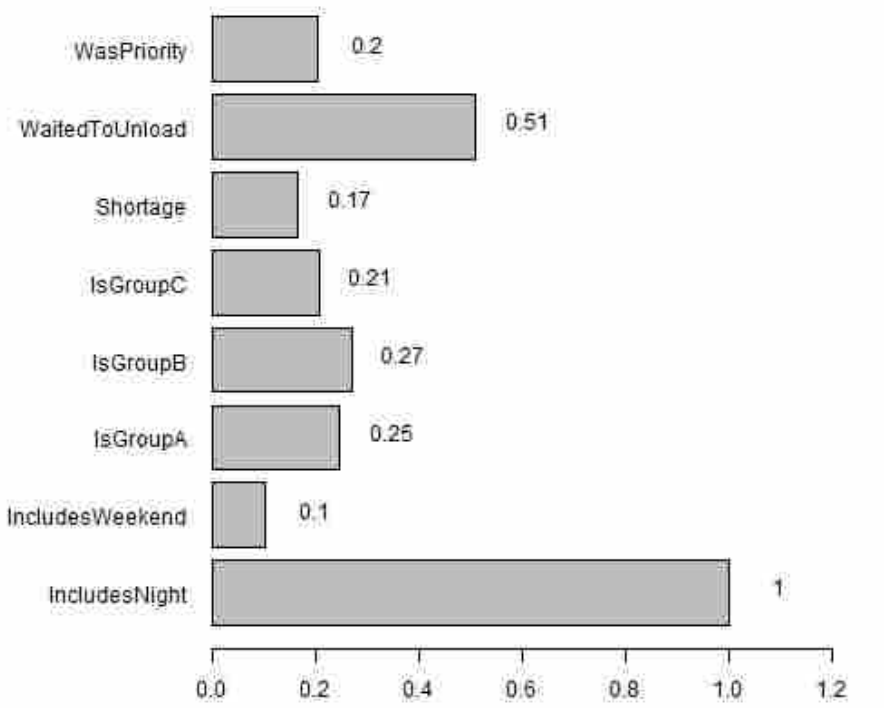

Figure 4-8: Average Importance Scores Excluding TrucksInLine from Second Test

Through Figure 4-8 it is possible to see a clear distinction between the first and the second variables, which are IncludesNight and WaitedToUnload. The other variables, however, have very similar scores, making it hard again to accurately make a distinction between them.

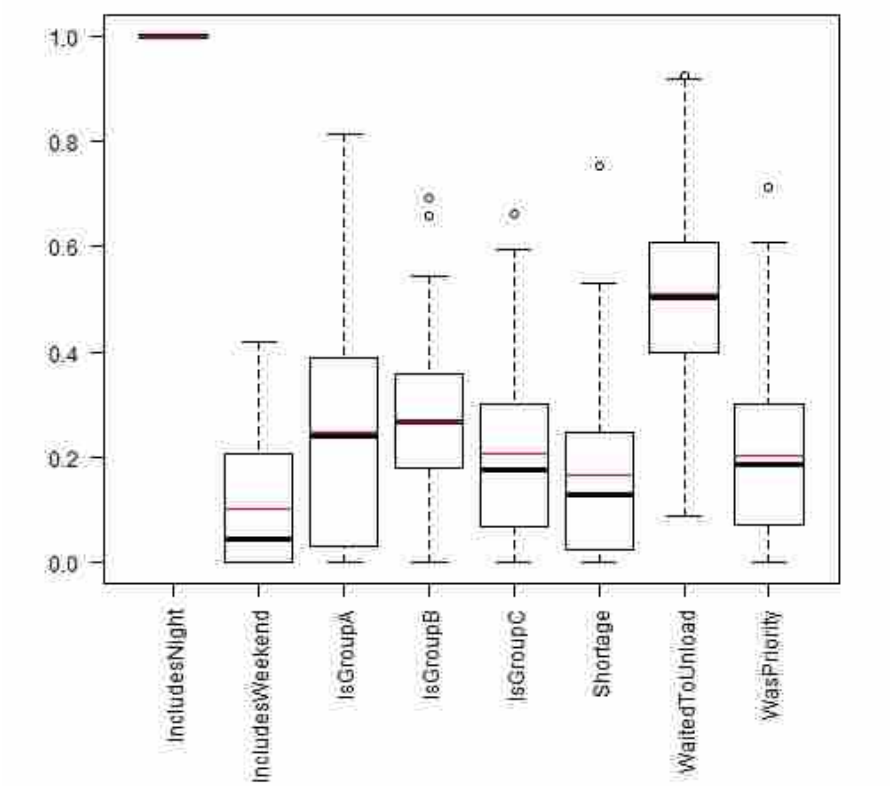

Figure 4-9: Importance Scores Excluding TrucksInLine from Second Test

Through the results of the three tests it was possible to observe that the connection weight approach tends to predict the best variable among the ones studied very accurately. While the next importance scores tend to be similar, making it hard to make a clear distinction between them. When the best variable is taken out of the models it is possible to see another variable that stands out.

When many variables are included in the model it is hard to determine an accurate ranking of the variables. And through the tests it is possible to see that the first ranking created was not accurate, as variable IncludesNight was ranked as number four, while in the following tests it was ranked as the second most important variable. This indicates that an iterative process to rank variables might be beneficial, as it will make possible for variables to stand out and not be hindered by the scored of the most important variable.

4.4 Simulation Improvement

To further test whether or not the most meaningful variables impacted on the results of the output of the system, each variable was analyzed. I split instances into two groups based on the values of the output across these two groups. I found that those variables considered important impacted the output of the system, while those considered not important had little or no impact.

The highest ranked variable listed in the study was TrucksInLine. I created two groups. One contained trucks that arrived with a below average number of trucks in line. The other group contained trucks that arrived with an above average number of trucks in line. As shown in Figure 4-10 trucks that arrived when the number of trucks was above the average number of trucks in line had to wait more to unload materials.

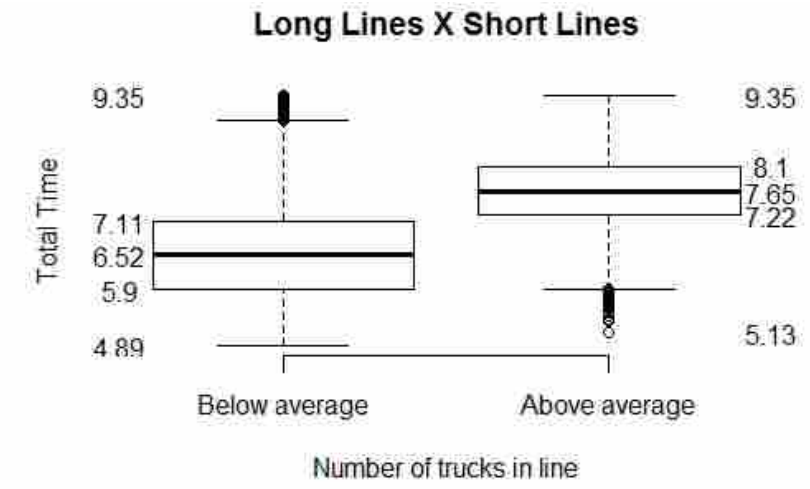

Figure 4-10: Total Time for Long and Short Lines

The second highest ranked variables listed in the study was IncludesNight. The same test was performed. Those trucks that had to wait over night to have their material unloaded stayed longer in the system. This is observed in Figure 4-11.

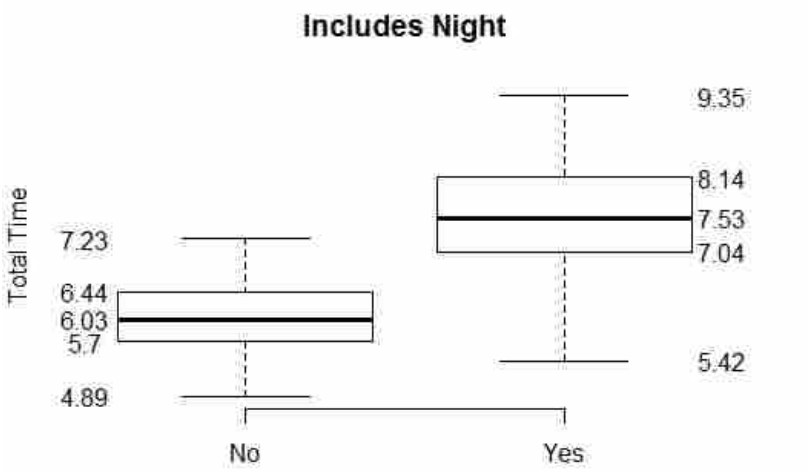

Figure 4-11: Total Time for Trucks that Waited Versus Did Not Wait Overnight

The third most meaningful variable is IsGroupB. The graph indicates that if the unload material in the truck belongs to Group B it will take less time in the facility than trucks containing other groups. This is shown in Figure 4-12.

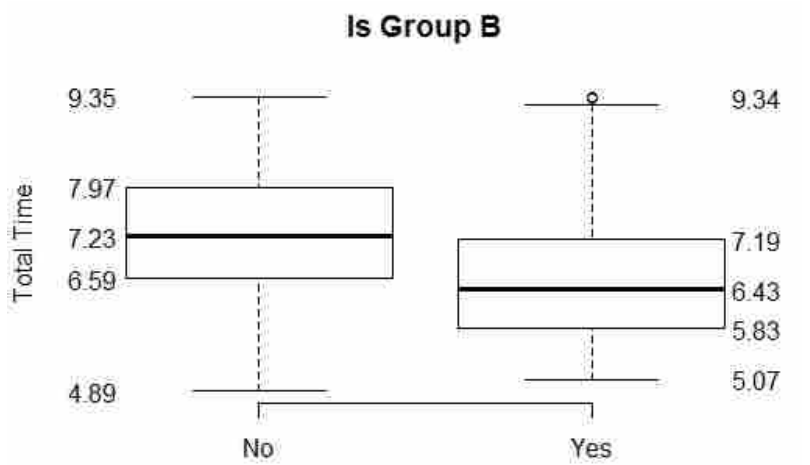

Figure 4-12: Total Time Group B Versus Other Groups

The fourth ranked variable that was considered meaningful was WaitedToUnload. The study indicates that those trucks that had to wait for available space so they would be able to unload its materials tended to stay longer in the facility. This is shown in Figure 4-13.

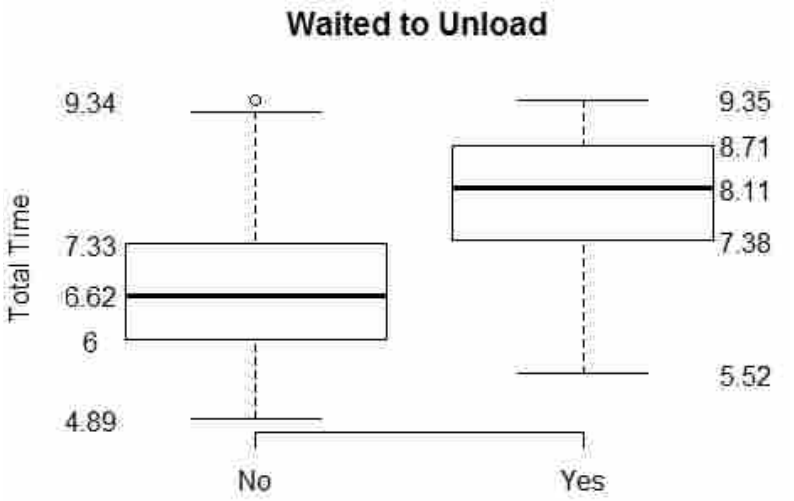

Figure 4-13: Total Time for Trucks that Waited Versus Did Not Wait to Unload

It can be observed that those variables that had the highest importance scores tended to have significant differences on the output of the system when the variable values changed. The differences observed in the median of those variables are listed in Table 4-7.

Table 4-7: Median Differences of Most Meaningful Variables

	Median Model 1	Median Model 2	Differences
TrucksInLine	6.52	7.65	1.13
IncludesNight	6.03	7.53	1.50
IsGroupB	7.23	6.43	0.80
WaitedToUnload	6.62	8.11	1.49

On the other hand, those variables that received small importance scores did not have much impact on the output of the system. The variable that was considered the least meaningful in the first test was TimeUnload. Through the analysis it was possible to see in fact that different values of this variable did not have an impact on the output of the system. As it can be seen in Figure 4-14 the total time a truck stays in the system is not impacted by the time it takes to unload its material.

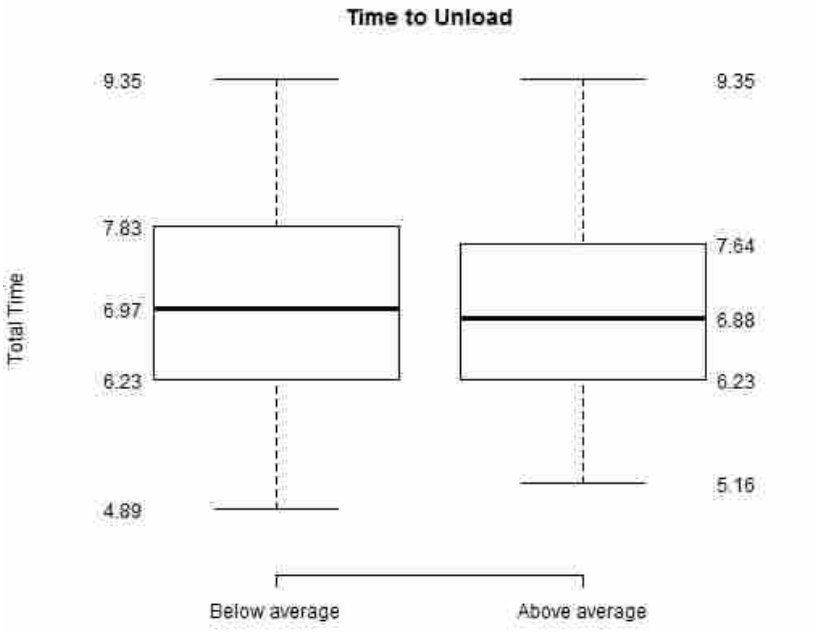

Figure 4-14: Impact of WaitedToUnload on Total Time

The same can be observed with the other variables that were taken from the artificial neural network models. The boxplots for variables TimeEntrance, UnloadQuantity and TimeCollection are shown in Figure 4-15, Figure 4-16 and Figure 4-17 respectively. A table containing their median and differences is shown in Table 4-8. Comparing Table 4-7 and Table 4-8 it is possible to see that those variables that were considered meaningful have higher differences in the medians of the different distributions, while those variables considered less meaningful had the lower differences in the medians. This indicates that when the values of those variables that are considered more meaningful were changed, there was a higher impact on the output variable. On the other hand, those variables that were considered less meaningful had very similar median values, indicating that a change in a variable that is less meaningful to the system does not impact much the output variable.

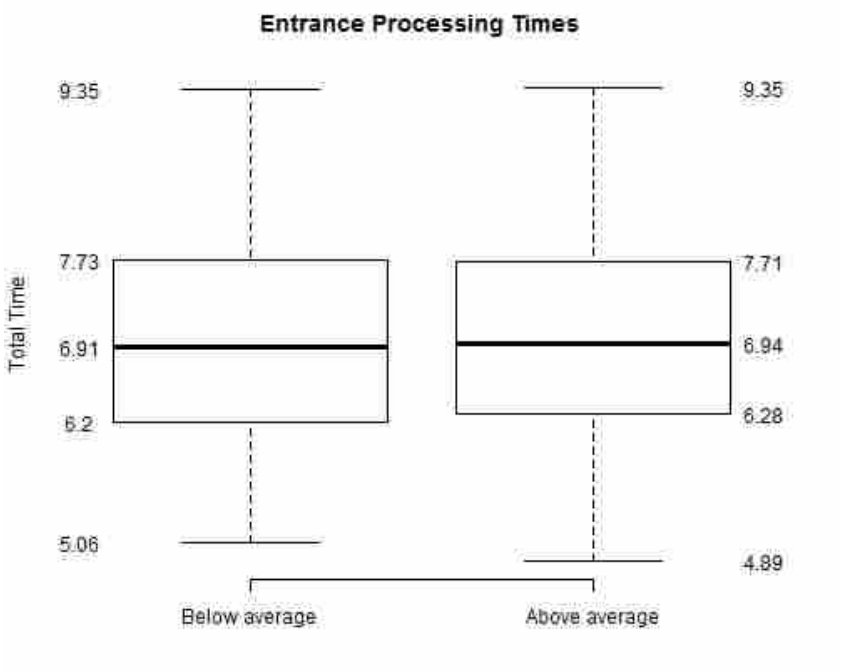

Figure 4-15: Impact of Entrance Time on Total Time

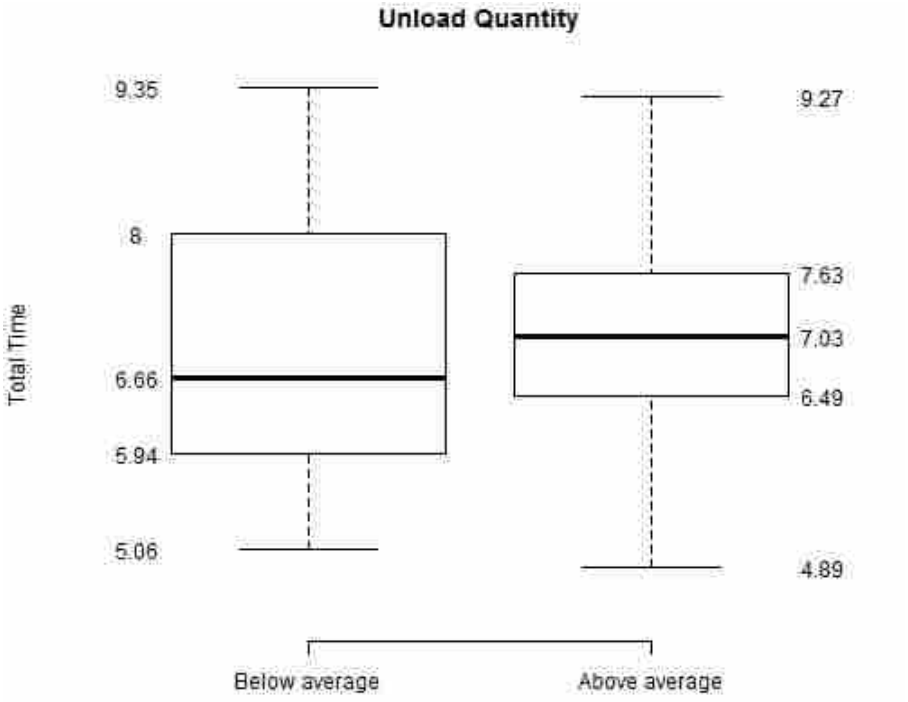

Figure 4-16: Impact of Unload Quantity on Total Time

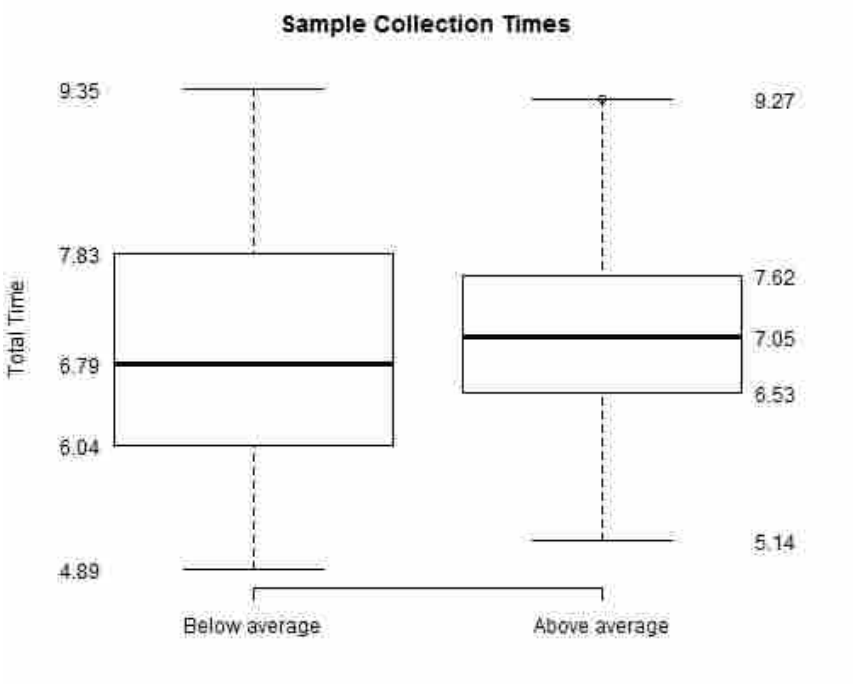

Figure 4-17: Impact of Collection Time on Total Time

Table 4-8: Median Differences of Least Meaningful Variables

	Median Model 1	Median Model 2	Differences
TimeUnload	6.97	6.88	0.09
TimeEntrance	6.91	6.94	0.03
UnloadQuantity	6.66	7.03	0.37
TimeCollection	6.79	7.05	0.26

One important observation is that if the manufacturer is looking specifically to improve efficiency of the stations existent in his process, it is important to have a model that will only take into consideration the variables that account for that. This is shown in Figure 4-18 and Figure 4-19. The figures show that the most important operation that should be improved in order to have better efficiency is the unloading operation time. The next one is the analysis time. These two are the operations that have the most impact. However, it is important to notice that other factors tend to impact much more the system them these operations, such as the number of

trucks in line, whether or not a truck arrives close to the time when no unload is performed and whether or not the truck has to wait for the storage to open space or not. These variables are related to the scheduling process and are the most impactful to the process and should be controlled in order to reduce the total time a truck stays in the system.

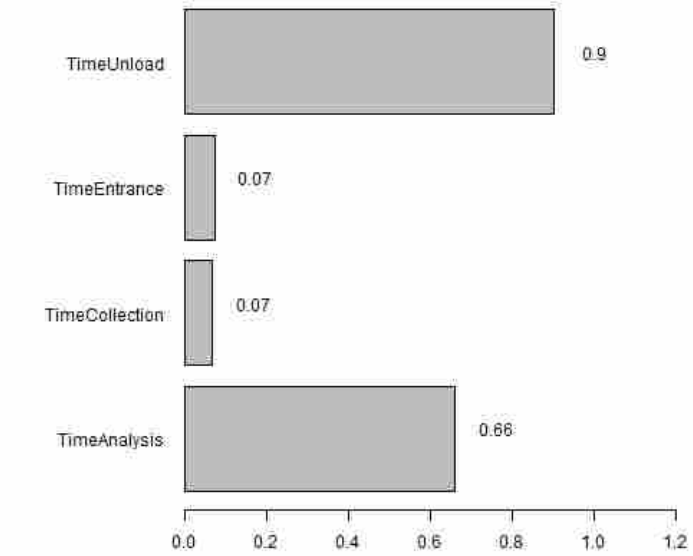

Figure 4-18: Average Importance Scores for Process Variables

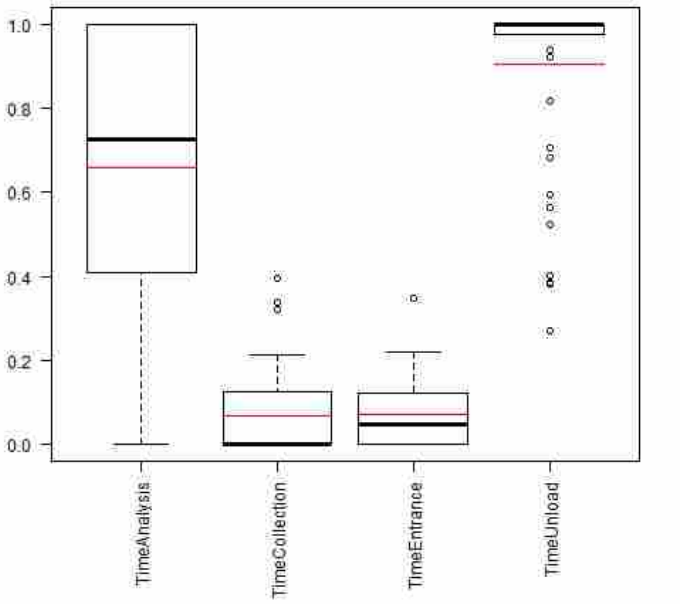

Figure 4-19: Importance Scores for Process Variables

5 SUMMARY AND FUTURE WORK

The present research applied the connection weight approach to artificial neural network models to interpret these models and find which variables are the most meaningful to it. This was done to a dataset generated by a discrete event simulation model that represented the operation of the receiving sector of a manufacturing. Defining which variables were the most important to the simulation model made it possible to improve it and use this information as a guide to efficiency improvement.

Through the study it was possible to observe that the data generated from a simulation model was useful and beneficial to data mining applications. In the case study there was not enough data to describe the system. This issue was solved by creating a simulation model that accurately represented the system to generate data. The data was useful to create artificial neural network models that could be used to make accurate predictions about the system. This indicated that in situations where there is a lack of data available for the creation of data mining models, simulation can be a good substitute.

Moreover, data mining was useful to a discrete event simulation model. This happened because through the interpretation of the artificial neural network models it was possible to guide the process of improving efficiency in a simulation model. Thus, combining discrete event simulation with data mining techniques is beneficial and can bring more insights about the process than by using just one of the methods by itself.

Discrete event simulation optimization can be very complex. In this research data mining supported the process of improving the simulation model by prioritizing which variables when adjusted will cause the highest improvement in the system. A relevant topic for future research would be to use this methodology as an optimization tool to not only prioritize, but also automatically optimize a simulation model.

Another conclusion of this study is that artificial neural network models can also be applied to a manufacturing environment, making good predictions about the outcomes of the system and bringing insights about the relationships of the variables involved in the process.

Furthermore, despite the high complexity of artificial neural network models, it is possible to interpret them. Although this is not an intuitive process, it can be performed and good insights can be extracted from it, such as a better understanding of the relationships that exist between variables.

The connection weight approach was useful in determining which variables are the most important to the output of the artificial neural network models. However, the ordinal ranking approach by itself does not seem to provide enough information about the real importance of the input variables. This occurs because some variables have very similar scores, making it hard to create a distinction between them. The ranking approach used in this study of normalizing the importance scores and using these scores to rank input variables according to their importance revealed and improved representation of how important the variable is in the model.

Moreover, the process of iteratively taking the most important variable out of the model and ranking the remnant variables can be beneficial. This process makes it possible for variables

that might have their scored hindered by the most important variables to stand out. In this research the process of ranking variables iteratively created a more accurate rank.

The high instability existent in artificial neural network models, due to its randomness, makes the interpretation process difficult. The creators of the connection weight approach determined that the ranking process should be done a large number of times. However, they did not specify either an acceptable range or number. In this research, this ranking process was repeated fifty times, which proved useful in terms of differentiating the impact of the ranked input variables. The histograms produced on fifty repetitions provided enough information to produce a useful box-and-whisker plot of the relative importance scores. This provides a graphical representation of the distribution of this data, which is helpful in determining hoe distributed there results were. Future research could further explore how the number of repetitions impacts the ranking stability. If it does, what would be a good number of models and how should this number be determined so that the ranking is accurate and consistent?

Moreover, each artificial neural network model tends to have different error rates. It would be interesting for future research to understand how much the errors of the model can impact the final ranking capabilities. Does a model with low error rate tend to rank variables more accurately than a model with a high error rate?

In this research, as part of the data cleaning process to create artificial neural network models, a trial and error approach was used to select which variables should be in the model. This approach can be time consuming and inefficient. In future research it would be interesting to apply more scientific methods for this process.

The use of artificial neural network models to prioritize variables was helpful in this study. One more question that arises is: could other data mining techniques such as linear regression and support vector machines do the same thing? Would other algorithms be better at ranking variables? Future research making a comparison of the different algorithms would be very interesting.

REFERENCES

Amaran, S., Sahinidis, N. V., Sharda, B., & Bury, S. J. (2016). Simulation optimization: a review of algorithms and applications. *Annals of Operations Research, 240*(1), 351-380. doi:10.1007/s10479-015-2019-x

Amato, F., López, A., Peña-Méndez, E. M., Vaňhara, P., Hampl, A., & Havel, J. (2013). Artificial neural networks in medical diagnosis. *Journal of Applied Biomedicine, 11*(2), 47-58. doi:https://doi.org/10.2478/v10136-012-0031-x

Anderson, R. K. (2012). *Visual Data Mining: The VisMiner Approach*: John Wiley & Sons.

Banks, J., Carson II, J. S., & Barry, L. (2005). *Discrete-event system simulationfourth edition*: Pearson.

Better, M., Glover, F., Kochenberger, G., & Wang, H. (2008). SIMULATION OPTIMIZATION: APPLICATIONS IN RISK MANAGEMENT. *International Journal of Information Technology & Decision Making, 7*(4), 571-587.

Better, M., Glover, F., & Laguna, M. (2007). Advances in analytics: Integrating dynamic data mining with simulation optimization. *IBM Journal of Research and Development, 51*(3.4), 477-487. doi:10.1147/rd.513.0477

Brady, T. F., & Yellig, E. (2005). *Simulation data mining: a new form of computer simulation output*. Paper presented at the Proceedings of the 37th conference on Winter simulation, Orlando, Florida.

Carson, Y., & Maria, A. (1997). *Simulation optimization: methods and applications*. Paper presented at the Proceedings of the 29th conference on Winter simulation, Atlanta, Georgia, USA.

Çiflikli, C., & Kahya-Özyirmidokuz, E. (2010). Implementing a data mining solution for enhancing carpet manufacturing productivity. *Knowledge-Based Systems, 23*(8), 783-788. doi:http://dx.doi.org/10.1016/j.knosys.2010.05.001

Cigolini, R., Pero, M., Rossi, T., & Sianesi, A. (2014). Linking supply chain configuration to supply chain perfrmance: A discrete event simulation model. *Simulation Modelling Practice and Theory, 40*, 1-11. doi:http://dx.doi.org/10.1016/j.simpat.2013.08.002

Diaz-Elsayed, N., Jondral, A., Greinacher, S., Dornfeld, D., & Lanza, G. (2013). Assessment of lean and green strategies by simulation of manufacturing systems in discrete production environments. *CIRP Annals - Manufacturing Technology, 62*(1), 475-478. doi:http://dx.doi.org/10.1016/j.cirp.2013.03.066

Dil, E. A., Ghaedi, M., Ghaedi, A. M., Asfaram, A., Goudarzi, A., Hajati, S., . . . Gupta, V. K. (2016). Modeling of quaternary dyes adsorption onto ZnO–NR–AC artificial neural network: Analysis by derivative spectrophotometry. *Journal of Industrial and Engineering Chemistry, 34*, 186-197. doi:https://doi.org/10.1016/j.jiec.2015.11.010

Dimopoulos, I., Chronopoulos, J., Chronopoulou-Sereli, A., & Lek, S. (1999). Neural network models to study relationships between lead concentration in grasses and permanent urban descriptors in Athens city (Greece). *Ecological Modelling, 120*(2–3), 157-165. doi:https://doi.org/10.1016/S0304-3800(99)00099-X

Dimopoulos, Y., Bourret, P., & Lek, S. (1995). Use of some sensitivity criteria for choosing networks with good generalization ability. *Neural Processing Letters, 2*(6), 1-4. doi:10.1007/bf02309007

Djanatliev, A., & German, R. (2013). *Prospective healthcare decision-making by combined system dynamics, discrete-event and agent-based simulation*. Paper presented at the Proceedings of the 2013 Winter Simulation Conference: Simulation: Making Decisions in a Complex World, Washington, D.C.

Fishman, G. (2013). *Discrete-event simulation: modeling, programming, and analysis*: Springer Science & Business Media.

Fonseca, D. J., Navaresse, D. O., & Moynihan, G. P. (2003). Simulation metamodeling through artificial neural networks. *Engineering Applications of Artificial Intelligence, 16*(3), 177-183. doi:http://dx.doi.org/10.1016/S0952-1976(03)00043-5

Freiberg, F., & Scholz, P. (2015). Evaluation of Investment in Modern Manufacturing Equipment Using Discrete Event Simulation. *Procedia Economics and Finance, 34*, 217-224. doi:http://dx.doi.org/10.1016/S2212-5671(15)01622-6

Fu, M. C., Glover, F. W., & April, J. (2005, 4-7 Dec. 2005). *Simulation optimization: a review, new developments, and applications.* Paper presented at the Proceedings of the Winter Simulation Conference, 2005.

Garson, G. D. (1991). Interpreting Neural-Network Connection Weights. *AI Expert, 6*, 47-51.

Gevrey, M., Dimopoulos, I., & Lek, S. (2003). Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecological Modelling, 160*(3), 249-264. doi:http://dx.doi.org/10.1016/S0304-3800(02)00257-0

Ghasemi, S., Ghasemi, M., & Ghasemi, M. (2011). Knowledge Discovery in Discrete Event Simulation Output Analysis. In P. Pichappan, H. Ahmadi, & E. Ariwa (Eds.), *Innovative Computing Technology: First International Conference, INCT 2011, Tehran, Iran, December 13-15, 2011. Proceedings* (pp. 108-120). Berlin, Heidelberg: Springer Berlin Heidelberg.

Gröger, C., Niedermann, F., & Mitschang, B. (2012). *Data mining-driven manufacturing process optimization.* Paper presented at the Proceedings of the world congress on engineering.

Harding, J. A., Shahbaz, M., Srinivas, & Kusiak, A. (2005). Data Mining in Manufacturing: A Review. *Journal of Manufacturing Science and Engineering, 128*(4), 969-976. doi:10.1115/1.2194554

Harrell, C., Ghosh, B. K., & Bowden, R. O. (2011). *Simulation using promodel*: Boston: McGraw-Hil.

Hemmat Esfe, M., Saedodin, S., Sina, N., Afrand, M., & Rostami, S. (2015). Designing an artificial neural network to predict thermal conductivity and dynamic viscosity of ferromagnetic nanofluid. *International Communications in Heat and Mass Transfer, 68*, 50-57. doi:https://doi.org/10.1016/j.icheatmasstransfer.2015.06.013

Hsieh, N.-C., & Chu, K.-C. (2009). Enhancing consumer behavior analysis by data mining techniques. *International Journal of Information and Management Sciences, 20*(1), 39-53.

Johnston, M. J., Paige, J. T., Aggarwal, R., Stefanidis, D., Tsuda, S., Khajuria, A., & Arora, S. (2016). An overview of research priorities in surgical simulation: what the literature shows has been achieved during the 21st century and what remains. *The American Journal of Surgery, 211*(1), 214-225. doi:http://dx.doi.org/10.1016/j.amjsurg.2015.06.014

Joshi, D. M., Rana, N. K., & Misra, V. M. (2010, 7-10 May 2010). *Classification of Brain Cancer using Artificial Neural Network.* Paper presented at the 2010 2nd International Conference on Electronic Computer Technology.

Jun, J. B., Jacobson, S. H., & Swisher, J. R. (1999). Application of discrete-event simulation in health care clinics: A survey. *Journal of the Operational Research Society, 50*(2), 109-123. doi:10.1057/palgrave.jors.2600669

Kusiak, A., & Smith, M. (2007). Data mining in design of products and production systems. *Annual Reviews in Control, 31*(1), 147-156. doi:https://doi.org/10.1016/j.arcontrol.2007.03.003

Lek, S., Belaud, A., Baran, P., Dimopoulos, I., & Delacoste, M. (1996). Role of some environmental variables in trout abundance models using neural networks. *Aquat. Living Resour., 9*(1), 23-29.

Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J., & Aulagnier, S. (1996). Application of neural networks to modelling nonlinear relationships in ecology. *Ecological Modelling, 90*(1), 39-52. doi:http://dx.doi.org/10.1016/0304-3800(95)00142-5

Liao, S.-H., Chu, P.-H., & Hsiao, P.-Y. (2012). Data mining techniques and applications – A decade review from 2000 to 2011. *Expert Systems with Applications, 39*(12), 11303-11311. doi:http://dx.doi.org/10.1016/j.eswa.2012.02.063

Manufacturers, N. A. o. Top 20 Facts About Manufacturing. Retrieved from http://www.nam.org/Newsroom/Top-20-Facts-About-Manufacturing/

Marsland, S. (2015). *Machine learning: an algorithmic perspective*: CRC press.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics, 5*(4), 115-133. doi:10.1007/bf02478259

Ngai, E. W. T., Xiu, L., & Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications, 36*(2, Part 2), 2592-2602. doi:http://dx.doi.org/10.1016/j.eswa.2008.02.021

Olafsson, S., Li, X., & Wu, S. (2008). Operations research and data mining. *European Journal of Operational Research, 187*(3), 1429-1448. doi:http://dx.doi.org/10.1016/j.ejor.2006.09.023

Olden, J. D., & Jackson, D. A. (2002). Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. *Ecological Modelling, 154*(1–2), 135-150. doi:http://dx.doi.org/10.1016/S0304-3800(02)00064-9

Olden, J. D., Joy, M. K., & Death, R. G. (2004). An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modelling, 178*(3–4), 389-397. doi:http://dx.doi.org/10.1016/j.ecolmodel.2004.03.013

Oña, J. d., & Garrido, C. (2014). Extracting the contribution of independent variables in neural network models: a new approach to handle instability. *Neural Computing and Applications, 25*(3), 859-869. doi:10.1007/s00521-014-1573-5

Öztürk, A., Kayalıgil, S., & Özdemirel, N. E. (2006). Manufacturing lead time estimation using data mining. *European Journal of Operational Research, 173*(2), 683-700. doi:http://dx.doi.org/10.1016/j.ejor.2005.03.015

Painter, M. K., Erraguntla, M., Gary L. Hogg, J., & Beachkofski, B. (2006). *Using simulation, data mining, and knowledge discovery techniques for optimized aircraft engine fleet management*. Paper presented at the Proceedings of the 38th conference on Winter simulation, Monterey, California.

Panayiotou, C. G., Cassandras, C. G., & Wei-Bo, G. (2000, 2000). *Model abstraction for discrete event systems using neural networks and sensitivity information.* Paper presented at the 2000 Winter Simulation Conference Proceedings (Cat. No.00CH37165).

Robinson, S. (2005). Discrete-event simulation: from the pioneers to the present, what next? *Journal of the Operational Research Society, 56*(6), 619-629. doi:10.1057/palgrave.jors.2601864

Sargent, R. G. (2005). *Verification and validation of simulation models*. Paper presented at the Proceedings of the 37th conference on Winter simulation, Orlando, Florida.

Sargent, R. G. (2013). Verification and validation of simulation models. *Journal of Simulation, 7*(1), 12-24. doi:10.1057/jos.2012.20

Scardi, M., & Harding Jr, L. W. (1999). Developing an empirical model of phytoplankton primary production: a neural network case study. *Ecological Modelling, 120*(2–3), 213-223. doi:https://doi.org/10.1016/S0304-3800(99)00103-9

Seng, J.-L., & Chen, T. C. (2010). An analytic approach to select data mining for business decision. *Expert Systems with Applications, 37*(12), 8042-8057. doi:http://dx.doi.org/10.1016/j.eswa.2010.05.083

Shao, G., Shin, S.-J., & Jain, S. (2014). *Data analytics using simulation for smart manufacturing*. Paper presented at the Proceedings of the 2014 Winter Simulation Conference, Savannah, Georgia.

Sharma, P. (2015). Discrete-event simulation. *International journal of scientific & technology research, 4*(4), 136-140.

Shmueli, G., Patel, N. R., & Bruce, P. C. (2016). *Data Mining for Business Analytics: Concepts, Techniques, and Applications with XLMiner*: John Wiley & Sons.

Silva, I. N. d., Spatti, D. H., Flauzino, R. A., Liboni, L. H. B., & dos Reis Alves, S. F. (2016). *Artificial Neural Networks: A Practical Course*: Springer.

Sokolowski, J. A., & Banks, C. M. (2010). *Modeling and simulation fundamentals: theoretical underpinnings and practical domains*: John Wiley & Sons.

Tekin, E., & Sabuncouglu, I. (2004). Simulation optimization: A comprehensive review on theory and applications. *IIE Transactions, 36*(11), 1067-1081. doi:10.1080/07408170490500654

Thiede, S., Seow, Y., Andersson, J., & Johansson, B. (2013). Environmental aspects in manufacturing system modelling and simulation—State of the art and research perspectives. *CIRP Journal of Manufacturing Science and Technology, 6*(1), 78-87. doi:http://dx.doi.org/10.1016/j.cirpj.2012.10.004

Yegnanarayana, B. (2009). *Artificial neural networks*: PHI Learning Pvt. Ltd.

Zaji, A. H., & Bonakdari, H. (2015). Application of artificial neural network and genetic programming models for estimating the longitudinal velocity field in open channel junctions. *Flow Measurement and Instrumentation, 41*, 81-89. doi:https://doi.org/10.1016/j.flowmeasinst.2014.10.011

APPENDIX A PROCESSING TABLES

Group A Processing Information

Location	Activity Time	Activity Resource	Next Location	Move Trigger	Move Time	Move Resource
Arrival Line	None	None	Entrance	When Entrance is available	N(5,5) min	None
Entrance	$1.94 + L(6.93, 3.65)$	Door man	Unload Line	When operation is finished	N(5,5) min	None
Unload Line	None	None	Mill Hopper	When Mill Hopper is available	N(10,10) min	None
Mill Hopper	$L(0.605, 0.371) * loadQuantity$ On weekends it is increased by 15%	Operator	Analysis Line	When operation is finished	N(10,10) min	Operator
Analysis Line	None	None	Laboratory	When Laboratory is available	None	Analyst
Laboratory	$28 + E(13)$ – 6.12% of the time $14 + L(20.2, 10.1)$ – The rest of the time	Analyst	Unload Line	When operation is finished	N(10,10) min	None
Unload Line	None	None	Scale	When scale is available	N(10,10) min	None
Scale	2 min	None	Mill Hopper	When Mill Hopper is available	N(5,5) min	None
Mill Hopper	$0.03 * loadQuantity$ – 6.12% of the time $L(0.835, 0.25) * loadQuantity$ – The rest of the time On weekends this time is increased by 15%	Operator	Exit	When operation is finished	None	None

Group B Processing Information

Location	Activity Time	Activity Resource	Next Location	Move Trigger	Move Time	Move Resource
Arrival Line	None	None	Entrance	When Entrance is available	N(5,5) min	None
Entrance	$1.94 + L(6.93, 3.65)$	Door man	Unload Line	When operation is finished	N(5,5) min	None
Unload Line	None	None	Mill Hopper	When Mill Hopper is available	N(10,10) min	None
Mill Hopper	$T(0, 0.653, 0.707) * loadQuantity$ On weekends this time is increased by 15%	Operator	Analysis Line	When operation is finished	N(10,10) min	Operator
Analysis Line	None	None	Laboratory	When Laboratory is available	None	Analyst
Laboratory	$56.3 + L(19.6, 38.9)$	Analyst	Unload Line	When operation is finished	N(10,10) min	None
Unload Line	None	None	Scale	When scale is available	N(10,10) min	None
Scale	2 min	None	Mill Hopper	When Mill Hopper is available	N(5,5) min	None
Mill Hopper	$L(1.24, 0.221) * loadQuantity$ 9.59% of the time there is an increase of 50% and on top of that on weekends this time is increased by 15%.	Operator	Exit	When operation is finished	None	None

Group C Processing Information

Location	Activity Time	Activity Resource	Next Location	Move Trigger	Move Time	Move Resource
Arrival Line	None	None	Entrance	When Entrance is available	N(5,5) min	None
Entrance	$1.94 + L(6.93, 3.65)$	Door man	Unload Line	When operation is finished	N(5,5) min	None
Unload Line	None	None	Mill Hopper	When Mill Hopper is available	N(10,10) min	None
Mill Hopper	$T(0, 0.653, 0.707) * loadQuantity$ On weekends this time is increased by 15%	Operator	Analysis Line	When operation is finished	N(10,10) min	Operator
Analysis Line	None	None	Laboratory	When Laboratory is available	None	Analyst
Laboratory	$-3030 + L(3100, 32.2)$	Analyst	Unload Line	When operation is finished	N(10,10) min	None
Unload Line	None	None	Scale	When scale is available	N(10,10) min	None
Scale	2 min	None	Mill Hopper	When Mill Hopper is available	N(5,5) min	None
Mill Hopper	$L(1.24, 0.221) * loadQuantity$ On weekends this time is increased by 15%.	Operator	Exit	When operation is finished	None	None

Group D Processing Information

Location	Activity Time	Activity Resource	Next Location	Move Trigger	Move Time	Move Resource
Arrival Line	None	None	Entrance	When Entrance is available	N(5,5) min	None
Entrance	$1.94 + L(6.93, 3.65)$	Door man	Unload Line	When operation is finished	N(5,5) min	None
Unload Line	None	None	Mill Hopper	When Mill Hopper is available	N(10,10) min	None
Mill Hopper	$T(0, 0.653, 0.707) * loadQuantity$ On weekends this time is increased by 15%	Operator	Analysis Line	When operation is finished	N(10,10) min	Operator
Analysis Line	None	None	Laboratory	When Laboratory is available	None	Analyst
Laboratory	$-3030 + L(3100, 32.2)$	Analyst	Unload Line	When operation is finished	N(10,10) min	None
Unload Line	None	None	Scale	When scale is available	N(10,10) min	None
Scale	2 min	None	Mill Hopper	When Mill Hopper is available	N(5,5) min	None
Mill Hopper	$L(0.835, 0.25) * loadQuantity$ On weekends this time is increased by 15%.	Operator	Exit	When operation is finished	None	None