University of Alberta

Productivity Analysis of Earthmoving Operations

by

Oscar Javier Gomez Rueda

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Science
in
Construction Engineering and Management

Department of Civil and Environmental Engineering

Examining Committee

Dr. Mohamed Al-Hussein, Department of Civil and Environmental Engineering,
University of Alberta

Dr. Simaan Abourizk, Department of Civil and Environmental Engineering,
University of Alberta

Dr. Tim Joseph, School of Mining and Petroleum Engineering,
Department of Civil and Environmental Engineering,
University of Alberta

Dr. Ahmed Bouferguene, Campus Saint-Jean,
University of Alberta

DEDICATION

This thesis is dedicated to my dear son Oscar and my beloved wife Johanna.

ABSTRACT

Heavy construction and mining general contractors record on a daily basis large amount of operational data. Nevertheless, this information is rarely used to enhance the knowledge and capabilities of the companies that spent great amount of money and resources recording it. This research presents different approaches on how to process this data to convert it in useful information. The prime goal of this analysis is to determine a suitable and convenient method to obtain and present historical productivities of key equipment, in order to provide a tool to aid estimating and generate reference information to support decision making.

Estimating construction operation productivity is mostly experience-based due to the complexity involved. However, predominantly empirical practices do not secure a reliable estimate because of the absence of a binding mechanism that relates the present case to past patterns (Chao and Skibniewski 1994). This study involved the analysis of the historical productivity of more than 230 hauling units, 160 excavator units, and 150 units of support equipment. The historical data has been recorded for about three years and represent the operations of one of the largest contractors on the Alberta Oil Sands in eleven different projects.

Data mining, artificial neural networks and summarization tools proved to assist effectively in the assessment of historical productivities and in the identification of the attributes that most influence the results. Multiple ANN configurations were evaluated in the determination of hauling trucks and excavators productivities. Ward net architectures that include different activation functions applied to hidden layer slabs performed better than standard backpropagation nets since they are able to detect different features in a pattern processed through a network.

ACKNOWLEDGEMENTS

This research was funded by the Organization of American States (OAS), North American Construction Group (NACG), the Natural Science and Engineering Research Council of Canada (NSERC) and by the University of Alberta, Construction Engineering and Management Group.

I would like to give my sincere appreciation to Dr. Mohamed Al-Hussein and Dr. Simaan M. AbouRizk for their mentorship during the course of my MSc studies. My special thanks to Mitch Holte and Anthony Bayduza from NACG for their support and valuable input on this analysis.

TABLE OF CONTENT

LIST OF TABLES

TABLE OF FIGURES

CHAPTER 1 INTRODUCTION

1.1 Motivation

In western Canada as in the rest of the world, heavy construction and mining general contractors record on a daily basis large amount of operational data. Nevertheless, this information is rarely used to enhance the knowledge and capabilities of the companies that spent great amount of money and resources recording it.

Several elements impact the productivity of heavy construction equipment. The production rate of a construction operation is constrained by not only the applied technology's capacities that are subject to the physical job conditions such as work dimensions and environment factors, but also its utilization rate or operating efficiency that is influenced by management circumstances. Estimating construction operation productivity is predominantly an experience-based task due to the complexity involved. According to experience, a contractor can intuitively adjust the standard rates in productivity handbooks to estimate for an operation in given project conditions. Nonetheless, such empirical practices do not guarantee a solid estimate because of the absence of a binding mechanism that relates the present case to past patterns (Chao and Skibniewski 1994).

Figure 1-1 shows a basic earthmoving operation where a set of hauling trucks are loaded by a single excavator. The cycle time of the process involves four main activities: loading, hauling, dumping and returning.

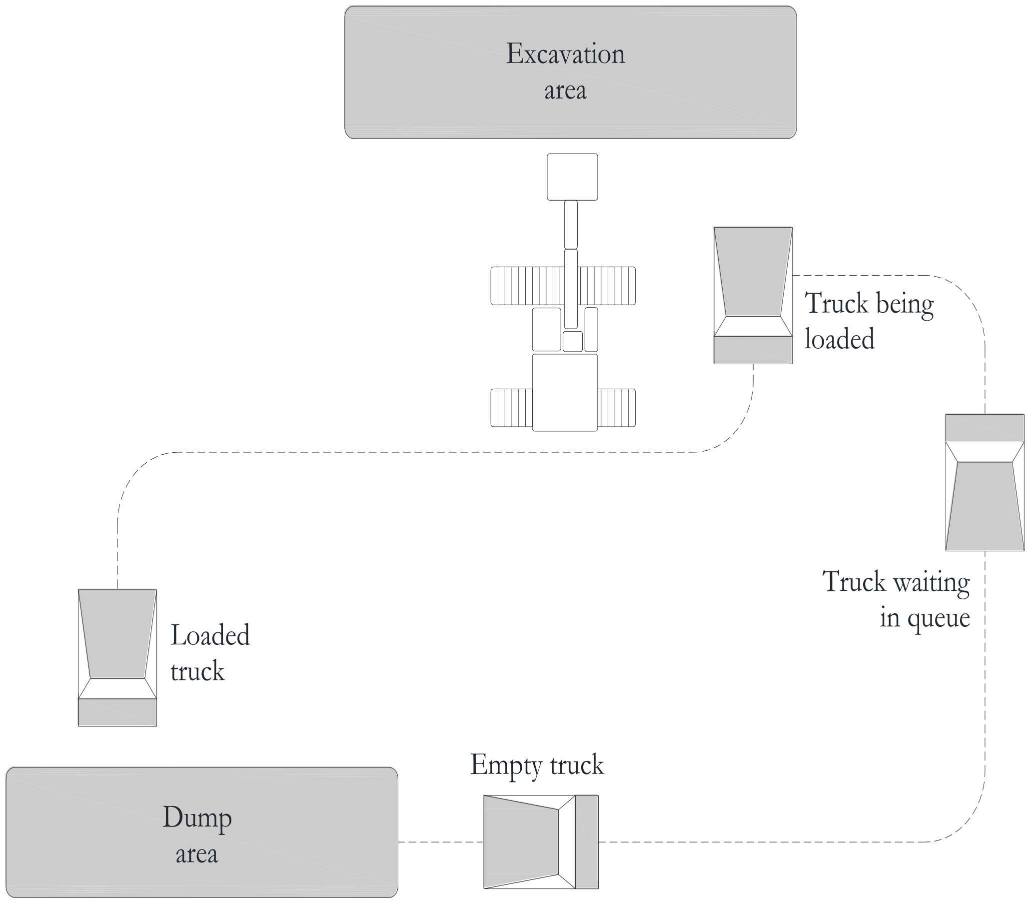


Figure 1-1 Excavation and Hauling Operation

1.2 Scope of research

This research focuses on the evaluation of historical operational data for the purpose of assessing the productivity of heavy construction and mining equipment. It presents different approaches on how to convert raw data into useful information.

1.3 Research Objectives

The aim of this research is to determine a suitable and convenient method to obtain and present historical productivities of key equipment in order to provide a tool to aid estimating and generate reference information to support decision making. To realize this goal, the following objectives have been attained:

- Implement, combine and compare different analysis tools and procedures to assess historical productivities of earthmoving operations.

- Identify common factors that affect the quality of the data being recorded and propose approaches to mitigate their effects.

- Analyze the sensitivity of different parameters of artificial neural networks in the evaluation of earthmoving operational data.

1.4 Thesis Organization

This thesis is composed of five chapters. Chapter 1 presents the motivation, general goal, objectives and the scope of the research. Chapter 2 gives a comprehensive overview of the current methodology to estimate earthmoving operations, and the state of the art of data mining and artificial neural networks used for productivity prediction. Chapter 3 summarizes an extensive analysis of earthmoving operations using summarization tools. Chapter 4 describes the use of data mining techniques and artificial neural networks in the assessment of earthmoving operations' productivities. Chapter 5 provides a summary report of activities performed, as well as a set of conclusions and contributions made by this MSc research.

1.5 Confidentiality

Confidential information has been used in the development of this research. Nevertheless, none of this information is released or published on this document. The author has taken special care in removing or modifying confidential data from tables, charts and other results. As part of this investigation a great amount of appendixes was generated for the general contractor, these appendixes has been declared confidential and not included as part of this thesis.

CHAPTER 2 LITERATURE REVIEW

2.1 Introduction

Equipment productivities of earthmoving operations have been deeply studied during the last fifty years. Current equipment specifications include sophisticated charts to compute their productivities in dependence of multiple external factors. There are numerous analytical methods that can be used to plan or analyze a construction operation. However, in most cases, analytical techniques require abstractions which tend to reduce confidence in model predictions (AbouRizk, Halpin and Hill 1991).

In today's increasingly competitive market, the contractor is required to plan and estimate an earthmoving contract as accurately as possible at tender stage and, if the contract is awarded, focus in control site operations to minimize cost. However, as it is widely recognized, production estimates at a tender stage are difficult to calculate accurately and at a construction stage, production targets are hard to maintain (Smith, Osborne and Forde 1995).

2.1.1 Simulation

In recent years the use of special purpose simulation systems in the analysis of earthmoving operations has expanded its popularity. Special purpose simulation (SPS) can be defined as a computer-based environment built to enable a practitioner who is knowledgeable in a given domain, but not necessarily in simulation, to model a project within that domain in a manner where symbolic representations, navigation schemes within the environment, creation of model specifications and reporting are completed in a format native to the domain itself (AbouRizk and Hajjar 1998).

Simulation studies are appropriate for the analysis of earthmoving operations for numerous reasons including: repetition of given operations, dynamics of resource interactions, external factors that need to be included in the analysis and the randomness associated with such systems (Hajjar and AbouRizk 1996). During the coursework of his master studies, the author was in contact with these simulation techniques, specifically with the use of the simulation tool Simphony. Simphony is a simulation platform for building general and special purpose simulation models. It is a Microsoft Windows based computer system developed with the objective of providing a standard, consistent and intelligent environment for both the development as well as the utilization of the construction of Special Purpose Simulation (SPS) tools (Hajjar and AbouRizk 1999). This research does not cover at this point any simulation technique as one of the tool to approach the analysis of the historical construction operational information.

2.2 Earthmoving Operations

Earthwork projects involve moving specific amounts of earth from a set of source locations to a set of destinations. Construction contractors use diverse methods and equipment to move earth depending primarily on their equipment availability and hauling distance (Kannan, Martinez and Vorster 1997).

The problem of accurate estimation of earthmoving productivity has intrigued many researches for decades; however a model that predicts the output of such operations with a satisfactory degree of confidence for all situations is not yet available (S. Smith 1999). Figure 2-1 presents a simplified overview of the estimation process of an earthmoving operation.

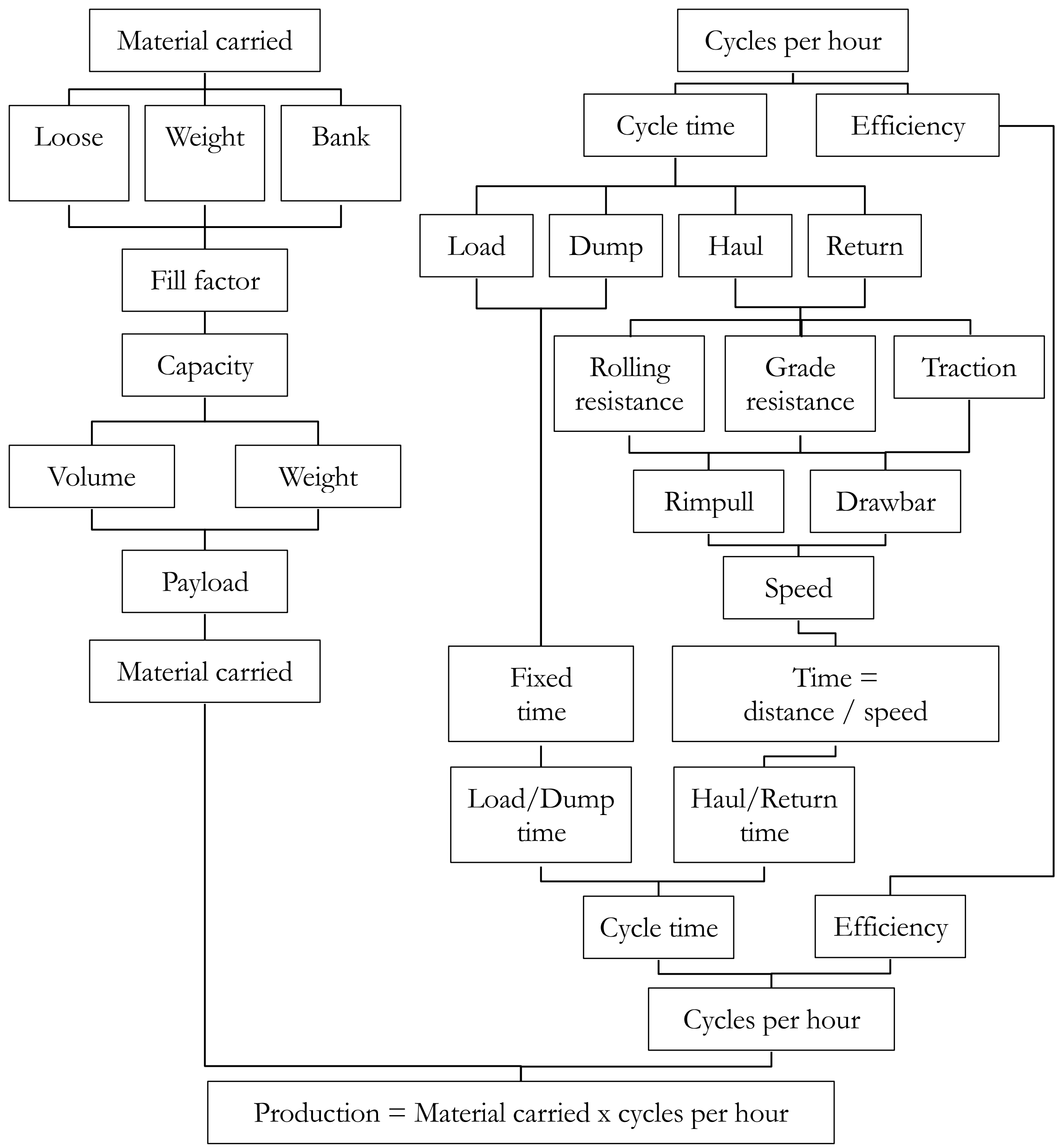


Figure 2-1 Earthmoving production process.
(Peurifoy, Schexnayder and Shapira 2006).

2.2.1 Excavators

The current practice to estimate the production rate of an excavator considered as an independent machine can be simplified in six steps (Peurifoy, Schexnayder and Shapira 2006).

- **Step1.** Obtain the heaped bucket load volume from manufacturers' data sheet. This would be a loose volume (lcm) value.
- **Step2.** Apply a bucket fill factor based on the type of machine and the class of material being excavated.
- **Step3.** Estimate a peak cycle time. This is a function of machine type and job conditions to include angle of swing, depth or height of cut, and in the case of loaders, travel distance.
- **Step4.** Apply an efficiency factor.
- **Step5.** Conform the production units to the desired volume or weight units (lcm to bcm or t).
- **Step6.** Calculate the production rate.

Production in this case is material carried per load * cycles per hour. In the case of excavators, this formula can be refined and written as per Equation 2-1.

$$Production = \frac{3{,}600\, sec\ \times Q \times F \times (AS{:}D)}{t} \times \frac{E}{60\ \text{min}\, hr} \times \frac{1}{volume\ correction} \quad \textbf{2-1}$$

where,

Q = bucket capacity (lcm)

F = bucket fill factor

AS:D = angle of swing and depth (height) of cut correction

t = cycle time in seconds

E = efficiency (min per hour)

2.2.2 Hauling trucks

Hauling trucks provide relatively low hauling costs because of their high travel speeds. The productive capacity of a truck depends on the size of its load and the number of trips it can make in an hour. The number of trips completed per hour is a function of cycle time. Truck cycle time has four components: (1) load time, (2) haul time, (3) dump time, and (4) return time. Examining a match between truck cargo body size and excavator bucket size yields the size of the load and the load time. The haul and return cycle times will depend on the weight of the truck, the horsepower of the engine, the haul and return distances, and the condition of the roads traversed. Dump time is a function of the type of equipment and conditions in the dump area (Peurifoy, Schexnayder and Shapira 2006).

2.2.2.1 Calculating truck production

Step 1. Number of bucket loads. The first step in analyzing truck production is to determine the number of excavator bucket loads it takes to load the truck (Equation 2-2).

$$Balanced\ number\ of\ bucket\ loads = \frac{Truck\ capacity\ (lcm)}{Bucket\ capacity\ (lcm)} \tag{2-2}$$

Step 2. Load time. The actual number of bucket loads placed on the truck must be an integer number. If the number of bucket loads is rounded down to an integer lower than the balanced number of loads (subscript LI), the loading time will be reduced; but the load on the truck is also reduced (Equation 2-3). The truckload in such cases will equal the bucket volume multiplied by the number of bucket loads (Equation 2-4).

$$Load\ time_{LI} = Number\ of\ bucket\ loads_{LI} \times Bucket\ cycle\ time \tag{2-3}$$

$$Truckload_{LI}(volumetric) = Number\ of\ bucket\ loads_{LI} \times Bucket\ volume \tag{2-4}$$

If the division of the truck cargo body volume by the bucket volume is rounded to the next higher integer (subscript HI) and that higher number of loads is placed on the truck, excess material will spill off the truck. In such case, the loading duration equals the bucket cycle time multiplied by the number of bucket swings (Equation 2-5). But the volume of the load on the truck equals the truck capacity, not the number of bucket swings multiplied by the bucket volume (Equation 2-6).

$$Load\ time_{HI} = Number\ of\ bucket\ loads_{HI} \times Bucket\ cycle\ time \tag{2-5}$$

$$Truckload_{HI}(volumetric) = Truckvolumetric\ capacity \tag{2-6}$$

Always check the load weight against the gravimetric capacity of the truck (Equations 2-7 and 2-8).

$$\begin{aligned} &Truckload\ (gravimetric) \\ &\qquad = Vol.load\ (lcm) \times Unit\ weight\ (loose\ vol.t/lcm) \end{aligned} \tag{2-7}$$

$$Truckload\ (gravimetric) < Rated\ gravimetric\ payload? \tag{2-8}$$

Step 3. Haul time. Based on the gross weight of the truck with the load, and considering the rolling and grade resistance from the loading area to the dump point, haul travel speeds can be estimated using the truck manufacturer's performance chart, see Figure 2-2. While performance charts indicates the maximum speed at which a vehicle can travel, the vehicle will not necessarily travel at this speed. A performance chart makes no allowance for acceleration or deceleration. In addition, other travel route conditions and safety can control travel speed (Peurifoy, Schexnayder and Shapira 2006). The anticipated effective speed is what should be used in calculating travel time (Equation 2-9).

$$Haul\ time\ (\text{min}) = \frac{Haul\ distance\ (km)}{Haul\ speed\ (km/h)} \times 60\ (min/hr) \tag{2-9}$$

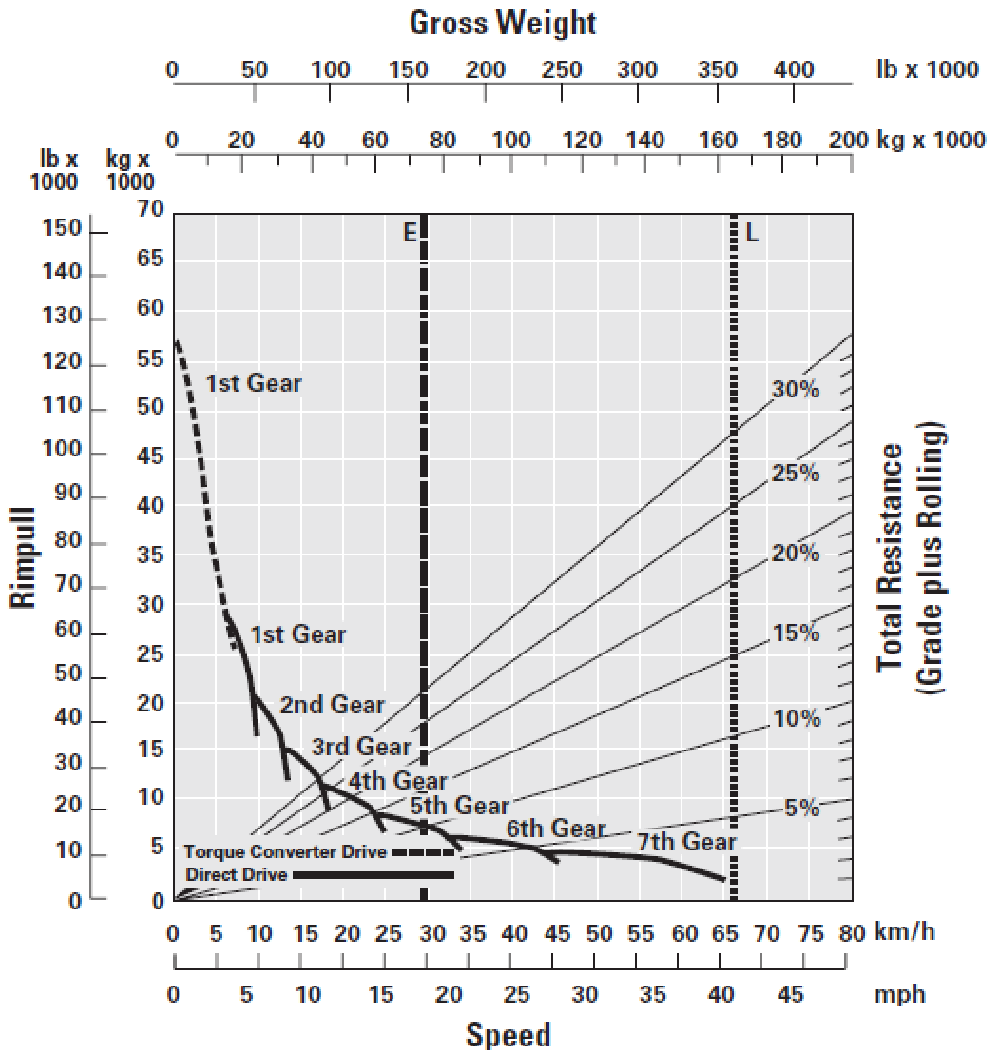


Figure 2-2 Performance Chart for a Caterpillar 777F Off-Highway Truck
(Taken from: AEHQ5749-01 (5-07) - 777F Off-Highway Truck specifications – © 2007 Caterpillar)

Step 4. Return time. Based on the empty vehicle weight, and the rolling and grade resistance from the dump point to the loading area, return travel speeds can be estimated using the truck manufacturer's performance chart. Return time can be computed using Equation 2-10.

$$Return\ time\ (\text{min}) = \frac{Haul\ distance\ (km)}{Haul\ speed\ (km/h)} \times 60\ (min/hr) \qquad \textbf{2-10}$$

Step 5. Dump time. Dump time will depend on the type of hauling unit and congestion in the dump area.

Step 6. Truck cycle time. The cycle time of a truck is the sum of the load time, the haul time, the dump time, and the return time (Equation 2-11).

$$Truck\ cycle\ time\ = Load_{time} + Haul_{time} + Dump_{time} + Return_{time} \quad \text{2-11}$$

Step 7. Number of trucks required. The number of trucks required to keep the loading equipment working at capacity is given by Equation 2-12

$$Balanced\ number\ of\ trucks = \frac{Truck\ cycle\ time\ (\text{min})}{Excavator\ cycle\ time\ (\text{min})} \quad \text{2-12}$$

Step 8. Production. The number of trucks must be an integer number. If an integer number of trucks lower than the balanced number of trucks is chosen, then the trucks will control production (Equation 2-13).

$$\begin{aligned} Production\ (lcm/h) \\ &= Truck\ load\ (lcm) \times Number\ of\ trucks \\ &\times \frac{60\ \text{min}}{Truck\ cycle\ time\ (\text{min})} \end{aligned} \quad \text{2-13}$$

When an integer number of trucks greater than the balanced number of trucks is selected, production is controlled by loading equipment (Equation 2-14).

$$Production\ (lcm/h) = Truck\ load\ (LCM) \times \frac{60\ \text{min}}{Excavator\ cycle\ time\ (\text{min})} \quad \text{2-14}$$

As a rule, it is better to never keep the loading equipment waiting (Peurifoy, Schexnayder and Shapira 2006).

Step 9. Efficiency. The production calculated before is based on a 60-min working hour. That production should be adjusted by an efficiency factor (Equation 2-15).

$$Adjusted\ production = Production \times \frac{Working\ time\ (\text{min/hour})}{60\ \text{min}} \quad \text{2-15}$$

Step 10. Production in desired units. Finally, this production can be converted as necessary into bank cubic meters or tonnes by using material property information specific to the job.

2.3 Data mining

Today's world is overwhelmed with data. The volume of data in our lives continues to increase and there's no end in sight. As the volume of data increases, inexorably, the proportion of it that people understand decreases, alarmingly. Lying hidden in all this data is information, potentially useful information, which is rarely made explicit or taken advantage of.

Data mining is defined as the process of discovering patterns in data. The process must be automatic or (more usually) semiautomatic. The patterns discovered must be meaningful in that they lead to some advantage, usually an economic advantage. The data is invariably present in substantial quantities (Chakrabarti, et al. 2009).

2.3.1 Data preprocessing

Real-world databases are full of noisy, missing, and inconsistent data because of their typically huge size and their likely origin from multiple, heterogeneous sources. There are a number of data preprocessing techniques. Data cleaning can be applied to remove noise and correct inconsistences in the data. Data integration merges data from multiple sources into a coherent data store, such a data warehouse. Data transformations, such as normalization, may be applied. Data reduction can reduce the data size by aggregating, eliminating redundant features, or clustering, for instance. Figure 2-3 summarize some of the data preprocessing forms.

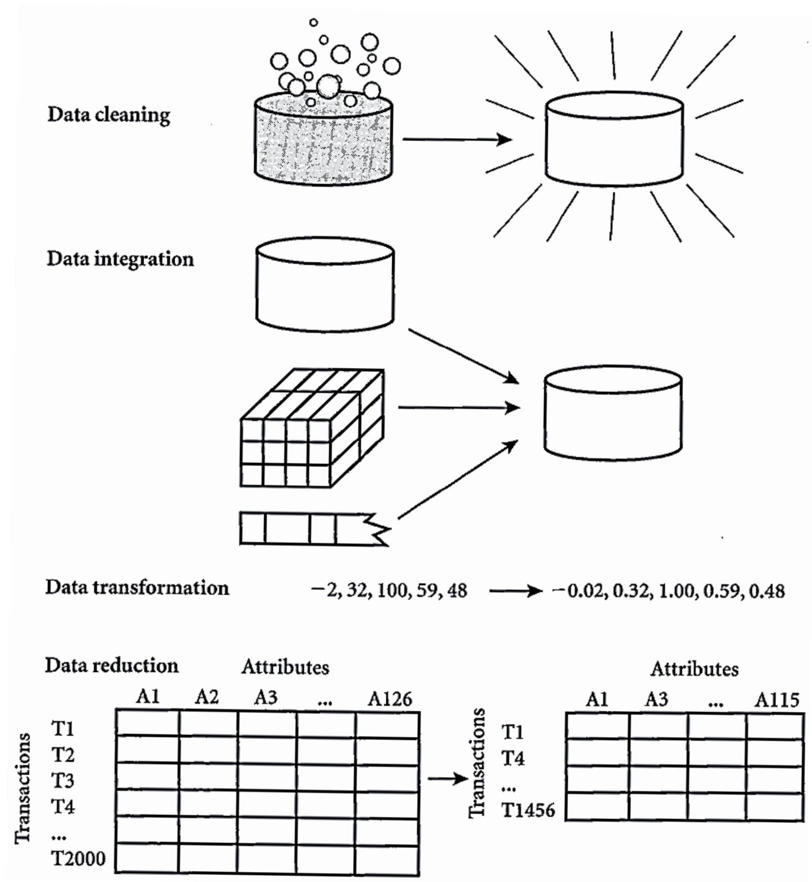


Figure 2-3 Forms of data preprocessing (Chakrabarti, et al. 2009).

2.4 Artificial Neuronal Networks

Artificial neural networks (ANNs) represent a different computational approach. In contrast to more conventional analytic methods, ANNs are an information processing technology that attempt to simulate, in a gross manner, the networks of nerve cell (neurons) of the biological (human or animal) central nervous system (Graupe 2007). As the brain, ANNs learn from experience, generalize from previous examples to new ones and abstract essential characteristics from inputs containing irrelevant data. A large variety of possible ANN applications now exist for non-computer specialist. Therefore, with only a very modest knowledge of the theory behind ANNs, it is possible to tackle complicated problems in a researcher's own area of specialty with the ANN technique. There is currently available a wide range of ANN models, in terms of topology and mode of operation. According to (Boussabaine 1996), each model can be specified by the following seven major concepts.

1. A set of processing neurons.
2. A state of activation for each neuron.
3. A pattern of connectivity among the neurons or topology of the network.
4. A propagation method to propagate the activities of the neurons through the network.
5. An activation rule to update the activities of each node.
6. An external environment that provides information to the network and interacts with it.
7. A learning method to modify the pattern of connectivity by using information provided by the external environment.

Figure 2-4 presents a three-layered ANN. The input and Output layers are responsible for communication with the outside world.

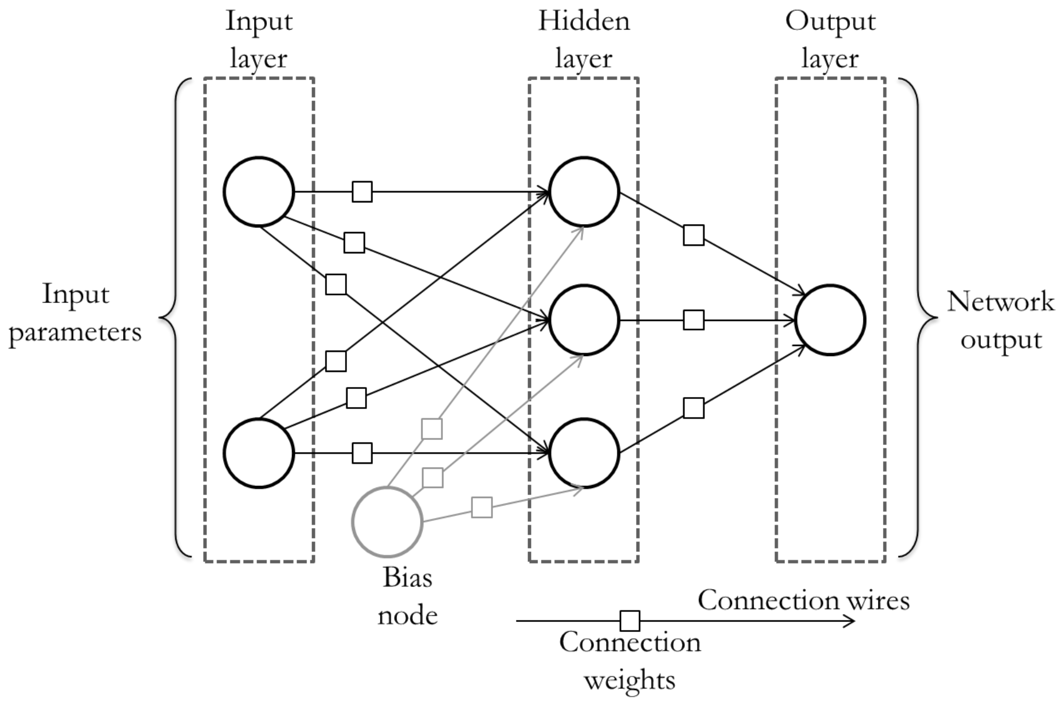


Figure 2-4 Example of an ANN with three layers

Figure 2-5 illustrates the functioning of the nodes. In a very simple scenario, a node receives inputs from two previous nodes X(1) and X(2), respectively modified by weight factors W(1) and W(2). The node computes the addition of these values X(1)*W(1) + X(2)*W(2), and delivers an output using a transfer function which can take a variety of forms.

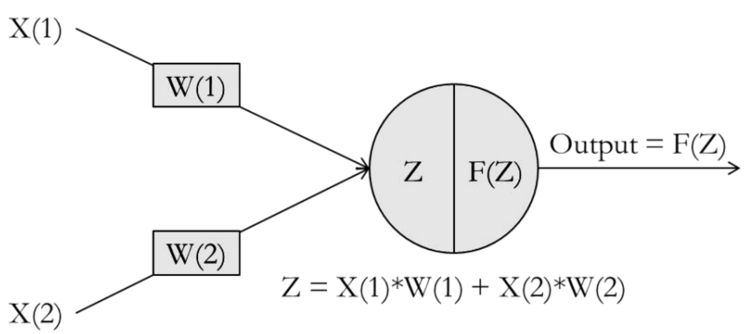


Figure 2-5 Typical ANN node

ANNs can be supervised or unsupervised. Supervised networks predict output based on the patterns observed in the input and output data that has been used for “learning” or “training”. Unsupervised networks are used to classify sets of data into a specific number of categories, without learning from other data sets (Mather 1998). The work within this research only deals with supervised networks. In this type of ANN, the input layer broadcast a pattern to all the hidden nodes, the system then computes an output using the procedure presented in Figure 2-5. The final output is compared with the target value that the trainer has previously specified. The difference yields the output error, and it is time to decide if further learning is necessary. If more learning is required, the output nodes calculate the derivatives of the error with respect to the weights and the result is sent back through the systems and the connection weights are corrected. Once the weights has been upgraded the feed-forward computation start over again.

Chapter four on this document presents the implementation of artificial neural networks on the analysis of the historical earthmoving operational information.

CHAPTER 3 EARTHMOVING PRODUCTIVITY ANALYSIS BASED IN HISTORICAL DATA

3.1 Introduction

This chapter summarizes how the historical operational data of the one biggest construction contractors in the oil sands industry in northern Alberta was analyzed through about seven months of research assistantship. The analyzed historical data had been recorded for more than three years in eleven different projects.

The specific objectives of the analysis were:

- Determine truck productivities as a function of the haul distance, categorized by truck model and project.
- Determine excavators and major loading equipment productivities categorized by excavator model and project.
- Obtain historical ratios between number of trucks and number of graders, as well as the historical relation between total project production in bcm and number of dozers required.
- Analyze the overall quality and consistency of the data contained in the databases.
- Recommend further analysis of operational data, as well as new and better procedures to perform the analysis.
- Suggest novel beneficial approaches for data collection.

3.2 Hauling truck productivity analysis

This section describes how the analysis of the historical information related to hauling operations of more than 230 hauling units was performed. The main goal of this analysis was to produce productivity curves as a function of the hauling distance.

3.2.1 Equipment models

This research cover ten trucks models grouped in four categories see Table 3-1. Figure 3-2 shows a picture of a 777 Hauling Truck.

Table 3-1 Hauling truck models and categories

Equipment Function	Equipment Category	Equipment Model
Hauling trucks	Trucks - 280t+	EH4500
		EH5000
		930E-AC
	Trucks - 220-280t	793C
		793D
		830E
	Trucks - 120-220t	785C
		785D
	Trucks - 80-120t	777D
		777F

Figure 3-1 Hauling truck models
(Picture from: AEHQ5749-01 (5-07) - 777F Off-Highway Truck specifications – © 2007 Caterpillar)

3.2.2 Methodology

Different actions were carried out to clean, combine, condense, and organize the data in order to obtain the productivity information for the main truck models analyzed.

The operational data in its basic state was contained in 47 tables. The first step was to identify what relevant data was needed for the analysis. With this in mind, certain information was grouped into three main tables (master tables). These tables represented data of three different natures: 1) hauling production data, 2) equipment timing data from availability and utilization records and 3) equipment timing data coming from accounting sources. This particular assembly assisted in the comparison of analogous information that was being recorded on different places and/or for different purposes; it also helped in the identification of data collection issues.

3.2.2.1 Pivot Tables

Pivot tables are the key engine in this analysis. They are used as the main tool to organize, classify and group the information in a meaningful way. Four Pivot tables were built for the hauling productivity analysis of each truck model. The three master tables described before were the data source of these pivot tables.

3.2.2.2 Shift weighted hauling distance

Most of the information related to earthmoving operations is recorded by the foreman at the end of the shift. The information is then input into a data management system. This schema contributes to the fabrication of information; principally, the durations of individual tasks within a shift. For instance, when one hauling unit performs different tasks during one shift (e.g. different hauling distance, material type, loading unit, etc.) the hauling time recorded into the databases for each different task is often a not very accurate estimation of the duration of the truck operations, see Figure 3-2. If those individual times are used to produce the ratio between the number of hauled loads and the time spent, the results will not be a real representation of the actual truck productivity.

Even if the recorded individual times are not reliable, the information could still be useful being that the summation of all these durations closely represents the total operating hours for the shift.

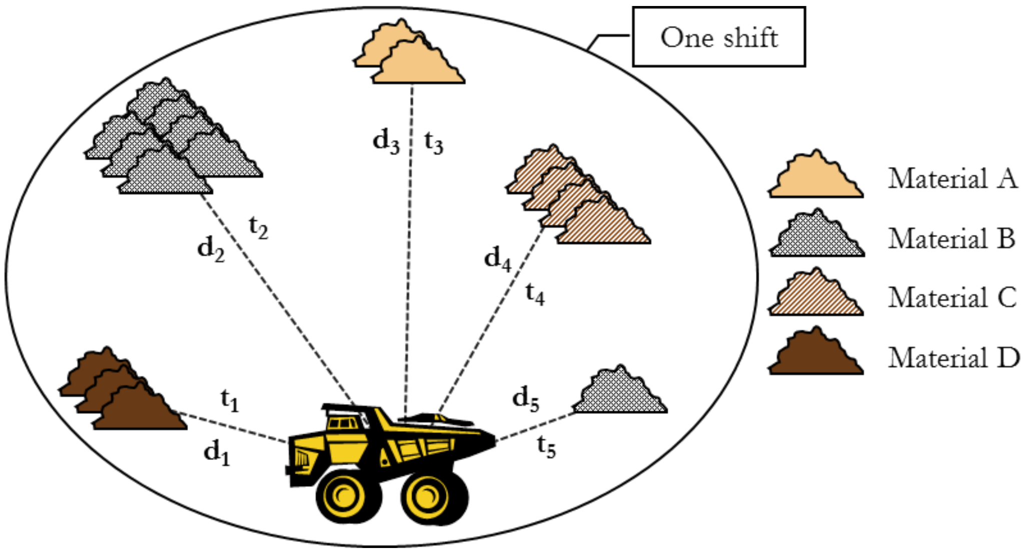


Figure 3-2 Different tasks performed by a truck unit during one shift.

To mitigate this issue, all the activities performed by a truck unit during a given shift were compressed on one productivity datum, as the relation between the total number of loads and the operating hours during the shift (Figure 3-3). The correspondent hauling distance for this productivity is called Shift weighted hauling distance (d_w) and it is computed as per Equation 3-1.

$$d_W = \frac{\sum_{i=1}^{n} L_i * d_i}{\sum_{i=1}^{n} L_i} \qquad \textbf{3-1}$$

where,

Li = Number of loads for task i performed by a truck unit during a specific shift.

di = haul distance for task i

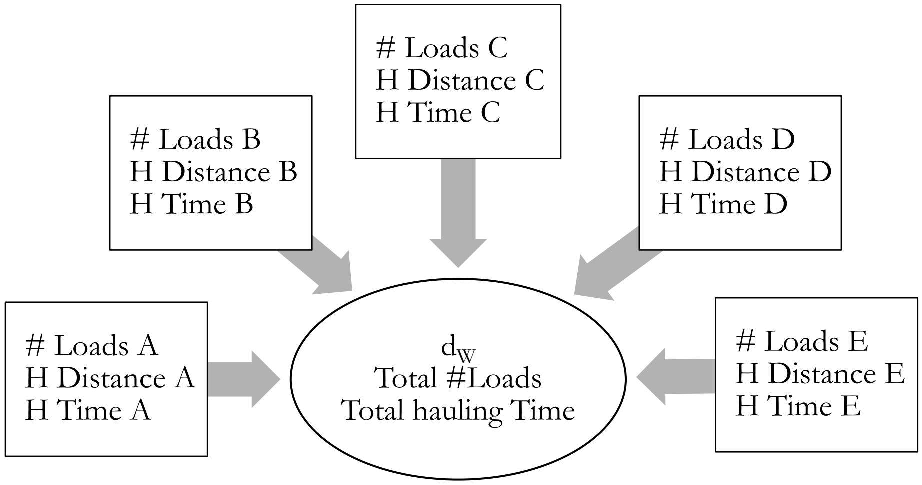


Figure 3-3 Shift weighted hauling distance (d_W).

Figure 3-4 shows how the use of the shift weighted hauling distance substantially reduces the dispersion on the results.

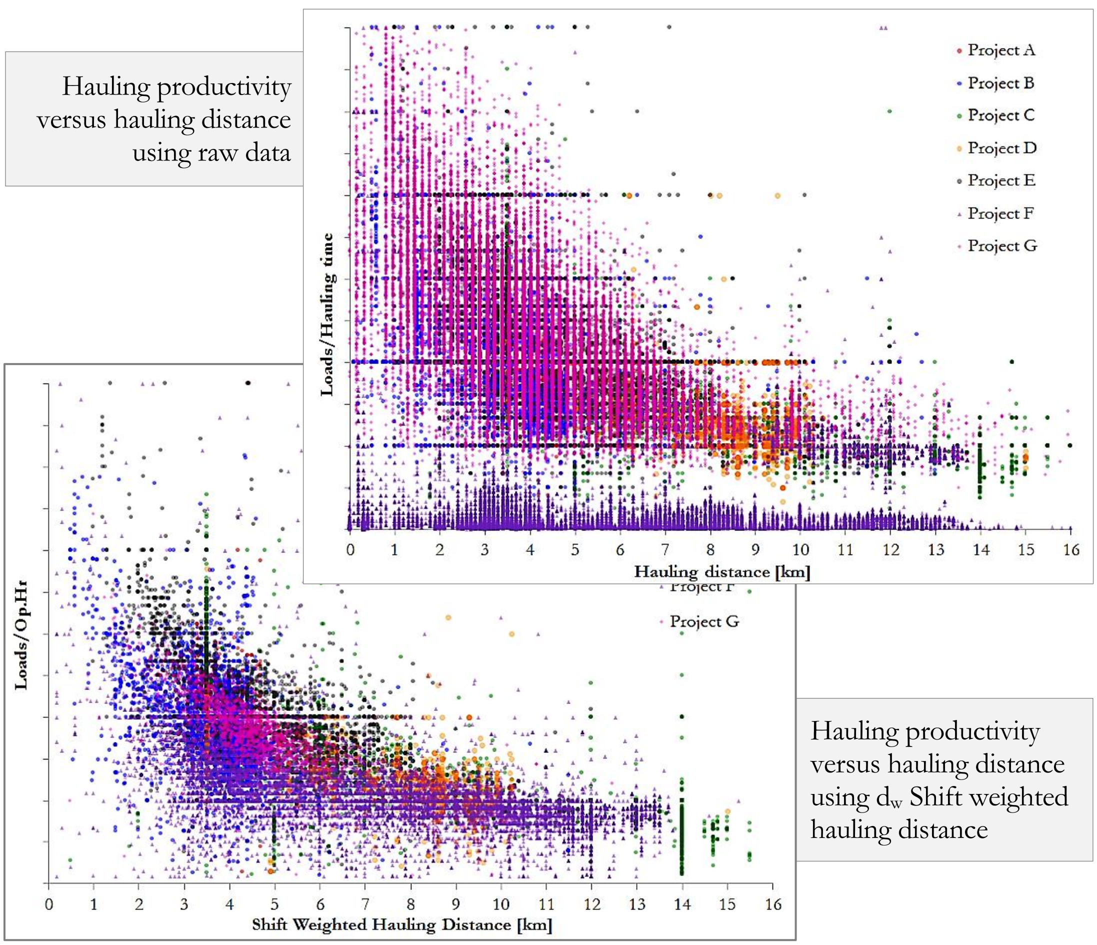


Figure 3-4 Condensing productivity information using dW.

3.2.2.3 Data clean up

As with most real-life datasets, the recorded operational data contains abundant erroneous values and typing mistakes. Before any conclusion can be reached, the data has to be reviewed, cleaned up and filtered, to improve its quality. The sources of the inconsistencies are often very difficult to identify. However, the pivot tables created for this analysis facilitate the identification of unusual and/or unrealistic values. The following are examples of the logic behind the set of checks used to clean up the data.

a) Look for unusual and or unrealistic hauling distances. *e.g. 0 km, 100 km, 10000 km, etc.*

b) Unrealistic values for the ratio Loads/hr. *e.g. Loads/hr≥ 10, 10 loads/hr is equivalent to one load every 6 minutes; allowing 2 minutes for loading and 2 minutes for dumping and maneuver in load area and dump point, only two minutes remain for hauling and returning, something extremely difficult to achieve even for short hauling distances.*

c) Unrealistic values for the ratio Hr/Loads (e.g. Hr/Loads ≥ 5). *Cycle times are rarely higher than 2 hours.*

d) Unusual values of the ratio Loads/hr for the correspondent hauling distance. *e.g. Loads/hr≥ 5, when hauling distance ≥ 10. 5 loads/hr is equivalent to one load every 12 minutes; allowing 2 minutes for loading and 2 minutes for dumping and maneuver in load area and dump point, 8 minutes remain for hauling and returning, which could be achieved on 10 km with an average speed of 150km/hr.*

e) Unusual values of the ratio Hr/Loads for the correspondent hauling distance. *e.g. Op.Hr/load≥2, when hauling distance ≤ 5.0 km.*

f) Artificial rules of correspondent hauling times for low values of number of loads *(e.g if load count = 1 then hauling time = 0.25hr or 0.5hr)*

Even though it was possible to correct some of this erroneous information by chasing the right value in the neighborhood of the entry, for most of the cases the correct value cannot be inferred and the haul records of the unit during the corrupted shift had to be discarded.

Since most of the erroneous entries or corrupted shifts were identified, it was possible to generate a summary of the quality improvement operations. The clean-up chart summarizes where and when problems in the data were found. It serves as an assessment of the performance of different projects regarding data collection. As an example, the Figure 3-5 presents the quality improvement summary corresponding to the 777 truck model.

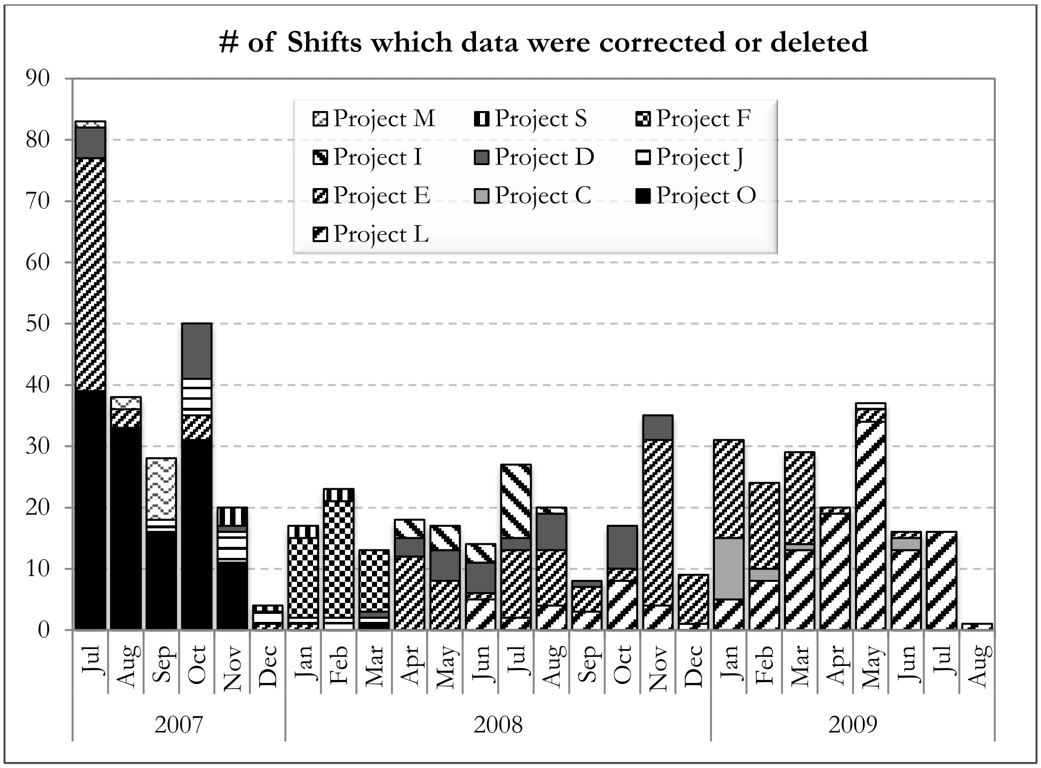


Figure 3-5 Clean up summary Truck Model 777.

3.2.2.4 Hauling Productivity Equation

Once the clean-up actions have been performed, productivity of the truck model was plotted in an XY scatter chart in terms of loads per hour versus hauling distance. At this point, it is possible to make a regression and extract an equation that represents the productivity of the truck model as a function of the hauling distance. Most of the existing commercial software allows several regression forms: exponential, linear, logarithmic, polynomial, power. Among these types, the logarithmic regression was found to be usually the best fit to represent the haul truck productivity, see Figure 3-6.

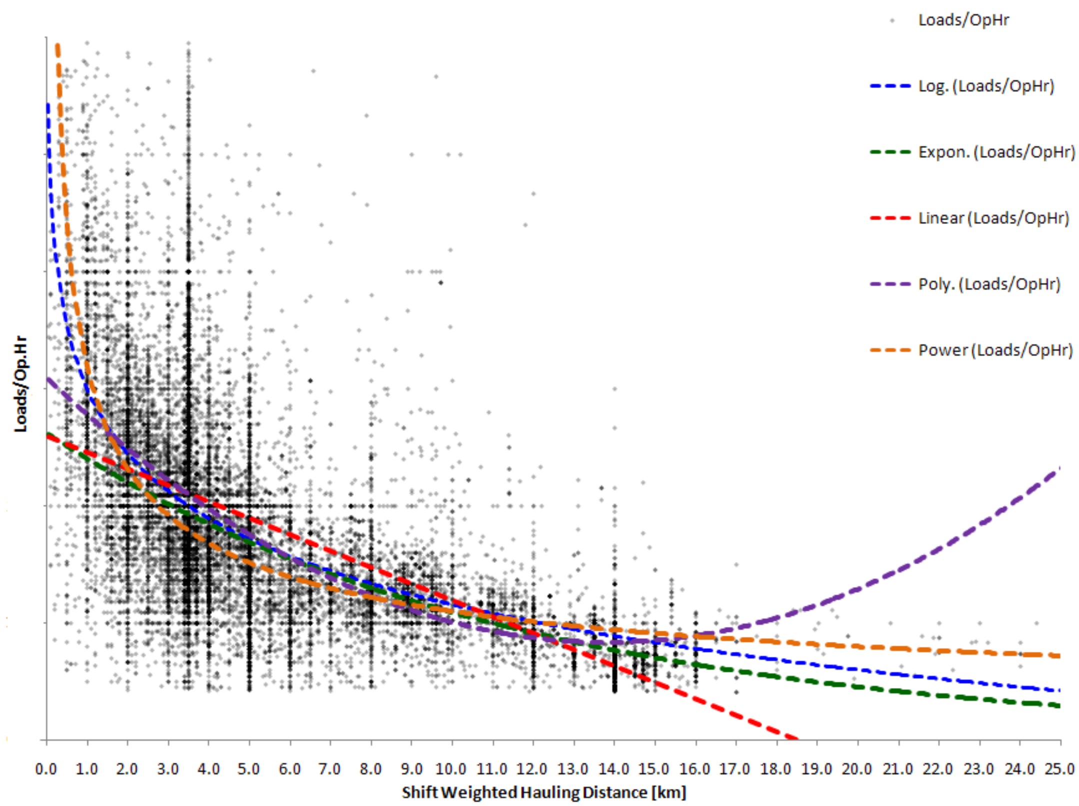


Figure 3-6 Different regression forms for hauling truck productivity.

On the other hand, the representation of an earthmoving operation through a logarithmic equation seemed to be quite suspicious (the logarithmic equation form does not intercept the y axis and cannot reflect fixed time component of the cycle time). Seeking for a more satisfactory equation form, a reconsideration of the analysis was made.

Total cycle time for a hauling unit is generally a combination of:

1. Fixed time
2. Hauling time (Loaded)
3. Return time (Empty)

In general, hauling and returning durations depend of the distance, while the durations of loading and dumping activities are independent of it, and could be fairly assumed to be constant value, see Figure 3-7.

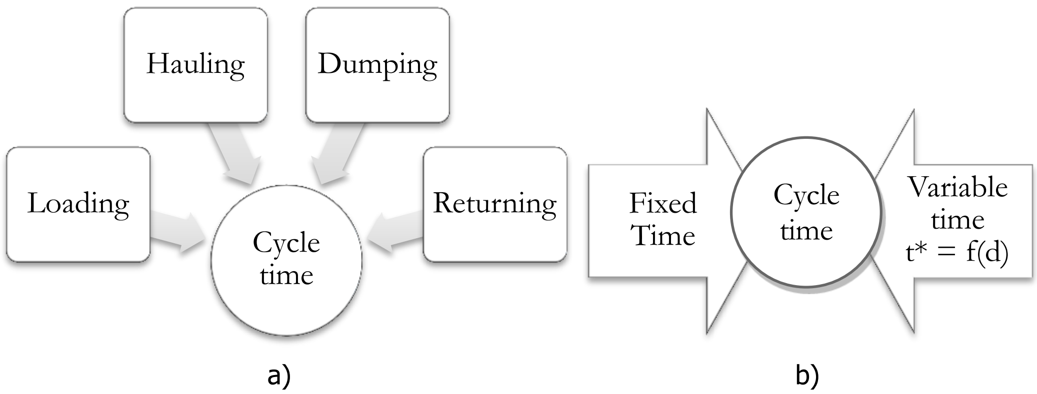


Figure 3-7 a) Cycle time main components - b) Type components of the cycle time.

Therefore, the following formulation for Productivity (P) is developed, see Equations 3-2 to 3-8.

$$P = \frac{1\,(Load)}{Cycle\ time} \tag{3-2}$$

$$P = \frac{1\,(Load)}{Fixed\ \ time + Variable\ time} \tag{3-3}$$

$$P = \frac{1}{C_1 + t(d)} \tag{3-4}$$

where, C1 is a constant that represent the fixed time component of the cycle time.

$$P = \frac{1}{C_1 + d\,/\,S(d)} \tag{3-5}$$

where, S(d) is the average truck speed and d is the haul distance.

Several forms for the average speed were proposed. However, the collected information did not support a very detailed analysis regarding the average speed; therefore the average speed was conveniently assumed to be a constant.

Then,

$$P = \frac{1}{C_1 + d / C_2} \qquad \textbf{3-6}$$

Where, C2 is a constant that represent the average speed.

The last equation form is not normally available in conventional software packages. Nevertheless, the following modification offers a convenient approach to compute the required constants.

$$\frac{1}{P} = C_1 + d / C_2 \qquad \textbf{3-7}$$

$$\frac{1}{P} = C_1 + C_3.d \qquad \textbf{3-8}$$

Thus, the inverse of the productivity in hours per load has the form of a linear equation.

Finally, the productivity data could be plotted as operating hours per load (Op.Hr/Load) versus shift weighted hauling distance (d_W) allowing a linear regression in which the needed constants could be computed, see Figures 3-8 and 3-9.

In general, this new equation offers better correlation than the previous logarithmic one, and in addition, the results at short distances are more accurate using this novel approach. The developed equation is also more intuitive given that its two constants keep a close relation to the fixed time component of the cycle time and the average speed of hauling and returning.

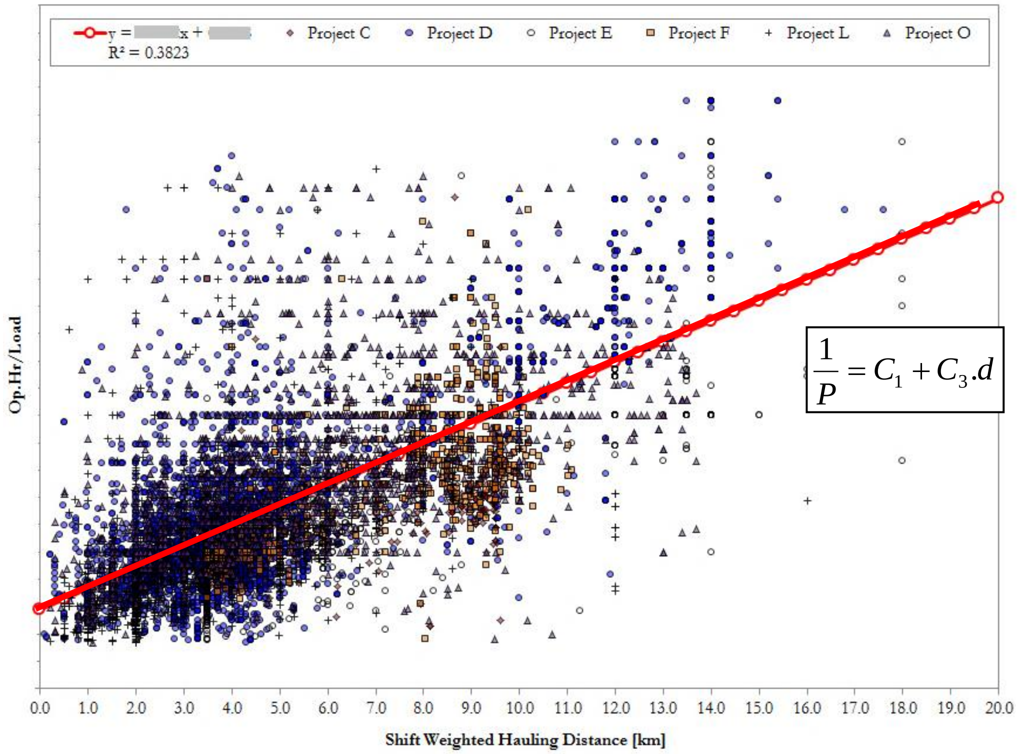


Figure 3-8 Operational hours per load versus shift weighted hauling distance.

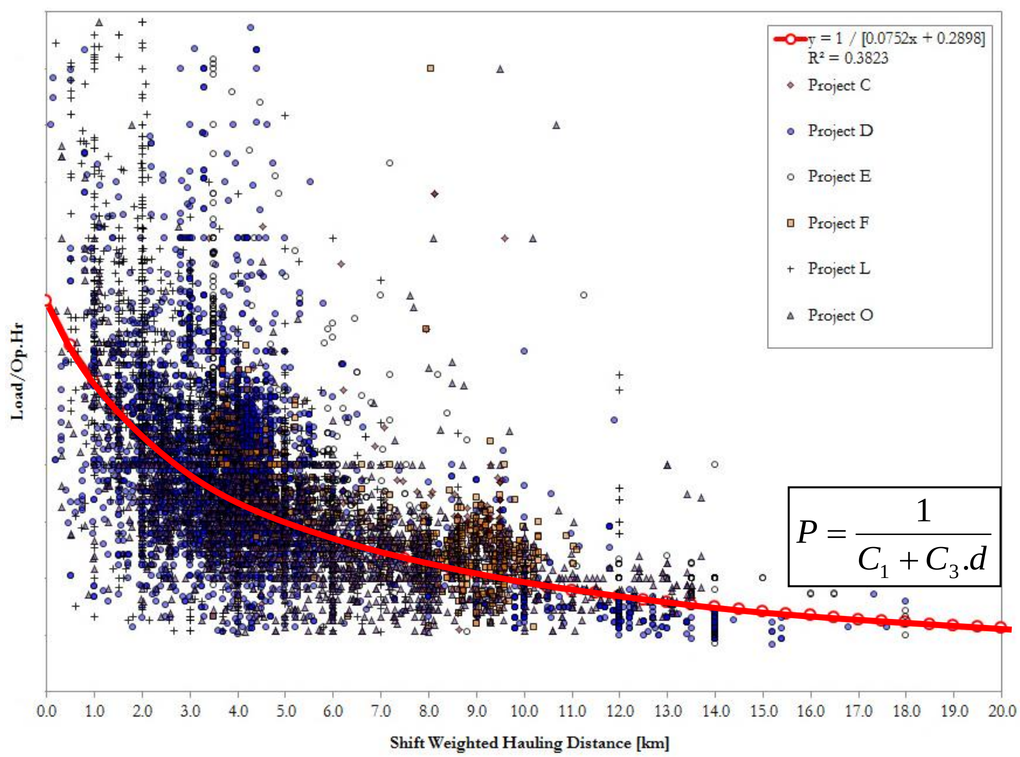


Figure 3-9 Productivity equation of hauling trucks

3.2.3 Results

The analysis procedure and main results were summarized in a single page layout for each individual truck model. These layouts contain:

- Loads/hauling time versus hauling distances using raw information.
- Loads/operating hour versus shift weighted hauling distance (d_W) using raw information.
- Quality improvement summary.
- Operating hours per load (input information for a linear regression).
- Final productivity information colored by project and general productivity equation.

As an example, the results of the productivity analysis for the truck model 785 are presented in Figure 3-10. In some cases detailed results by projects are also offered. A graph that shows this situation for the truck model 785 is presented in Figure 3-11. Please note that to maintain the confidentiality of this information Vertical axes values are omitted and project names have been changed.

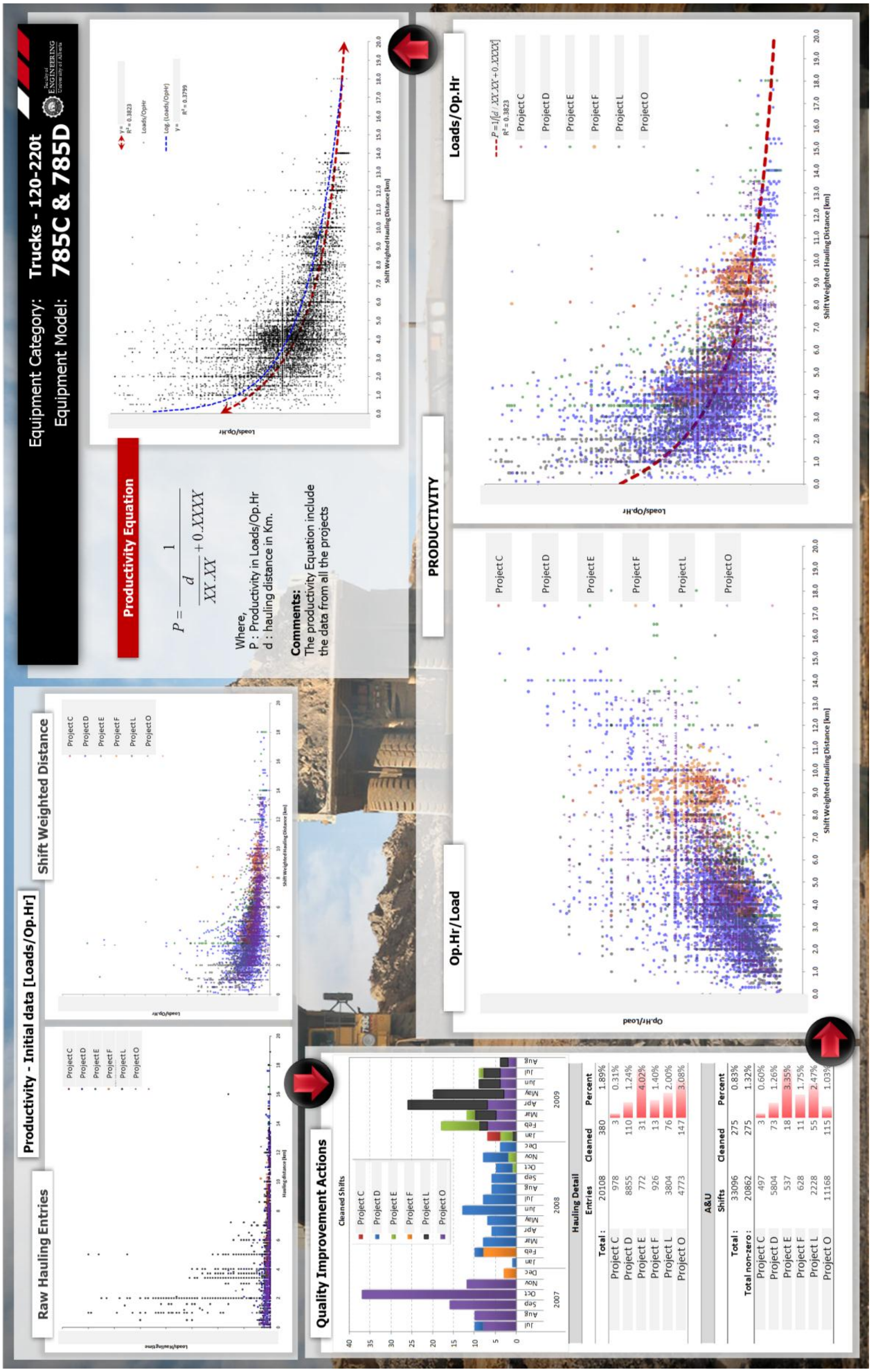


Figure 3-10 Hauling truck productivity summary – Models: 785C & 785D.

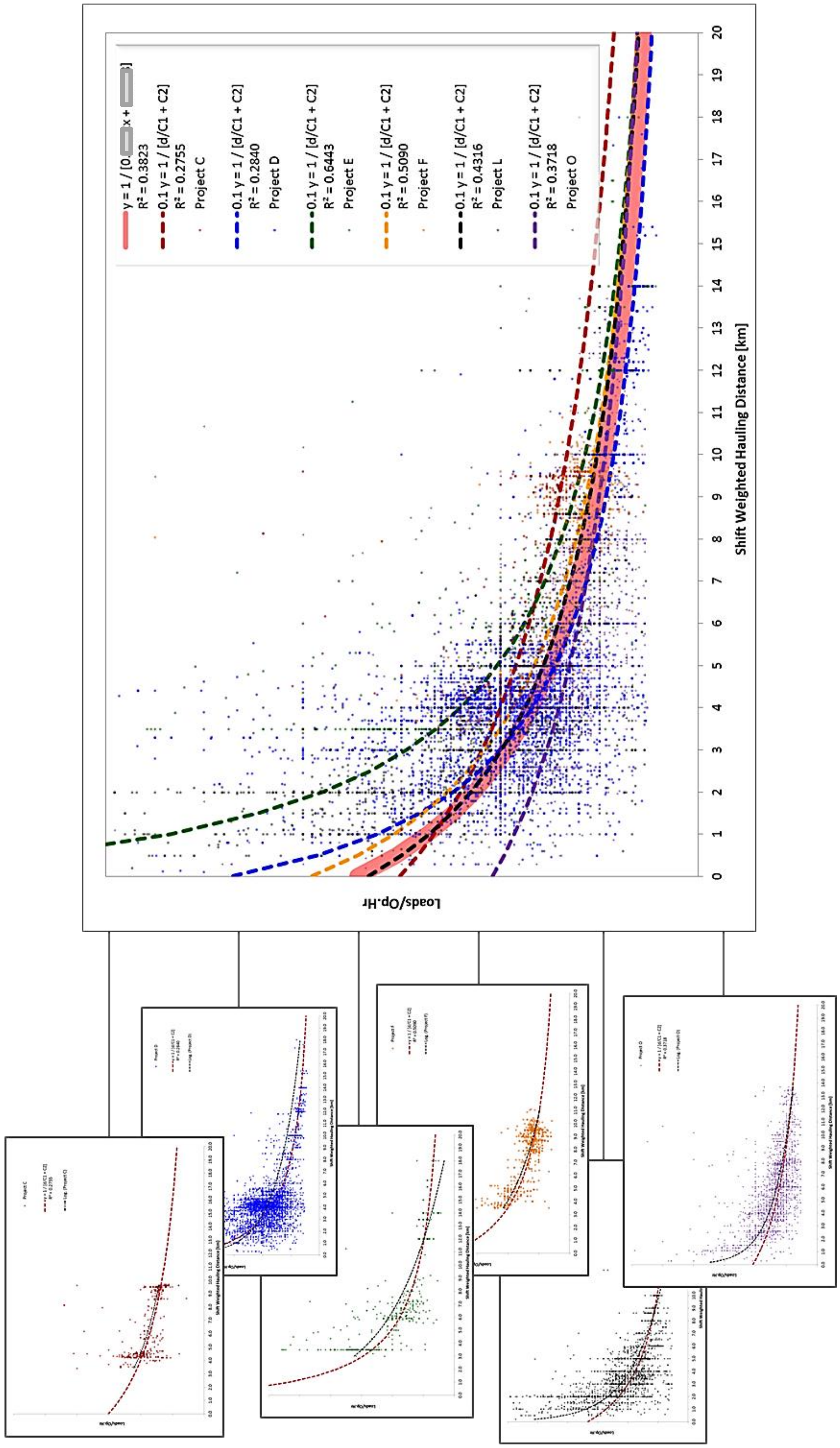


Figure 3-11 Hauling truck productivity - Models: 785C & 785D by project

3.2.4 Weighted linear regression

The results presented before in section 3.23 were obtained assuming that each data point provides equally precise information. This assumption, however, is debateable for this specific application. It is believed that the level of accuracy of the productivity ratio (loads/hour) increases when larger numbers of loads are used.

This section aims to illustrate how a modification of the previously used linear regression approach could lead into better values of the square of the correlation coefficient (r squared).

Section 3.2.2 described how a linear regression on scatter chart of the ratio hours per load versus the hauling distance could be used to find the intercept and the slope of the equation 3.9.

$$\frac{1}{P} = C_1 + C_3.d \qquad \textbf{3-9}$$

If the "1/P" values are called "y", the hauling distances values "d" are called "x" and the constants "C_1" and "C_3" are called "a" and "b" respectively, then equations 3.9, 3.10 and 3.11 could be used to obtain the intercept "b", slope "a" and correlation coefficient "r" of the linear regression.

$$b = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} \qquad \textbf{3-10}$$

$$a = \bar{y} - b\bar{x} \qquad \textbf{3-11}$$

$$\mathrm{r} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}} \qquad \textbf{3-12}$$

Where, $\bar{x}$ is the average of the hauling distances "d" and $\bar{y}$ is the average of the hours per load values "1/P".

As mentioned before, if we assume that the level of accuracy of the productivity ratio loads per hour increases when larger numbers of loads are used, and if we accept that the increment in this truthfulness is proportional to the number of loads used to compute the productivity ratio; then the equations 3-10, 3-11 and 3-12 used to compute the intercept "b", slope "a" and correlation coefficient "r" of the linear regression can be replaced for equations 3-13, 3-14 and 3-15.

$$b = \frac{\sum i(x-\bar{x})(y-\bar{y})}{\sum i(x-\bar{x})^2} \qquad \textbf{3-13}$$

$$a = \bar{y} - b\bar{x} \qquad \textbf{3-14}$$

$$r = \frac{\sum i(x-\bar{x})(y-\bar{y})}{\sqrt{\sum i(x-\bar{x})^2 \sum i(y-\bar{y})^2}} \qquad \textbf{3-15}$$

Where "i" is the number of loads associated with the hours per load ratio.

The use of equations 3-10 to 3-12 on the historical information of hauling trucks resulted on larger values for the square of the correlation coefficients (r squared), revealing the advantage of using a weighted linear regression with the ability to attach a level of quality (proportional to the number of loads hauled during the shift) to each productivity ratio data point.

3.2.5 Hauling truck productivity range

Due to the high dispersion on the hauling truck productivity charts, it could be desirable to represent these results using a range instead of a single productivity equation. With the productivity equation as foundation, it is possible to find a pair of "sister equations" that will define a productivity range. The range can be built so it limits encompasses a given percentage of data points.

The methodology to find these "sister equations" is simple. Using Goal Seek, which is part of a suite of commands sometimes called what-if analysis tools, is possible to find a factor that modify the slope and intercept of the of the original linear regression (or only the intercept) so that the number of data points between the sister equations is a given percentage of the total number of data points on the scatter chart. For example, Figure 3-12 presents the linear regression of a hours-per-load versus hauling distance chart and the corresponding sister equations. The goal was that 75% of the data points were inside the range; this was achieved by modifying the intercept with the y axes of the original linear regression by 56.22%. In Figure 3-12 the intercept of the original productivity equation was 0.2595, thus the intercept of the sister equation 1 is then 0.2595*(1+0.5622) = 0.4051, and the intercept of the sister equation 2 is 0.2595*(1-0.5622) = 0.1139.

Figure 3-13 presents the results and the productivity range in loads per hour versus hauling distance.

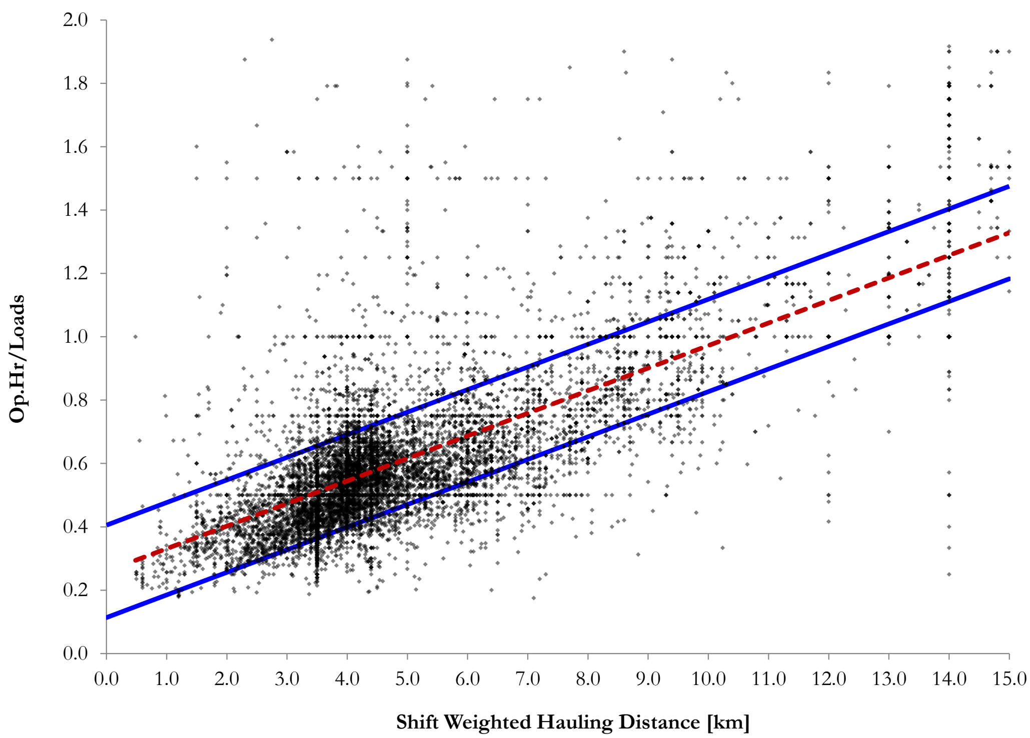


Figure 3-12 Hauling productivity range hours/load vs. hauling distance – a.

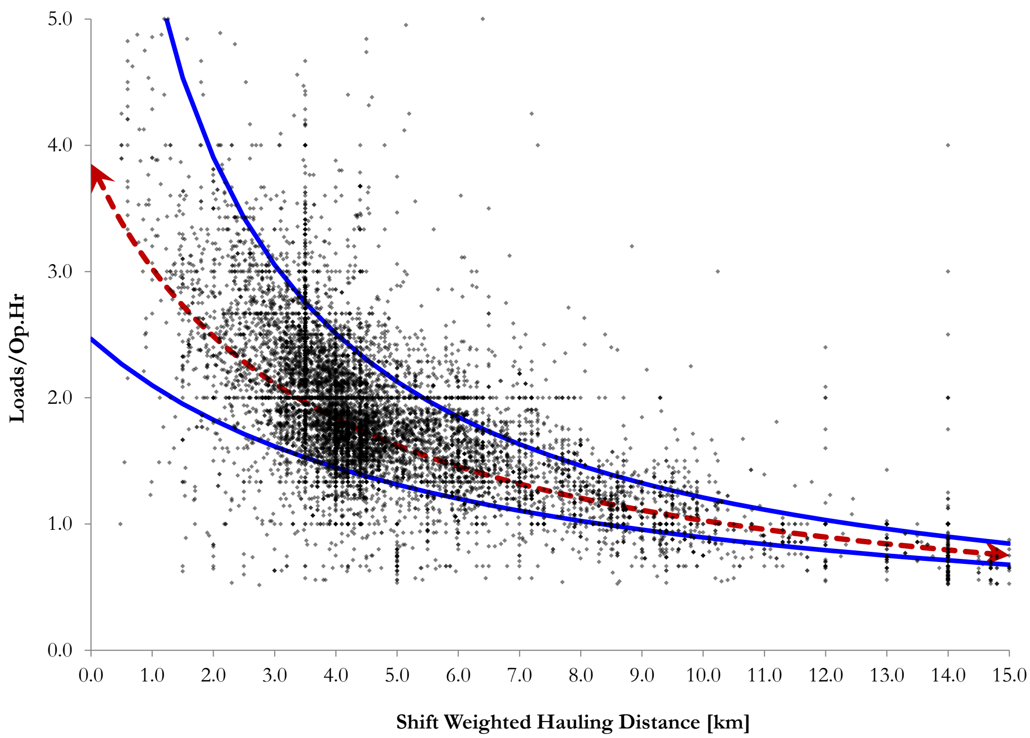


Figure 3-13 Hauling productivity range loads/hour vs. hauling distance – a.

It is also possible to obtain this productivity range and the sister equations by modifying both the intercept and the slope of the original linear regression.

Figure 3-14 the same linear regression of a hours-per-load versus hauling distance chart and the corresponding new sister equations. The goal was also that 75% of the data points were inside the range; this was achieved by modifying the intercept with the y axes and the slope of the original linear regression by 23.74%. In Figure 3-14 the intercept of the original productivity equation was 0.2595, thus the intercept of the sister equation 1 is then 0.2595*(1+0.2374) = 0.3211, and the intercept of the sister equation 2 is 0.2595*(1-0.2374) = 0.1979. Likewise, the slope of the original productivity equation was 0.0713, therefore the slope of the sister equation 1 is then 0.0713*(1+0.2374) = 0.0882, and the slope of the sister equation 2 is 0.2595*(1-0.0713) = 0.0544. Figure 3-15 presents the results and the productivity range in loads per hour versus hauling distance.

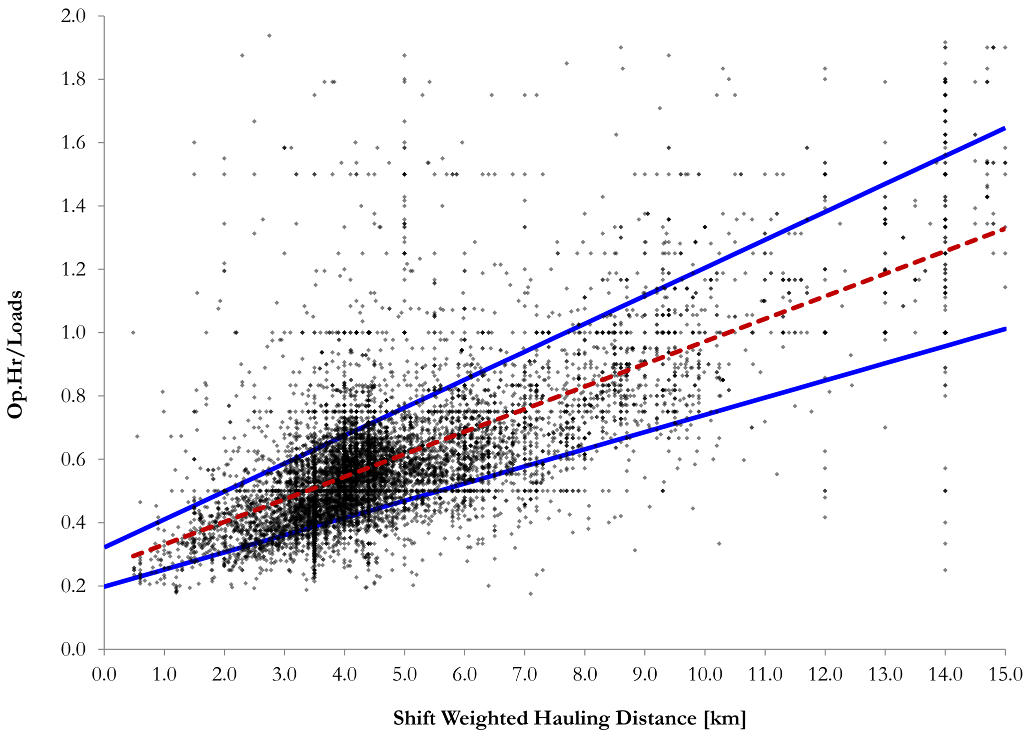


Figure 3-14 Hauling productivity range hours/load vs. hauling distance – b.

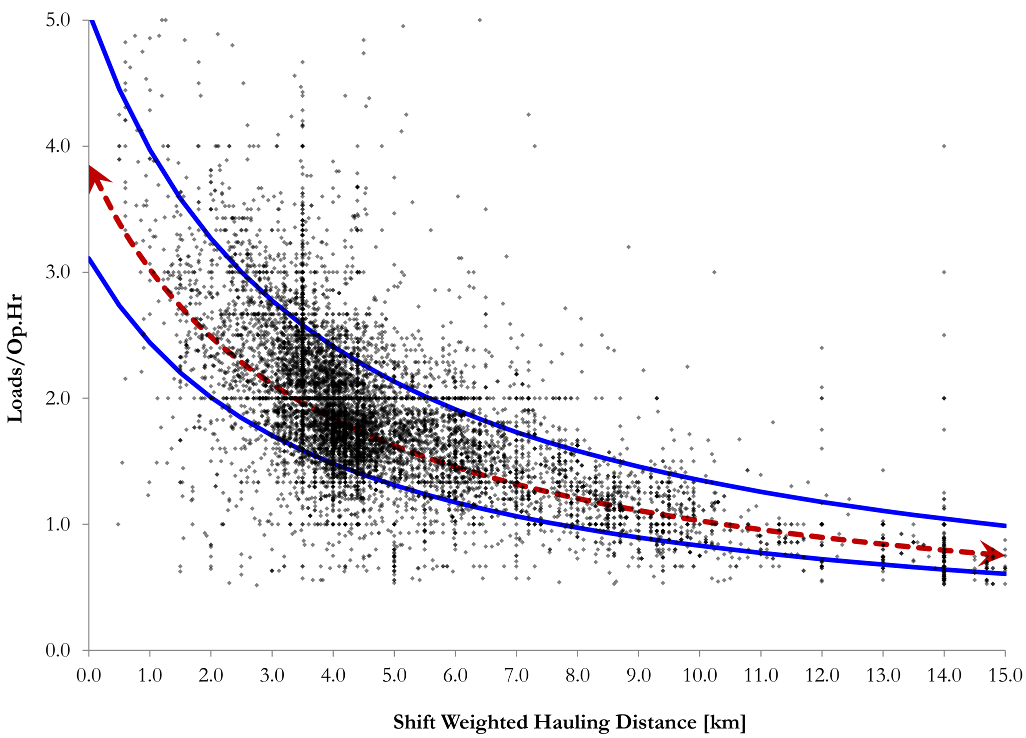


Figure 3-15 Hauling productivity range hours/load vs. hauling distance – b.

3.3 Excavator productivity analysis

3.3.1 Introduction

This section presents the analysis of the historical information related to the operation of excavator units. Cumulative frequency productivity curves are generated for the major loading equipment models and classified by project.

3.3.2 Equipment models

This research cover fifteen different excavator models grouped in four categories for a total of more than 150 individual units, see Table 3-2. Figure 3-16 shows pictures of a Hydraulic shovel EX5500 and an excavator EX1900.

Table 3-2 Excavator models and categories

Equipment Function	Equipment Category	Equipment Model
LOADING UNITS	Shovels - Electric	495HF
	Shovels - Hydraulic	EX8000
		EX5500
	Excavators - Medium	EX3600
		EX2500
		EX1900
		EX1800
		EX1200
	Excavators - Small	EX850
		EX800
		EX750
		EX600
		EX550
		EX450
		ZX450LC-3

Figure 3-16 Hydraulic shovel EX5500 and Excavator EX1900
Hitachi. Retrieved March 13, 2011, from:
http://www.hitachi-c-m.com/global/products/excavator/large/ex5500-6/index.html
http://www.hitachi-c-m.com/global/products/excavator/large/ex1900-6/index.html

3.3.3 Methodology

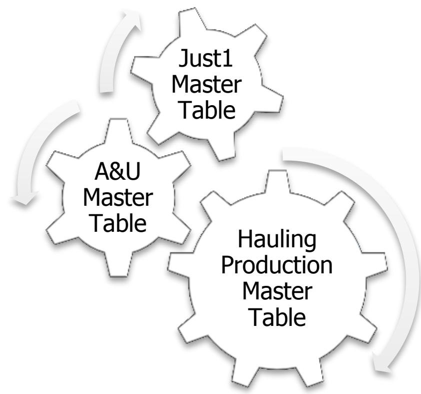
The methodology implemented for the analysis of excavator models is very similar to the one described for the analysis of hauling equipment.

3.3.3.1 Master tables

The analysis uses the same three "master" tables that were created in the hauling analysis. These tables group the data in: 1) hauling production information, 2) equipment timing data from availability and utilization recordings and 3) equipment timing data coming from accounting sources. As it was described before, these three master tables connect relevant information coming from the original 47 tables in the DBMS.

3.3.3.2 Excavator Model Information

In this step the information from the three master tables is filtered by excavator model and combined into a single source. The purposes of this breakdown are classification and reduction of the amount of data to be processed to a quantity suitable to manipulate with the available computer resources.



3.3.3.3 Pivot Tables

Pivot tables again are the key engine in the analysis. They are used as the main tool to organize, classify and group the information in a meaningful way. The source data of the pivot table used for the loading productivity analysis are the three master tables. These pivot tables classify the information by excavator unit, date, shift and project. They are also oriented to obtain the relation between the total number of bcm performed by the excavator unit during a specific shift, and the operating hours extracted from the Availability and Utilization (A&U) information, all this through the inclusion of calculated fields.

3.3.3.4 Arranging the results

After the classification and grouping made by the pivot tables, the data is ranked from the lowest to the highest productivity value and the correspondent cumulative frequencies are computed.

3.3.3.5 Cumulative frequency

Cumulative frequency tells how often the value of the variable is less than or equal to a particular reference value.

A cumulative frequency graph is a very convenient way to present information visually, it also allows other information to be inferred. For example, from a cumulative frequency graph, we can obtain the median (or middle) mark. The median is the mark which half of all computed productivities exceed and half do not reach.

It is also possible to find the upper and lower quartile marks from the graph, as well as different percentiles.

As an example, Figures 3-17 and 3-18 present the results of the productivity analysis for Electric Shovel - 495HF and Hydraulic shovel - EX8000 models.

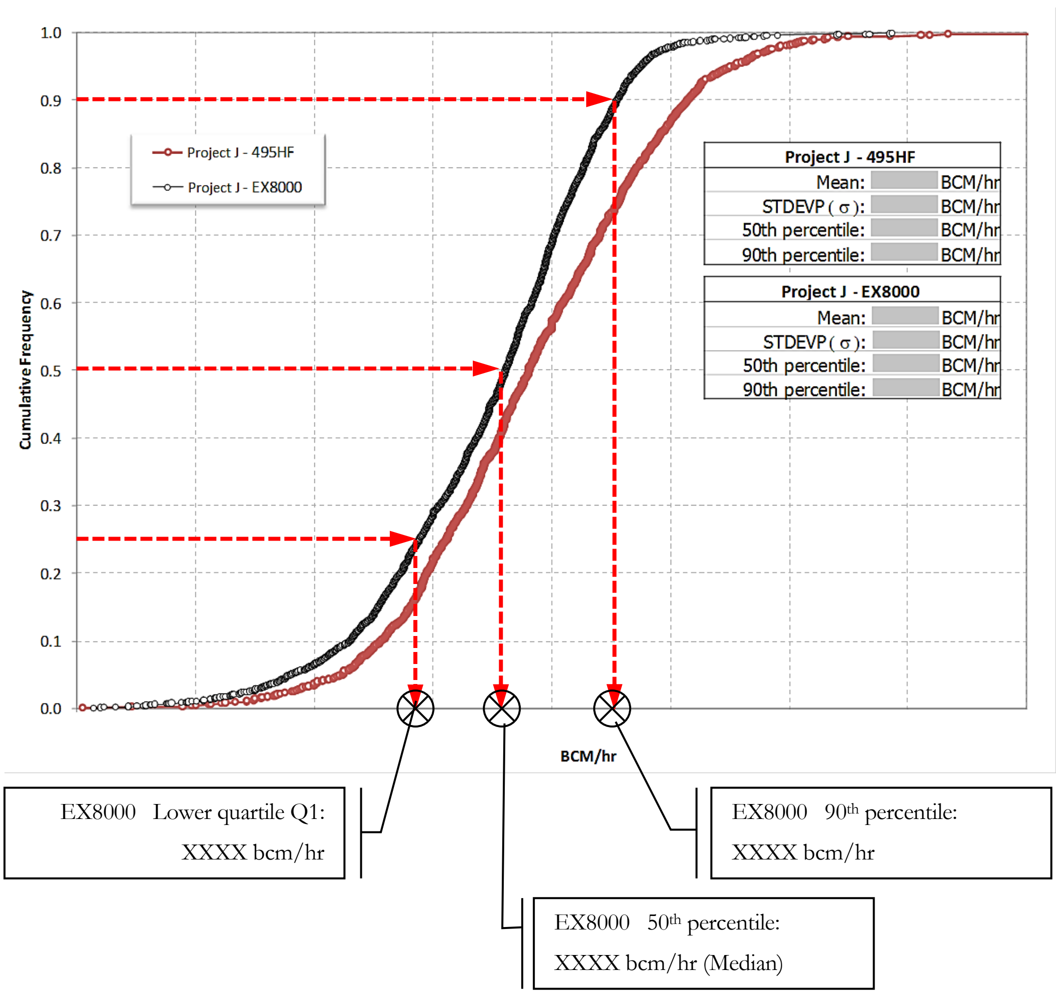


Figure 3-17 Productivity cumulative frequency excavators 495HF and EX8000 a).

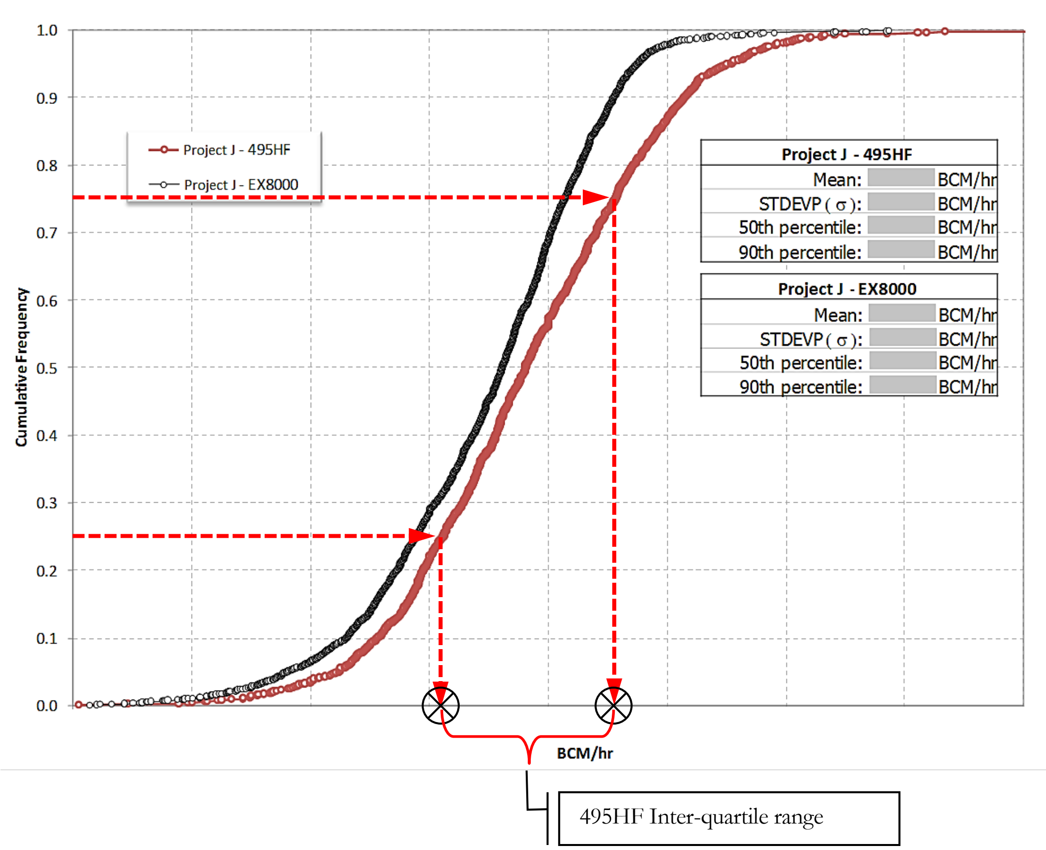


Figure 3-18 Productivity cumulative frequency excavators 495HF and EX8000 b).

3.3.4 Results

The results of the analysis are represented by a graph of cumulative frequency versus bcm (bank cubic meter) per operating hour, for every excavator model. Additionally, cumulative frequency curves are also produced for every project in which units of the excavator model are or were present.

3.4 Performance history of equipment units

3.4.1 Introduction

The purpose of this analysis was to offers an overall performance picture of each individual piece of equipment analyzed during the research. The information contained in these summary charts is grouped on a monthly basis and cover the following three modules:

1. Comparison of operating hours coming from three sources:
    a. Hauling detail information
    b. Timing data from availability and utilization information, and
    c. Timing data from accounting sources.
2. Location of the unit. This module is a diagram that summarizes where and when the unit has been operating. Information regarding the transfers of the unit from one project to another and the fraction of the month that the unit was working for each project is captured using three different series (Max, Min and Average).
3. Total bcm transported or excavated (depending if the performance history summary refers to a hauling truck or to an excavator) during the month.

These three sets of information together, provide an overall idea of the equipment unit life and it is a useful tool to identify trends and problems in productivity and/or data collection.

3.4.2 Results

Figure 3-19 presents the performance history for a 793 truck unit. In this specific case, a problem regarding operating hours is clearly identified at the top of the graph; a zoom into the operating hours was convenient for a better visualization of the values. Performance history summaries for each analyzed haul and excavator unit were produced using the same pivot tables described before for excavator and truck productivities and following a systematic procedure.

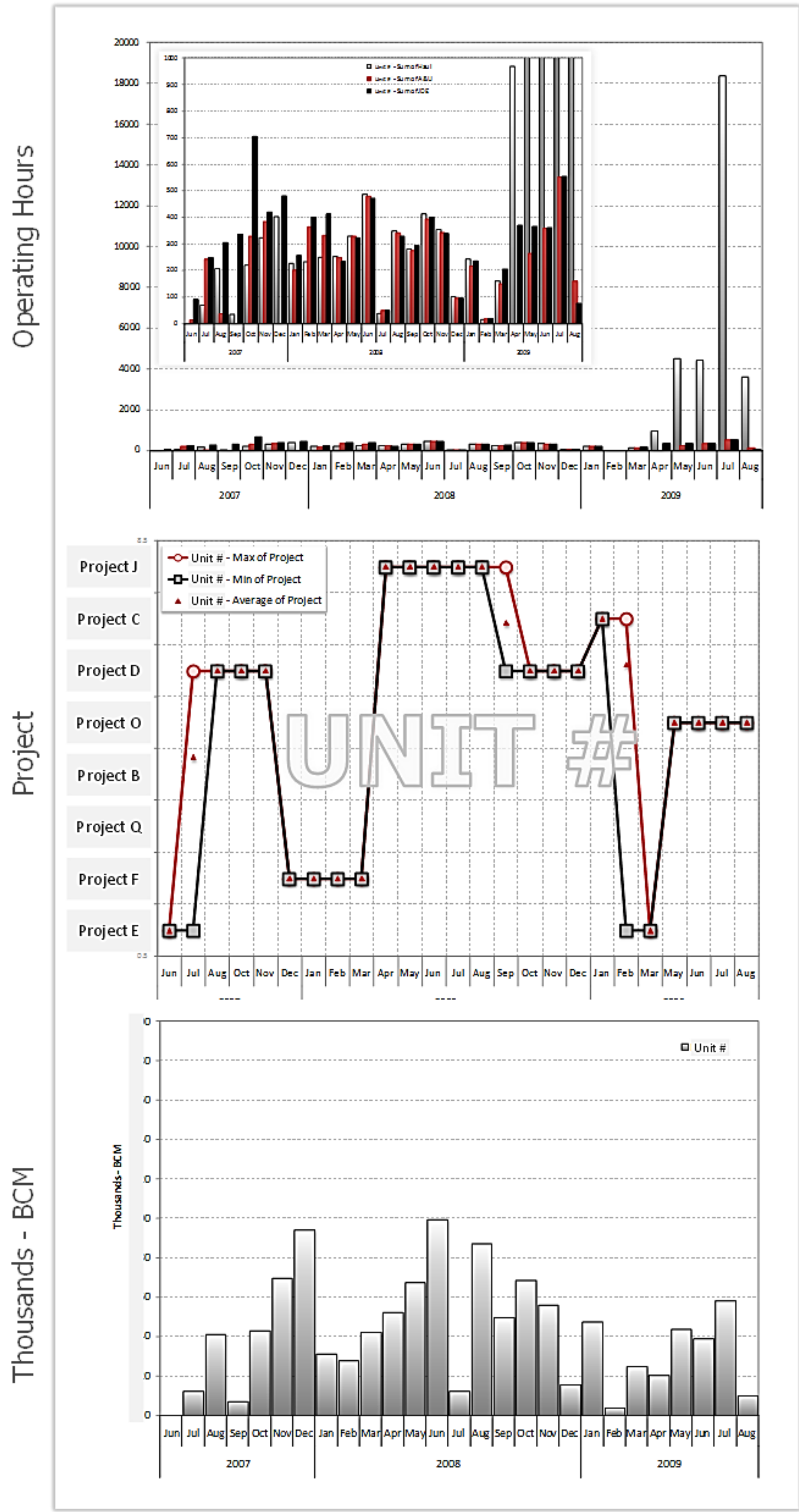


Figure 3-19 Example of cumulative frequency excavator productivity curve a)

3.5 Support Equipment

This section presents the analysis of the historical information related to the operations of support equipment (Graders and Dozers) and their relation with the operations of hauling or loading equipment, as well as with the overall production. The analysis is divided in two modules: Graders-Trucks and Dozers-Excavators.

3.5.1 Equipment

This analysis cover the information from the totality of trucks and excavators studied before plus the historical operational data from the dozers and graders models that support the operations. Around 43 grader units from eight different models and more than 200 dozer units grouped on seven different model types are part of the present analysis. Table 3-3 and 3-4 organize respectively these grader and dozer models. Note that the last column of these tables contains a multiplier factor used to convert a specific model equipment unit into the equivalent number of base model units. Figure 3-20 shows pictures of a 16M Grader and a D10T Dozer.

Table 3-3 Grader models and categories

Equipment Function	Equipment Category	Model Group	Equipment Model	Blade Lenght	% base model
GRADERS	Graders - Small	14 ft	14G	14 ft	88%
			14H	14 ft	88%
			976	14 ft	88%
	Graders - Medium	16 ft	16G	16 ft	100%
			16H	16 ft	100%
			16M	16 ft	100%
	Graders - Large	24 ft	24H	24 ft	150%
			24M	24 ft	150%

Table 3-4 Dozer models and categories

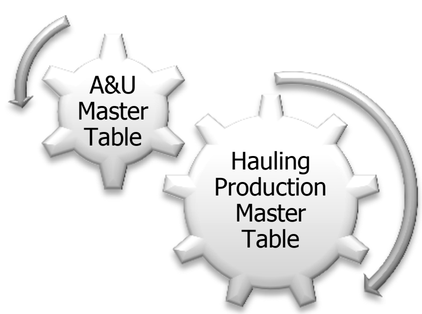
Equipment Function	Equipment Category	Model Group	Equipment Model	Flywheel Power		% base model
DOZERS	Dozers Small	D5 & Less	450	70 hp	(52 kW)	23%
			450 LGP	74 hp	(55 kW)	24%
			D32	80 hp	(60 kW)	26%
			D37	85 hp	(63 kW)	27%
			D41	110 hp	(82 kW)	35%
			D5H	96 hp	(72 kW)	31%
			D5H LGP	96 hp	(72 kW)	31%
			D5G	96 hp	(72 kW)	31%
			D5N	96 hp	(72 kW)	31%
		D6	D6D	140 hp	(104 kW)	45%
			D6H LGP	140 hp	(104 kW)	45%
			D6M	140 hp	(104 kW)	45%
			D6M LGP	140 hp	(104 kW)	45%
			D6N LGP	150 hp	(112 kW)	48%
			D6R	165 hp	(123 kW)	53%
			D6R LGP	165 hp	(123 kW)	53%
			D6T LGP	165 hp	(123 kW)	53%
			D6T XW	165 hp	(123 kW)	53%
			850	185 hp	(138 kW)	60%
		D7	D7R LGP	240 hp	(179 kW)	77%
	Dozers Medium	D8	D8N	285 hp	(213 kW)	92%
			D8R	310 hp	(231 kW)	100%
			D8T	310 hp	(231 kW)	100%
		D9	D9R	410 hp	(306 kW)	132%
			D9T	410 hp	(306 kW)	132%
	Dozers Large	D10	D10N	520 hp	(388 kW)	168%
			D10R	580 hp	(433 kW)	187%
			D10T	580 hp	(433 kW)	187%
			D375A5	606 hp	(452 kW)	195%
		D11	D11R	850 hp	(634 kW)	274%
			D475A5	900 hp	(671 kW)	290%
			D11T	850 hp	(634 kW)	274%

Figure 3-20 Grader 16M and Dozer D10T
(Pictures from: AEHQ5734-01 (1-07) - 16M Motor Grader specifications - © 2007 Caterpillar
AEHQ5592-01 (7-07) - D10T Track-Type Tractor specifications - © 2007 Caterpillar)

3.5.2 Methodology

The methodology implemented for the analysis of support equipment follows the path of loading and hauling productivity analyses with some variations. The analysis in this section focuses only on those projects in which the activities of the secondary equipment represent support of the operations of the primary equipment instead of general mining or heavy construction tasks.

The new input information of the analysis is the combination of the hauling data contained in the hauling production master table and the information of the availability and utilization broken-down by equipment category.



Two main pivot tables are used in this analysis. The first one generates monthly categorized information related to operating hours of graders and trucks model groups, as well as bcm production and average hauling distances, while the second generates monthly categorized information related to operating hours of dozers and excavator model groups, as well as bcm production.

The information is then extracted from the pivot tables and prepared to plot. Line, column and staked area charts are the final result of the analysis. The use of an analysis built template allows this process to be semi-automatic. In order to present an appropriate comparison of equipment operating hours, different equipment models are transferred into a base equipment model. This transformation is based on the main feature of the equipment, as follows: tons capacity for trucks, blade width for graders, bucket size for excavators, and horse-power for dozers.

3.5.3 Results

3.5.3.1 Truck-Graders

The results summary for the interaction between trucks and graders project-specific includes:

- Total grader operating hours by model group (staked area series).
- Total grader operating hours in equivalent 16ft grader units (line series).
- Total truck operating hours by model group (staked area series).
- Total truck operating hours in equivalent 777 truck model units (line series).
- Ratio between equivalent operating hours of trucks and graders.
- Total production in bcm.
- Average hauling distance.

Figure 3-21 describes these results for a specific project, while Figure 3-22 presents an example of the summary chart for project J.

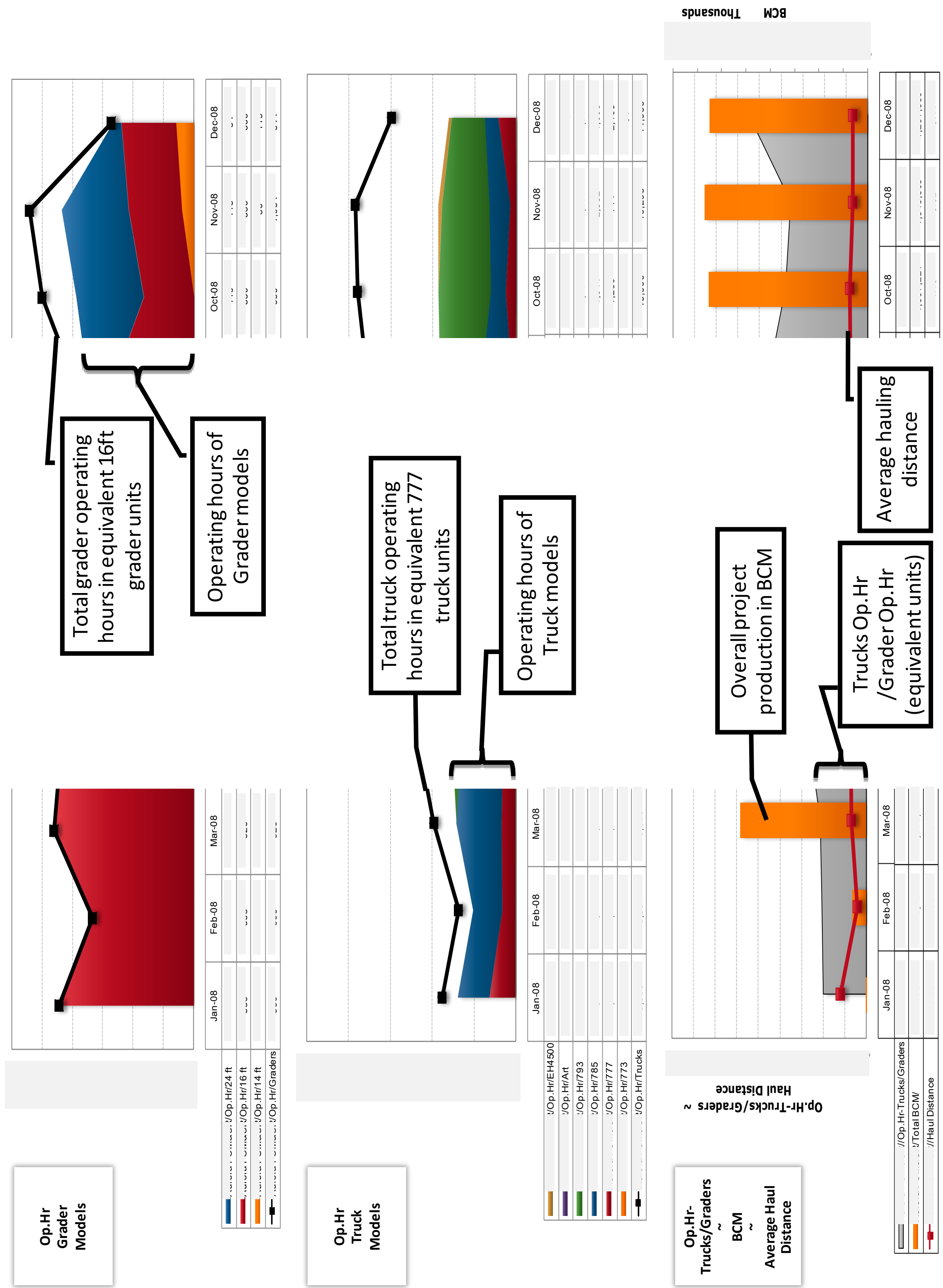


Figure 3-21 Support equipment analysis Truck vs. Graders

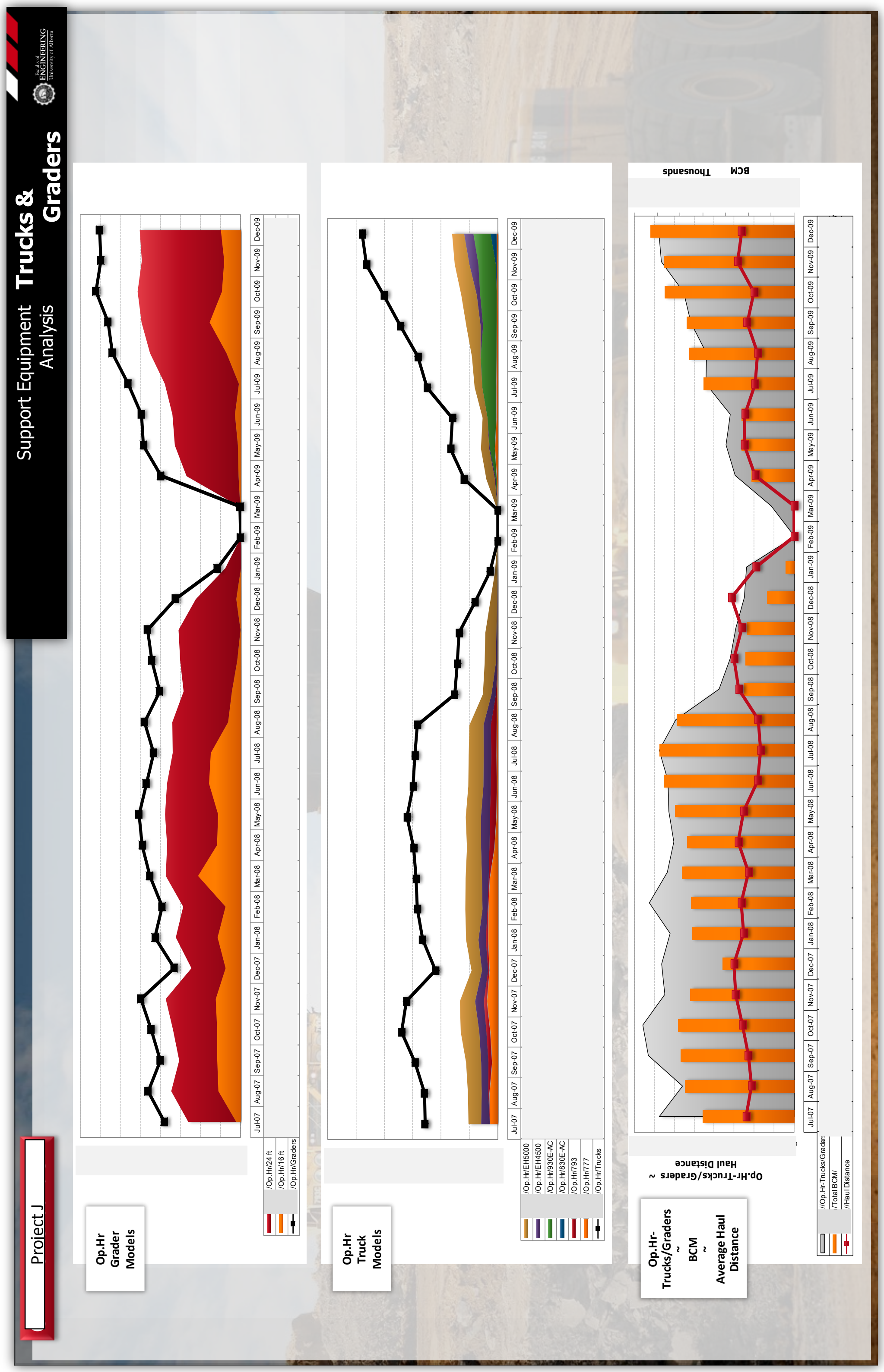


Figure 3-22 Support equipment analysis Truck vs. Graders

3.5.3.2 Excavators-Dozers

The results summary chart for the interaction between excavators, dozer and project production includes:

- Total dozer operating hours by model group (staked area series).
- Total dozer operating hours in equivalent D8 dozer units (line series).
- Total excavator operating hours by model group (staked area series).
- Total excavator operating hours in equivalent $10m^3$ bucket size excavator model units (line series).
- Total production in bcm.
- Ratio between equivalent operating hours of excavators and dozers based on equivalent units.
- Ratio between bcm production and operating hours of dozers.

Figure 3-23 presents an example of the summary chart for project J.

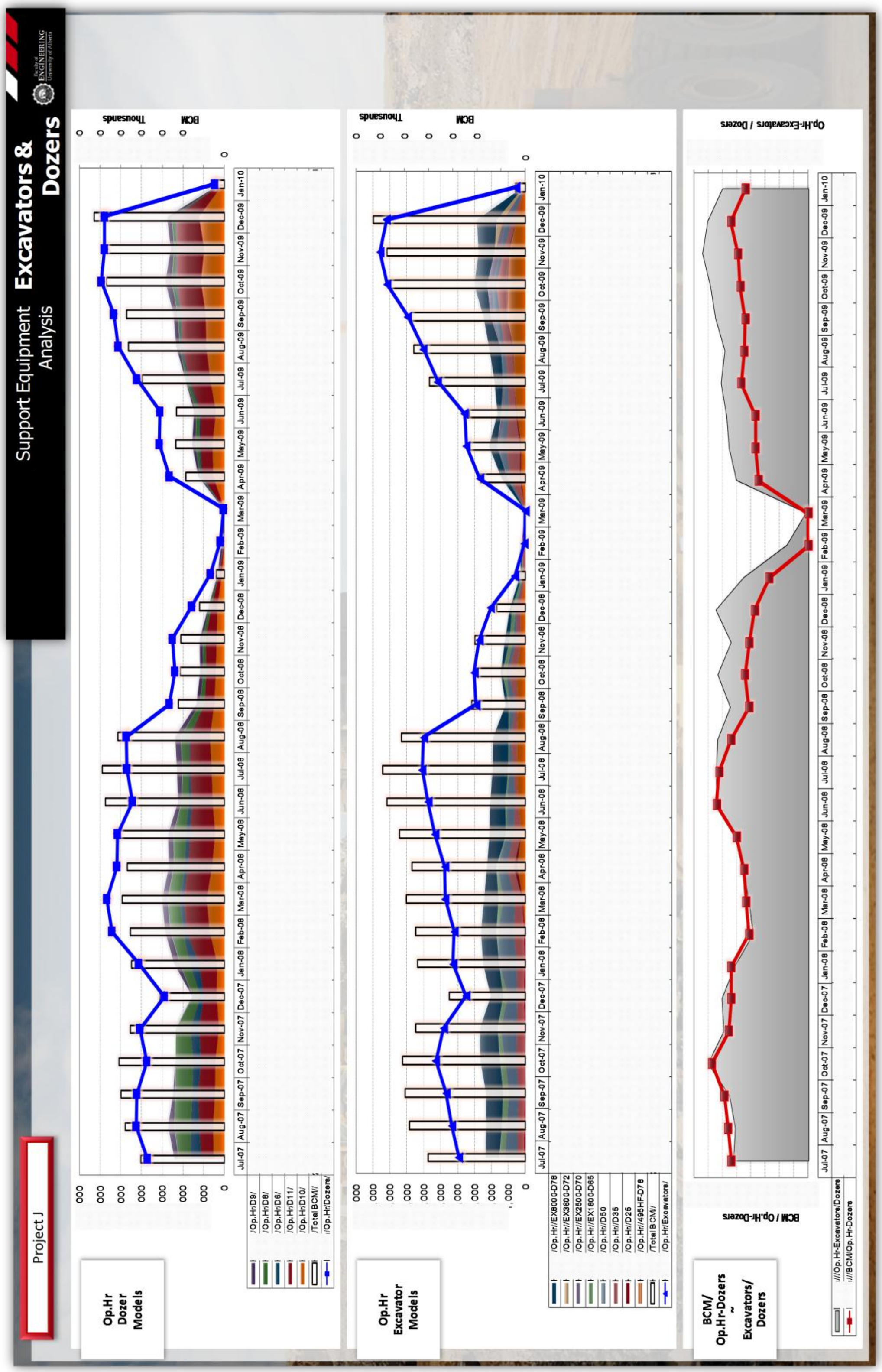


Figure 3-23 Support equipment analysis Truck vs. Graders

3.6 Conclusions and recommendations

This Chapter has described the analysis of historical information regarding the operations of major hauling and loading equipment as well as support equipment. Productivity curves and equations were built and presented as the output of the analysis. It also presented a methodology to represent the performance summaries of equipment units involved in the earthmoving operations. All these results represent a valuable tool for estimating.

Practical knowledge contained in the results includes:

- Differences in project performances (easily recognized from the productivity curves for hauling and loading).
- Project characteristics (e.g. hauling distance ranges, project scale, equipment models involved, etc.)
- Problems in the data collection at different sites and periods (contained in the quality improvement summary).
- Productivity charts (can be used to forecast hauling and loading units required in new projects).
- Historical ratios between primary and support equipment compared to accomplished production.

This analysis involved detailed reviews of the data that is contained in a large DBMS of a major construction contractor. Through the analysis, multiple sources of errors and inconsistencies were identified. The errors range from erroneous data being entered (such us 10,000 km haul distances) to duplicate data (same haul unit entered at two sites).

Results from this analysis could be combined to extract handy information e.g. how many units of a given truck model should be used for a specific type of excavator. This is possible with the combination of the average hauling distance, the productivity information of trucks and excavators, and certainly, expert knowledge.

CHAPTER 4 ANALYZING EARTHMOVING PRODUCTIVITY USING DATA MINING AND NEURONAL NETWORKS

4.1 Introduction

In the construction management field, data mining and ANNs will perhaps look as components of complicated systems that use expert-given rules or statistical inference techniques to provide decision support for experts, help decision makers perform at a higher level, assist in the training of inexperienced personnel and help scenario planning (i.e. what if?) by managers (Boussabaine 1996). This research aims to incorporate Neural Networks into a system oriented to the enhancement of managerial decision making on the heavy construction and superficial mining earthmoving operations field.

With the use of data mining techniques and the utilization of artificial neural network tools this chapter presents the study of the influence of the following variables on the productivity of haul trucks: hauling distance, excavator model utilize for loading, material type, temperature or season and average slope of the hauling path. Likewise, it studies the influence of the material type, project, temperature or season, and the size of the truck that is being served, on the productivity of the excavators.

4.2 Nominal vs. numerical variables

Nominal, or categorical, variables contain values that lack the properties of order, scale, or distance between them. If these variables will be used in any kind of algorithms, it is important to retain the lack of order or scale in categorical variables. Consequently, it is not desirable that a nominal variable be converted into a series of integers. Ordinal variables are categorical variables with the notion of order added to them (e.g. low, medium, high). Real measures, or continues variables, are the easiest to use and interpret as they have all desirable properties of variables: order, scale, and distance (Chakrabarti, et al. 2009).

One particular goal for this chapter is to analyze the convenience of using numerical variables instead of nominal ones as the main input type of ANNs. For example, will it be better to use the mean temperature of the day instead of the season of the year as an input of an ANN in order to represent surrounding conditions? Figure 4-1 shows the Daily Max, Min and Mean temperatures in Fort McMurray from Jun-07 to Feb-09.

Table 4-1 and 4-2 presents the different nominal attributes extracted from the historical operational information that is being analyzed in this research, and the correspondent numerical attributes that could be used instead in the analysis of hauling trucks and excavator productivities.

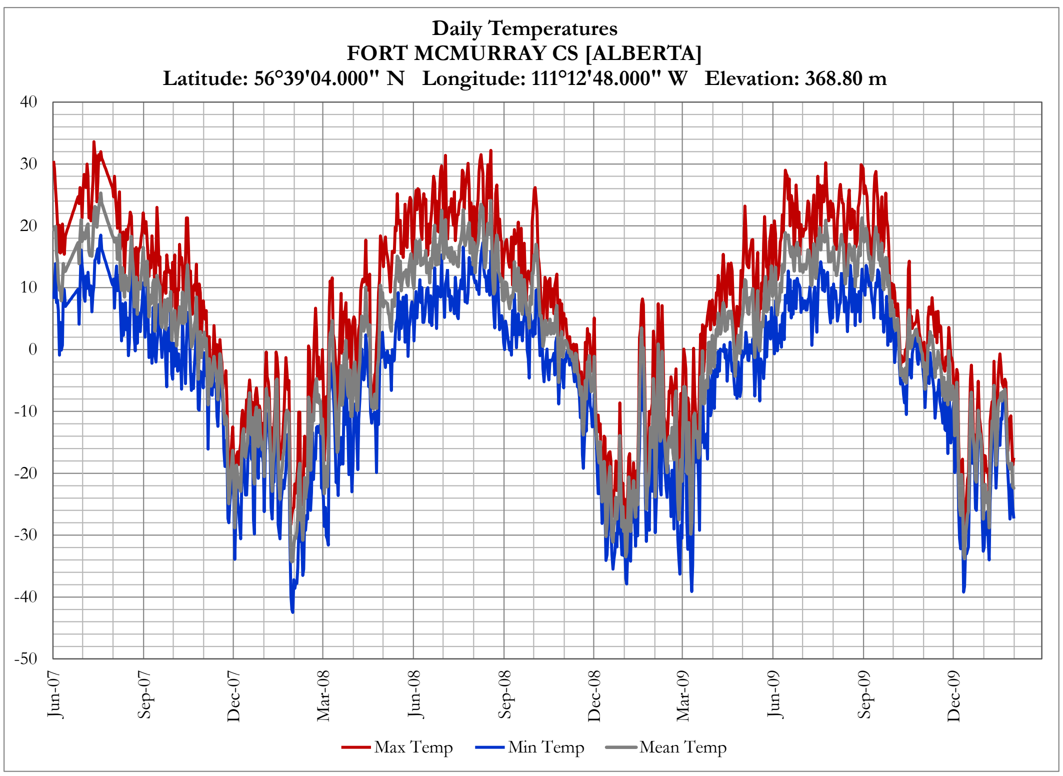


Figure 4-1 Daily Max, Min and Mean temperatures in Fort McMurray.

Table 4-1 Input type approaches for the evaluation of hauling truck productivities.

Attribute description	Input type #1 Nominal variables as main type of input	Input type #2 Numerical variables as main type of input
Distance	Distance	Distance
Excavator model	495HF, EX8000, EX5500, EX3600, EX2500, EX1900, EX1800, EX1200, EX850, EX800, EX750, EX600, EX550, EX450, EX330/350, EX300, EX200	Excavator bucket size
Material type	Granular, Muskeg, Oilsand, Overburden, Rock - High, Rock - Medium, Slop	Material density
Project	C, D, E, F, H, I, J, L, M, O, S	C, D, E, F, H, I, J, L, M, O, S
Time of year	Fall, Winter, Spring, Summer	Daily mean temperature
Hauling conditions*	Average path slope*	Average path slope*

* Available for the data coming from only one project.

Table 4-2 Input type approaches for the evaluation of excavator productivities.

Attribute description	Input type #1 Nominal variables as main type of input	Input type #2 Numerical variables as main type of input
Material type	Granular, Muskeg, Oilsand, Overburden, Rock - High, Rock - Medium, Slop	Material density
Project	C, D, E, F, H, I, J, L, M, O, S	C, D, E, F, H, I, J, L, M, O, S
Time of year	Fall, Winter, Spring, Summer	Daily mean temperature
Truck category	Trucks - 280t+ Trucks - 220-280t Trucks - 120-220t Trucks - 80-120t Trucks - 40-80t	Truck model size

4.3 Data mining

Data mining is defined as the process of discovering patterns in data. Moreover, data mining is not only used for predictions, but it is frequently used to gain knowledge from data, which it certainly sounds like a good idea if you can do it (Chakrabarti, et al. 2009).

The system used for data mining in this research is called WEKA. The Waikato Environment for Knowledge Analysis (WEKA) is recognized as a landmark system in data mining and machine learning as it has achieved widespread acceptance within academia and business circles, and has become a widely used tool for data mining research (Hall, et al. 2009).

4.3.1 Visualizing

A scatter plot is one of the most effective graphical methods for determining if there appears to be a relationship, pattern, or trend between two numeric attributes (Chakrabarti, et al. 2009) plus it gives the ability to easily identify outliers. Figure 4-2 shows different examples of scatter plots. The productivity information for hauling trucks and excavator models was plotted using the selected data mining tool, which offers the opportunity to visualize the interaction of every input attribute with the rest and with the output, see Figure 4-3.

It was also possible to plot with a single click the way in which different input attributes affect the output values. See Figure 4-4, where histograms of loads per hour values of one model of truck are plotted with different colors representing several input attributes (project, excavator model being used, and material type). Figure 4-5 presents similar information, but instead of histograms, a set of scatter plots of productivity versus hauling distance is presented colored by project, excavator model and material type.

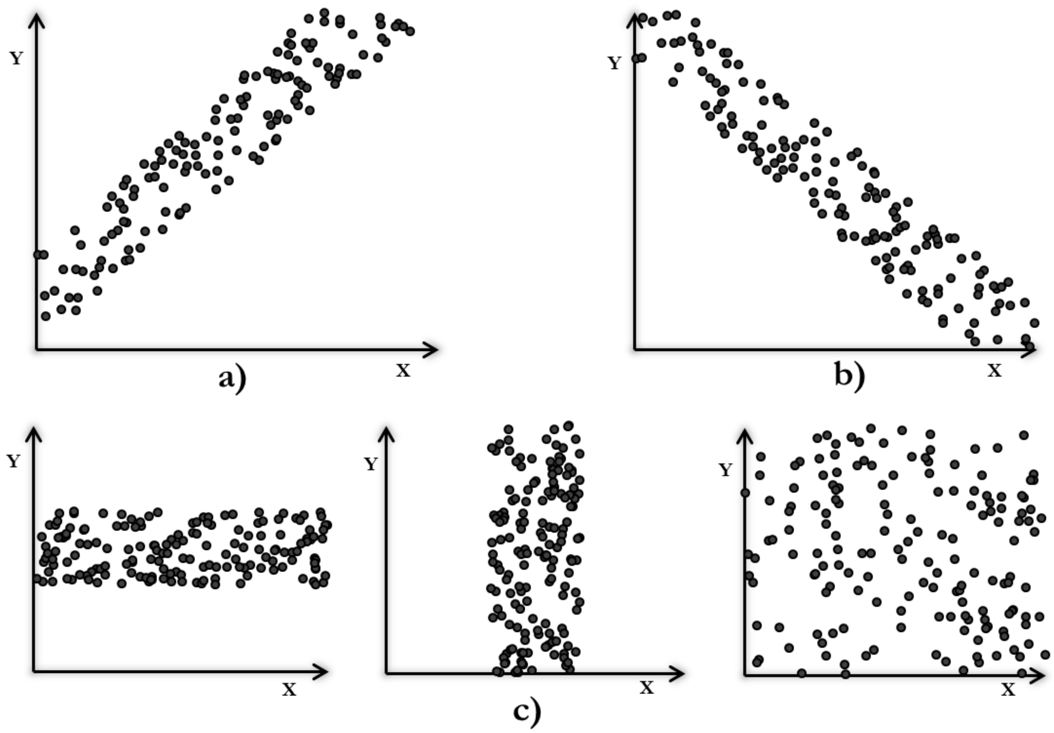


Figure 4-2 Examples of scatter plots.
a) Positive correlation b) Negative correlation c) No observed correlation.

4.3.2 Filtering and transforming

Data mining tools were also useful to filter the data and get rid of outliers. The analysis of different attributes was made only for those attributes in which the number of instances was higher than the 0.1% of the total number of events (i.e. generally more than ten instances). Transforming nominal or categorical attributes into a set of binary fields was possible using a simple nominal-to-binary filter. After this transformation, the data sets were ready for ANN implementation.

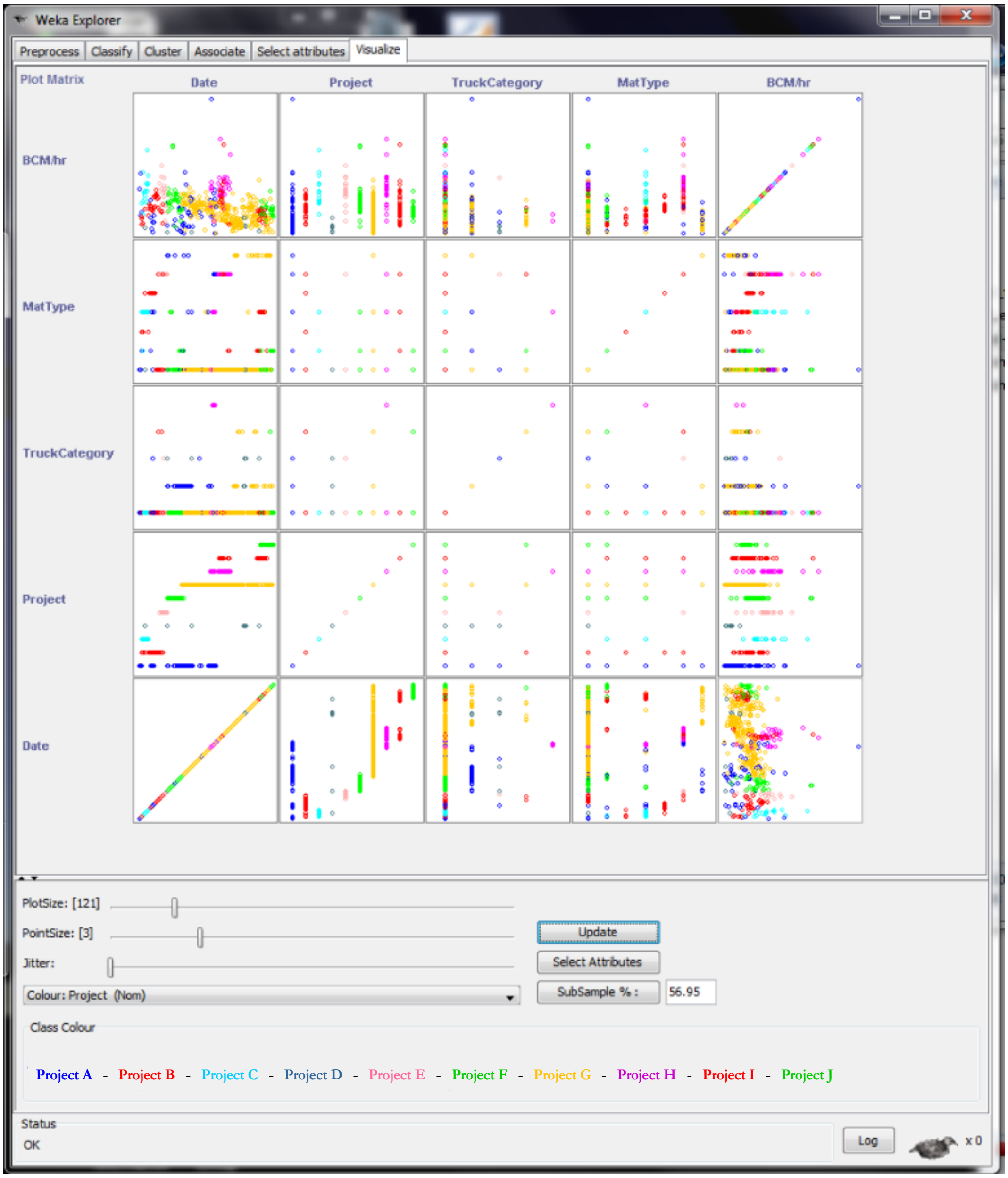


Figure 4-3 **Scatter plots - productivity parameters of a particular excavator model.**

Truck 793
Loads/hr By:

- Project
- Ex Model
- Material Type

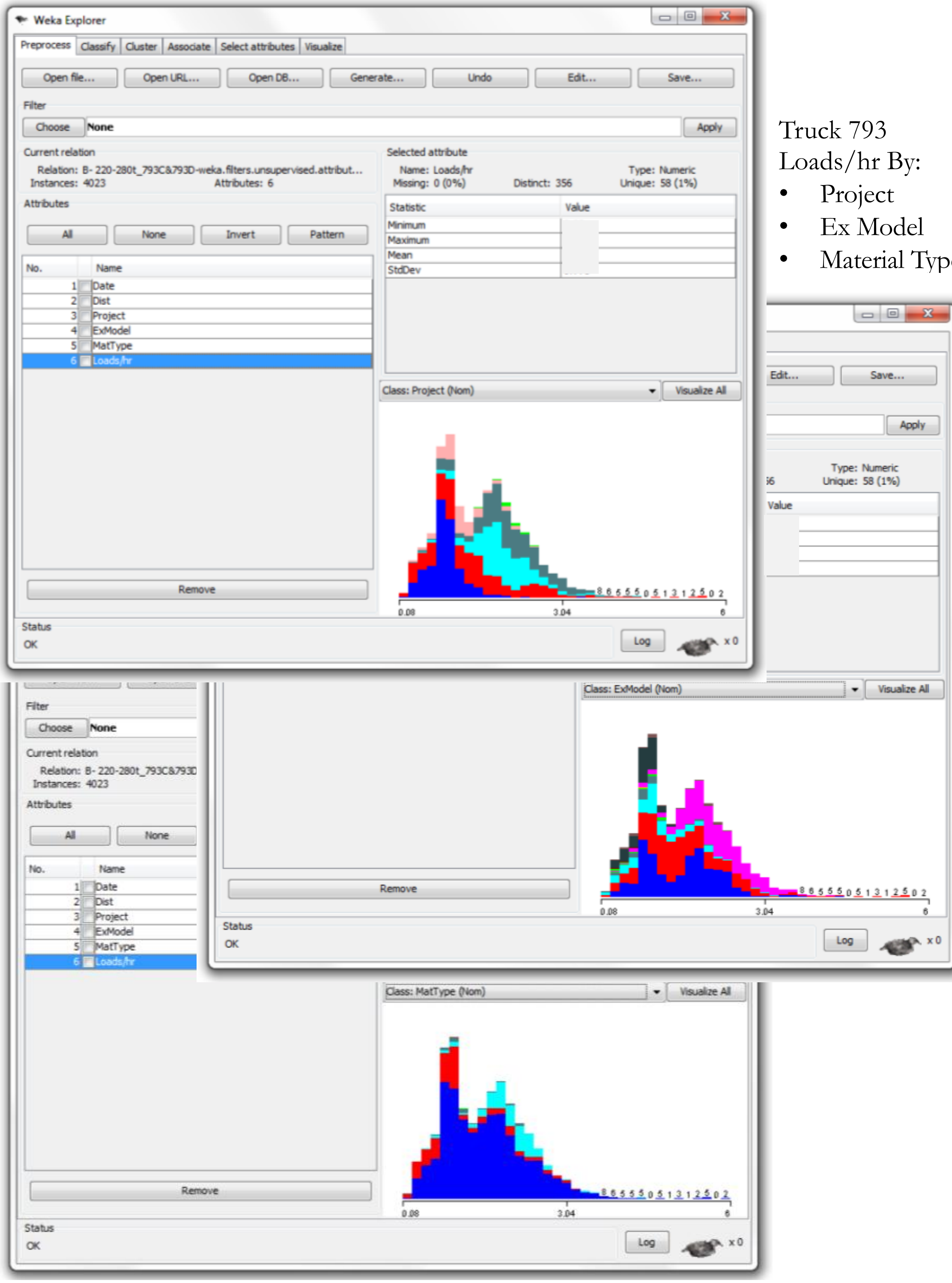


Figure 4-4 Histograms of loads/hour -793 Truck model.
by project, excavator model, and material type.

Truck 793
Loads/hr vs.
hauling distance
By:

- Project
- Ex Model
- Material Type

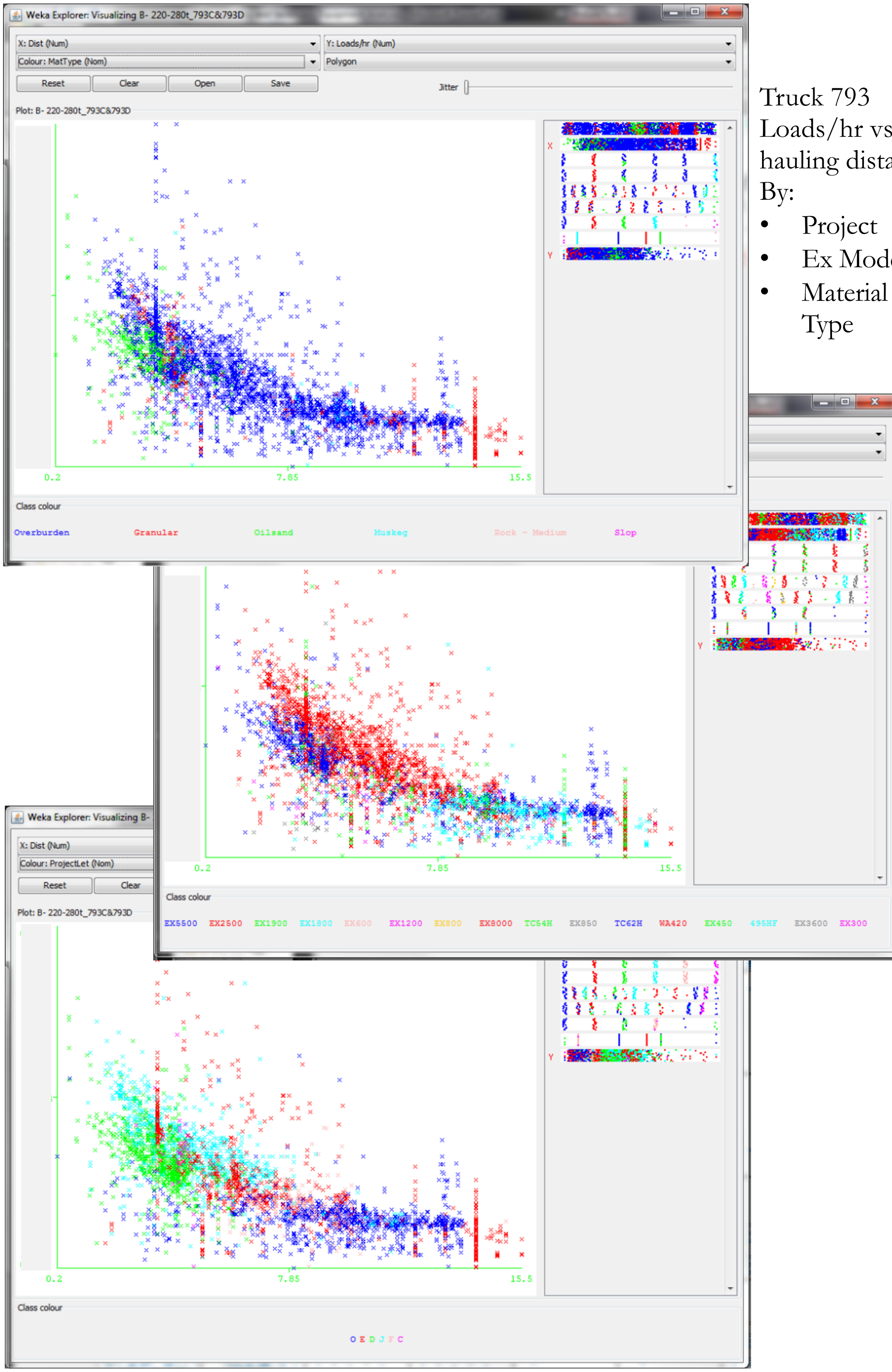


Figure 4-5 Loads/hour vs. distance - 793 Truck model by project, excavator model, and material type.

4.4 Artificial Neuronal Networks

This section studies the development and implementation of artificial neural networks (ANNs) as a mean of improving the abilities of an estimator to predict heavy construction equipment productivity rates. ANNs are information processing technologies that attempt to simulate, in a gross manner, the networks of nerve cell (neurons) of the biological (human or animal) central nervous system (Graupe 2007). As the brain, ANNs learn from experience, generalize from previous examples to new ones and abstract essential characteristics from inputs containing irrelevant data.

The system used on this research to develop ANNs is Neuroshell ® 2. NeuroShell 2 is a software program developed by Ward Systems Group®, Inc. that enables you to build sophisticated custom problem solving applications without programming. You tell the network what you are trying to predict or classify, and NeuroShell 2 will be able to "learn" patterns from training data and be able to make its own classifications, predictions, or decisions when presented with new data (NeuroShell 2 Help n.d.). Figure 4-6 presents the display of the NeuroShell 2 advanced options screen in which the independent modules that may be used to create a neural network application are shown.

The software also allows the user to obtain the relative contribution of every input parameter, as well as to track graphically the training average error. Figures 4-7 and 4-8 respectively show the relative contribution factors and training average error evolution for one of the ANN configurations that evaluate the productivity for the hauling truck model 785.

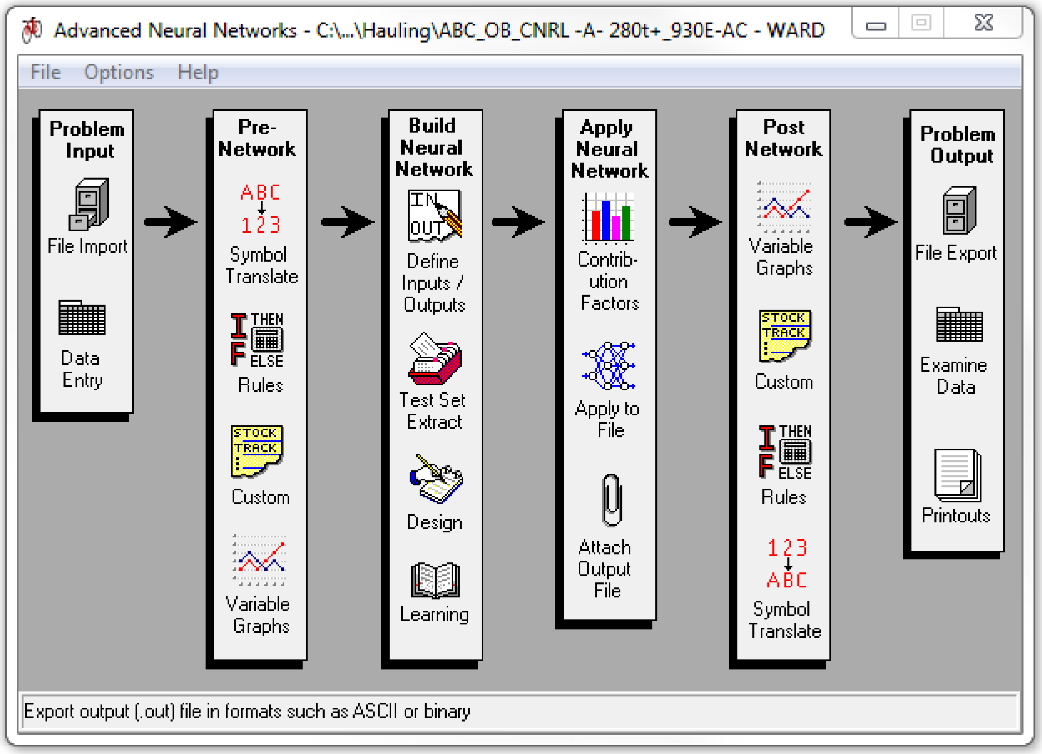


Figure 4-6 The NeuroShell 2 Advanced Options screen display

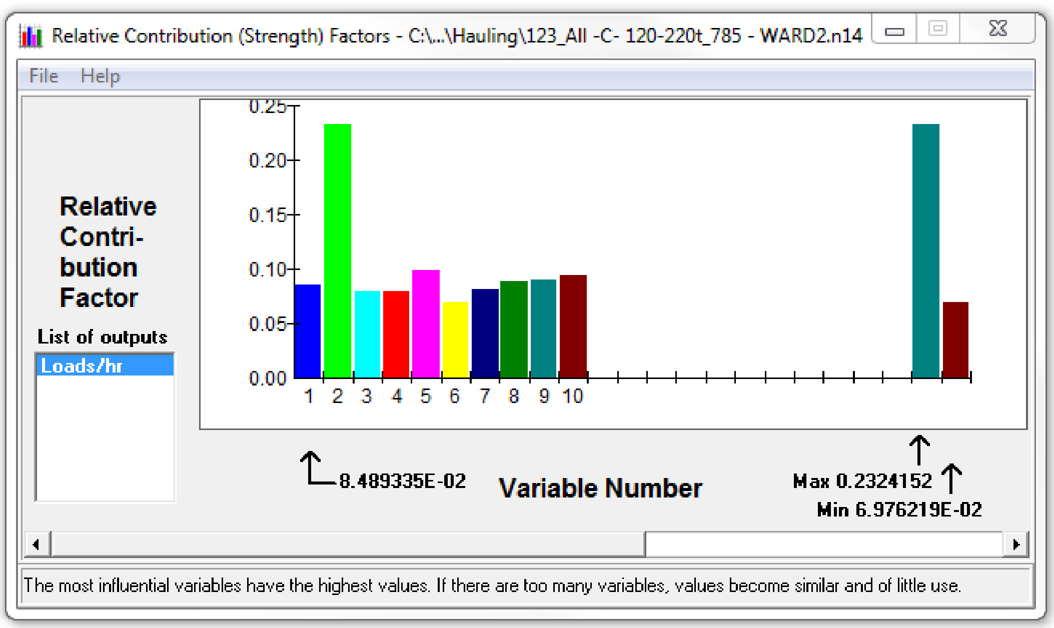


Figure 4-7 Relative contribution factors - Truck model 785.

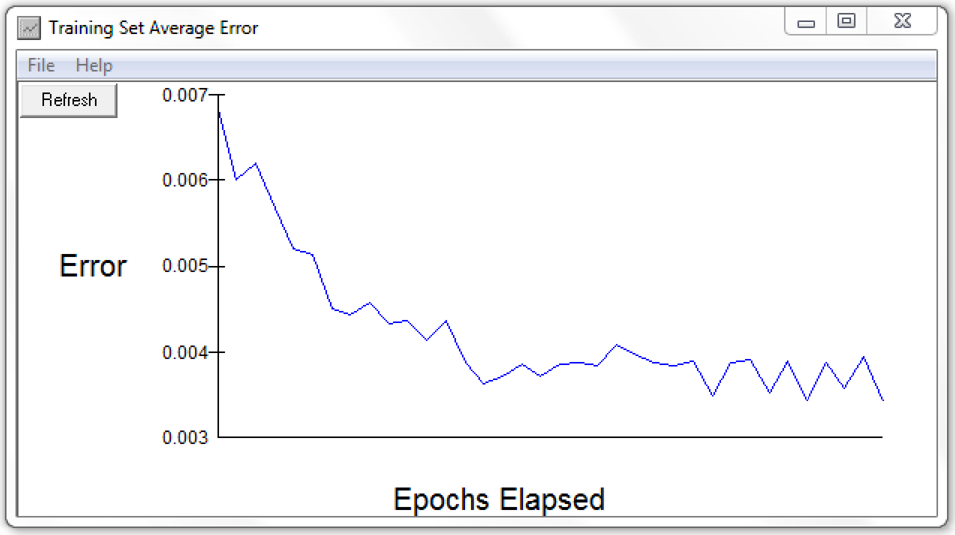


Figure 4-8 Training average error evolution - Truck model 785.

The hauling truck and excavator operational information that was pre-processed using data mining techniques is the foundation matter of the ANNs that are developed on this section. Note again that this information represents the operation information of more than 380 pieces of equipment between hauling trucks and excavators, collected for more than three years in eleven different projects.

One of the objectives of this research is to analyze the sensitivity of different parameters of artificial neural networks in the evaluation of earthmoving operational data. With this in mind, the influence of different input types, ANN architectures, and number of hidden nodes on the performance of the ANN are evaluated in this section. Other ANN parameters such as learning rate, momentum, and initial weights will not be evaluated on this research.

4.4.1 Architectures

Neuroshell® 2 offers to the ANN developer a wide range of network architecture options (see Figure 4-9). After a fairly varied evaluation, and using the rule of thumb, two different architectures were selected to be assessed on this research: Four layers – Standard connections and Ward net with three hidden slabs and different activation functions. Both of these architectures are backpropagation networks. Backpropagation networks are known for their ability to generalize well on a wide variety of problems. That is why they are used for the vast majority of working neural network applications. (NeuroShell 2 Help n.d.)

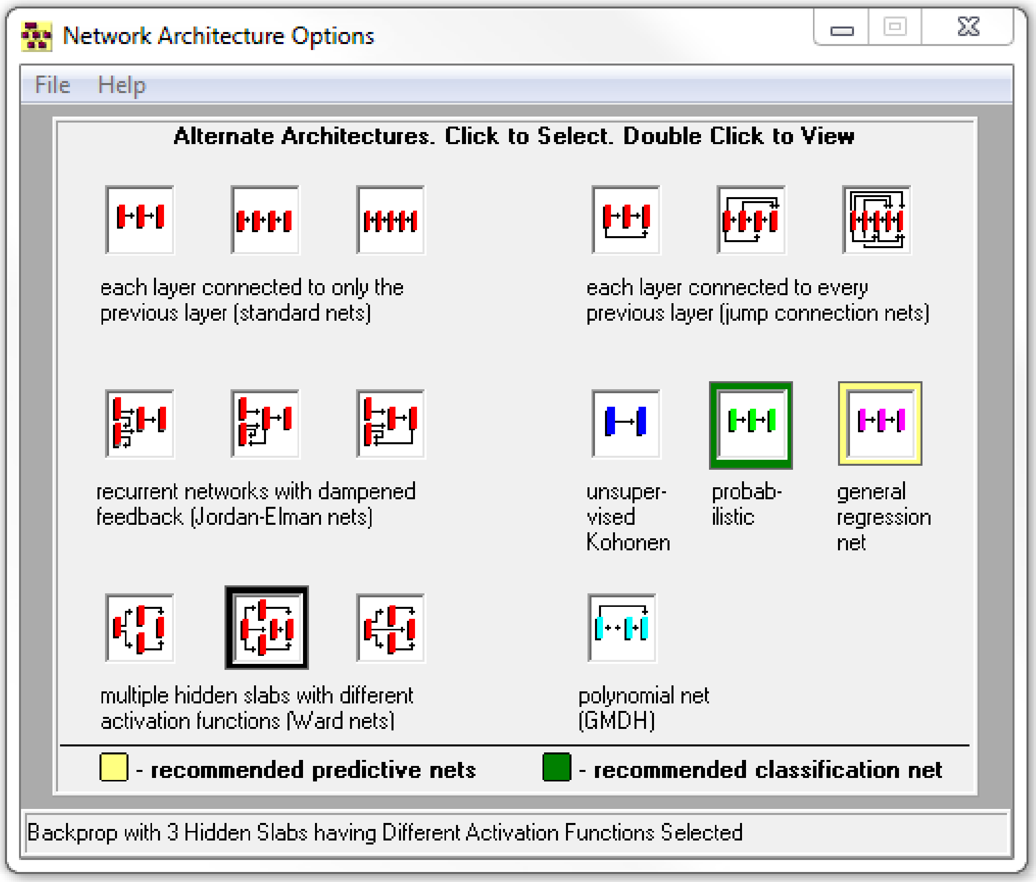


Figure 4-9 Network architecture options

4.4.1.1 Four layers – Standard connections

This is the standard type of backpropagation network in which every layer is connected or linked to the immediately previous layer. It has four different layers including one input layer, one output layer and two hidden layers; see Figure 4-10. The number of nodes on the input layer depends on the model of equipment analyzed and the type of input, the number of nodes on the hidden layers varies and is a parameter that will also be analyzed. The output layer has only one node (equipment productivity).

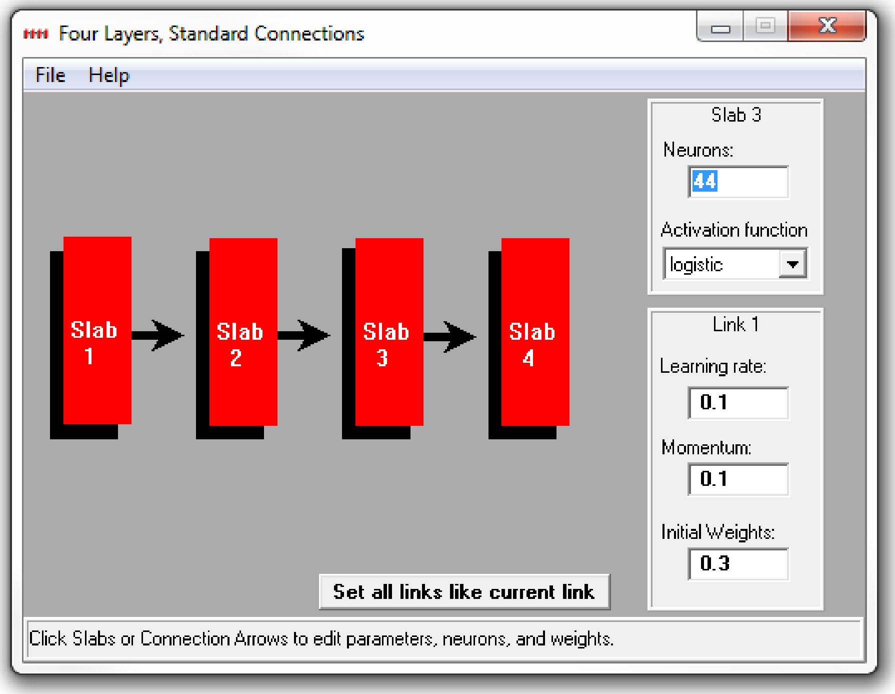


Figure 4-10 Standard backpropagation network with four layers.

4.4.1.2 Ward net with three hidden slabs and different activation functions

Hidden layers in a neural network are known as feature detectors. Ward Systems Group invented three different backpropagation network architectures with multiple hidden layers. Different activation functions applied to hidden layer slabs detect different features in a pattern processed through a network. For example, a network design may use a Gaussian function on one hidden slab to detect features in the mid-range of the data and use a Gaussian complement in another hidden slab to detect features from the upper and lower extremes of the data. Thus, the output layer will get different "views of the data." Combining the two feature sets in the output layer may lead to a better prediction (NeuroShell 2 Help n.d.). This section will assess the performance of the 3 hidden slabs ward neural network architecture, see Figure 4-11.

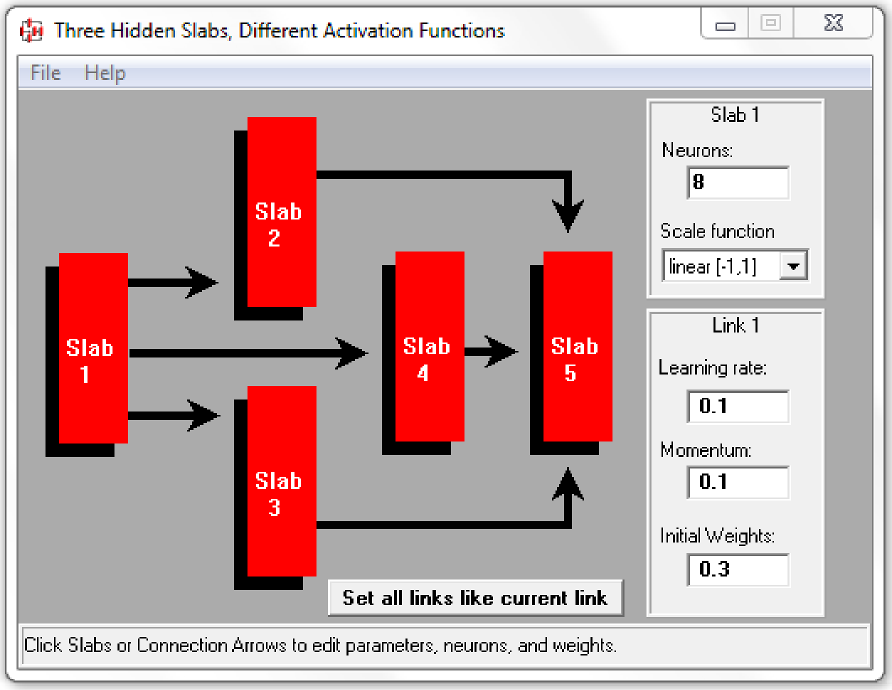


Figure 4-11 Ward backpropagation net with three hidden slabs.

4.4.2 Number of Hidden Neurons

In Backpropagation networks, the number of hidden neurons determines how well a problem can be learned. If you use too many, the network will tend to try to memorize the problem, and thus not generalize well later. If you use too few, the network will generalize well but may not have enough "power" to learn the patterns well. Getting the right number of hidden neurons is a matter or trial and error, since there is no science to it. The software default number of hidden neurons for a 3 layer network is computed following Equation 4-1. For more hidden slabs, divide the number above by the number of hidden slabs.

$$\#\ of\ hidden\ neurons\ = 1/2\,(Inputs\ +\ Outputs)\ +(\#\ of\ patterns\ used\ for\ training)^{1/2} \qquad \text{4-1}$$

This default manner to compute the number of hidden neurons is the first approach of number of hidden nodes to be analyzed and it is denoted by the letter A, as STD-A and WARD-A for the different architectures using this approach. The second manner of computing the number of hidden neurons was selected using the rule of thumb, and use the product of the number of inputs parameter and the number six. This approach is denoted with the letter B (e.g. STD-B and WARD-B).

4.5 Results

This section presents a comparative summary of the performances of different ANN configurations that aim to evaluate hauling truck and excavator productivities. The parameter used in order to perform this comparison is the square of the correlation coefficient (r squared). The correlation coefficient r (Pearson's Linear Correlation Coefficient) is a statistical measure of the strength of the relationship between the actual vs predicted outputs. The r coefficient ranges from -1 to +1. The closer r is to 1, the stronger the positive linear relationship, and the closer r is to -1, the stronger the negative linear relationship. It is possible to get the same results by using the Correlation Scatter Plot and graphing actual vs predicted outputs. Another comparison using the coefficient of multiple determination R Squared is offered as an appendix, see Appendix A. The formula used for the correlation coefficient r is given by Equation 4-2.

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}} \quad \text{4-2}$$

where,

$$SS_{xy} = \sum XY - \frac{(\sum X)(\sum Y)}{\text{n}} \quad \text{4-3}$$

$$SS_{xx} = \sum x^2 - \frac{(\sum x)^2}{\text{n}} \quad \text{4-4}$$

$$SS_{yy} = \sum y^2 - \frac{(\sum y)^2}{\mathrm{n}} \qquad \textbf{4-5}$$

Where n equals the number of patterns, x refers to the set of actual outputs, and y refers to the predicted outputs.

4.5.1 Haul truck productivity

4.5.1.1 r-squared - ANN – different architectures

Figure 4-12 presents the comparison summary between different analyzed architectures. Four different ANN configurations were generated for twenty six sets of data. On the data set description, the first three characters are a hint of which type of input is used, ABC represents nominal input type and 123 refers to numerical input type (see section 4.2). The second part of the name string has two different values All or OB, this refers to the number of projects that the set of data contain. OB means that the data contains only one project in which material type is overburden; in those data sets the hauling path average slope parameter was available. The last part of the name string refers to the equipment category and model.

The results clearly show the low performance of every analyzed ANN configuration. Nevertheless, it can be noted that ward net architectures got higher correlation values compared with four layer architectures. The difference in the ANN performance depending on how the number of hidden neurons is computed (-A or -B) is minor. Figure 4-12 presents these results graphically.

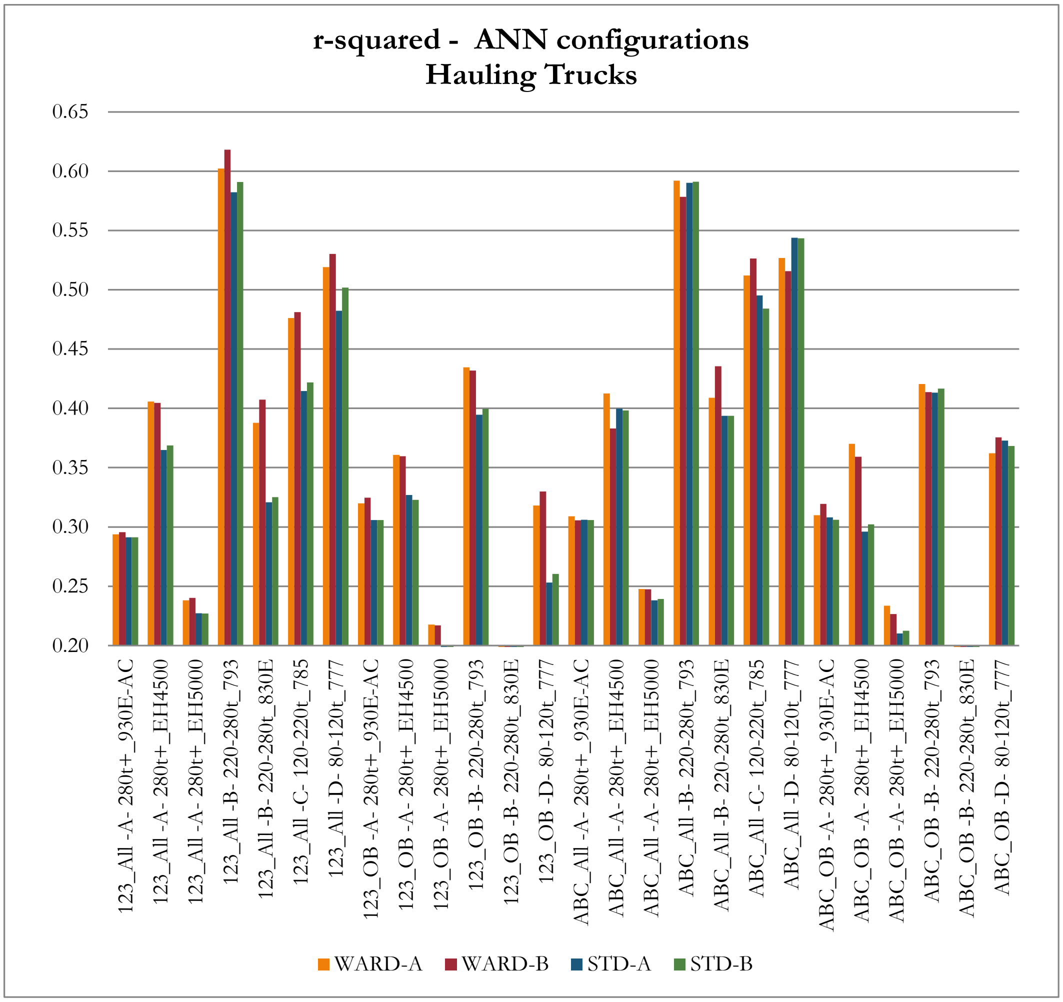


Figure 4-12 r-squared comparison summary - ANN architectures – Hauling trucks.

4.5.1.2 r-squared - ANN – different input types

Figure 4-13 aims to summarize which type of input (nominal or numerical) result in better ANN performance. The values in this figure come from the Ward-A configuration. As it is shown, ANN configurations that include all the projects resulted in higher r-square values. Nominal inputs perform slightly better than numerical ones.

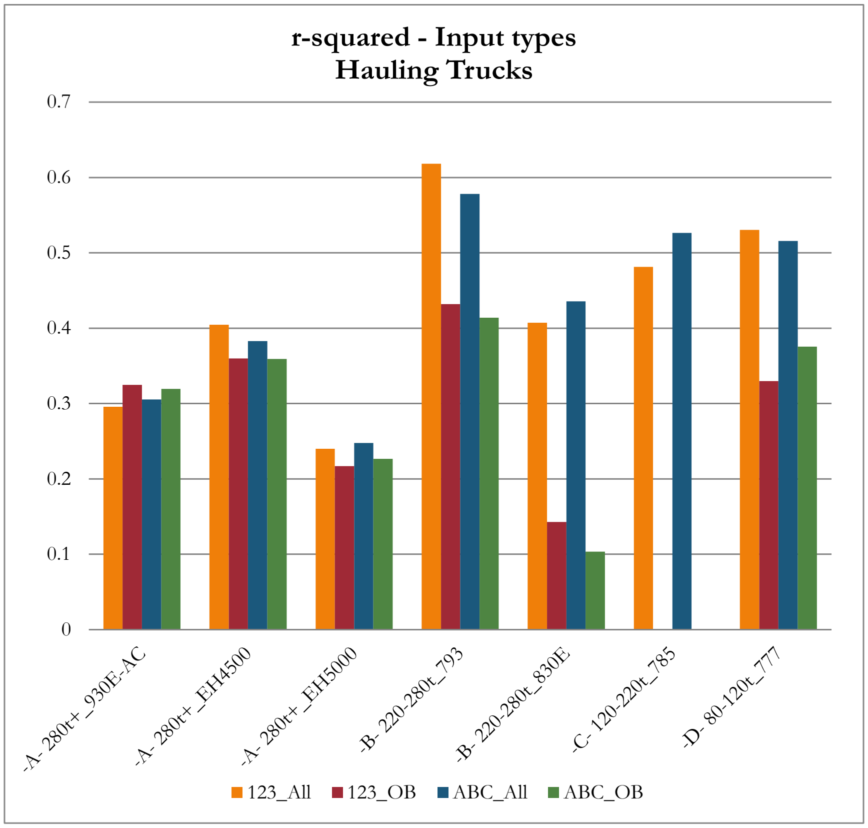


Figure 4-13 r-squared Input type comparison summary – Hauling trucks.

4.5.1.3 ANN vs. summarization tools analysis

Figure 4-14 presents a comparison between the results obtained using ANN and Data mining versus the results obtained in chapter three using summarization tools and only the hauling distance as a productivity feature. In general the analysis using ANN reached better correlations, and there is no a tangible influence in which type of input is preferred (numerical or nominal).

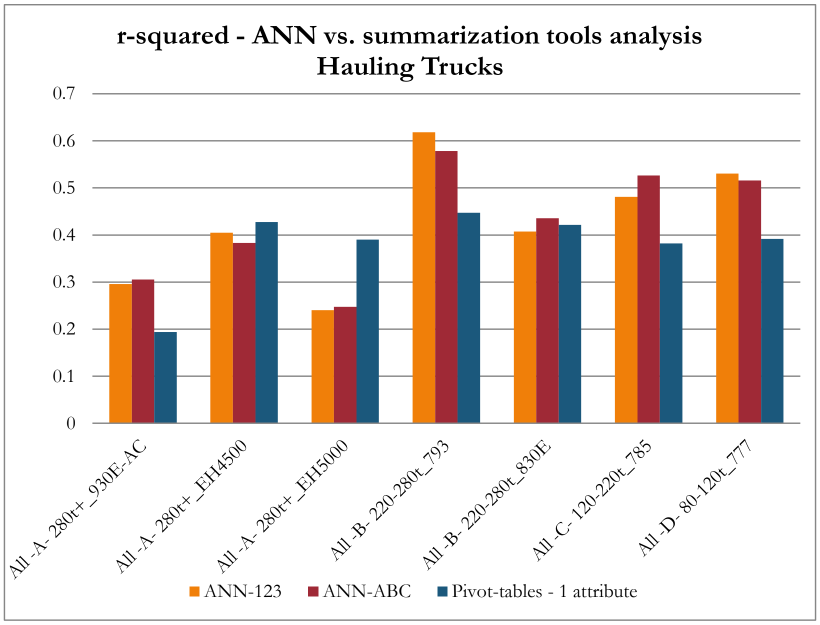


Figure 4-14 r-squared Input type comparison summary – Hauling trucks.

4.5.2 Excavator productivity

A similar analysis than the one produced for hauling trucks was made for the excavator models.

4.5.2.1 r-squared - ANN – different architectures

Figure 4-15 presents the comparison summary between the analyzed ANN architectures (Ward and Standard). The results clearly show very low correlation values, nevertheless, it can be noted that ward net architectures are associated with higher r-squared values than standard four layer architectures.

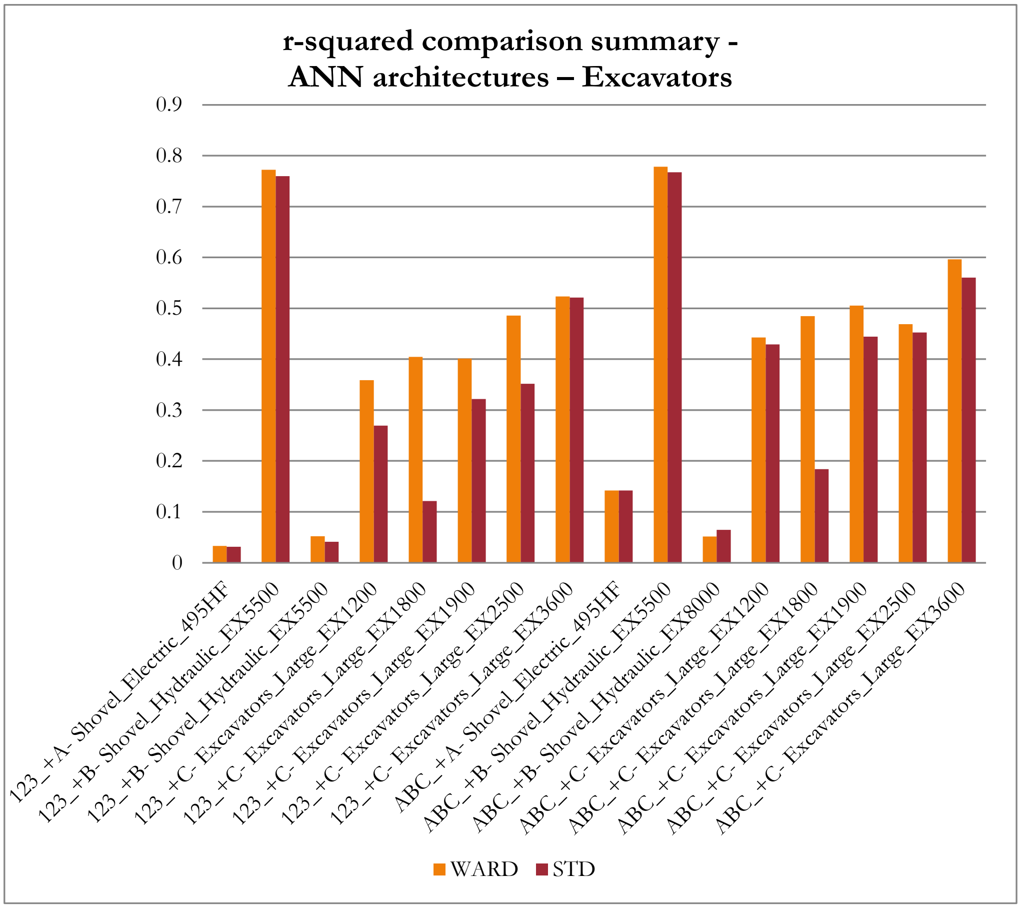


Figure 4-15 r-squared comparison summary - ANN architectures – Excavators.

4.5.2.2 r-squared - ANN – different input types

Figure 4-16 aims to summarize which type of input (nominal or numerical) result in better ANN performance when analyzing excavator productivities. The values in this table come from the Ward-A configuration. As it is shown, nominal inputs perform somewhat better than numerical ones.

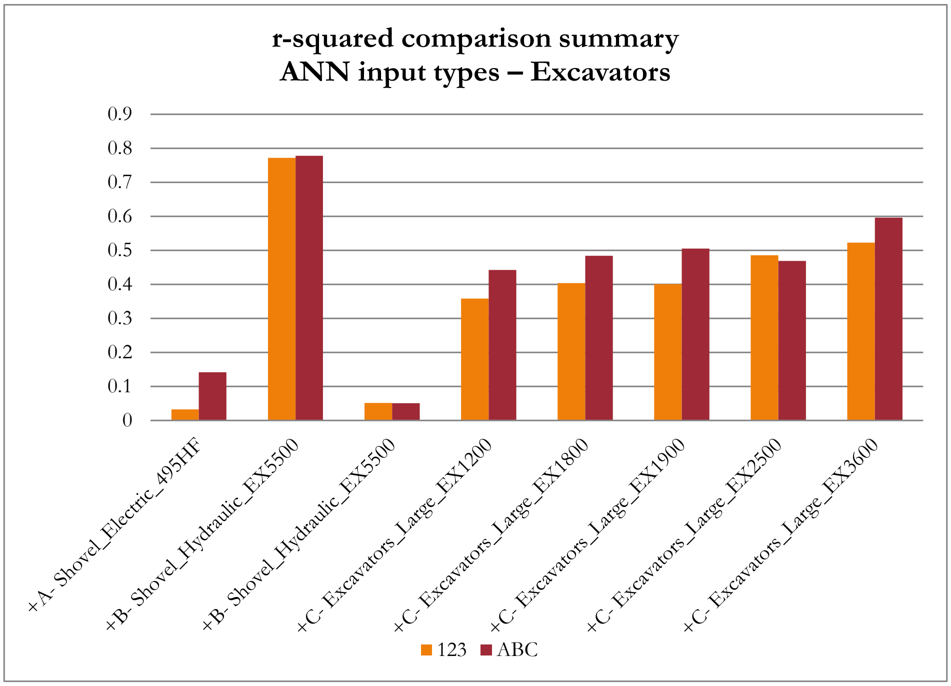


Figure 4-16 r-squared comparison summary – ANN input types – Excavators.

CHAPTER 5 CONCLUSIONS AND RECOMENDATIONS

This study involved the analysis of the historical operational data of more than 230 hauling units, 160 excavator units, and 150 units of support equipment. The data was recorded for more than three years and represents the operations of one of the largest contractors on the Alberta Oil Sands in eleven projects.

Multiple analysis tools were implemented throughout the analysis. The use of them was more complementary than competitive. Data mining, artificial neural networks and summarization tools proved to assist effectively in the assessment of historical productivities and in the identification of the attributes that most influence the results.

Most of the information related to earthmoving operations is recorded by the foreman at the end of the shift. The information is then input into a data management system. This schema contributes to fabricate information, e.g. durations of individual task within a shift. Consolidating the shift operations information into a single datum could improve the accuracy of the results as it was shown in this study with the use of the shift weighted hauling distance.

This research has described the analysis of historical information regarding the operations of hauling and loading equipment as well as support equipment. Productivity curves, equations and ranges were built and presented as one of the outputs of the analysis. The research also proposed a methodology to represent performance summaries of equipment units involved in the earthmoving operations; those summaries offer a valuable overall picture to executive staff and project managers. A novel approach was presented on how to compute historical ratios between primary and support equipment compared to accomplished production, the results could certainly assist in project resource allocation.

The research involved a detailed review of the data contained in a large DBMS of a major heavy construction contractor. The quality of the recorded information is affected by multiple sources of errors and inconsistencies. It is paramount to establish standard forms and better procedures for data collection. Only valuable information should be recorded and used. Multiple recording should be avoided by unifying input platforms.

Results of the poor data quality are the very low correlations that were obtained in the analysis. Through the use of data mining techniques and artificial neural networks it was possible to include more variables into the analysis, leading into slightly better results.

Multiple ANN configurations were evaluated in the determination of hauling trucks and excavators productivities. Ward net architectures that include different activation functions applied to hidden layer slabs performed better than standard backpropagation nets because they are able to detect different features in a pattern processed through a network. There was not a strong effect on the way in which the number of hidden neurons was computed. Nominal input type reached slightly better correlations than numerical type when evaluating the productivity of excavators; but still the obtained correlations were very low.

REFERENCES

AbouRizk, Simaan M., and Daniel W Halpin. "Statistical Properties of Construction Duration Data." *Journal of Construction Engineering and Management* 118, no. 3 (1992): 525-544.

AbouRizk, Simaan M., and Dany Hajjar. "A Framework For Applying Simulation." *Canadian Journal of Civil Engineering* 25 (1998): 604-617.

AbouRizk, Simaan M., Daniel W. Halpin, and James R. Wilson. "Visual Interactive Fitting of Beta Distributions." *Journal of Construction Engineering and Management* 117, no. 4 (1991): 589-605.

AbouRizk, Simaan M., Daniel W. Halpin, and Stephen L. Hill. "Measuring Productivity and Validating Microcomputer." *Microcomputers in Civil Engineering* 6 (1991): 205-215.

Boussabaine, A. Halim. "The use of artificial neural networks in construction management: a review." *Construction Management and Economics* 14 (1996): 427-436.

Boussabaine, A. Halim. "The use of artificial neural networks in construction management: a review." *Construction Management and Economics* 14 (1996): 427-436.

Chakrabarti, Soumen, et al. *Data Mining Know It All.* Burlington, MA: Morgan Kaufmann Publishers, 2009.

Chao, Li-Chung, and Miroslaw J. Skibniewski. "Estimating Construction Productivity Neural-Network-Based Approach." *Journal of Computing Engineering, ASCE* 8, no. 2 (1994): 234-251.

El-Moslmani, Khalil. *Fleet selection for earthmoving operations using queuing method.* MSc Thesis, Montreal: Concordia University, 2002.

Graupe, Daniel. *Principles of Artificial Neural Networks.* 2nd Edition. Singapore: World Scientific, 2007.

Hajjar, Dany, and Simaan AbouRizk. "Building a special purposes simulation tool for earthmoving operations." *Winter simulation conference.* P. A. Farrington, H. B. Nembhard, D. T. Sturrock, and G. W. Evans, eds., 1996. 1313 - 1320.

—. "Simphony: an environment for building special purpose Construction simulation tools." *Proceedings of the 1999 Winter Simulation Conference.* P. A. Farrington, H. B. Nembhard, D. T. Sturrock, and G. W. Evans, eds., 1999. 998-1006.

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. "The WEKA Data Mining Software: An Update." *SIGKDD Explorations* 11, no. 1 (2009): 10-18.

Kannan, Govindan, Julio C. Martinez, and Michael C. Vorster. "A framework for incorporating dynamic strategies in earth-moving simulation." *Proceedings of the 1997 Winter Simulation Conference.* S. Andradóttir, K. J. Healy, D. H. Withers, and B. L. Nelson, 1997. 1119-1126.

Mather, Kevin. *A CAD-based simulation modeling methodology for construction.* MSc Thesis, Edmonton, AB: University of Alberta, 1998.

NeuroShell 2 Help. n.d. http://www.wardsystems.com/manuals/neuroshell2/ (accessed January 08, 2011).

Peurifoy, Robert L., Clifford J. Schexnayder, and Aviad Shapira. *Construction Planning, Equipment, and Methods.* 7th Edition. Boston: McGraw Hill, 2006.

Smith, S. D., J. R. Osborne, and M. C. Forde. "Analysis of Earth-Moving System Using Discrete-Event Simulation." *Journal of Construction Engineering and Management* 121, no. 4 (1995): 388-396.

Smith, Simon. "Earthmoving productivity estimation using linear regression techniques." *Journal of Construction Engineering and Management* 125, no. 3 (1999): 133-141.

APENDIX A – ANN COMPARISON RESULTS USING R-SQUARED

The coefficient of multiple determination, **R Squared,** is a statistical indicator usually applied to multiple regression analysis. It compares the accuracy of the model to the accuracy of a trivial benchmark model wherein the prediction is just the mean of all of the samples. A perfect fit would result in an R squared value of 1, a very good fit near 1, and a very poor fit less than 0. If the ANN predictions are worse than what could be predicted by just using the mean of the sample case outputs, the R squared value will be less than 0. Equation A-1 is used for the coefficient of multiple determination R squared.

$$R^2 = \frac{SSE}{SS_{YY}} \tag{A-1}$$

where,

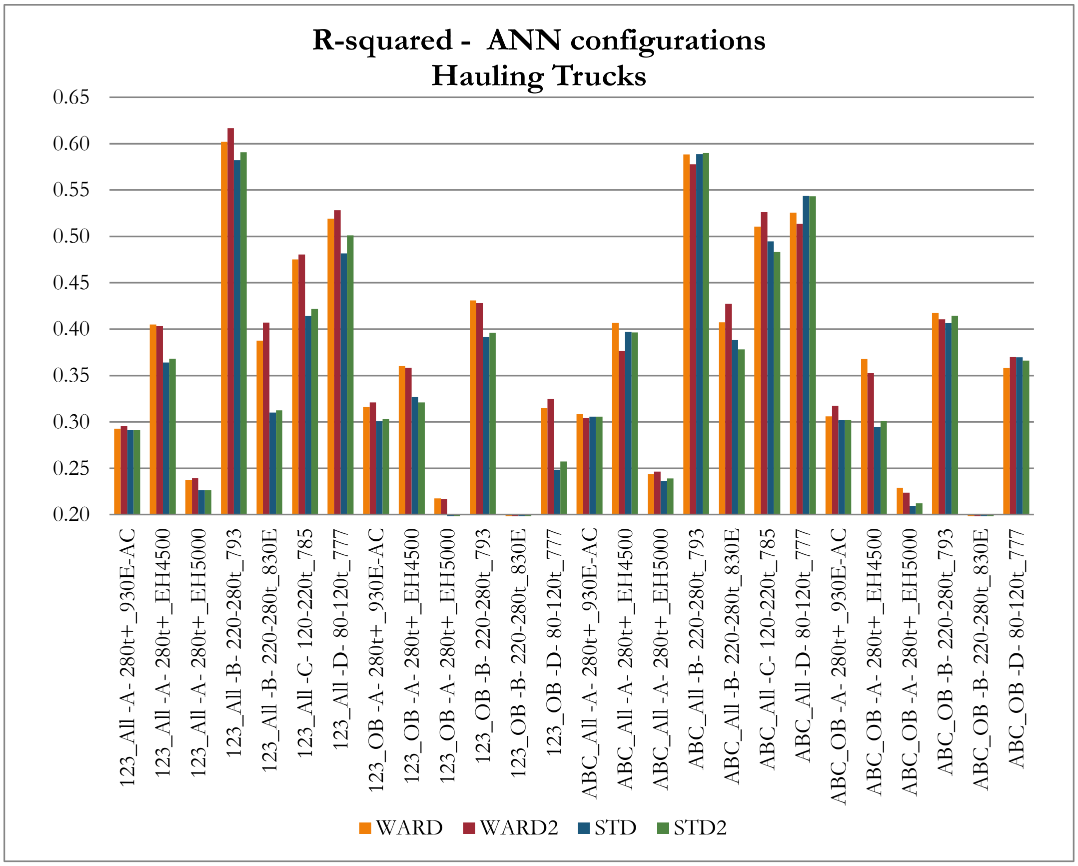
$$SSE = \sum (y - \hat{y})^2 \tag{A-2}$$

$$SS_{YY} = \sum (y - \bar{y})^2 \tag{A-3}$$

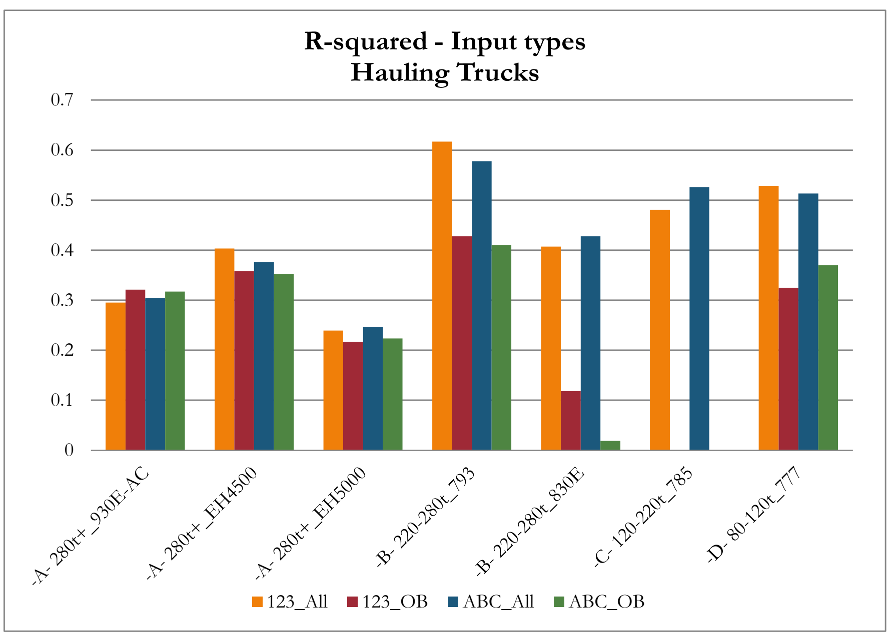
Where y is the actual value, $\hat{y}$ is the predicted value of y, and $\bar{y}$ is the mean of the y values.

Hauling truck productivities

R-squared - ANN – different architectures

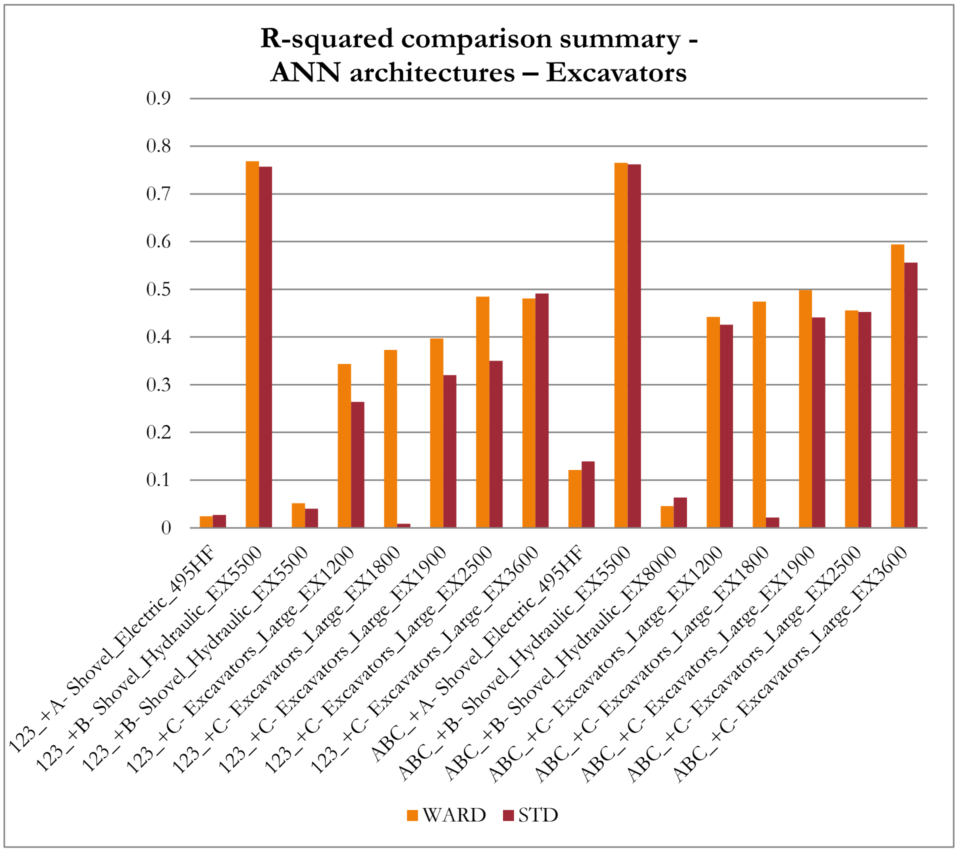


R-squared - ANN – different input types



Excavator productivity

R-squared - ANN – different architectures



R-squared - ANN – different input types