University of Alberta

APPLYING CORRELATION ANALYSIS TO SHOVEL CABLE LIFESPAN RESEARCH

by

Maidong Hu ©

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

Master of Science

Department of Computing Science

Edmonton, Alberta
Spring 2008

Library and
Archives Canada

Published Heritage
Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et
Archives Canada

Direction du
Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-45824-2
Our file Notre référence
ISBN: 978-0-494-45824-2

Abstract

This thesis addresses the research of shovel cable lifespan. This is motivated by the unexpectedly shortened lifespan of shovel cables in Syncrude, where an average of lifespan is only $1,000$ hours against the expected $2,000$ hours. Power-related features are used to analyze the operation patterns which may be harmful to the cables and reduce their lifespan greatly. Correlation analysis is applied to time series of the power-related features and some value segments of the feature distributions are identified as negatively correlated to the cable lifespan. We found some material and some teams of operators have more negative influence on the cable lifespan than other ones. The conclusions may help optimize the working schedule of shovels and the training of operators. In the field of data mining, the procedures used in our work can scan time series for non-sequential patterns which are correlated to an external event.

Acknowledgements

Many thanks to my supervisors Dr. Hong Zhang and Dr. Joerg Sander for their time, encouragement and support throughout this research study. Special thanks to Dr. Ian Parsons, from Syncrude Canada Ltd., at Syncrude Research, Edmonton, Alberta, for his time and contribution to this work.

Table of Contents

List of Figures

Chapter 1

Introduction

1.1 Problem Definition

The truck and shovel mining method is used for both overburden and ore removal at two mines of Syncrude. Some of the largest available mining shovels such as P&H 4100 series electric cable shovels are used. A shovel cable is an important component in the operation of a shovel, and its lifespan is rated for $2,000$ hours. However most cables used in Syncrude mines are not meeting this expectation now. Different factors, such as operators, geology, and weather, may have negative effects on the cables singly or simultaneously; therefore the effectiveness of cable usage with these factors must be optimized. To attain this optimization, the relationship between these factors and the cable lifespan should be understood.

The cable lifespan now in Syncrude ranges from 400 to $1,600$ hours with few cables approaching the $2,000$ hours mark, and the average lifespan is only about 1,000 hours. A cable replacement costs about $\$25,000$; therefore, extending the average cable lifespan would lead to economic benefit: 50% longer lifespan means about $1/3$ fewer cables being used. Moreover, fewer cables used would improve the shovel's working efficiency through fewer unplanned shutdowns due to early cable failure.

The telemetry data from hoist, crowd and swing motors are collected for eight different shovels. The operations of three motors are monitored by voltage and current sensors. The data set we used includes two years of telemetry data for the eight shovels. Given that data comprising nine parameters are collected at one-second time intervals, the quantity of data handled is huge. The corresponding dispatch data are also collected for the eight shovels. The dispatch data are useful to identify the information related to shovel operations, including working status, teams of operators, and quantities of materials dug in a certain working period.

Using telemetry data, we can obtain the time series of values for powers issued by crowd, hoist, and swing motors, which apply energy directly to the cables. Previous research pointed out that the three types of power or energy are key performance indicators, which can be used to analyze the influences of operators, geology, and weather [27]. We will use the power time series to find how these different types of power or energy affect the cable lifespan, and then attempt to distinguish the different effects on the cable lifespan between operators or materials. Therefore, what we are concerned most about in our work is what patterns of energy application to the cables are most harmful to the cable.

1.2 Initial Research on Prediction Model

No literature on previous research on shovel cable lifespans is found; however, heterogeneous system failure has been widely studied in material science. Unfortunately, no reasonable explanation has been provided to describe the underlying physical mechanism of heterogeneous system failure so far. Therefore some researchers believe that heterogeneous system failure is a random process which can not be predicted accurately using only a few simple features. Our early studies were focused on a prediction model with several input parameters related to energy, material, and operator practices. We applied some of the recognized machine learning algorithms using these features. The testing results, however, show that almost no cable failure can be predicted. From these results and the research in material science, we believe that a simple prediction model may not be feasible for our problem.

The probability distribution of an event is a possible descriptive method to predict on the condition that such an event occurs randomly. J. Anderson and D. Sornette proposed models of conditional probability prediction of earthquakes and fibre bundle breaking[1]. In our work, lifespan history data include only 77 occurrences of cable failure, therefore, may not be large enough to construct the probability distribution, let alone the probability distribution conditioning on certain temporal or space features. However, the analysis of whether or how the cable lifespan is conditioned on cable early working status is feasible and useful to our problem.

To simplify the problem, we built two models with some features of energy or materials. The first model was constructed with the following procedure. First, the 77 cables were separated into two groups: *long* group and *short* group. The long group included the cables whose lifespans are longer than 800 hours, while the short group includes the others. The ratio of sizes of the long group to the short group is $38 : 39$, thus, the probability that a cable belongs to the long group or the short group is around 0.5. Second, we considered the cables' first 100 hours of working time as the initial conditions since no cable's lifespan is shorter than 100 hours. For the first 100 hours, some features were extracted to represent the initial conditions. The features included the energy of crowd, hoist and swing applied to the cables, the sum of the product of crowd and hoist power, the sum of the product of crowd and swing power, the sum of the product of hoist and swing power, the sum of the product of crowd and hoist and swing power, and the dug quantity of two different materials: overburden and oil sand.

In the first model, we tested if the features for the first 100 hours of working time can decide whether cables would last longer ($\geq$ 800 hours) or not ($<$ 800 hours). The testing results show that the cables' lifespans depend heavily on their initial working status. In a model built with SVM, almost 73% (cross verification) of 77 cables are correctly classified into the long or short group with all features for their first 100 hours of working times. Further study shows that using only the sum of the product of hoist power and crowd power in the first 100 hours can classify over 57% (cross verification) cables correctly, the highest compared to the results using other single features. Therefore, in this model, this feature are very important to classify cables.

The second simplified model resulted from an observation that the cable lifespan distributions are different for each shovel, and that some shovels have an average cable lifespan of about 1000 hours, longer than the other shovels with an average of only about 600 hours. In this model, the cables were separated into two groups: *long* group and *short* group. The long group includes cables which belong to the shovels whose cable lifespans are longer (with an average of 1000 hours), while the short group includes the other cables which belong to the shovels whose cable life spans are shorter (with an average of 600 hours).

In the second model, we tested if the features for the first 100 hours of working time can decide whether cables would be long with an average lifespan of about 1000 hours or be *short* with an average lifespan of about 600 hours. The testing results show that the amount of different materials dug in the first 100 hours of working time may affect the cable lifespan. If using the rule: the cables with over $180,000$ tonnes of overburden dug in the first 100 hours would be long, 91% (cross verification) of cables are classified correctly. If using the rule: the cables with over $6,000$ tonnes oilsand dug in the first 100 hours would be short, 93% (cross verification) of cables are classified correctly.

For the present, there is no supporting theory to explain the testing results of the above two simplified models. Due to the small set of sampling cables, testing of more cables may be needed later.

1.3 Mining of PDF Patterns

Our initial research mentioned above has shown that the prediction of cable failures accurately using some features is a difficult task for the present. However, we have noticed the fact that time series for each of three types of power values, or for any product of two or three of these power values, can be formed from the raw telemetry data, and that it is possible to search in these time series for patterns which may be associated with reduced cable lifespan. The key is how to find such patterns using a proper method.

Due to the huge samples of raw telemetry data, using statistical methods, the PDFs (Probability Density Function) of hoist power, crowd power, swing power, product of hoist and crowd power, product of crowd and swing power, product of hoist and swing power, and product of hoist and crowd and swing power can be estimated for a certain working time period, for example, for a shovel's working time, for a cable's working time, or for a team's working time. For each shovel, there are four teams that take turns to operate the shovel in 12-hour shifts. All four teams work alternatively during a cable lifespan if the cable works long enough, although their working times are different. The statistical testings show that for each of the above seven features, its distribution for each shovel differs among the different materials (oilsand and overburden), and among four teams who jointly operate the same shovel.

To resolve the problem in our work, a systematic method should be designed to locate in each of the above PDFs the related segments which are the most harmful to the cable if any. Based on these related segments, we would attempt to answer the question: whether the different materials or different teams have different harmful effects on the cable?

The method can be done with the following steps:

1. For each feature, estimate its PDFs for all 77 cables in the same range of the feature value. Separate the range into a number of, say 100, segments and get the estimated probabilities for these segments.

2. For each feature, test some correlation coefficient between each segment's estimated probabilities of all cables and the cable lifespans. Get a distribution curve of the correlation coefficient for each segment. Find the segments with the largest negative correlation coefficients.

3. For each segment obtained in the previous step, divide the values in it according to materials or teams. Identify the material or the teams which contribute most harmful effects on the cables.

According to the results from the described method, we can propose some recommendations to the operators: which range of these feature values are harmful to the cables and should be avoided in operation as much as possible, or provide some suggestions to the management of the optimal working schedule of a shovel.

With respect to the method, this thesis will focus on the following challenges:

- How to estimate efficiently PDFs for all ranges or each segment?

- How to test the correlation coefficients or how to find the related segments if any?

- How to identify the material or teams which are most harmful to the cables?

1.4 Thesis Contribution and Outline

The contribution of this dissertation is a framework to search in time series data for patterns which are correlated with an external event, by testing with segmental distribution of the time series. The use of PDF of time series data allows the correlation mining phase to overcome many preprocessing problems such as cleaning noises, fixing missing data in a sample of raw data, if the sample is good enough to form a proper PDF. The use of PDF also allows to ignore the actual sequence of data. Thus, the thesis underlying this dissertation is that some specific patterns in time series can be mined by testing correlation of time series with an external event using a proper segmentation of PDF of the time series.

The implementation of the framework was tested for the research on shovel cable lifespan. The results from the experiments provided us with some applicable conclusions about our problem raised in the first section of this chapter. These conclusions include:

- Some segments of the power-related feature values are really harmful to the cables, making the cable lifespan shortened greatly.
- Different materials have different influences on the cables. Oilsand may more be harmful than overburden on average.
- Different operators have different influences on the cables. Some teams of operators have more harmful operation than other teams on average.

Chapter 1 has introduced the problem of the unexpected shortened lifespan of oil shovel cables, along with our early research, challenges, and a high-level description of the contribution of this thesis.

- Chapter 2 will survey previous related research in material science and data mining.
- Chapter 3 will present the correlation mining framework, highlight the purpose of each step in the process, describe in detail the implementation of this framework in our work.
- Chapter 4 will presents experimental results with respect to our problem of cable lifespan.
- Chapter 5 will present a summary of this work, and discuss potential future directions of research.

Chapter 2

Related Work

Unfortunately, there is no published literature addressing the problem of oil-shovel cable lifespan. However, the researches in material science and data mining provide many practical models and methods to predict heterogeneous system failures and to find patterns in time series data. Although this section will not cover in detail all of the approaches proposed, it will provide a survey of many of the proposed approaches for probabilistic prediction and special pattern mining.

The remainder of this chapter will be divided into three sections. The first one we examine is the recent methods in material science on probabilistic prediction models that incorporate spatio-temporal information of the damage. The next two are pattern mining studies on time series data: correlation mining and anomaly detection.

2.1 Research on System Failure in Material Science

2.1.1 Difficulty in Prediction of Failure

Heterogeneous system failures are widely studied in material science due to their economic and human cost. A wide range of phenomena may be considered heterogeneous system failure, like, cracking of glass, breaking of fibre bundles, collapse of concrete structure, earthquake, etc. Many models or strategies have been proposed to predict the remaining lifetime of a complex structure or the precise time of failure; however, the physical mechanisms remain unclear [32][8][15].

J. Anderson and D. Sornette argue in their papers [1][34] that due to no comprehensive understanding of the underlying mechanism, all the current prediction models have limitations and are restricted to relatively simple situations, and these models are not reliable in most situations. The reasons leading to the situation may lie in the complex interaction between various elements of the system under consideration and modes of damage.

To bypass the difficulty, Anderson and Sornette formulate in their papers a new general probabilistic prediction scheme. The hidden assumption in the procedure is: failure of a system is highly history and sample-dependent. This assumption means that the system failure is a random and gradual process, conditional on its damage history. By incorporating new on-going damage information into probabilistic models, different levels of probability distributions conditional on the current damage modes and spatial-temporal information can be constructed. Thus, the prediction problem can be converted into a problem of calculating probability distribution using posterior probability models in statistics.

Section 2.1.1 and Section 2.1.2 will discuss several models proposed by Anderson and Sornette to solve different kinds of system failures: predictions of percolation threshold and a hierarchical fibre rupture model with time dependence.

2.1.2 Predictions of Percolation Threshold

Roux *et al.* [28] have proved that, in heterogeneous systems, rupture processes are equivalent to the correlated site percolation problems. Percolation theory originates from filtering fluids through porous material from one side to the opposite one. In mathematics, it is typical to model the percolation problems as a three-dimensional network of $D \times D \times D$ points (or vertices), in which the connection (edge) between two neighbors is open with a probability of p. When $D \rightarrow \infty$, there is a critical p (p_c) below which the probability that an open path exists from the top to the bottom is always 0 and above which the probability is always 1 [9]. A site percolation is a percolation that occurs on a lattice of $L \times L$ sites where each site would be filled or not. If p denotes the fraction of filled sites, any percolation realization C_L will be related to some percolation threshold $p_c(C_L)$. For any given L, $p_c(C_L)$ has a random distribution decided by a probability density function $P(p_c)$. To study prediction methods for earthquakes, Anderson and Sornette used a two-dimension site percolation model as the base on which a rupture or an earthquake can be predicted [1]. In their model, p is also the running time of a lattice which is damaged by one site per unit time. Therefore, $p_c(C_L)$ is the time when a rupture happens, and $P(p_c)$ decides the distribution of the critical time. The purpose of the model is to find the $P(p_c)$ in a lattice with a given L.

Anderson's model is a hierarchical one composed of levels of prediction based on the increasing available information and implements the first three levels. The whole hierarchical structure is described as follows:

- **Level 1**: The first prediction level is the probability distribution function (PDF) $P_L(p_c)$ of the percolation thresholds. For a given L, the distribution can be attained by statistics. This level is actually the unconditioned probability distribution of the percolation threshold.

- **Level 2**: From this level up, each prediction level $P_L(p_c)$ is conditional on the spatiotemporal knowledge of damaged sites. The second level is the PDF based on the fact the lattice has a fraction p of damaged sites and is still not percolating. Using this information improves the first level.

- **Level 3**: The third prediction level incorporates additional information about the spatial localization of damaged sites. Anderson and Sornette implement it using two similar measurements of the localization. One is $p_\xi(p)$, the fraction of sites belonging to the largest cluster, and the other is $\xi(p)$ the largest of the linear size projected on the x and y axes of the largest cluster. The calculated conditional PDFs match the human intuitive prediction on percolation.

- **Other Levels**: Other information on the damaged clusters' shape, position, etc can be used as additional conditions of the PDF. The last level should include all the available knowledge of the damaged sites, and the PDF calculated at this level should have the best prediction accuracy.

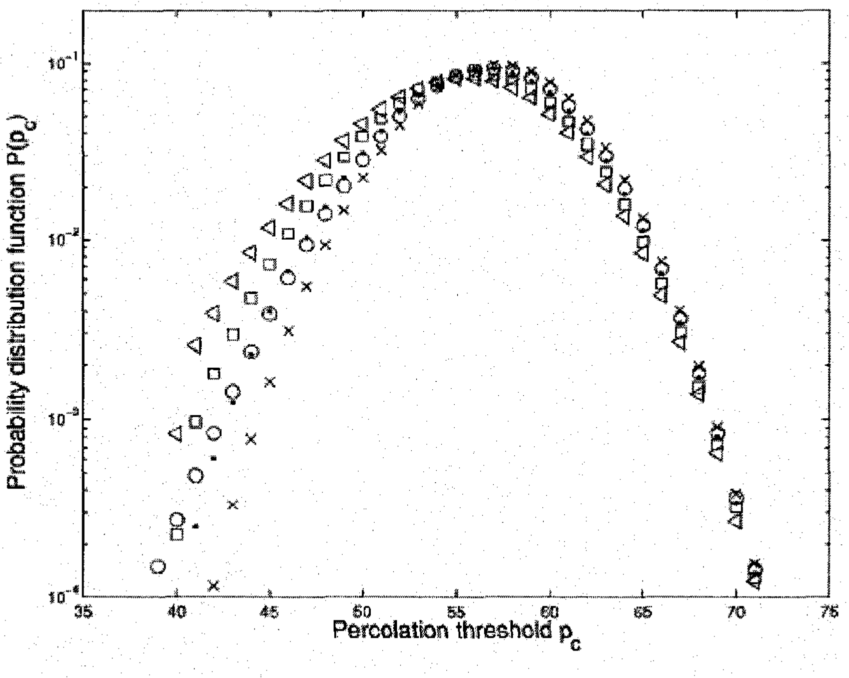


Figure 2.1: An example of different PDFs of percolation threshold.
PDF $P_L(p_c|p,\xi)$ is a function of percolation threshold p_c (in percent) conditioned on both p and ξ, where ξ is the largest of the linear size projected on the x and y axes of the largest cluster within the system, for the fixed $p = 40\%$ and different values of ξ: $\xi/L = 0.04$ (cross), $\xi/L = 0.06$ (dots), $\xi/L = 0.08$ (squares), and $\xi/L = 0.1$ (triangles). The open circles represent the unconditional PDF $P_L(p_c)$ for reference [1].

Figure 2.1 shows an example of different probability distribution functions P_L of percolation threshold p_c. In the figure, there are four different probability distribution functions. One is an unconditional PDF $P_L(p_c)$ (open circles). The other three are PDFs $P_L(p_c|p,\xi)$ conditioned on both the fraction of filled sites and the largest of the linear size projected on the x and y axes of the largest cluster, for a fixed $p = 40\%$ and different ξ: $0.04L$ (cross), $0.06L$ (dots), $0.08L$ (squares), and $0.1L$ (triangles). For a fixed p, the conditional PDFs shift and widen gradually to smaller values of p_c for increasing ξ. This means that if one observes in two different systems for different mass of the largest cluster with the same p, the system with the larger mass of the largest cluster is more likely to percolate at a earlier time.

The described structure above is simple and feasible; however, we should know how much the additional information upgrades the quality of prediction and how well the model predicts. Anderson and Sornette use the information gain to measure the improvement of forecasts from the first level to the third level. From the distribution of the information gain of p for various p_ξ and ξ, they find that the distribution is not uniform for p, and can be divided into three regions according to the values of p. Some information gain is obtained when increasing p_ξ and ξ in the region of the small p, while is limited for the intermediate values of p. For the larger values of p, the information gain increases drastically even with the small values of p_ξ and ξ. In other words, the information gain is most useful for large values of p and helps a little for small p, but is almost no use for intermediate p.

2.1.3 Hierarchical Fibre Rupture Model

Using the general principles of the above model, Anderson and Sornette propose another hierarchical model to predict fibre ruptures [1][34]. They give a definition of a hierarchical model (studied by earlier researchers [20][21][29]) made of interactive elements with multi-levels, in which each element is loaded with a stable stress σ. All single elements form the first level. These elements are associated to each other into pairs of elements that form the second level. Then, the pairs of pairs form the third level and so on. In such a topology, the dynamics of fibre rupture proceed in the following way. The whole rupture process always begins from the first level; one element will fail at a random time. When an element fails, its load σ is transferred to the remaining element in the pair so that its load is doubled. When the whole pair fails, the surviving pair associated to it will be double loaded. When the whole pair at this level fails, its load is transferred to the pair at the next hierarchical level. The time of a single element failure is still a probability decided by a cumulative distribution function:

$$P_0(t) \equiv \int_0^t p_0(t')\,dt' = 1 - \exp\left\{-\kappa \int_0^t [\sigma(t')]^\rho\,dt'\right\}. \tag{2.1}$$

Here, $\sigma(t')$ is the given stress history, and $\rho > 0$ is the stress amplification exponent.

In a system of four elements, if t_i is the time when the fibre i is broken ($i = 1, 2, 3, 4$), the pairs are (1,2) and (3,4), and $t_1 < t_2$, then

$$t_{(1,2)} = t_1 + \alpha(t_2 - t_1) < t_2, \qquad \alpha = 2^{-\rho}. \tag{2.2}$$

Thus the hierarchy model can be described as follows:

- **Level 1**: Actual prediction in the absence of revealed damage of any of the four elements. Since all we know is PDFs of t_i, the PDF of $t_{(1,2,3,4)}$ (denoted by $P_{(1,2,3,4),t}(t_{(1,2,3,4)})$) can be obtained from the equation (2.1):

$$\frac{1}{\alpha}\int_t^{t_{(1,2,3,4)}} dt_{(1,2)}\,\tilde{P}_{(1,2),t}(t_{(1,2)})\tilde{P}_{(3,4),t_{(1,2)}}\left(\left[t_{(1,2,3,4)} - (1-\alpha)t_{(1,2)}\right]/\alpha\right) + ((1,2) \leftrightarrow (3,4)), \tag{2.3}$$

where $\tilde{P}_{(m,n),t}(t_{(m,n)}) = \frac{P_{(m.n)}(t_{(m.n)})}{\int_t^\infty P_{(m,n)}(x)\,dx}$ is the PDF of the pair (1,2) or (3,4), conditional on that it has not yet been broken at t. In formula (2.3), the first contribution is corresponding to $t_{(1,2)} < t_{(3,4)}$, while the second one $((1,2) \leftrightarrow (3,4))$ is corresponding to $t_{(1,2)} > t_{(3,4)}$. In the first contribution, $\tilde{P}_{(1,2),t}(t_{(1,2)})\tilde{P}_{(3,4),t_{(1,2)}}\left(\left[t_{(1,2,3,4)} - (1-\alpha)t_{(1,2)}\right]/\alpha\right)$ is the integrant. $((1,2) \leftrightarrow (3,4))$ denotes an integral formula similar to one in the first contribution, in which the positions of $(1,2)$ and $(3,4)$ are interchanged.

- **Level 2**: This level of prediction is conditional on the revealed damage information of the elements. The PDF of the system will vary with the number of the broken fibres. The whole failure process can be described as this procedure: one element fails, then two elements in the same or each pair are broken, and finally three elements are broken. For each scenario of two broken elements, there are different formulas of PDF conditional on three broken elements. The formula corresponding to the two broken elements in the same pair first is:

$$P_{(1,2,3,4),t^*,t^\dagger}\,(t^{*,\dagger}_{(1,2,3,4)}) = \frac{\tilde{P}_{2,t^*,t^\dagger}\left(\frac{t^{*,\dagger}_{(1,2,3,4)}}{\alpha^2} - \frac{1-\alpha}{\alpha^2}t^\dagger - \frac{1-\alpha}{\alpha}t^*\right)}{\int_{t^\dagger}^\infty dx\,\tilde{P}_{2,t^*,t^\dagger}\,\left(\frac{x}{\alpha^2} - \frac{1-\alpha}{\alpha^2}t^\dagger - \frac{1-\alpha}{\alpha}t^*\right)}, \quad \text{for } t \geq t^\dagger, \quad (2.4)$$

where $t_1 = t^*, t_{(3,4)} = t^\dagger$ and $t^* < t^\dagger < t_2$ The formula corresponding to the two broken elements in each pair is:

$$P_{(1,2,3,4),t^*,t^\dagger}\,(t^{*,\dagger}_{(1,2,3,4)}) = \frac{P_{(3,4),t^\dagger}\left(\left[t^{*,\dagger}_{(1,2,3,4)} - (1-\alpha)t^\dagger\right]/\alpha\right)}{\int_{t^\dagger}^\infty dx\,P_{(3,4),t^\dagger}([x-(1-\alpha)t^\dagger]/\alpha)}, \quad \text{for } t \geq t^\dagger. \quad (2.5)$$

where $t_1 = t^*, t_2 = t^\dagger$ and $t^* < t^\dagger < t_{(3,4)}$

Up to the second level, the PDF of system failure time can be given by a formula; and the formula can be adapted to any system with 2^n fibres. When the model is applied in systems with millions of elements, an *element* is often a bundle of fibres or *coarse-grained* [34]. There are some optimized algorithms to calculate PDFs with thousands of elements.

- **Level 3**: Anderson and Sornette tested the third level prediction conditional on the space-time of the largest crack in a system with $2^8 = 256$. In this testing, all available damage information is only about the failure time of the largest crack with size of 2^m and the broken fibres within this crack. The testing results show an improvement of the prediction for the global failure, conditional on the partial knowledge [1].

Figure 2.2 shows an example of performance improvement with information of the largest crack for a given realization. The thick curve is the unconditional cumulative number of broken fibers, while the thin curve shows, as a function of time t, the cumulative number of broken fibers, which broke either within the largest crack or within its complement in their pair within the hierarchy. One can observe that with the conditioning on the largest crack and its complement, a linear increase is

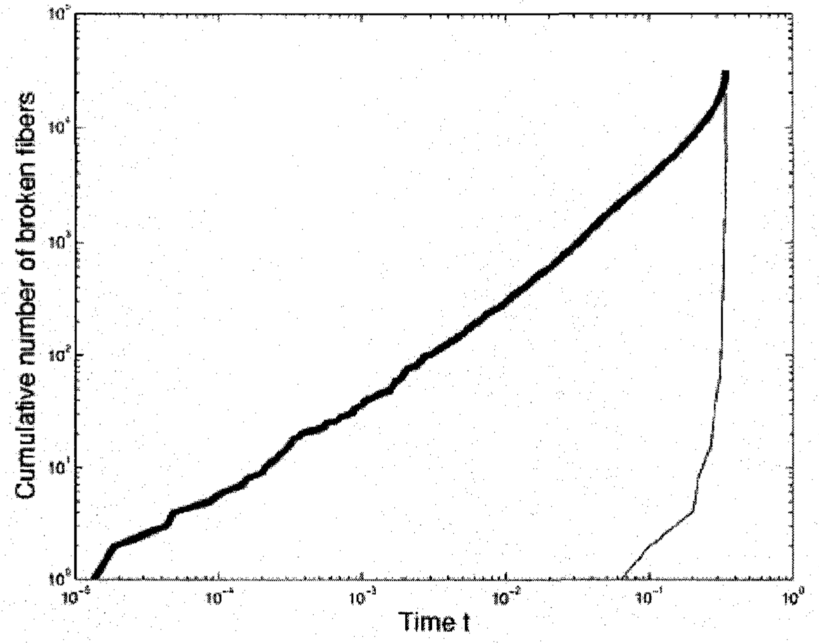


Figure 2.2: One measure of the cumulative number of broken fibers as a function of time t for a given realization.
The thick curve shows the unconditional cumulative number of broken fibers. The thin curve shows the (conditional) cumulative number of broken fibers, which broke either within the largest crack identified up to time t or within its complement in their pair within the hierarchy [1].

accelerated very early while the absence of the condition makes a linear increase occur only at the very end.

Anderson and Sornette opened a new way to predict the heterogeneous system failure based on the updated knowledge of damage history. Their models show the gains of the prediction from the information of the space and time organization of damage. Applying their idea can solve many other rupture problems. However for some rupture problems, in which only limited samples can be tested or there is less or no access to information of damage history, the conditional probabilistic model is difficult to build, and even the unconditional PDFs cannot be obtained in some scenarios.

2.2 Correlation Mining

Performing statistical data correlation analysis is not a new topic in data mining. Some association patterns can be found by search for correlated items or attribute values [2][22][16]; the correlation analysis can also be used to classify and cluster; the data can be aggregated by evaluating the correlation between their attributes; and the dimensions can be reduced in the case of high dimensional data using correlation analysis [10]. In time series data, the correlation relationship may exist between elements in a single time series or between two time series [37]. Most of the current methods are used to mine the correlation inside a data set, of different attributes, items, elements, or time series; however, in our work, our concern is how to find the correlation between the data set and the external events, i.e. the cable lifespan.

Many different coefficients are used to measure correlations. *Lift* (originally *interest*, introduced by [3]) is usually used to estimate the correlation of two itemsets. For two itemsets I_1 and I_2,

$$lift(I_1, I_2) = \frac{P(I_1 \cup I_2)}{P(I_1)P(I_2)},$$

where $P(I_1)$ is the probability of occurrence of I_1, $P(I_2)$ is the probability of occurrence of I_2, and $P(I_1 \cup I_2)$ is the probability of occurrence of both I_1 and I_2. This definition originates from the dependence of two events. If $P(I_1 \cup I_2) = P(I_1)P(I_2)$, I_1 and I_2 are independent to each other; otherwise, they are dependent or correlated to each other. If the value of lift is less than 1, then the occurrence of one itemset is negatively correlated with the occurrence of another one. If the value is greater than 1, then the two itemsets are positively correlated. If the value is equal to 1, then the two itemsets are independent and there is no correlation between them.

Another correlation measure is referred to as χ^2(chi-square) measure, which measures the correlation relationship between categorial data sets. In statistics, one use of the χ^2 test is to evaluate if the paired observations on two variables are independent of each other or not. For the two categorial data sets C_1 and C_2 with respectively c and r elements, a table named contingency table may be constructed in this way: the values in C_1 denote the different columns, while the values in C_2 denote the different rows. Each and every possible joint event of the occurrence of one value in C_1 and one value in C_2 has its own cell (or slot) in the table. The χ^2 is calculated as the following formula:

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}},$$

where o_{ij} is the observed frequency (i.e., actual count) of the joint event of the occurrence of the ith value in C_1 and the jth value in C_2, and e_{ij} is the expected frequency of this event, which can be calculated as

$$e_{ij} = \frac{count(\text{the ith value in } C_1) \times count(\text{the jth value in } C_2)}{N},$$

where N is the total number of joint events with respect to C_1 and C_2, $count$(the ith value in C_1) is the number of occurrence of the joint events having the ith value in C_1, and $count$(the jth value in C_2) is the number of occurrence of the joint events having the jth value in C_2. For a given significant level and degree of freedom, the hypothesis that C_1 and C_2 are independent will be rejected or not according to the statistic value and a corresponding table of the χ^2 distribution.

All-confidence [23] for an itemset I is defined as

$$all\_conf(I) = \frac{supp(I)}{max\_item\_supp(I)},$$

where $max\_item\_supp(I)$ is the maximum item support for all items in I. The *all-confidence* is the minimal confidence for all the rules: $i_1 \rightarrow i_2$ where $i_1, i_2 \in I$. Therefore, if a rule is called *all-confident*, it satisfies the minimal *all confidence* threshold.

Cosine [30]is a variant of lift, and its definition has two formulae. For two itemsets I_1 and I_2, their cosine is

$$cosine(I_1, I_2) = \frac{P(I_1 \cup I_2)}{\sqrt{P(I_1) \times P(I_2)}} = \frac{supp(I_1 \cup I_2)}{\sqrt{supp(I_1) \times supp(I_2)}}$$

.

To evaluate the correlation of two continuous data sets, Pearson product-moment correlation coefficient (or shortly, Pearson correlation coefficient) named for its developer Karl Pearson, is used most widely. This is, given two data sets M and N,

$$r_{MN} = \frac{\sum_{i=1}^{N}(m_i - \bar{M})(n_i - \bar{N})}{N\sigma_M\sigma_N} = \frac{\sum_{i=1}^{N}(m_i n_i) - N\bar{M}\bar{N}}{N\sigma_M\sigma_N},$$

where N is the total number of pairs (m_i, n_i), m_i and n_i are the ith values of M and N respectively, $\bar{M}$ and $\bar{N}$ are the means of M and N respectively, σ_M and σ_N are the standard deviations of M and N respectively, and $\sum_{i=1}^{N}(m_i n_i)$ is the sum of the MN cross-product. The value of r_{MN} ranges from -1 to $+1$. $r_{MN} = 0$ means that M and N are independent and there is no correlation between them. If $r_{MN} > 0$, then M and N are positively correlated, meaning that the values of M increase as the values of N increase. The higher the value of r_{MN}, the stronger the correlation. If $r_{MN} < 0$, then M and N are negatively correlated, where the values of M increase as the values of N decrease. Although Pearson correlation coefficient has the underlying assumption of normality of both M and N, it can be used for some unspecified distributions and may get satisfying results.

There are other correlation measures for continuous data in statistics, such as, *correlation ratio* [14], *mutual information/total correlation [7], Point biserial correlation [4], Spearman's ρ [17]* and *Kendall's τ [13]*. Some of them (*correlation ratio, mutual information/total correlation*) can be used to measure more general dependencies (including nonlinear); while some are non-parametric methods to measure for unknown distributions.

2.3 Anomaly Detection in Multiple Time Series

The aim of anomaly detection is to learn what the normal behavior is and determine when an event occurs that differs significantly from the normal behavior. A time series is always an underlying process in a system. The anomaly detection of a time series needs to learn the normal behavior using some "normal" series data input in advance; however, when used to detect anomalies, only segments of series data mixed with "normal" and "abnormal" behavior are available. Moreover, in the case of multiple time series, the anomaly detection will include analyzing the hidden multi-process which may contribute to "abnormal" events.

The applications of anomaly detection in time series are widespread nowadays. A majority of these applications is in the medical field. Using multiple stream data obtained by sensors or monitors, doctors should find "abnormal" conditions of humans so that some diseases may be found as soon as possible. Another major application is on network monitoring. Network administration needs anomaly detection to block "malicious" attacks. Anomaly detection is also used in the study of climate and geography conditions. The biggest advantage of application of anomaly detection in these fields is: it can not only detect the known anomaly, but also predict the possible and unknown anomaly.

Anomaly detection is the opposite of signature detection. Signature detection is explicitly given information on what is "bad", and simply attempts to detect it when it happens. Because anomaly detection in single time series is relatively simple, this section focuses on only literature on anomaly detection in multiple time series. First it provides an overview and comparison of the current models and their methods in anomaly detection in time series, and then a summary of its applications in medicine, network administration, climate and geology. At last, some limitations and possible directions of future research are pointed out.

2.3.1 Models

There are several definitions of "anomaly" [19]. Although these definitions reflected the diverse interests of the people, "anomaly" is generally believed as events or observations deviated from normal or usual things. Very few models have been developed specially to detect anomalies in multiple time series up to now. Some models for single time series may be extended to be used for multiple time series.

A. Box Modeling

Chan *et al.* have tried to address the problem of online detection of anomalous modes in a series of their papers [31][18][6]. Two efficient techniques are proposed: path modeling and box modeling, both of which allow online scoring. The predecessor is their earlier work named Gecko and RIPPER [31]. Gecko and RIPPER are applied to a single time series, and may be extended to handle multiple time series by aligning and averaging multiple series as single one; similarly, when handling multiple time series, path modeling needs extending by finding the nearest path of testing point or specifying whether the testing pointing lies "between" the training paths.

Box modeling is developed specially to generate a model directly from multiple time series. In this method, a model characterizing the behavior of multiple "normal" time series is produced; given a test time series and a model, anomaly scores are generated in a real-time manner; depending on whether the model is order-dependent, a stateful or stateless test is performed. In state testing, the anomaly score is zero if the current data point fits the current or next state/box. Otherwise, the anomaly score is the squared distance to the surface of the current or next box, whichever is closer. In stateless testing all boxes/states are compared. The anomaly score is zero if the data point is within a box. Otherwise, the closest box is located and the squared distance to the box surface is the anomaly score.

Chan *et al.* proposed three box-based algorithms to construct comprehensible models from multiple training time series. A "box" is a constraint with minimum and maximum values in a feature space. The model consists of such "boxes" along the axis in the feature space. Before a model for multiple series is produced, a model for each single series is constructed. Greedy-Split algorithm is used to generate a model for a single series data. That is, a sequence of n points is approximated by $(n-1)$ boxes, each including a pair of adjacent points. Then these points are merged by replacing a pair of adjacent points with a box that minimizes the increase in volume. This process will be ended when the specified k boxes are obtained, where k is user-specified.

Three algorithms to deal with multiple series data are: order-dependent, order-independent, and order-independent without all-series constraint.

The order-dependent algorithm is the basis of the other two. For each series, Greedy-Split is used to find k boxes, and then expand the boxes to include the closest points in other series. Because each series will be processed, the model produced by this algorithm is affected by the order of the set of time series.

The order-independent algorithm first approximate each series by boxes and merge the closest boxes. During merging, each merged box will include at least one box from each time series.

The order-independent algorithm without the all-series constraint is the same as the order-dependent algorithm except that the merging does not need to include boxes from all series.

The above three algorithms give us three possible models from multiple time series. The best one may be chosen according to the testing.

B. Process Models

Unlike the methods to produce a general model from multiple time series, the other way to detect anomaly in multiple time series is to try to find various reasons for anomalies, i.e., underlying or hidden multiple processes or variables [12][24][36].

Mixture Model [12]. In this model, it is assumed that any data set is generated by a mixture of finite underlying probability distributions. When applied to time series, a data sample at time t is generated from a mixture of K autogressive (AR) models. After the parameters of AR models are calculated using maximum likelihood methods, an augmented probability vector of data can be obtained. The vector is the combination of the probability of belonging to each AR model and not belonging to any AR model. If a sample does not belong to any model, it is assumed to be an anomaly. One advantage of mixture model is the augmented probability vector can be visualized in 3-D after using dimension reduction. In such a graph, normal samples that belong to the same model are clustered together, while an anomaly will be far away from these normal samples.

SPIRIT (Streaming Pattern dIscoveRy in multIple Time-series)[24]. It is a comprehensive approach to discover correlations that effectively and efficiently summarize large collections of stream. In SPIRIT proposed by Sun *et al.*, PCA(Principal Components Analysis) is used to capture correlations and find hidden variables corresponding to trends in the collection of streams, which are assumed to be co-evolving. The approach used in SPIRIT can dynamically detect changes (both gradual and sudden) in the input streams, and automatically determines the number of hidden variables.

When applied to detect anomaly, a change in the number of hidden variables indicates a sudden change of system behavior (anomaly). Although domain-knowledge is needed to give some explanation to the sudden changes, this way can capture anomalies beyond traditional threshold-based scheme. That means, it will issue an alarm when it can not explain the present stream reasonably using the past.

TensorStream [36][35]. In this model, Sun *et al.* extended PCA analysis to a multi-dimensional matrix, in which the values of each stream include domain-dependent environmental information, for example, a sample value of network flow at time t may have “source-IP“, “destination-IP“, and “port“. In TensorStream, such a stream is called tensor stream. The aim of streaming tensor analysis is to adjust projection matrix smoothly as the new tensor comes in. In anomaly detection, abnormal tensors are first picked up from a sequence of tensors, then in the suspicious tensors, abnormal modes are located, and last the abnormal dimensions in the given modes are found. In each level, the abrupt large reconstruction error is used to decide whether an anomaly has occurred.

The process-based models are not only special for anomaly detection, but also for mining hidden main trends and other patterns in multiple time series. These trends and patterns may be used to capture “roots“ of anomaly. Also, these models can help to understand the correlations among the streams and thus be expected to capture anomaly with better precision and recall.

2.3.2 Applications

Most researchers are interested in addressing a particular application problem that involves detecting anomalies, rather than in attempting to address a large or general class of different anomaly detection tasks. Researchers generally use the methods they believe best for the specific problem according to their past experience.

Classical methods used for disease outbreaks rely on statistical models. However, in many cases of syndromic data, outbreak patterns are unknown. Although signatures are always indicators for disease outbreaks, anomaly detection may be a better method for biosurveillance in such a case. Galit Shmueli *et al.* proposed a wavelet-based monitoring for biosurveillance [33]. The main idea is to decompose a series into a time-frequency domain and then monitor the different scale and level for abnormalities. So far, wavelet is used to analyze a single series in Shmueli’s method. They believe wavelet-based generalization for multiple series data is a potential powerful tool.

Anomaly detection in network monitoring is another important application. In a network, it is impossible to know all possible attacks, especially new ones. Jake D. Brutlag proposed WebTV to detect aberrant behavior in network time series [5]. In this model, the mechanism to detect anomaly is: if the number of observations that fall outside the confidence band is over the threshold, an alarm is triggered. The problem in the model is: each series is processed independently on others. The knowledge of the relationship between these series may be helpful to catch more network attacks.

InteMon, a prototype monitoring and mining system for large scale clusters and data centers, use

SPIRIT to analyze many stream data [11]. One application of InteMon is to monitor computer room air-condition systems. Detection of anomalous fluctuation of sensor data is useful for administrators of data center to tackle troubles before the data center is damaged.

Generally speaking, there are two kinds of ways to detect anomaly in multiple time series: one is to produce a general model from multiple series; another is to find hidden processes or variables in multiple series. More research is on the latter till now since it can give the underlying reasons for anomaly besides the internal properties of the systems; while the former is special to detect anomaly.

Process models are concise and are easily read and modified by humans, but their generation requires parameters to be set by a human who must have knowledge of the underlying processes that produce the time series. The general model, however, does not need to know much about the generation of time series. All models or algorithms are transparent; this property is appealing since humans may tune the parameters to fit the models or algorithms to a specified domain.

Compared with signature detection, the false positive rate of anomaly detection may be higher. However, anomaly detection may catch unknown anomalies which signature detection may not catch. The trade-off between false positive and true positive is often a needed standard of anomaly detection models.

The previous sections have introduced the popular approaches related to our work in material science and data mining, and discussed several efforts to employ these methodologies in prediction of the heterogeneous system failures, correlation analysis, and mining patterns in time series data. These includes predicting *rupture* as a PDF conditional on the knowledge of spatio-temporal distribution of the damage, finding correlation between attributes in data sets, detecting anomaly as outliers or utilizing the *normal* information to account for *abnormal* behavior. In our opinion, the researches in material science provide us the theory basis on which our work were conducted, the correlation analysis methods show us the idea on using statistical theory to search in time series for patterns correlated to external events, and anomaly detection tell us the possibility of predicting *unknown* unwanted observations by generalizing the time series. In this thesis, we make use of these ideas by presenting a system that integrates the use of PDF information of telemetry data, the use of simple and convenient statistical techniques, and finally the use of a large amount of dispatch data from oilsand-shovel on-site records to analyze at different levels.

Chapter 3

Framework

The previous chapter discussed a variety of approaches proposed in the literature for tasks of prediction of *rupture* in a heterogeneous system and special pattern mining in time series. This chapter will first motivate our approach for correlation pattern mining, and then present different steps of the framework. We will begin by briefly reviewing the important conclusions about a system we should have, based on the previous work. This will be followed by a presentation of the features of our problem. This leads naturally into the motivation for developing our approach, which incorporates many of the ideas presented in the previous works, but further explores the utility of PDF (Probability Density Function) in order to take advantage of the huge set of time series data used by human to monitor the shovel operation. The sections following this motivation will present the individual components in the framework.

Based on the existing literature, several conclusions can be drawn with respect to elements of a framework that can be used to mining a pattern in time series.

- A classic classifier is not feasible to predict shovel cable failures. Preliminary results that we obtained using such an approach were not promising (see Chapter 1).
- The probabilistic models proposed in material science may not be used in our work. Due to a small sample of used cables and no access to the spatio-temporal distribution of damage to the cables, the approaches in material science are hard to implement in our work.
- The approaches available to search patterns in time series in data mining field are not appropriate to our work. The time series patterns we are interested in are not trends, periodic oscillations, or abnormal behaviors. For us, the cable failure is a random process conditioned on some physical features, such as energy, properties of material, etc. We need within the time series of power values to search for some ranges of values correlated to an external event.

Before building the framework, we will first briefly review some features about our problem and the huge data set. These can be summarized as follows:

1. Energy generated by the motors in shovels is applied directly to the shovel cables. Three motors (hoist, crowd, and swing) in a shovel control the moving of the dipper through the cable. Therefore, despite the fact that most of the produced energy is converted into the dipper's motions, some of the energy is used to make the cable drag the dipper. The energy is the most immediate factor which does harm to the cable. Other factors, such as materials and operators, exert influence on the cables through the applied energy.

2. Colossal set of telemetry data are accumulated as signals of the shovel operations. Syncrude provides over 30 Gb of data recorded from the sensors mounted in the motors of the shovels. These data are sampled at the rate of one second. For each cable with an average lifespan of 1000 hours, we have 3.6×10^6 samples of the electrical current or voltage values. These samples should be enough to estimate a probability density function for a certain long time period, such as a cable lifespan.

3. The dispatch data are synchronized with the telemetry data. The dispatch data are sets of information about shovel operation recorded on-site. The information includes the shovel working status, the dug materials, and the working shifts. This information is recorded at an unfixed interval with a timestamp. Using the dispatch data, we can know whether a shovel is working or under maintenance, which team is operating it, and which material it is digging, at a time.

Obviously, the purpose of our framework is to utilize the telemetry data and the dispatch data to analyze the relationships between the three kinds of energy , different teams of operators, different materials (oilsand and overburden) and the cable lifespan.

Although a probability distribution of cable failures like in the probabilistic models in material science is difficult to build in our work, estimation of the probability density function of three types of power (hoist, crowd, and swing) values is easier to implement. Thus, it is feasible to design a new method to find whether the probability distribution of each segment of the whole range of a power value is correlated with the cable lifespan. Our goal is to find the segments with maximum negative correlation with the cable lifespan for the 77 cables.

The remaining sections of this chapter will outline and present our correlation mining framework, while the next chapter will analyze the testing results using our framework.

3.1 Overview

The steps in the framework are as follows (and are illustrated in Figure 3.1):

- Data Preprocessing: The removal of errors and imputation of missing values in the raw telemetry data.

- Feature Extraction: The construction of the time series of the power-related features from the raw electrical current and voltage data for each shovel.
- Merging: The incorporation of the information from the dispatch data sets, including material, teams, loading elevation, etc.
- Time series Separation: The division of the constructed time series into time series for different teams, different materials, and different cables.
- PDF Generation: The estimation of the probability density function in a proper scale for each time series of the features separated in the previous step.
- Correlation Mining: The process of searching the features' PDFs of all cables for certain ranges of the feature values that have maximum negative correlation with the cable lifespan.

The first 4 steps actually belong to the Sample File Creation and the last one is the core component in our framework.

3.2 Data Preprocessing

The first stage in the processing pipeline is data preprocessing. This stage aims to delete records that are not relevant for our analysis and impute missing values in the raw telemetry data. This step is not vital, since the raw telemetry data has already included enough information to produce the probability density functions of features we can use in the following steps. Skipping this step won't change our testing results greatly. As the *noisy* effects we should eliminate in this step will typically not be severe, an important property of the methods used for this step is that they do not introduce additional noise to change the results of the tests we will do later, and only seek to make minor corrections. Figure 3.2 is an example of the raw telemetry data. In the figure, each line is a record including both readings of currents or voltages values from the three motors at a time and the motors statuses. The readings from all armatures are what we need because they represent the energy directly applied to the cables, and have been already scaled to have the *ampere* as the unit of all currents and the volt as the unit of all voltages. All records are sampled per second despite the fact that several seconds' records are missing. Some records are incomplete due to no readings for some currents and voltages. We will discuss how to process the missing data in the raw data in this section.

3.2.1 Cleaning

The term Cleaning refers to a process of deleting from telemetry data some records corresponding to the timestamps when a shovel is not working, or under maintenance. Through checking the telemetry data and their corresponding dispatch data, when a shovel is not working, the motor armatures will

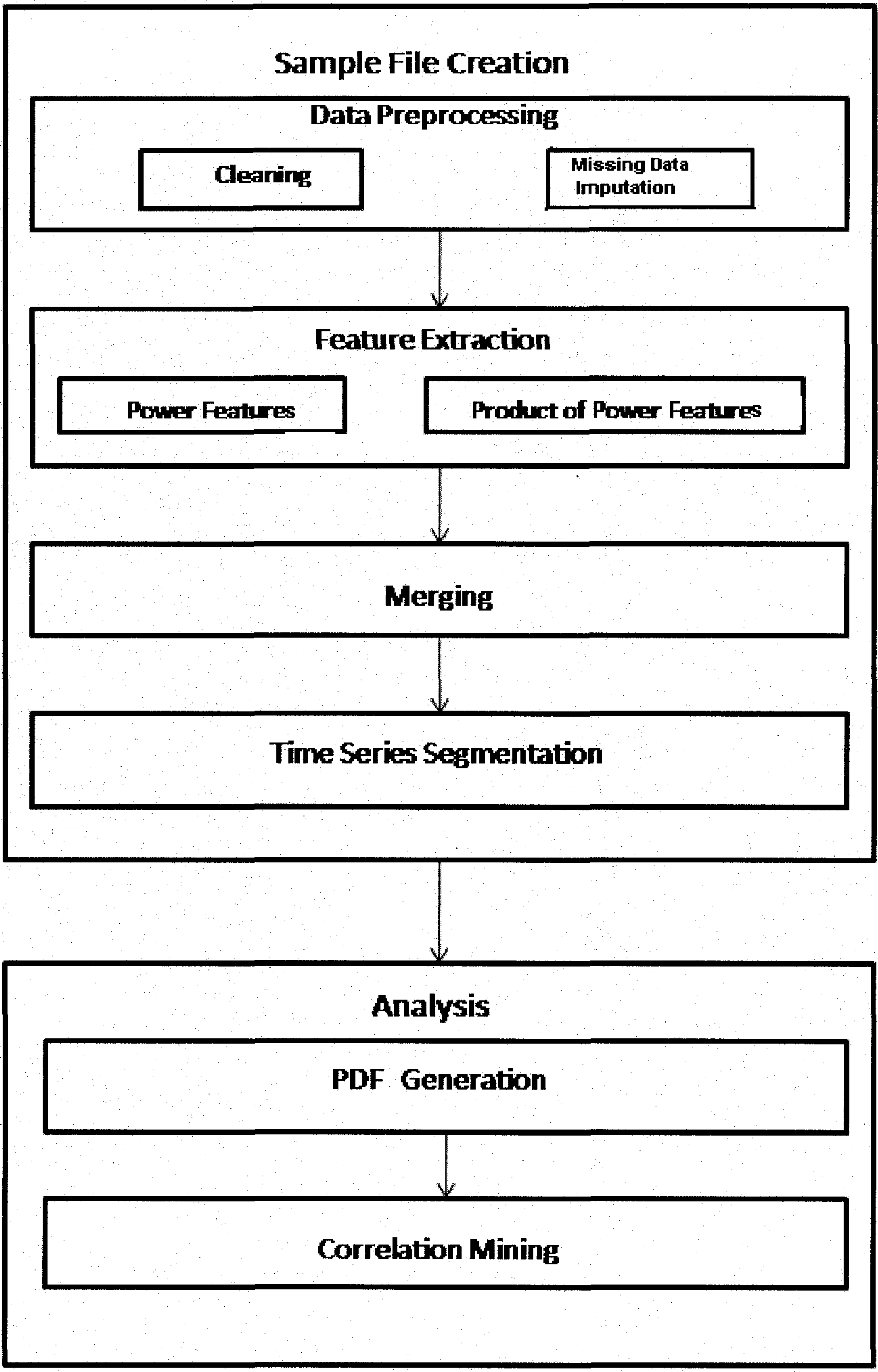
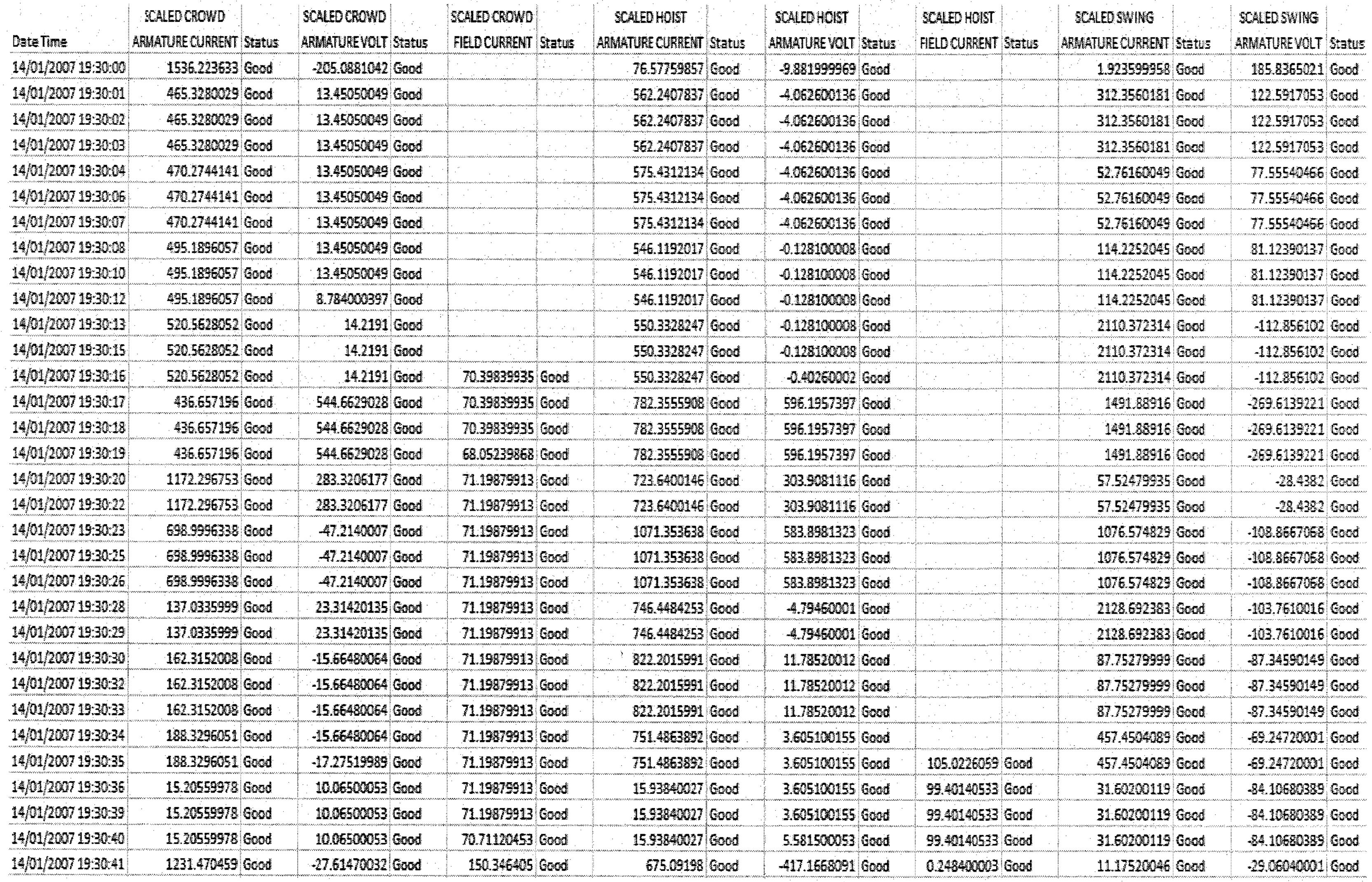


Figure 3.1: Overview of presented framework.



	SCALED CROWD		SCALED CROWD		SCALED CROWD		SCALED HOIST		SCALED HOIST		SCALED HOIST		SCALED SWING		SCALED SWING	
Date Time	ARMATURE CURRENT	Status	ARMATURE VOLT	Status	FIELD CURRENT	Status	ARMATURE CURRENT	Status	ARMATURE VOLT	Status	FIELD CURRENT	Status	ARMATURE CURRENT	Status	ARMATURE VOLT	Status
14/01/2007 19:30:00	1536.223633	Good	-205.0881042	Good			76.57759857	Good	-9.881999969	Good			1.923599958	Good	185.8365021	Good
14/01/2007 19:30:01	465.3280029	Good	13.45050049	Good			562.2407837	Good	-4.062600136	Good			312.3560181	Good	122.5917053	Good
14/01/2007 19:30:02	465.3280029	Good	13.45050049	Good			562.2407837	Good	-4.062600136	Good			312.3560181	Good	122.5917053	Good
14/01/2007 19:30:03	465.3280029	Good	13.45050049	Good			562.2407837	Good	-4.062600136	Good			312.3560181	Good	122.5917053	Good
14/01/2007 19:30:04	470.2744141	Good	13.45050049	Good			575.4312134	Good	-4.062600136	Good			52.76160049	Good	77.55540466	Good
14/01/2007 19:30:06	470.2744141	Good	13.45050049	Good			575.4312134	Good	-4.062600136	Good			52.76160049	Good	77.55540466	Good
14/01/2007 19:30:07	470.2744141	Good	13.45050049	Good			575.4312134	Good	-4.062600136	Good			52.76160049	Good	77.55540466	Good
14/01/2007 19:30:08	495.1896057	Good	13.45050049	Good			546.1192017	Good	-0.128100008	Good			114.2252045	Good	81.12390137	Good
14/01/2007 19:30:10	495.1896057	Good	13.45050049	Good			546.1192017	Good	-0.128100008	Good			114.2252045	Good	81.12390137	Good
14/01/2007 19:30:12	495.1896057	Good	8.784000397	Good			546.1192017	Good	-0.128100008	Good			114.2252045	Good	81.12390137	Good
14/01/2007 19:30:13	520.5628052	Good	14.2191	Good			550.3328247	Good	-0.128100008	Good			2110.372314	Good	-112.856102	Good
14/01/2007 19:30:15	520.5628052	Good	14.2191	Good			550.3328247	Good	-0.128100008	Good			2110.372314	Good	-112.856102	Good
14/01/2007 19:30:16	520.5628052	Good	14.2191	Good	70.39839935	Good	550.3328247	Good	-0.40260002	Good			2110.372314	Good	-112.856102	Good
14/01/2007 19:30:17	436.657196	Good	544.6629028	Good	70.39839935	Good	782.3555908	Good	596.1957397	Good			1491.88916	Good	-269.6139221	Good
14/01/2007 19:30:18	436.657196	Good	544.6629028	Good	70.39839935	Good	782.3555908	Good	596.1957397	Good			1491.88916	Good	-269.6139221	Good
14/01/2007 19:30:19	436.657196	Good	544.6629028	Good	68.05239868	Good	782.3555908	Good	596.1957397	Good			1491.88916	Good	-269.6139221	Good
14/01/2007 19:30:20	1172.296753	Good	283.3206177	Good	71.19879913	Good	723.6400146	Good	303.9081116	Good			57.52479935	Good	-28.4382	Good
14/01/2007 19:30:22	1172.296753	Good	283.3206177	Good	71.19879913	Good	723.6400146	Good	303.9081116	Good			57.52479935	Good	-28.4382	Good
14/01/2007 19:30:23	698.9996338	Good	-47.2140007	Good	71.19879913	Good	1071.353638	Good	583.8981323	Good			1076.574829	Good	-108.8667068	Good
14/01/2007 19:30:25	698.9996338	Good	-47.2140007	Good	71.19879913	Good	1071.353638	Good	583.8981323	Good			1076.574829	Good	-108.8667068	Good
14/01/2007 19:30:26	698.9996338	Good	-47.2140007	Good	71.19879913	Good	1071.353638	Good	583.8981323	Good			1076.574829	Good	-108.8667068	Good
14/01/2007 19:30:28	137.0335999	Good	23.31420135	Good	71.19879913	Good	746.4484253	Good	-4.79460001	Good			2128.692383	Good	-103.7610016	Good
14/01/2007 19:30:29	137.0335999	Good	23.31420135	Good	71.19879913	Good	746.4484253	Good	-4.79460001	Good			2128.692383	Good	-103.7610016	Good
14/01/2007 19:30:30	162.3152008	Good	-15.66480064	Good	71.19879913	Good	822.2015991	Good	11.78520012	Good			87.75279999	Good	-87.34590149	Good
14/01/2007 19:30:32	162.3152008	Good	-15.66480064	Good	71.19879913	Good	822.2015991	Good	11.78520012	Good			87.75279999	Good	-87.34590149	Good
14/01/2007 19:30:33	162.3152008	Good	-15.66480064	Good	71.19879913	Good	822.2015991	Good	11.78520012	Good			87.75279999	Good	-87.34590149	Good
14/01/2007 19:30:34	188.3296051	Good	-15.66480064	Good	71.19879913	Good	751.4863892	Good	3.605100155	Good			457.4504089	Good	-69.24720001	Good
14/01/2007 19:30:35	188.3296051	Good	-17.27519989	Good	71.19879913	Good	751.4863892	Good	3.605100155	Good	105.0226059	Good	457.4504089	Good	-69.24720001	Good
14/01/2007 19:30:36	15.20559978	Good	10.06500053	Good	71.19879913	Good	15.93840027	Good	3.605100155	Good	99.40140533	Good	31.60200119	Good	-84.10680389	Good
14/01/2007 19:30:39	15.20559978	Good	10.06500053	Good	71.19879913	Good	15.93840027	Good	3.605100155	Good	99.40140533	Good	31.60200119	Good	-84.10680389	Good
14/01/2007 19:30:40	15.20559978	Good	10.06500053	Good	70.71120453	Good	15.93840027	Good	5.581500053	Good	99.40140533	Good	31.60200119	Good	-84.10680389	Good
14/01/2007 19:30:41	1231.470459	Good	-27.61470032	Good	150.346405	Good	675.09198	Good	-417.1668091	Good	0.248400003	Good	11.17520046	Good	-29.06040001	Good

Figure 3.2: An Example of the Raw Telemetry Data.
Each line is a record of readings and status of the three motors for a second. The unit for all current is *ampere* and that for all voltages is *volt*.

always generate currents with very low values, and sometimes no data were recorded in the telemetry data. There are many empirical methods to decide if a shovel is working or not. An easy one among them is choosing a threshold as the upper limit of the low armature currents and a period of time as the duration of the low currents. However, we should not just ignore records with missing data since they are not only due to a period when a shovel is not working. Although most of the missing data with timestamps match the periods when a shovel is not working, some of them are actually generated by errors of the sensors or the GPS system. There is no way to distinguish these two types of the missing data 100% correctly, although the dispatch data may provide useful additional information. Since the timestamps in the dispatch data do not match perfectly with the timestamps in the telemetry data, we used a simple method to decide whether to skip the missing data or not: if the duration of the missing data is not shorter than 5 minutes, the missing data will be skipped assuming no working signals were generated by the motors. The complete rule that we choose in our work is the following: a period of time longer than 5 minutes with both the crowd motor armature current and the hoist motor armature current lower than 18 amp. will be regarded as a period when a shovel is not working. The next section will discuss how we deal with missing data when the duration is shorter than 5 minutes.

3.2.2 Missing Value Imputation

Missing value imputation is a process of interpolating the missing values in the telemetry data. Due to electrical, mechanical, or telecommunicating failures, the currents or voltage values of the motors in a shovel at some timestamps are not stored in the telemetry data. We introduced the method to tackle continuous missing data for a period not shorter than 5 minutes in the previous section. For missing data for a period shorter than 5 minutes, we also use a simple method of interpolation to replace these missing data with one value, which is the value just before the missing data occur. For example, if there are missing values of crowd armature voltage from 15:00:02, September, 2006 to 15:01:02, September, 2006, the missing values will be replaced with the value at 15:00:01, September, 2006.

Although there are many advanced methods to deal with missing data in time series, the simple interpolation may be strong enough in our framework. We used the same interpolation method as in a previous shovel performance study [27]. Moreover, the implementation of the simple methods is easy in any programming language. One weakness in our Java implementation of this step may lie in a great deal of time consumed for the huge data. How to improve the efficiency of the algorithms may be a further direction of this project.

3.3 Feature Extraction

After data preprocessing, the next stage in the framework is the construction of time series of the power-related features that will be used in correlation mining. At this point, the telemetry data have been preprocessed to include the data we need. However, the raw current and voltage values cannot be used directly in our framework, since the main component of our framework is to analyze the energy profiles for the cable lifespans. This section will highlight the features that have been implemented in our framework. The testing with the time series of these different features will be explored in the next chapter.

The main consideration when selecting features is that the features used should reflect the amount of energy which were applied directly to the cables. Furthermore, we need to take the combined effects of these features into account, since in the current framework there would be no method to obtain the combined effects in the following steps. The only source for the generation of useful features is the telemetry data, therefore, the features should be extracted directly from it.

3.3.1 Power Features

Power features can be used to represent the energy applied to the cables during one second. Using the power feature, we can also calculate the energy applied to the cables during longer periods, *e.g.*, a shift (12 hours), a whole cable lifespan, etc., but in our framework what we need is a time series of each type of power (crowd, hoist, and swing). The power applied to the cables can be calculated by the following formula:

$$P = V \times I$$

where, V is the voltage of the motor armature, while I is the current of the motor armature. We can obtain the three power values from the three pairs of armature current and voltage values. Therefore, the power features discussed in this section should be as follows [27]:

- Hoist power: hoist energy per second applied to the cable, produced by a hoist motor, which is responsible for pulling the dipper up or down through the face, or *up and down* motion of the dipper. The power values are obtained from the product of the current and voltage values of the armature in the hoist motor.

- Crowd power: crowd energy per second applied to the cable, produced by a crowd motor, which provides a thrust on the dipper to force it into the face or pulls it back, or *ahead and backward* motion of the dipper. The power values are obtained from the product of the current and voltage values of the armature in the crowd motor.

- Swing power: swing energy per second applied to the cable, produced by a swing motor, which controls the dipper to sway towards the truck, dump and sway back to the face, or *left and right* motion of the dipper. The power values are obtained from the production of the current and voltage values of the armature in the swing motor.

Using the time series of the armature currents and voltages, we can obtain a time series of each power feature for each shovel; we have three time series for each shovel now.

3.3.2 Product of Power Features

After calculating the different power features, an additional Feature Extraction stage can be used to construct a feature set that is more helpful to represent the combined effects of different types of energy applied to the cables.

When adding extra features, we noticed an observation that a dipper always conducts simultaneously hoist and crowd motion. Actually, the records in the raw telemetry data show that any two or three of the armatures applied simultaneously energy to the cables when the shovel is working. Like the individual power features, the features of the combined effects are easy to form time series.

We use the products of powers to represent the combined effects of the power features on the cables. The reason is that the high values of the three types of power should not occur frequently under "normal" operation. For the hoist, crowd, and swing power features, therefore, we used the following Product of Power Features:

- Product of Hoist and Crowd Power: This product measures the combined effects of hoist and crowd power simultaneously applied to the cables. We will use *product of hoist and crowd* to denote this feature.

- Product of Hoist and Swing Power: This product measures the combined effects of hoist and swing power simultaneously applied to the cables. We will use *product of hoist and swing* to denote this feature.

- Product of Crowd and Swing Power: This product measures the combined effects of crowd and swing power simultaneously applied to the cables. We will use *product of crowd and swing* to denote this feature.

- Product of Hoist, Crowd and Swing Power: This product measures the combined effects of hoist, crowd and swing power simultaneously applied to the cables. We will use *product of hoist and crowd and swing* to denote this feature.

The last three features are the combined effects of the energy for digging and the energy for swinging for dumping. We will use the four Products of Power features in our framework to analyze the relationship of their corresponding combined effects to the cable lifespan.

The computation of all features follows this order: for each record in the telemetry data, for example, for a record at 23:04:56, April 15, 2006, we calculate the three power features using the three pairs of the armature current and voltage values, then calculate the four products of power features using the three power values at the same time. Thus, we can extract all seven features at the same time and store seven time series of the features in one single file. In this file, each record (line) has seven values of hoist power, crowd power, swing power, product of hoist and crowd, product of hoist and swing, product of crowd and swing, and product of hoist and crowd and swing, with one timestamp, and all records are ordered as time goes on. At this point, we have constructed seven time series of features for each shovel.

Although two different types of features were explored in this work, it should be emphasized that there remains a considerable amount of exploration that can be made with respect to feature extraction, for example, any of the six armature currents and voltages, and any combination among these six are the indicators of how the power are applied to the cables. Since using these features may involve the knowledge of electro-magnetic mechanism in motors, for the present we leave these in the future work. The computation of each of the features discussed in this section was implemented in Java.

3.4 Merging Dispatch Information

After the time series of power-related features have been constructed, the dispatch information will then be linked to the time series of seven features and merged into the time series file created in the previous step. The merging stage has two components, *scanning* and *linking*. *Scanning* means that for each group of seven feature values with a timestamp, we find its corresponding dispatch information in the dispatch data set. Scanning should be done only once, since the dispatch data set is very large and very time-consuming to scan. *Linking* refers to the process of adding the lines of feature values with the corresponding dispatch information found in the scanning phase.

The dispatch data is a data set about shovel operations. It contains on-site records of working status, working shifts, materials, and loads with the corresponding timestamps. Figure 3.3 shows an example of the dispatch data set. A dispatch data set normally includes two small data sets. One contains the information about working status and shifts, the other contains the information about materials and loads. We must use the information provided by both two small data sets. The dispatch information is important to study the factors, such as materials and operators, which are related to the cable lifespan. We will use the dispatch information to do examinations at different levels. The dispatch information discussed in this section is summarized as follows:

- Working status: what a shovel is doing. A shovel spend most of its time digging and dumping, but sometimes it needs temporary maintenance, including mechanical repair, cable changing, etc. The information is shown in column TMT TIME IDENT (see the upper table in Figure 3.3). In the column, the codes with 'N' as the first letter denotes the working times of a shovel, while the codes with 'O' or 'M' as the first letters denotes the maintenance time of a shovel.

- Shifts: who is operating the shovel. A shovel in Syncrude works in two 12-hour shifts per day, daytime shift and night shift. There are always four teams that operate one shovel, and one team works for one shift at a time. The information is shown in column TMT SHIFT IDENT and TMT TEAM IDENT (see the upper table in Figure 3.3). '1' and '2' in TMT

SHIFT IDENT denote daytime shift and night shift respectively, while 'A', 'B', 'C' and 'D' in TMT TEAM IDENT denote different four teams.

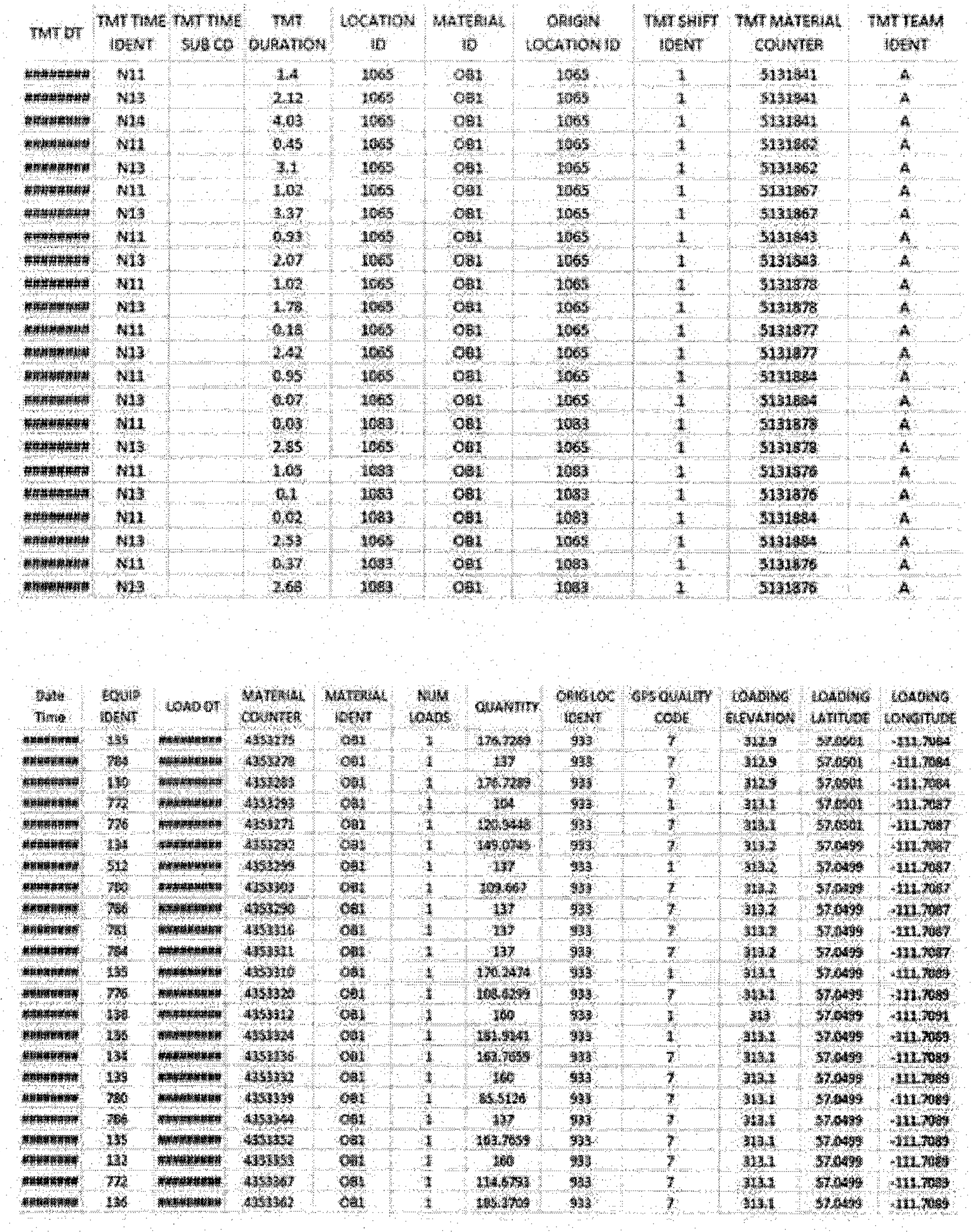
- Materials: what a shovel is digging. A shovel is not always digging oilsand. It is sometimes digging overburden, which refers to the waste on top of oilsand, or a mixture of oilsand and overburden. The dispatch data tells us which and how much material a shovel is digging in each load. In the dispatch data, 'TS', 'OB' and 'OTS' denotes oilsand, overburden and mixture respectively. The column MATERIAL IDENT in the lower table of Figure 3.3 is the codes for different material, while the column of QUANTITY is the quantity of material dug in each load.

- Loads: how a shovel is digging in each load. A load is a cycle of digging and dumping. Dispatch data of loads include their elevations, latitudes, longitudes, and completion times (see the corresponding columns in the lower table in Figure 3.3).

Although working status is included in the dispatch data, we did not include the information in the feature time series file because the generated time series should match the shovel working time if we preprocess the raw telemetry data in the first step. The working status information would be used only to separate the feature time series into parts for different cables later in our framework. We will discuss in detail the working Status in the next section.

Shifts will not be changed during the working time of a load record. In Syncrude, a working shift for shovels is 12 hours long. A daytime shift is from around 7:30 am to around 7:30 pm, while a night shift is from around 7:30 pm to around 7:30 am in the next day. During one shift, only one team of operators will work on the shovel. The four teams work alternatively during a 12-day cycle. However, we cannot know which team is operating for each shift because the raw dispatch data has only shift information (daytime or night) and not the corresponding team information. According to the schedule provided by Syncrude, we identify the different teams for each shovel using this procedure: from 7:30 am of a given day a few years ago, in each cycle, team A works in daytime shifts and team B works in night shifts for three days, then team C works in daytime shifts and team A works in night shifts for the following three days; in the next six days, team D works in daytime and team C works at night for the first three days and then team B works in daytime and team D works at night for the remaining three days. In such a cycle, each of the four teams will work continuously for six shifts, three daytime shifts first and then three night shifts. After finding the corresponding team information, each line of the feature values is assigned a shift number and a team number. We use 1 and 2 to denote the daytime shift and night shift respectively, and A, B, C, and D to denote the four teams. Naturally, in the time series file all lines of feature values belonging to the same shift will be assigned the same shift number and team number.

The information of materials and loads are closely related because the material information is always assigned to each load in the dispatch data set. In the dispatch data (see Figure 3.3), for each



TMT DT	TMT TIME IDENT	TMT TIME SUB CD	TMT DURATION	LOCATION ID	MATERIAL ID	ORIGIN LOCATION ID	TMT SHIFT IDENT	TMT MATERIAL COUNTER	TMT TEAM IDENT
########	N11		1.4	1065	OB1	1065	1	5131841	A
########	N13		2.12	1065	OB1	1065	1	5131841	A
########	N14		4.03	1065	OB1	1065	1	5131841	A
########	N11		0.45	1065	OB1	1065	1	5131862	A
########	N13		3.1	1065	OB1	1065	1	5131862	A
########	N11		1.02	1065	OB1	1065	1	5131867	A
########	N13		3.37	1065	OB1	1065	1	5131867	A
########	N11		0.93	1065	OB1	1065	1	5131843	A
########	N13		2.07	1065	OB1	1065	1	5131843	A
########	N11		1.02	1065	OB1	1065	1	5131878	A
########	N13		1.78	1065	OB1	1065	1	5131878	A
########	N11		0.18	1065	OB1	1065	1	5131877	A
########	N13		2.42	1065	OB1	1065	1	5131877	A
########	N11		0.95	1065	OB1	1065	1	5131884	A
########	N13		0.07	1065	OB1	1065	1	5131884	A
########	N11		0.03	1083	OB1	1083	1	5131878	A
########	N13		2.85	1065	OB1	1065	1	5131878	A
########	N11		1.05	1083	OB1	1083	1	5131876	A
########	N13		0.1	1083	OB1	1083	1	5131876	A
########	N11		0.02	1083	OB1	1083	1	5131884	A
########	N13		2.53	1065	OB1	1065	1	5131884	A
########	N11		0.37	1083	OB1	1083	1	5131876	A
########	N13		2.68	1083	OB1	1083	1	5131876	A

Date Time	EQUIP IDENT	LOAD DT	MATERIAL COUNTER	MATERIAL IDENT	NUM LOADS	QUANTITY	ORIG LOC IDENT	GPS QUALITY CODE	LOADING ELEVATION	LOADING LATITUDE	LOADING LONGITUDE
########	135	#########	4353275	OB1	1	176.7269	933	7	312.9	57.0501	-111.7084
########	784	#########	4353278	OB1	1	137	933	7	312.9	57.0501	-111.7084
########	130	#########	4353283	OB1	1	176.7289	933	7	312.9	57.0501	-111.7084
########	772	#########	4353293	OB1	1	104	933	1	313.1	57.0501	-111.7087
########	776	#########	4353271	OB1	1	120.9446	933	7	313.1	57.0501	-111.7087
########	134	#########	4353292	OB1	1	149.0745	933	7	313.2	57.0499	-111.7087
########	512	#########	4353299	OB1	1	137	933	1	313.2	57.0499	-111.7087
########	780	#########	4353303	OB1	1	109.667	933	7	313.2	57.0499	-111.7087
########	786	#########	4353290	OB1	1	137	933	7	313.2	57.0499	-111.7087
########	781	#########	4353316	OB1	1	137	933	7	313.2	57.0499	-111.7087
########	784	#########	4353311	OB1	1	137	933	7	313.2	57.0499	-111.7087
########	135	#########	4353310	OB1	1	170.2474	933	1	313.1	57.0499	-111.7089
########	776	#########	4353320	OB1	1	108.6299	933	7	313.1	57.0499	-111.7089
########	138	#########	4353312	OB1	1	160	933	1	313	57.0499	-111.7091
########	136	#########	4353324	OB1	1	181.9141	933	1	313.1	57.0499	-111.7089
########	134	#########	4353336	OB1	1	163.7659	933	7	313.1	57.0499	-111.7089
########	139	#########	4353332	OB1	1	160	933	7	313.1	57.0499	-111.7089
########	780	#########	4353339	OB1	1	85.5126	933	7	313.1	57.0499	-111.7089
########	786	#########	4353344	OB1	1	137	933	7	313.1	57.0499	-111.7089
########	135	#########	4353352	OB1	1	163.7659	933	7	313.1	57.0499	-111.7089
########	133	#########	4353353	OB1	1	160	933	7	313.1	57.0499	-111.7089
########	772	#########	4353367	OB1	1	114.6793	933	7	313.1	57.0499	-111.7089
########	136	#########	4353362	OB1	1	185.3709	933	7	313.1	57.0499	-111.7089

Figure 3.3: An Example of the Dispatch Data.

A raw dispatch dataset normally includes two small data sets. The upper one provides us with working status and shift information. Each line is a record with the timestamp at TMT DT (the first column). The other columns in the upper part we used in our work are: TMT TIME IDENT (working status codes), TMT DURATION (durations for the corresponding records, with *minute* as the unit), TMT SHIFT IDENT (shift No.), and TMT TEAM IDENT (team No.) The lower one includes the records on loads with completion timestamps at LOAD AT (the third column). The other columns in the lower part we included in the resulting time series files are: MATERIAL COUNTER, MATERIAL IDENT, QUANTITY(with *tonne* as the unit), LOADING ELEVATION (with centimeter as the unit). The MATERIAL ID in the upper part should be identical to the MATERIAL IDENT in the lower part for their respective timestamps are matched.

load, there is one line of record, including the types and amounts of material dug during one load, and the conditions of the load. Since a shovel needs several minutes to complete a load, there are hundreds of the lines of feature values for each load. The lines belonging to the same load would be assigned the same information about this load.

The procedure described up to now is used to form a time series file for one shovel. Syncrude provides the telemetry and dispatch data for eight shovels; therefore, we would have eight time series files altogether.

In our implementation of this step in Java, since both the records in the dispatch data set and the lines of feature values in the time series file are ordered as time goes on, we can do scanning and linking at the same time. After the dispatch information of some lines of feature values are found, this information will be merged with these lines of feature values; then the process will be repeated until all the needed dispatch information is merged into the time series file.

Figure 3.4 is an example of a resulting time series file with dispatch information merged.

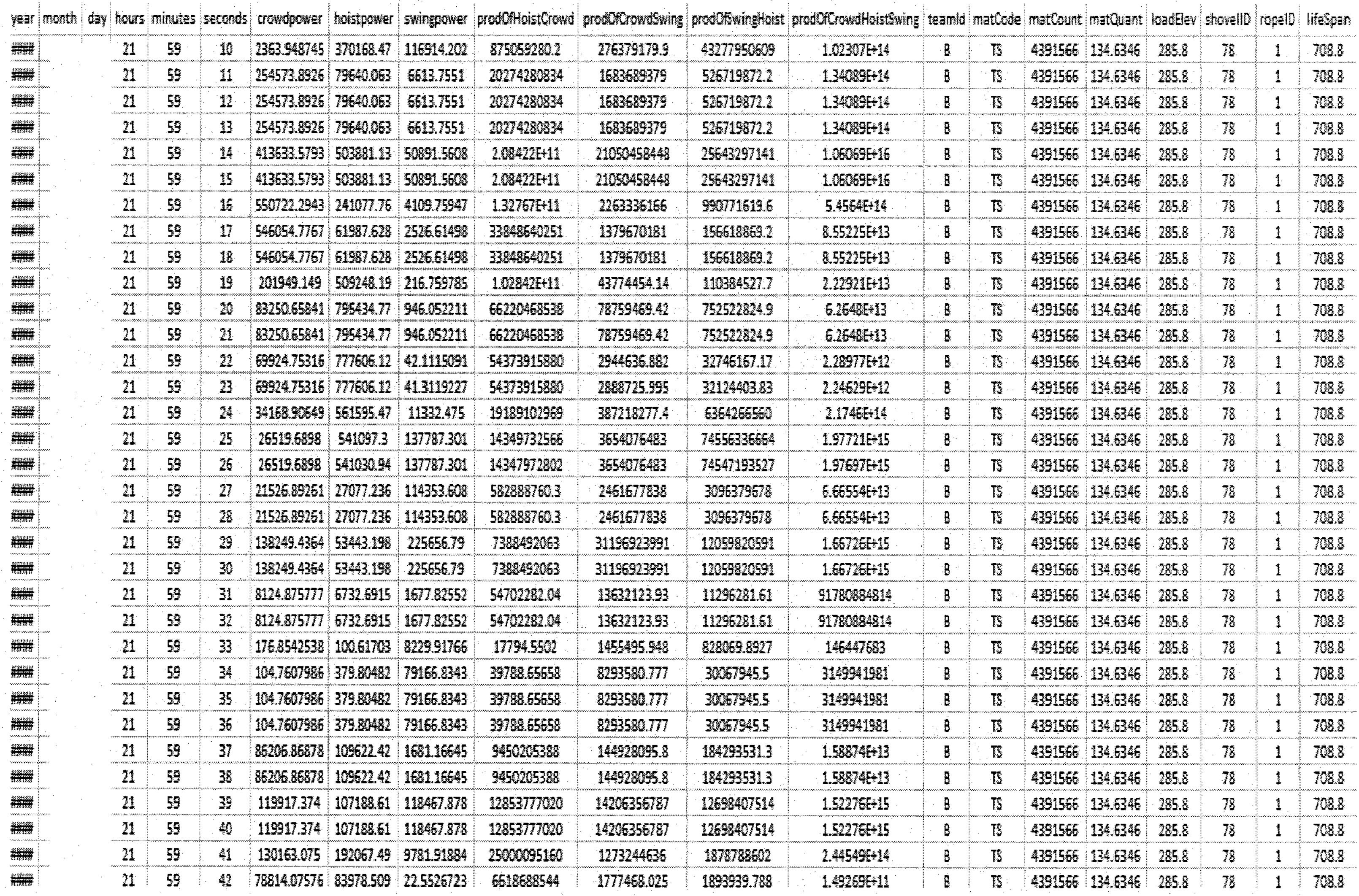
It should be noted that not all dispatch information we mentioned in this section would be used in our work. Testing with some of the dispatch information were not included in the experiments we will discussed in Chapter 4, and the unused information will be left for the future research.

3.5 PDFs Generation

The resulting time series files generated in the merging step can be used to estimate the probability density function of each feature. There are many different methods to estimate a PDF in the literature. These methods can be classified as *parametric* or *non-parametric*. Parametric methods are used to estimate the parameters of the PDF based on the assumption that the PDF is of a standard form (generally, Gaussian, Raleigh or uniform) or a combination of known form; while non-parametric methods apply to the PDF of a unknown form. The parameters of an assumed PDF can be estimated either using ML (Maximum Likelihood) estimation or Expectation Maximization (EM) algorithm. The non parametric methods include histogram based, kernel based methods and K nearest neighbor methods. Since the PDFs of the power-related features are unknown, the non-parametric methods are most appropriate for our work. We chose a histogram with a suitable number of bins (for example, 100) to estimate the PDFs. To make the histograms for different distributions comparable, we use the same fixed number of bins for all histograms. The advantage of using a histogram is that we can obtain the probability for each bin with ease for a given number of bins. We need the probabilities of all bins to test their correlations with the cable lifespan. We used Matlab to obtain and plot the PDFs. We chose 100 bins for all PDFs shown in thesis and tested in our work.

3.5.1 Feature PDFs for Different Shovels

We can obtain the histograms to plot PDF curves of seven features for each shovel. Figures 4.2 to 4.8 (page 38 – 41) show the PDF curves of the features for eight shovels.



year	month	day	hours	minutes	seconds	crowdpower	hoistpower	swingpower	prodOfHoistCrowd	prodOfCrowdSwing	prodOfSwingHoist	prodOfCrowdHoistSwing	teamId	matCode	matCount	matQuant	loadElev	shovelID	ropeID	lifeSpan
####			21	59	10	2363.948745	370168.47	116914.202	875059280.2	276379179.9	43277950609	1.02307E+14	B	TS	4391566	134.6346	285.8	78	1	708.8
####			21	59	11	254573.8926	79640.063	6613.7551	20274280834	1683689379	526719872.2	1.34089E+14	B	TS	4391566	134.6346	285.8	78	1	708.8
####			21	59	12	254573.8926	79640.063	6613.7551	20274280834	1683689379	526719872.2	1.34089E+14	B	TS	4391566	134.6346	285.8	78	1	708.8
####			21	59	13	254573.8926	79640.063	6613.7551	20274280834	1683689379	526719872.2	1.34089E+14	B	TS	4391566	134.6346	285.8	78	1	708.8
####			21	59	14	413633.5793	503881.13	50891.5608	2.08422E+11	21050458448	25643297141	1.06069E+16	B	TS	4391566	134.6346	285.8	78	1	708.8
####			21	59	15	413633.5793	503881.13	50891.5608	2.08422E+11	21050458448	25643297141	1.06069E+16	B	TS	4391566	134.6346	285.8	78	1	708.8
####			21	59	16	550722.2943	241077.76	4109.75947	1.32767E+11	2263336166	990771619.6	5.4564E+14	B	TS	4391566	134.6346	285.8	78	1	708.8
####			21	59	17	546054.7767	61987.628	2526.61498	33848640251	1379670181	156618869.2	8.55225E+13	B	TS	4391566	134.6346	285.8	78	1	708.8
####			21	59	18	546054.7767	61987.628	2526.61498	33848640251	1379670181	156618869.2	8.55225E+13	B	TS	4391566	134.6346	285.8	78	1	708.8
####			21	59	19	201949.149	509248.19	216.759785	1.02842E+11	43774454.14	110384527.7	2.22921E+13	B	TS	4391566	134.6346	285.8	78	1	708.8
####			21	59	20	83250.65841	795434.77	946.052211	66220468538	78759469.42	752522824.9	6.2648E+13	B	TS	4391566	134.6346	285.8	78	1	708.8
####			21	59	21	83250.65841	795434.77	946.052211	66220468538	78759469.42	752522824.9	6.2648E+13	B	TS	4391566	134.6346	285.8	78	1	708.8
####			21	59	22	69924.75316	777606.12	42.1115091	54373915880	2944636.882	32746167.17	2.28977E+12	B	TS	4391566	134.6346	285.8	78	1	708.8
####			21	59	23	69924.75316	777606.12	41.3119227	54373915880	2888725.995	32124403.83	2.24629E+12	B	TS	4391566	134.6346	285.8	78	1	708.8
####			21	59	24	34168.90649	561595.47	11332.475	19189102969	387218277.4	6364266560	2.1746E+14	B	TS	4391566	134.6346	285.8	78	1	708.8
####			21	59	25	26519.6898	541097.3	137787.301	14349732566	3654076483	74556336664	1.97721E+15	B	TS	4391566	134.6346	285.8	78	1	708.8
####			21	59	26	26519.6898	541030.94	137787.301	14347972802	3654076483	74547193527	1.97697E+15	B	TS	4391566	134.6346	285.8	78	1	708.8
####			21	59	27	21526.89261	27077.236	114353.608	582888760.3	2461677838	3096379678	6.66554E+13	B	TS	4391566	134.6346	285.8	78	1	708.8
####			21	59	28	21526.89261	27077.236	114353.608	582888760.3	2461677838	3096379678	6.66554E+13	B	TS	4391566	134.6346	285.8	78	1	708.8
####			21	59	29	138249.4364	53443.198	225656.79	7388492063	31196923991	12059820591	1.66726E+15	B	TS	4391566	134.6346	285.8	78	1	708.8
####			21	59	30	138249.4364	53443.198	225656.79	7388492063	31196923991	12059820591	1.66726E+15	B	TS	4391566	134.6346	285.8	78	1	708.8
####			21	59	31	8124.875777	6732.6915	1677.82552	54702282.04	13632123.93	11296281.61	91780884814	B	TS	4391566	134.6346	285.8	78	1	708.8
####			21	59	32	8124.875777	6732.6915	1677.82552	54702282.04	13632123.93	11296281.61	91780884814	B	TS	4391566	134.6346	285.8	78	1	708.8
####			21	59	33	176.8542538	100.61703	8229.91766	17794.5502	1455495.948	828069.8927	146447683	B	TS	4391566	134.6346	285.8	78	1	708.8
####			21	59	34	104.7607986	379.80482	79166.8343	39788.65658	8293580.777	30067945.5	3149941981	B	TS	4391566	134.6346	285.8	78	1	708.8
####			21	59	35	104.7607986	379.80482	79166.8343	39788.65658	8293580.777	30067945.5	3149941981	B	TS	4391566	134.6346	285.8	78	1	708.8
####			21	59	36	104.7607986	379.80482	79166.8343	39788.65658	8293580.777	30067945.5	3149941981	B	TS	4391566	134.6346	285.8	78	1	708.8
####			21	59	37	86206.86878	109622.42	1681.16645	9450205388	144928095.8	184293531.3	1.58874E+13	B	TS	4391566	134.6346	285.8	78	1	708.8
####			21	59	38	86206.86878	109622.42	1681.16645	9450205388	144928095.8	184293531.3	1.58874E+13	B	TS	4391566	134.6346	285.8	78	1	708.8
####			21	59	39	119917.374	107188.61	118467.878	12853777020	14206356787	12698407514	1.52276E+15	B	TS	4391566	134.6346	285.8	78	1	708.8
####			21	59	40	119917.374	107188.61	118467.878	12853777020	14206356787	12698407514	1.52276E+15	B	TS	4391566	134.6346	285.8	78	1	708.8
####			21	59	41	130163.075	192067.49	9781.91884	25000095160	1273244636	1878788602	2.44549E+14	B	TS	4391566	134.6346	285.8	78	1	708.8
####			21	59	42	78814.07576	83978.509	22.5526723	6618688544	1777468.025	1893939.788	1.49269E+11	B	TS	4391566	134.6346	285.8	78	1	708.8

Figure 3.4: An Example of A Final Feature Value Time Series File.

The units of all power features are *watt*, therefore, the units of product of power features are *watt*×*watt* or *watt*×*watt*×*watt*. The units of other columns of dispatch data are the same as in the original dispatch data sets.

Although the PDF curves will change with different number of bins, the trends of these curves should be similar for a wide range of bin sizes since our data are very large. We noticed that all PDFs are similar in shape, although the ranges of the seven features are different. All curves have the highest values at the leftmost ends, then drop drastically and close to zero in the following segments. However, there are still some differences among the PDFs. For example, the PDFs of product of power features drop more rapidly, and their following segments are flatter than those of power features. Some conclusions with respect to the shovel operation can be drawn from these PDF curves as follows:

- The operation patterns of all shovels is very similar with respect to the PDFs of the power-related features.
- For most of the working time of shovels, the hoist, crowd, and swing power are applied with low values to the cables.
- For most of the working time of shovels, the values of the product of power features are small; therefore, the combined effects of the three types of power applied to the cables are small.

One hypothesis in our work is that the different teams would have different influences on the cables due to their different operation patterns. Since the PDFs of the power-related features can be used as indicators of team operation, we can verify the assumption with these PDFs of different teams. Before testing, we should know if there is any significant difference between these PDFs of different teams. The estimation of the PDFs between the teams is feasible due to the merged team numbers in the time series files. The examples of some resulting PDFs of the power-related features for different teams of a shovel is in Figures 3.15 to 3.19 (page 41 – 43). The shapes of the four teams' PDF curves are overall similar to each other. There are also local differences between some of them although they are difficult to see in the scale the figures are plotted.

To find whether there is any significant difference between the PDFs of the power-related for different teams, a proper statistical test can be used for the PDFs. There are some Hypothesis Tests in statistics to test whether several samples belong to the same distribution or not. Two-sample t-test is used to determine if two populations are from normal distributions with equal means or with unequal means. Two-sample F-test tests whether the two populations are from normal distributions with equal variances or with unequal variances. Ansari-Bradley test is used to test the null hypothesis that the two population distribution functions corresponding to the two samples are identical against the alternative that they come from distributions that have the same median and shape but different dispersions (*e.g.* variances). Two-sample Kolmogorov-Smirnov test performs a test that whether two samples come from the same unspecified distribution. Two-sample Kolmogorov-Smirnov test is most suitable in our framework because it does not need the samples to have the same size and need does not assume a specific the distribution.

Two-sample Kolmogorov-Smirnov ($K-S$) test is based on the one sample Kolmogorov-Smirnov test, which tests whether an underlying probability distribution differs from a hypothesized distribution. The one-sample KS test compares the empirical distribution function with the cumulative distribution function specified by the specified distribution, which is often a normal distribution. The two-sample K-S test compares the the cumulative distribution functions of the two sample distributions. Let $S1(x)$ and $S2(x)$ be the cumulative distribution functions from the sample distributions X1 and X2, the test statistics are

$$max|S1(x) - S2(x)|.$$

We implemented the two-sample $K - S$ tests on the feature time series for different teams of each shovel. The results show that different teams of one shovel have different distributions of each of the power-related features at the significance level of 0.05. Based on this results, we can analyze the influences of different team operations on the cable lifespan.

Another hypothesis in our work is that different materials which a shovel is digging would cause the different operation patterns and these different patterns have different influences on the cable lifespan. Using the material id, the PDFs of the power-related features for different materials can be estimated with ease. Figures 3.20 to 3.21 (page 44) are examples of the feature distributions of two materials, TS (oilsand) and OB (overburden), for one shovel. To simplify the analysis, we included the mixture (OTS) in overburden (OB). The two-sample $K - S$ test was also performed to test whether any significant difference exists between the distributions of each feature for different materials. The results of the test show that there is a significant difference between these distributions for each shovel. Therefore, we can test whether oilsand or overburden would do more harm to the cables.

We did an additional test to find whether there is any significant difference between the operation patterns of different teams for the same material. Figures 3.22 to 3.23 (page 45) are some examples of the feature distributions of different teams for the same materials. The two-sample $K - S$ tests show that such differences exist.

We can summarize this section as follows:

- The different teams of one shovel have different operation patterns, and each team has different operation patterns for one material.
- A shovel has different overall operation patterns for different materials.

These conclusions justify our research of the different influences of different teams or different materials on the cable lifespan. We designed some related experiments and will discuss them in detail in Chapter 4.

3.5.2 Feature PDFs for Different Cables

At this point, we completed all the steps before the one for estimating the PDFs of the features for different cables. The following paragraphs will discuss the process in which we obtained the PDF curves for 77 cables. The process has two phases: in the first, the time series of the features for all cables are extracted from the time series files; in the second, the PDFs of these time series are estimated.

To segment the time series file for each shovel into the parts corresponding to the different cables, we need to find the time when an old cable is replaced by a new one before the old one fails. The information about the replacement of cables can be found in the dispatch data set. In the dispatch data set, there are some special codes denoting the maintenance time of the shovels. These codes include ones signifying a cable replacement. According to the time of the cable replacement codes in the dispatch data set, we can find the corresponding periods of time in the time series files.

We used the same histogram-based methods as in the previous section to estimate the PDFs of the features for the cables. Figures 3.24 to 3.31 (page 46 – 49) are some examples of the feature distributions for different cables. Naturally, the distributions are similar to those we obtained in the previous section. We will use these distributions for correlation mining in the next section.

3.6 Correlation Mining

The last step in our framework is Correlation Mining, which attempts to search in the feature distributions for the segments whose frequencies are correlated to the cable lifespan.

In this part of our framework, we use Pearson correlation coefficient as the criterion of correlation judgement due to its simplicity. What we are looking for is some ranges of a feature value whose probabilities in the whole cable lifespan have high negative Pearson correlation coefficients, ideally, less than -0.5, with the length of the lifespan.

When using Pearson correlation coefficient, we need two samples of values. The lengths of different cable lifespan naturally form one sample. The other sample should be the different probabilities of the same range of feature values for all cables. Using the histograms for PDFs of the cables, we can obtain the probabilities for each bin. Thus, we will get a sample of probabilities of all cables for each bin. Taking each sample of probabilities and the sample of cable lifespans, the different Pearson correlation coefficients for each bin can be calculated. By comparison of different Pearson correlation coefficients, the bins with highest negative Pearson correlation coefficients will be regarded as most "harmful" to the cables, or *relevant* as we will denote later in the thesis.

A histogram of a PDF is composed of the bars with the same widths and the different heights, distributing along the axis of a feature. The width of a bar is a small range of a feature value, while the height of the bar is the count of the feature value falling in this range. When estimating PDFs from the histogram of samples, the probability for each bar may be calculated by dividing its counts

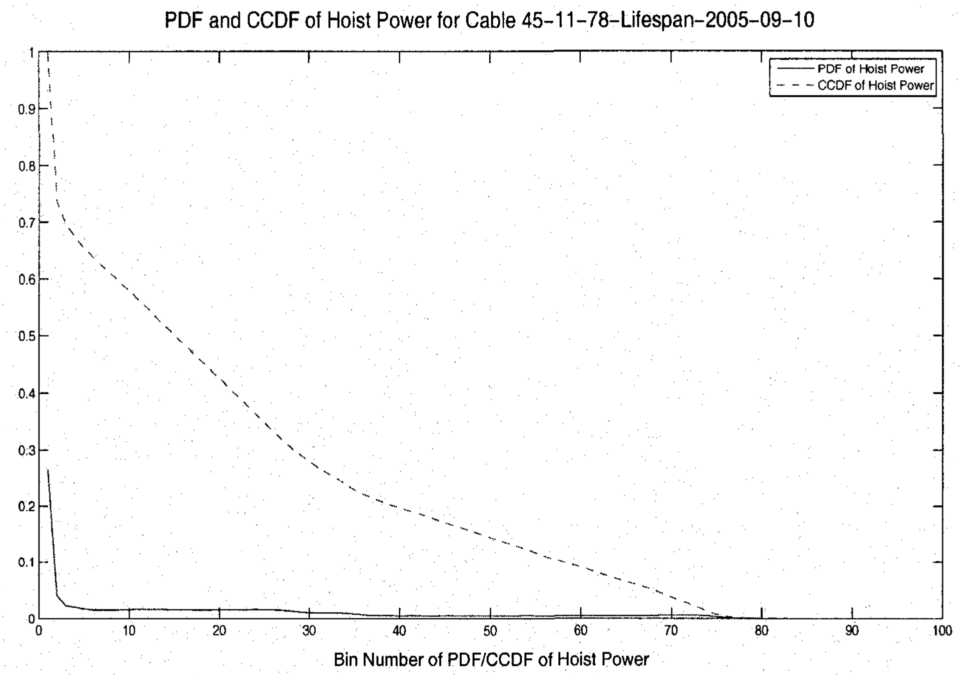


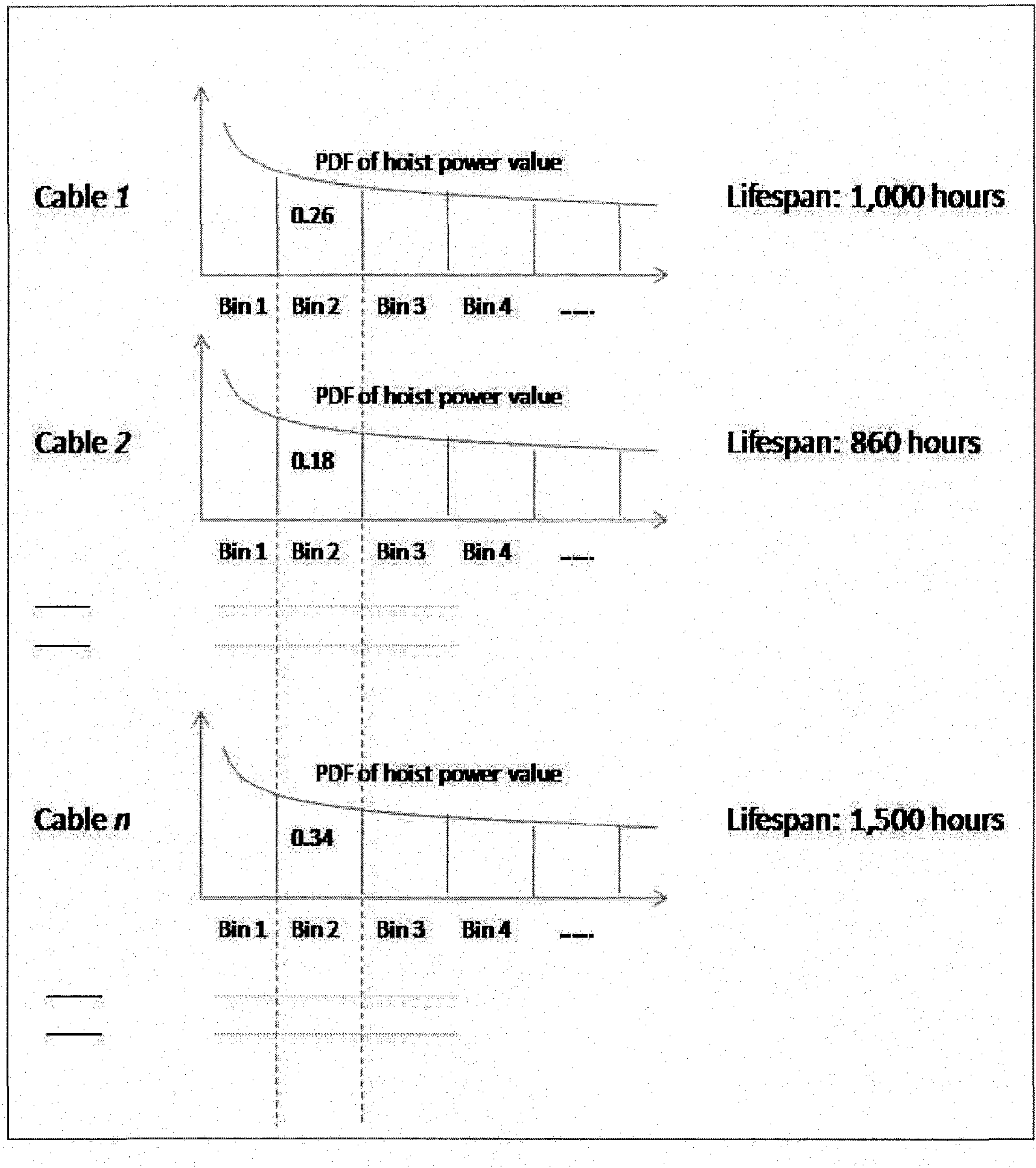
Figure 3.5: PDF and CCDF of crowd power for a cable of Shovel78 as functions of bin number.

by the sum of counts of all counts of all bars. To compare the different probabilities of the same range of a feature value for all cables, we must produce the histograms of a feature for all cables with one large range which contains all values of the feature for all cables. Some trivial coding in Matlab can produce the PDFs for all cables with one group of probabilities of each bar to calculate the Pearson correlation coefficients.

Another kind of testing with Pearson correlation coefficients is using the complementary cumulative distribution function (CCDF) of the feature values. CCDF is used to study how often the random variable is *above* a particular level. In our work, we utilize CCDFs of distributions of features to test the correlation between the cable lifespans and the probabilities when a feature value is greater than or equal to the left edge of a particular bin for all cables. Using such tests, we can answer the question: for which value, the tail of a feature's PDF starting at that value has the highest negative correlation with cable lifespan? Figure 3.5 is an example of PDF and CCDF of hoist power for a cable as functions of bin number.

Both PDFs and CCDFs tests are important to mine the correlation for us. PDFs are appropriate to find certain small relevant ranges; while CCDFs are convenient to ascertain a continuous range from a value on to the high value end. The latter may be more useful in practice because a feature value can be identified as dangerous only using a threshold. We used both two tests in our work and draw conclusions from the results.

Figure 3.6 illustrates the process of calculating the Pearson Correlation Coefficient for a feature



Step 1: Estimate the PDFs of the feature values (e.g., hoist power) for cables (Cable *1*, Cable *2*, ___, Cable *n*, ___) using histogram with a suitable number (e.g., 100) of bins.

Step 2: To test whether the probabilities for the corresponding bins (segments, value ranges), e.g., for bin 2, are correlated to the cable lifespans, calculate the Pearson Correlation Coefficient from two groups of samples *S1* and *S2*:
S1={0.26, 0.18, ___, 0.34,___}, including the probabilities for bin 2 of all cables.
S2={1,000, 860, ___, 1,500, ___}, including the cable lifespans of all cables.

Step3: Calculate the Pearson Correlation Coefficients for each bin using the same procedure as in the previous step.

Figure 3.6: The process of calculation the Pearson Correlation Coefficients for PDFs of a feature value for all cables.

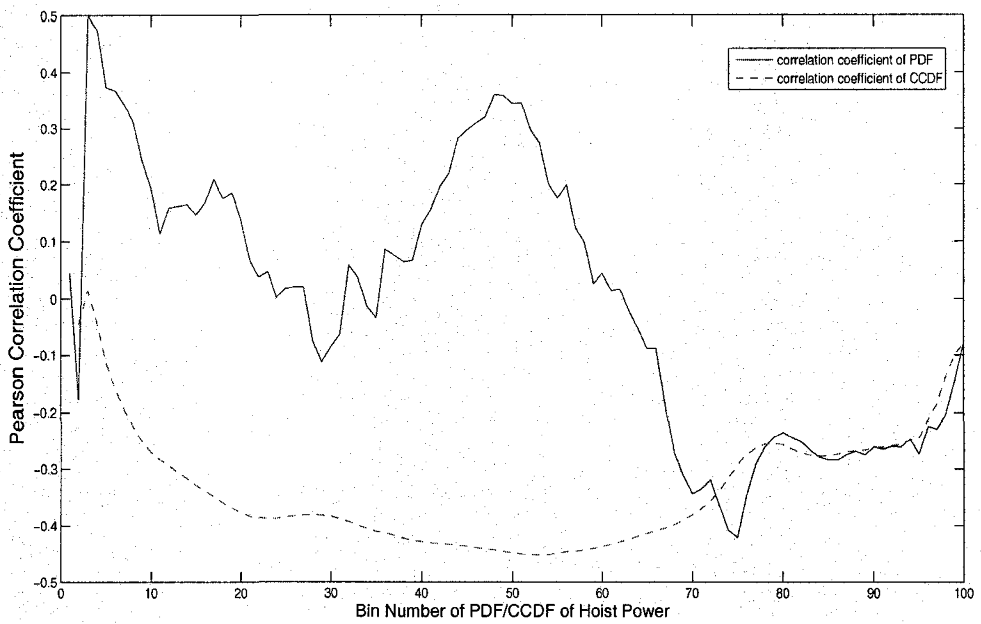


Figure 3.7: Pearson correlation coefficient of hoist power as functions of bin number.

value's PDFs for all cables. After obtaining the estimated PDFs, to test whether the probabilities of the corresponding bins (segment, value range), *e.g.*, bin 2, are correlated to the cable lifespans, the Pearson Correlation Coefficient are calculated from two groups of data. One group includes the probabilities of bin 2 for all cables $(0.26, 0.18, , 0.34,)$; the other group includes the lifespans of all cables $(1,000, 860, , 1,500,)$. Using this way, we obtained the Pearson Correlation Coefficients for all bins (1 to 100) of PDFs for each power-related feature. The same procedure was applied to obtain the Pearson Correlation Coefficients for all bins of CCDFs for each power-related feature. Figure 3.7 is the Pearson Correlation Coefficients for PDFs and CCDFs of hoist power as functions as bin number. In this figure, Pearson Correlation Coefficient has different values for different hoist power value segments (bins); some are positive and some negative. This means that different hoist power values have different impacts on the cable lifespan. From Pearson Correlation Coefficient for CCDF, its values for a segment from different bins to the highest end are also different. Using these two curves for each feature, we can find the segments whose distribution probabilities are most negative correlated with the cable lifespan.

We examined only Pearson correlation coefficient, which is used to mine only linear relationship. Other methods, especially the non-parametric ones, are still worth testing in our framework to find the relevant ranges which may not be found using Pearson correlation coefficient. With respect to the separation of the feature value space, we ultimately selected to use 100 bins rather than more due to the small time required for both the division of the feature values and the calculation of the probabilities. More number of bins, for example, 200, 500, or 1000, may be tested in future. However, too high number of bins may lead to unsmooth PDF curves and should be avoided.

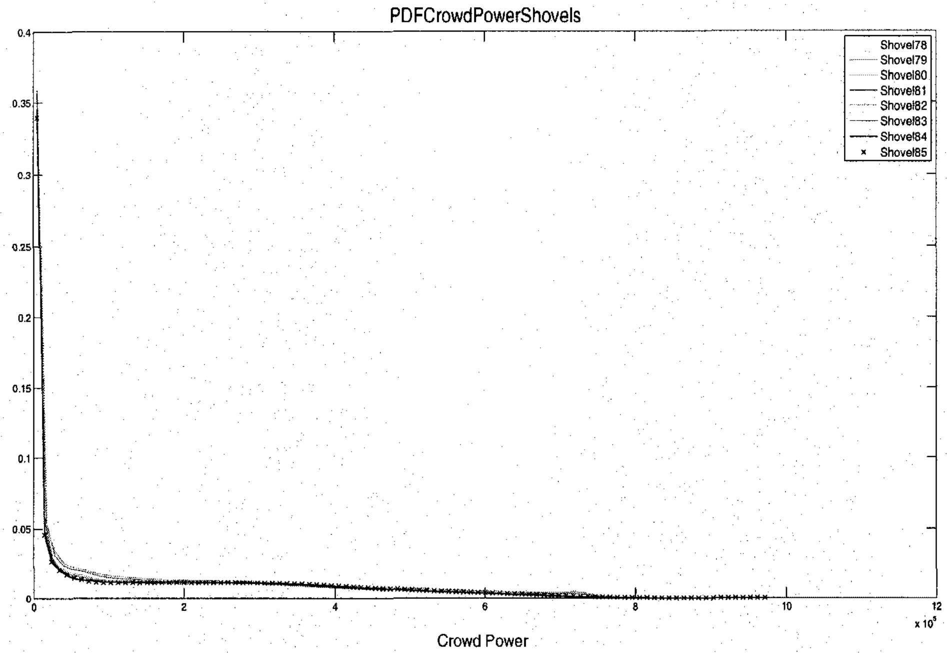


Figure 3.8: PDFs of crowd power for the eight shovels.

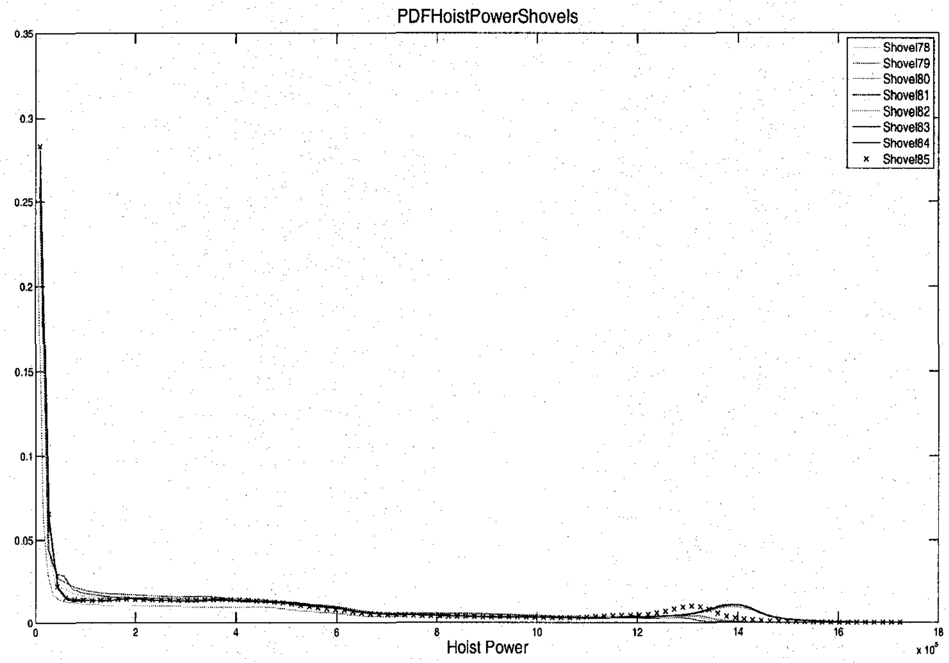


Figure 3.9: PDFs of hoist power for the eight shovels.

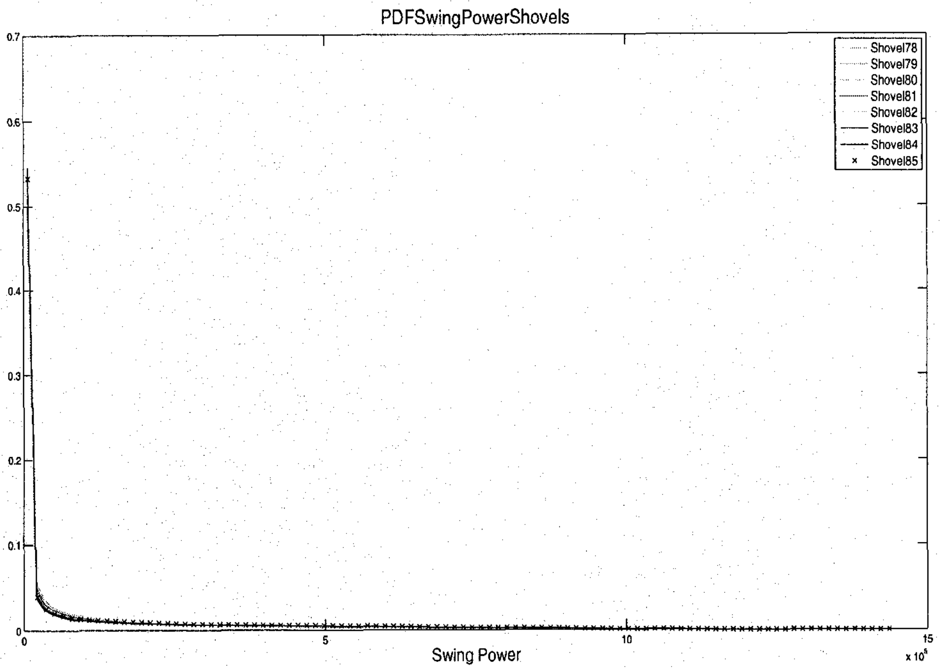


Figure 3.10: PDFs of swing power for the eight shovels.

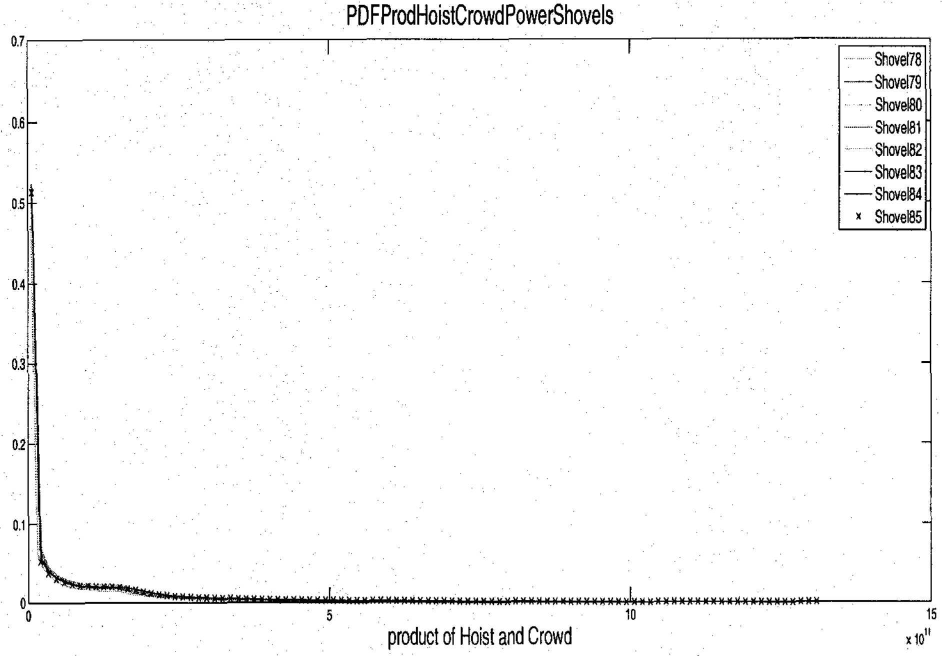


Figure 3.11: PDFs of product of hoist and crowd powers for the eight shovels.

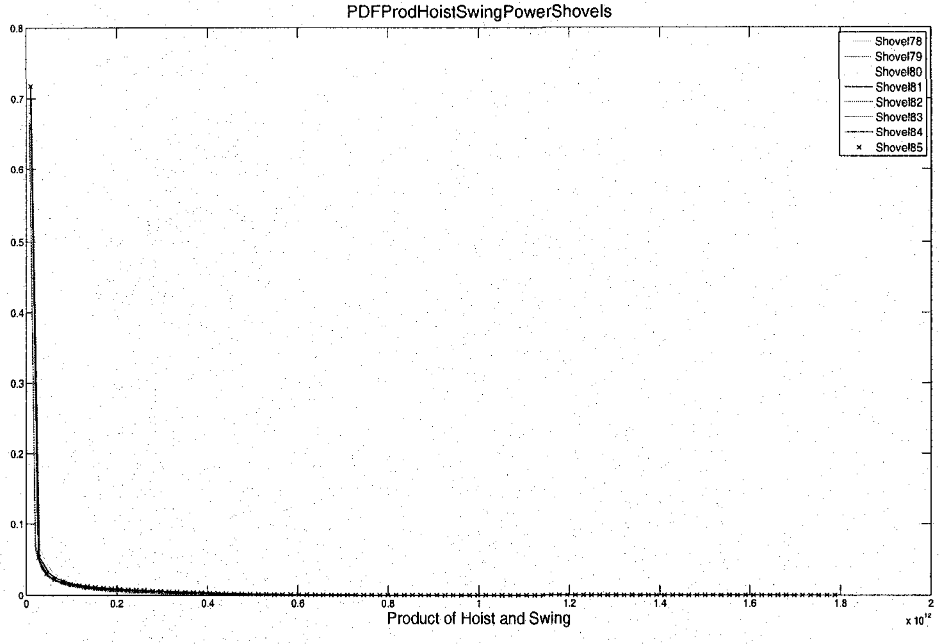


Figure 3.12: PDFs of product of hoist and swing powers for the eight shovels.

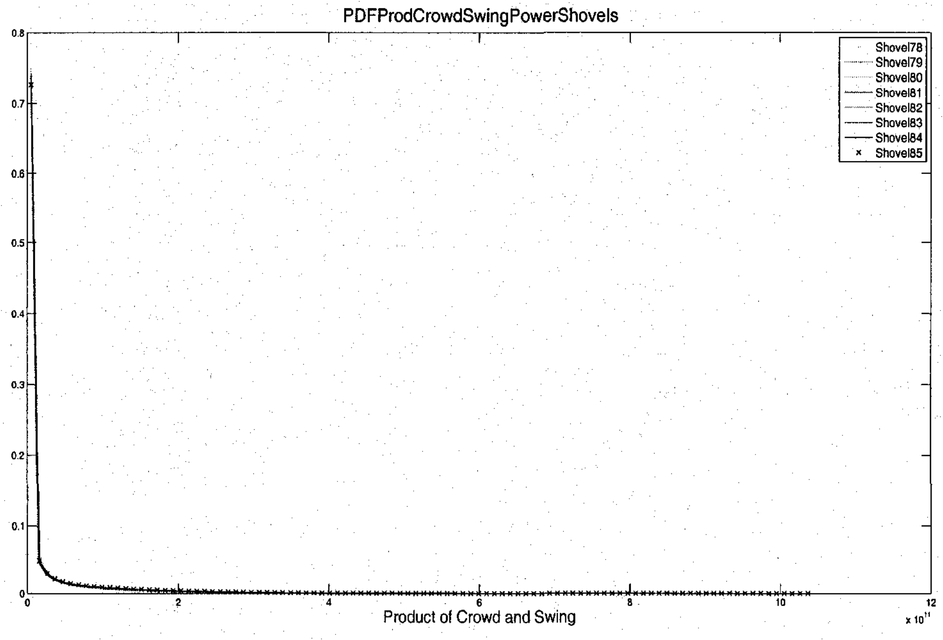


Figure 3.13: PDFs of product of crowd and swing powers for the eight shovels.

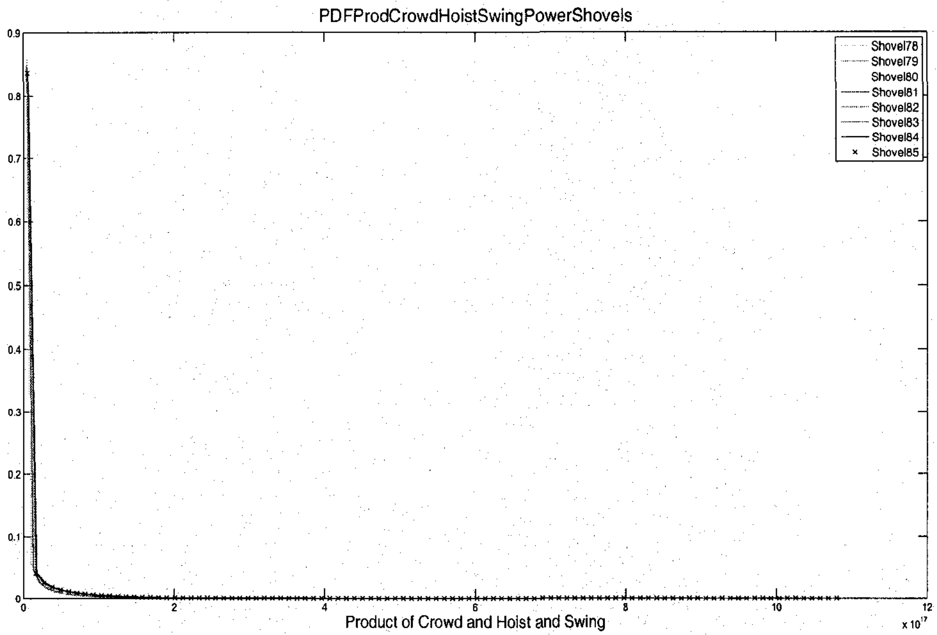


Figure 3.14: PDFs of product of crowd and hoist and swing powers for the eight shovels.

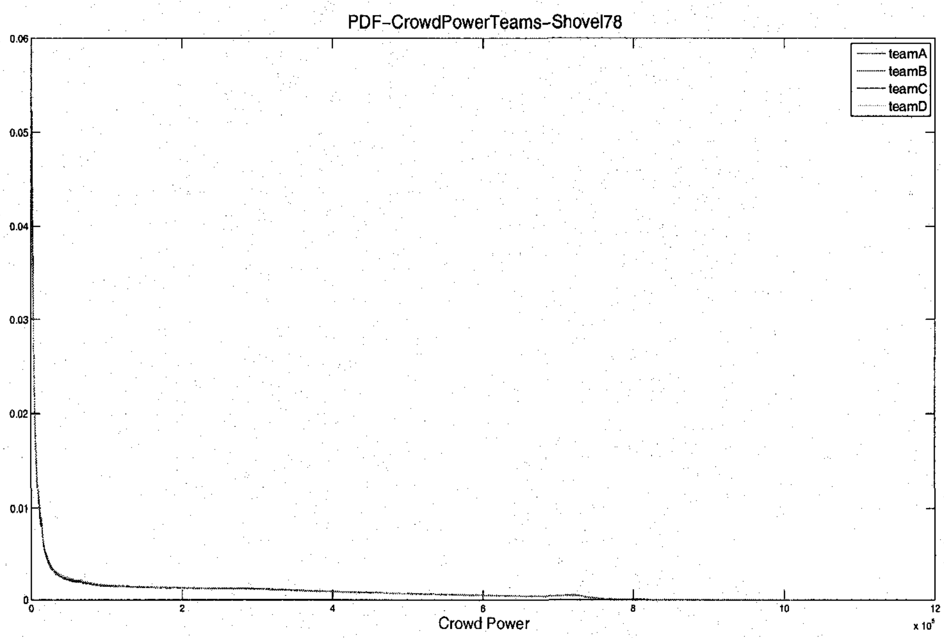


Figure 3.15: PDFs of crowd power for the four teams of Shovel78.

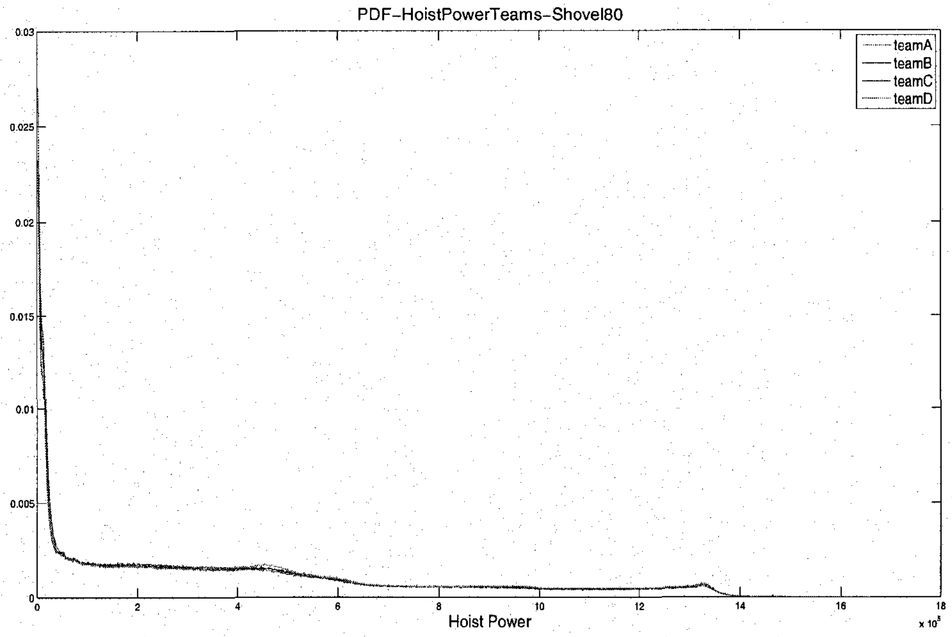


Figure 3.16: PDFs of hoist power for the four teams of Shovel80.

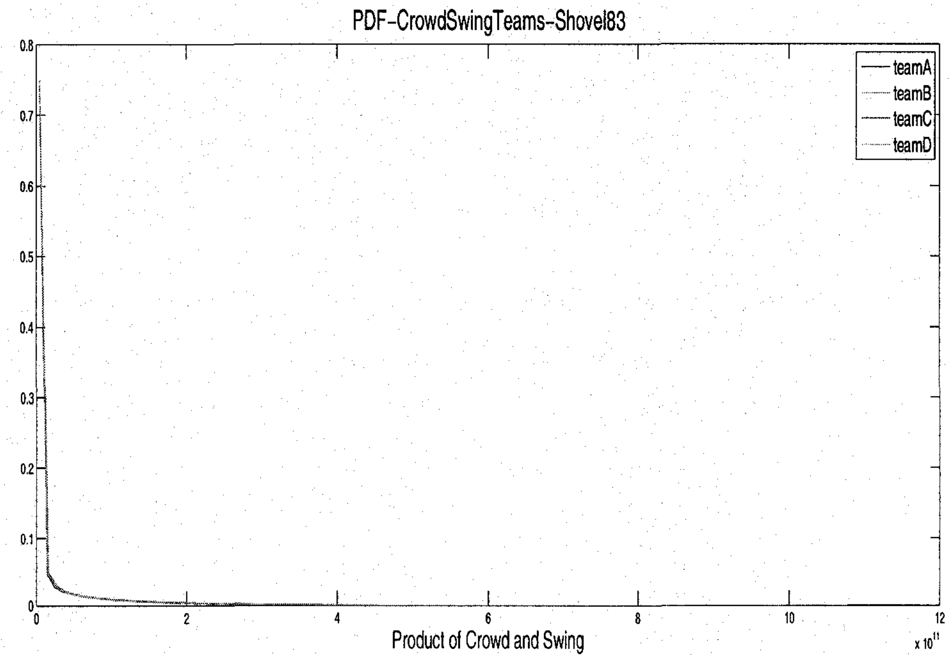


Figure 3.17: PDFs of product of crowd and swing powers for the four teams of Shovel83.

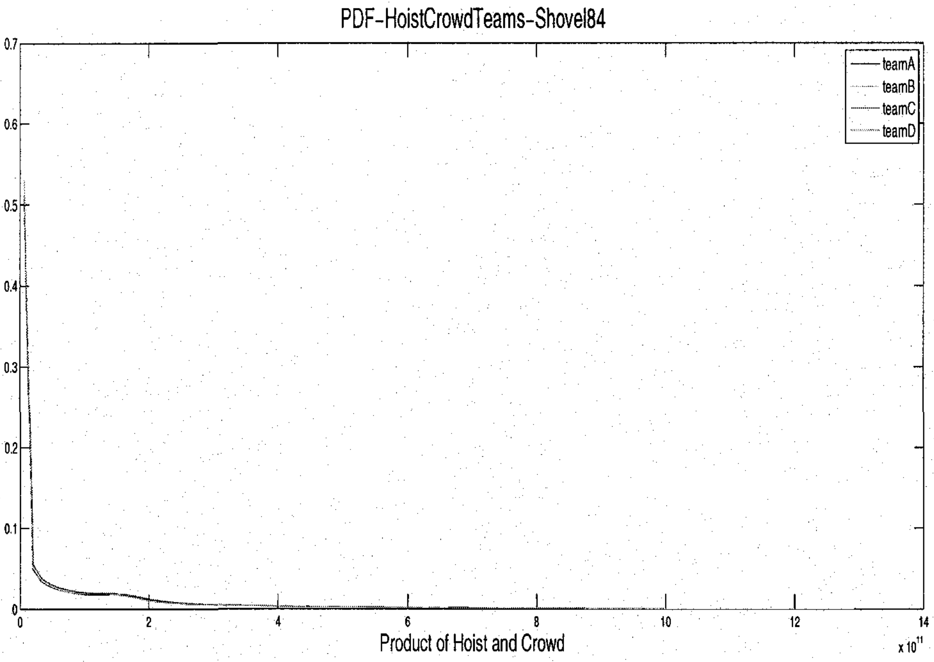


Figure 3.18: PDFs of product of hoist and crowd powers for the four teams of Shovel84.

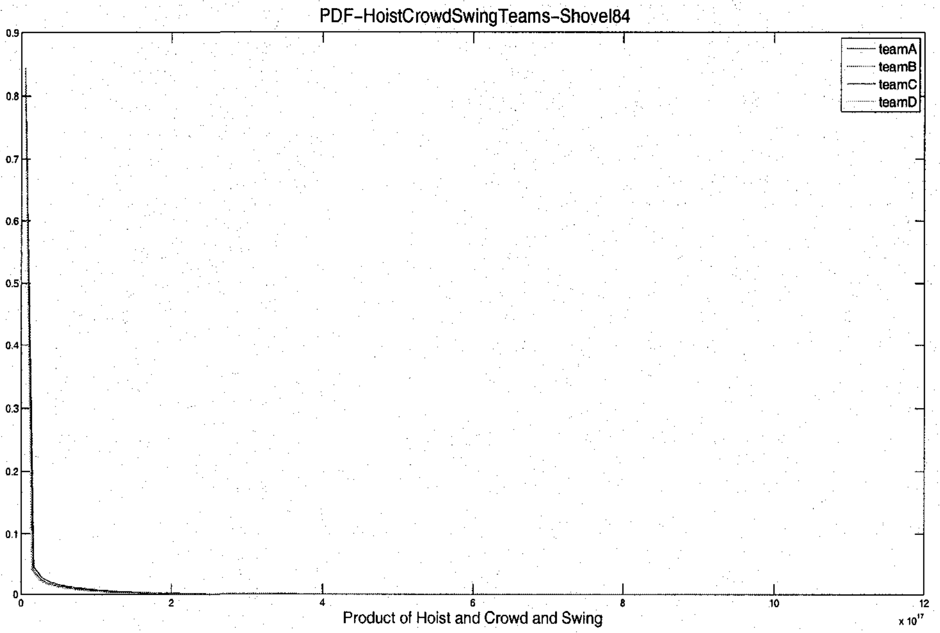


Figure 3.19: PDFs of product of hoist and crowd and swing powers for the four teams of Shovel84.

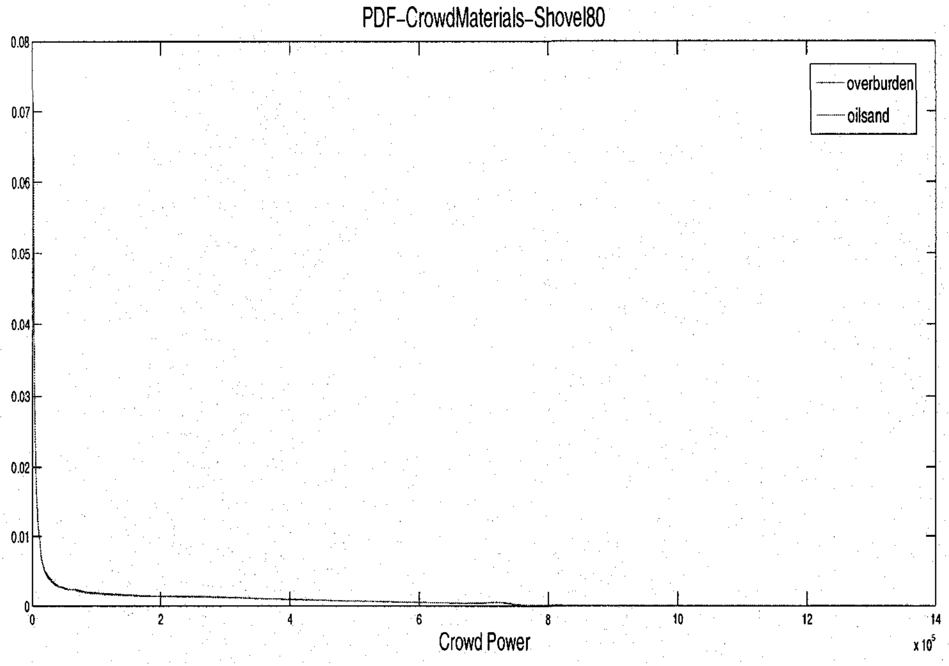


Figure 3.20: PDFs of crowd power for the two materials of Shovel80.

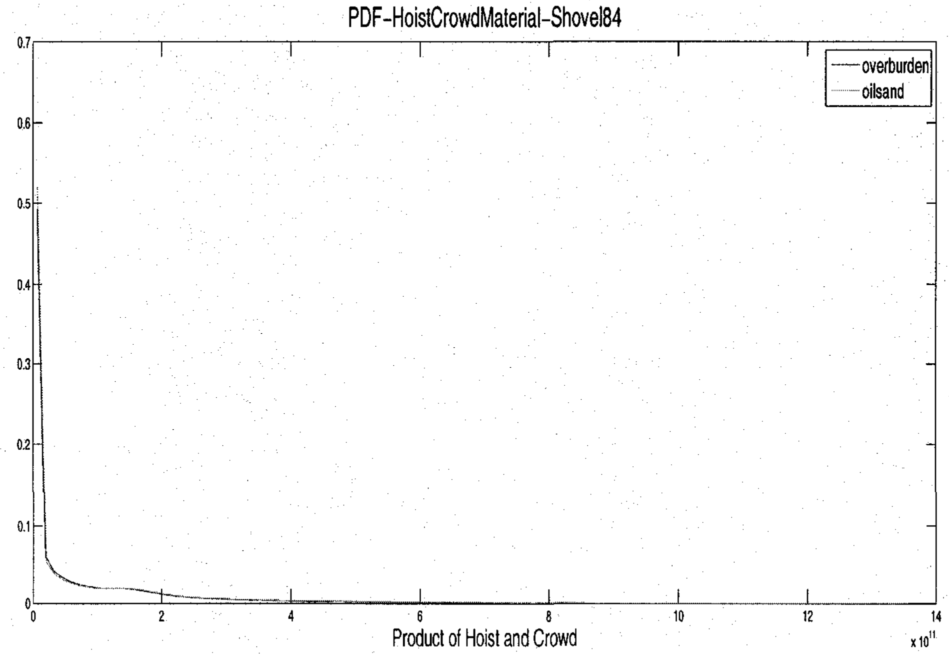


Figure 3.21: PDFs of product of hoist and crowd powers for the two materials of Shovel84.

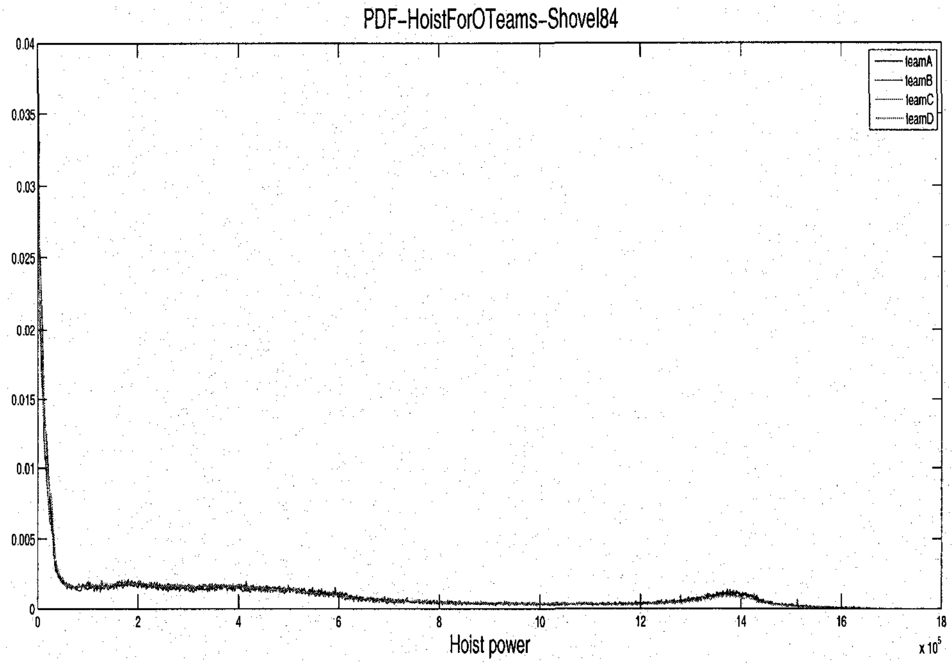


Figure 3.22: PDFs of hoist power of the four teams of Shovel84 for material of overburden.

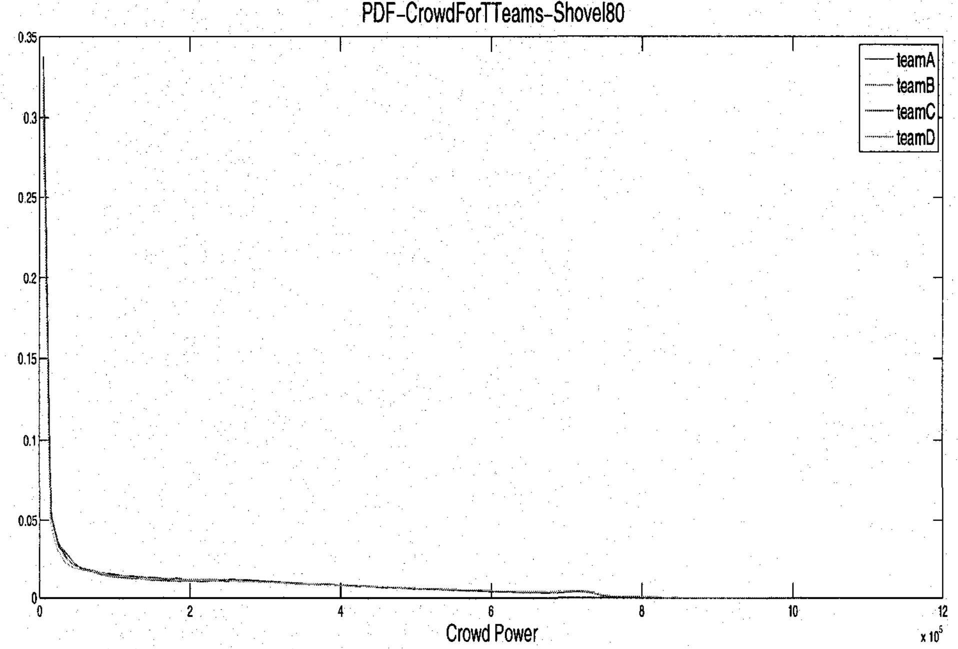


Figure 3.23: PDFs of crowd power of the four teams of Shovel80 for material of oilsand.

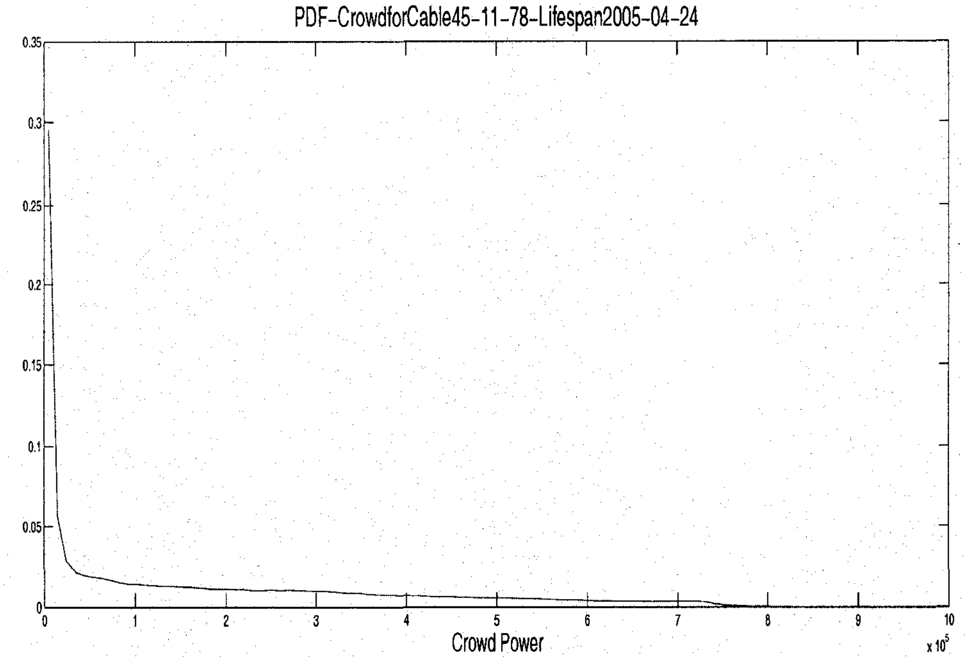


Figure 3.24: PDF of crowd power for a cable of Shovel78.

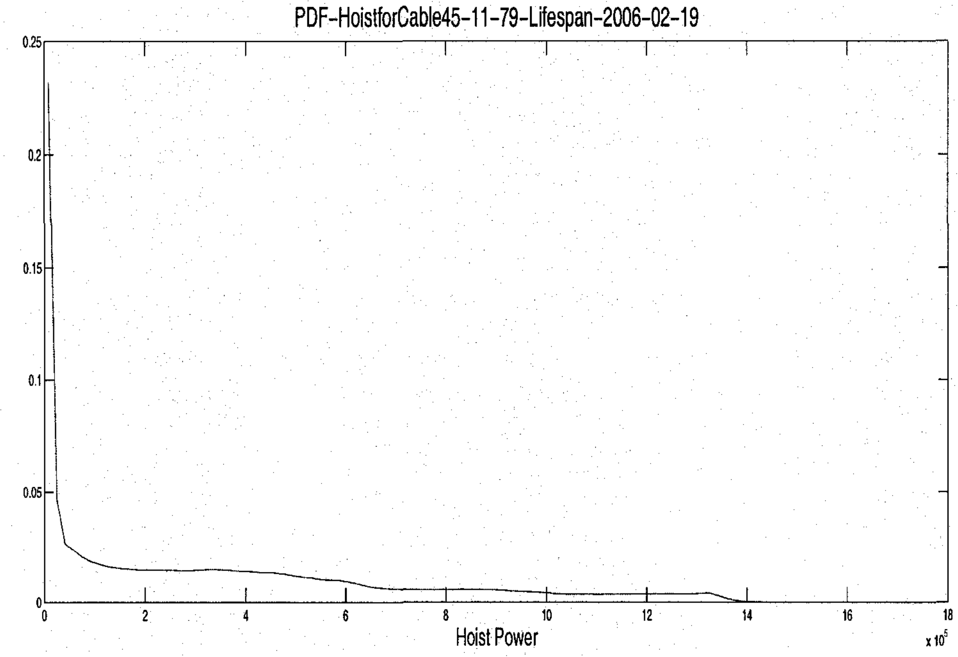


Figure 3.25: PDF of hoist power for a cable of Shovel79.

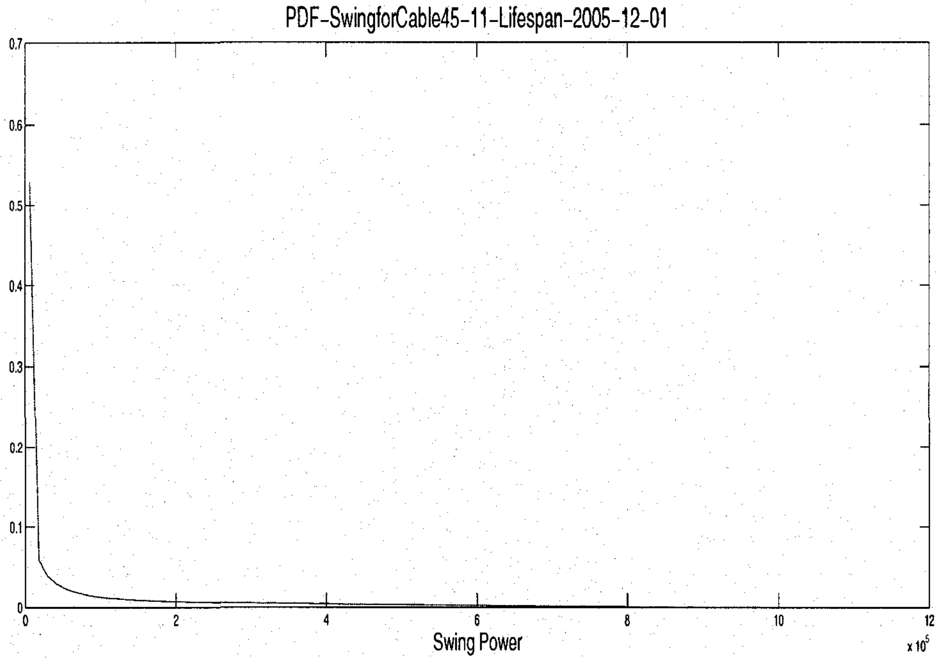


Figure 3.26: PDF of swing power for a cable of Shovel80.

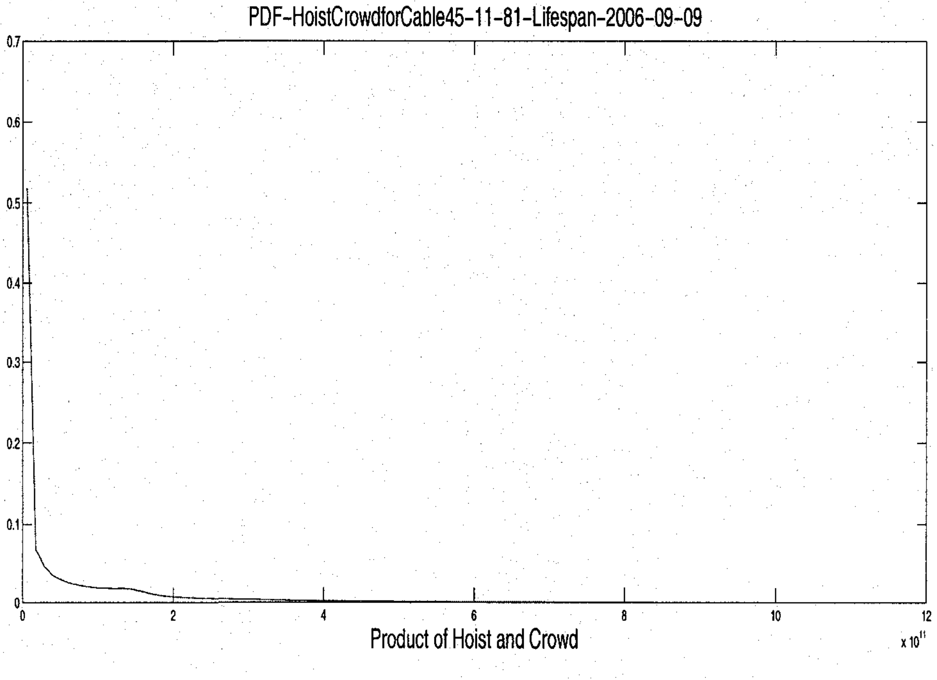


Figure 3.27: PDF of product of hoist and crowd power for a cable of Shovel81.

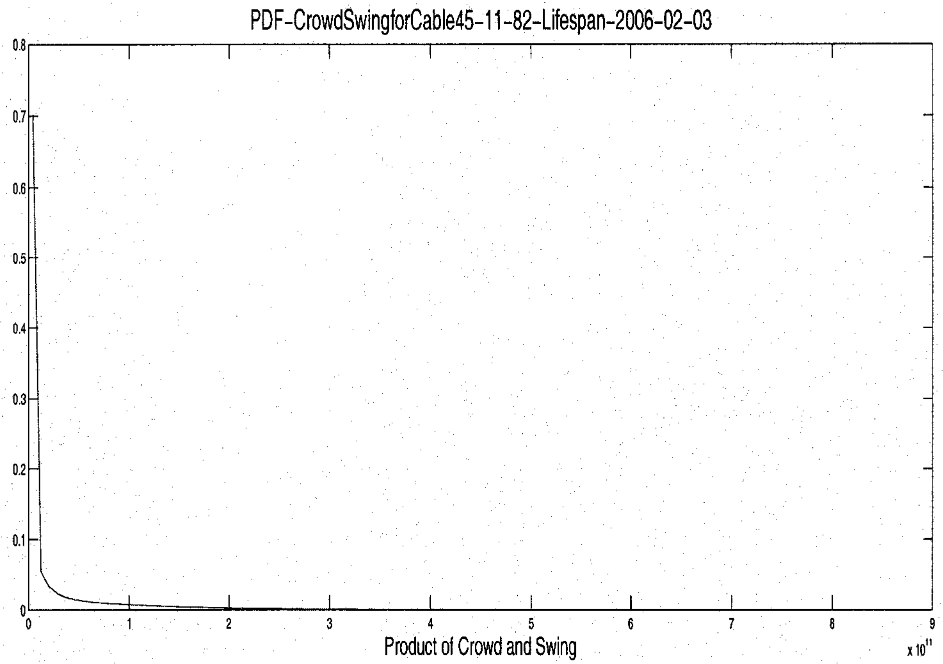


Figure 3.28: PDF of product of crowd and swing power for a cable of Shovel82.

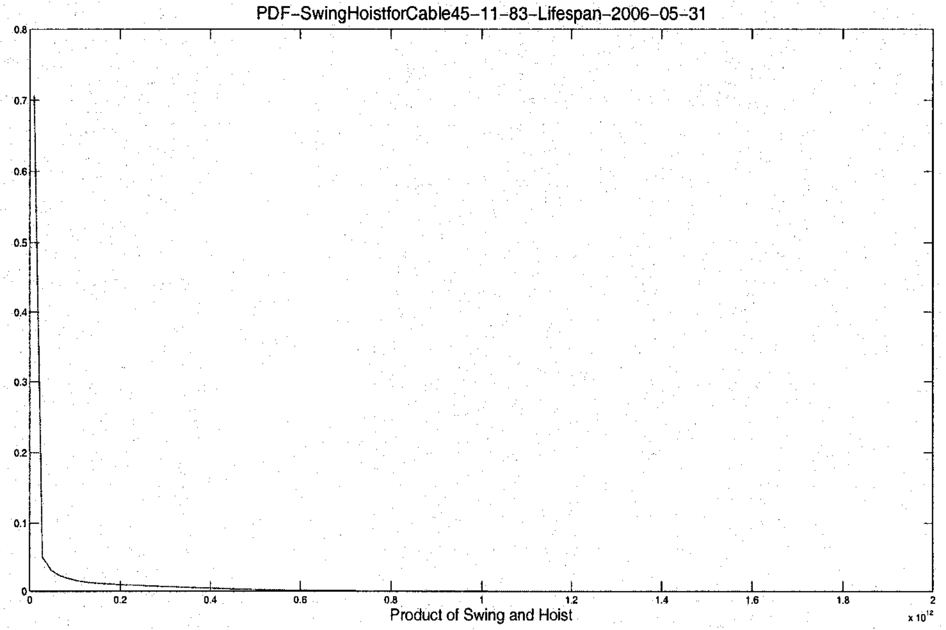


Figure 3.29: PDF of product of swing and hoist power for a cable of Shovel83.

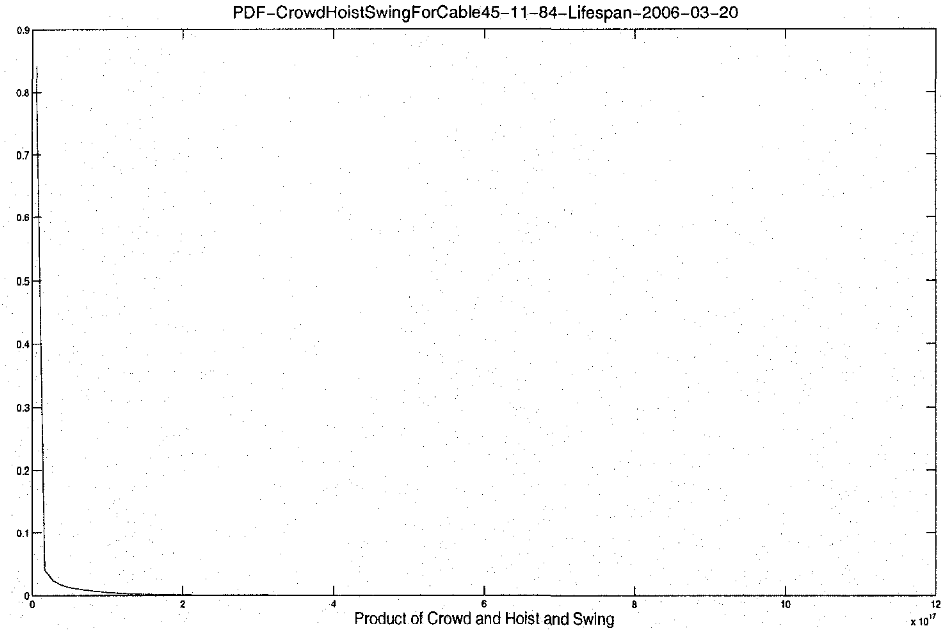


Figure 3.30: PDF of product of crowd and hoist and swing power for a cable of Shovel84.

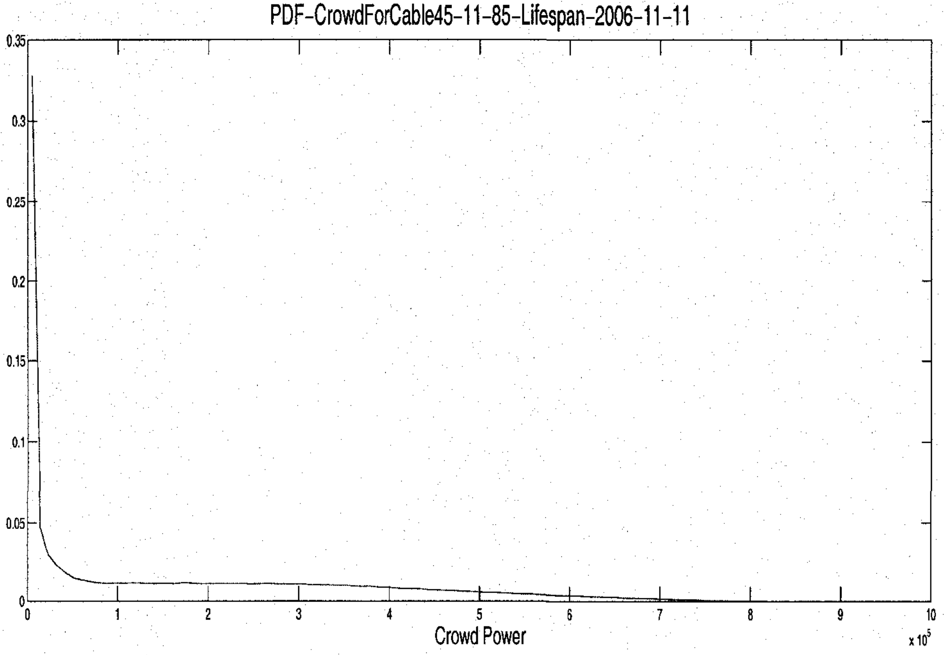


Figure 3.31: PDF of crowd power for a cable of Shovel85.

Chapter 4

Results

Assessing the effects of different factors such as power-related features, materials and operators, on the cable lifespan using the correlation method is not trivial. This is partially due to the lack of a standard data set, and the lack of previous works in the literature. In addition to these challenges, there are also the tasks of explaining the correlation values obtained in our framework, defining an evaluation measure, and choosing the sample data. Before presenting the experimental results, we will discuss the decision made with respect to these issues.

A major issue that must be addressed in scanning the ranges of the feature values with probabilities negatively correlated to the cable lifespan is the means through which the correlation is quantitatively assessed. The normal method in statistics has used the values of Pearson correlation coefficient that have been naturally defined as a measurement of the maximum negative correlation to the maximum positive correlation with its range of -1 to $+1$, and the values greater than 0.5 or less than -0.5 are considered acceptable indicators of high correlations. However, it represents a simplified ideal mathematical case. In practical cases, the mechanisms that cause positive or negative values are difficult to explain. In our work, the ranges with positive Pearson correlation coefficients may not be explained as 'beneficial' to the cable, because the power or energy applied to the cables would always be harmful to the cables even though the influences of power with low values may ignored. Another problem of the correlation coefficients is that in our experiments they are rarely beyond the scope of -0.5 to $+0.5$ although there are obvious differences among them. These problems cause explicit decisions hard to make. It was thus decided to ignore the ranges with positive coefficients and choose the ranges with the prominently highest negative coefficients less than -0.4.

In order to obtain an appropriate sample of cables, the specific cables should be chosen with prudence, especially for the cables with a great deal of time when the telemetry data were missing. It was noted that some periods with no data recorded in telemetry data were skipped as the shovels are not working. Although most of such periods are verified with the corresponding dispatch data, there are still a few of them for which it has been otherwise confirmed that the shovels were actually working; the durations of these intervals range from several minutes to several days. As the distributions of features in the skipped periods may have important information relevant to the cable lifespan, using the cables with too many of such periods may affect the test results. We chose the samples of cables using the following rule: if a cable has available telemetry data with total time not shorter than its lifespan by 15%, the cable will be considered as an acceptable sample.

In order to quantitatively assess whether there exist different effects on the cable lifespan between different teams or materials, we chose to use a simple descriptive statistic: sum of the counts of feature values for different teams and different materials for a certain range of feature values. First, we calculated the percentages of total counts for different teams or materials in the relevant segments; second, we compared these percentages to those for the whole range of the corresponding feature values. The first step may show differences between the contributions of the teams or materials to the relevant ranges; however, based on these differences, we cannot draw a conclusion that specific team or material with bigger contributions is the principal cause for the existence of the relevant ranges and thus more harmful to the cable lifespan, because the materials or teams could make the same contributions to the whole range as to the relevant segments. A material or team's higher contribution to the whole range may lead to its higher contribution to the relevant segments, since the most time of a cable lifespan was spent digging a material or operated by a team. Only though the comparison in the second step can we specify some team or material which are more harmful to the cables.

The resulting sample of the cables used in our work consisted of 77 cables with different lifespans. As we have noted, each shovel has four different teams, but one shovel's four teams are not the same as those of other shovels despite the same set of identifiers (TA, TB, TC, and TD) used for all shovels. Therefore, we conducted the team-specific experiments for each shovel not for all shovels; while we conducted the material-specific experiments both for individual shovels and for all shovels. To simplify the experiments and give the explanations on a uniform level, we used 100 bins to estimate all PDFs and CCDFs in the experiments.

This chapter is organized as a series of sections about related experiments, designed to learn more about our problem in their respective scenarios. Each section will describe one kind of experiment, which is motivated by a question, and then describe the experimental outline, the predicted results, the experimental results, and a discussion of the results.

4.1 Feature-Specific Correlation Experiments

Question: A first natural question to ask is, *For each feature, is there any segment (range of power-related feature values) which is really very harmful to the cables?*

Experimental Outline: To answer this question, the time series of all seven power-related features during the whole lifespans of 77 cables were extracted from the telemetry raw data and then their PDFs or CCDFs were estimated to be used for correlation mining described in the previous chapter. We obtained the values of the Pearson correlation coefficient for each feature for the whole cable lifespan using the built-in functions in Matlab.

Predicted Results: For each feature, the correlation coefficient is different for all segments of the feature distribution. Some segments of the feature have high negative correlation coefficients (< -0.5); while some have not. In other words, we can discern clearly the segments with high negative values from the correlation coefficient value curves.

Experimental Results: The Pearson correlation coefficients for PDF and CCDF of the seven features are shown in Figures 4.2 to 4.8 (page $58 - 61$) as a function of the bin number.

Discussion: The correlation coefficient is really different for all segments of each feature; however, only some features have segments (bins) with the correlation coefficients around -0.5.

For *product of hoist and crowd* (Figures 4.5), the negative correlation coefficients are around -0.5 from $20th$ to $34th$ bins with the highest negative of -0.5461 at $21th$ bin. From $20th$ to $100th$ bins, the value of correlation coefficient is increasing gradually from around -0.5 to 0.0; and all bins except the very leftmost ones before $20th$ bin have still the negative correlation coefficients. Therefore, the *product of hoist and crowd* should have the negative effects on the cable lifespan when bigger than some low value. This threshold should be around the $15th$ bin with the highest negative correlation coefficient of -0.4938, according to the value curve of the correlation coefficient for the CCDF. Therefore, we regard the value after the $15th$ bin, as a relevant segment of *product of hoist and crowd*.

The feature *product of crowd and swing* (Figures 4.6) has similar correlation coefficient value curves for both PDF and CCDF compared to the *product of hoist and crowd*. However, the fluctuation at the rightmost end of the curve for the PDF cause the curve for the CCDF in Figure 4.6 to be higher than -0.4 for each bin. For this reason, we could not find a relevant segment for *product of crowd and swing* in this experiment.

For the feature *hoist power* (Figures 4.3), the PDF's correlation coefficient values form an odd curve: most bins before the $50th$ bin have positive values with a global maximum at the $48th$ bin, while most bins thereafter have negative values with a distinct global minimum around the $75th$ bin (-0.4216, approaching -0.5). In this case, the positive coefficients should be ignored due to the reason we have given. For the negative coefficients, unlike the other features, *hoist power* has the highest negative correlation coefficients for only a narrow range around the $75th$ bin. Therefore, the bin 75 should be regarded as a *relevant* segment.

For the feature *crowd power* (Figures 4.2), the PDF's correlation coefficient values form a curve different from that of *hoist power*: most bins before the $50th$ bin have the negative values, while most bins thereafter have the positive values. However, there is no distinct global maximum or minimum on the curve. As we did for *hoist power*, only bins with negative coefficient values are scanned for the *relevant* ones. Finally, we chose the values from the $17th$ to $39th$ bins as a relevant segment since these bins have the highest negative coefficient values (all around -0.4).

For each of the other features, since no Pearson correlation coefficient for PDF and CCDF is lower than -0.4, we could not identify any segment of the feature value as *relevant* in this experiment.

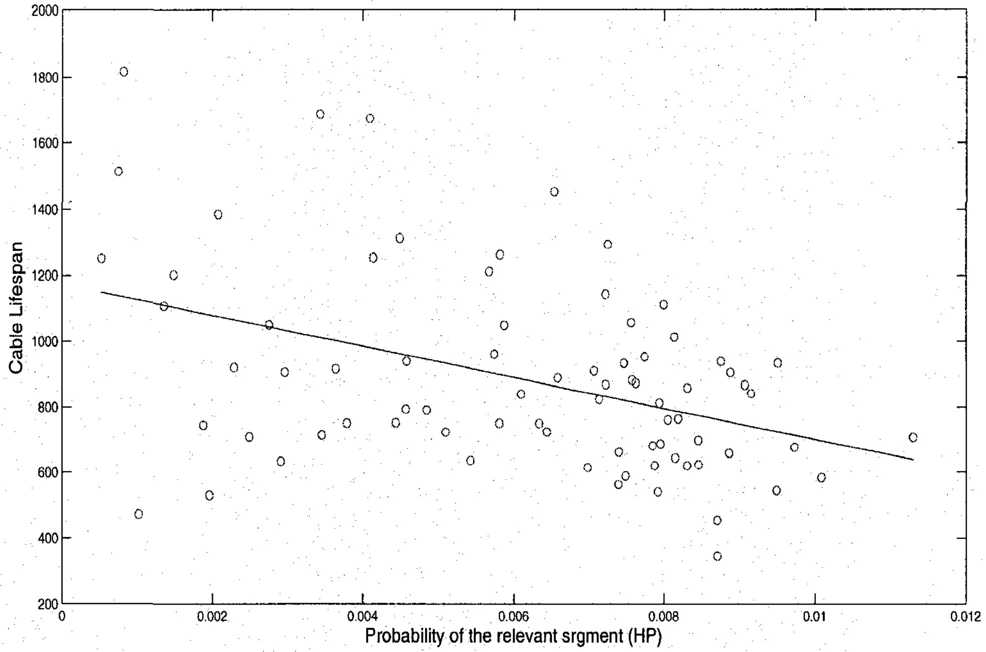


Figure 4.1: The relation of probability for 75th bin of hoist power distribution to the cable lifespan.

In this experiment, there are only few ranges of some feature values that have correlation coefficients lower than -0.5. For some features, there are a few segments whose correlation coefficients are the distinctly lowest even though they are only around -0.4. These segments could be meaningful and could be studied further. To include these segments into the *relevant* ones, we relax the threshold for negative correlation up to -0.4. Thus, the resulting *relevant* ranges of the power-related features are:

- CP: *crowd power* values from $1.9277 \times 10^5 watt$ to $4.6988 \times 10^5 watt$ (from the $17th$ to $39th$ bins).
- HP: *hoist Power* values from $1.3311 \times 10^6 watt$ to $1.3491 \times 10^6 watt$ (the $75th$ bin).
- HCP: *product of hoist and crowd* values from $1.9841 \times 10^{11} watt \times watt$ to $1.4172 \times 10^{12} watt \times watt$ (from the $15th$ to $100th$ bins).

Relevant means if the corresponding feature values fall in a relevant segment more often during the cable lifespan, then length of the cable lifespan would be shorter. Figure 4.1 shows the relation of the probability for a relevant segment (HP) with the cable lifespan. For this segment (HP), almost half points of the cables are around the regression line. Each material or team of operators may have different contributions to each relevant segment. By investigating these contributions and comparing them with the corresponding contributions to the whole cable lifespan, it is expected to find a specific material or team of operators which may be more harmful to the cables than the others. We used the above relevant segments in the following experiments to find the results with respect to the materials and the operators.

4.2 Material-Specific Experiments

4.2.1 Material-Specific Experiment 1

Question: *Is there a specific material (overburden, oilsand or the mixture of them) which have more negative effects on the cable lifespan than the other?*

Experimental Outline: In this experiment, we analyzed the contributions of different materials both to the relevant segments and to the whole cable lifespan using percentages of the total counts of the feature values in a range (a relevant segment or the whole lifespan) for different materials. The material with the percentage for a segment or a range of values (p_s) higher than the percentages of other materials was regarded as contributing more to the segment. Compared with its percentage of total counts for the whole cable (p_w), a material could produce the feature values more frequently falling in the relevant segment than falling in an *average* segment if its p_s greater than its p_w. For each material, we checked if any of its p_s was greater than its p_w. We chose the material with the distinctly largest positive difference between its p_s and its p_w (for $p_s > p_w$) for all relevant segments as the principal cause for the relevant segments overall, and therefore more harmful to the cables than other materials. Thus, the procedure used in this experiment is: for each of the relevant segments obtained in the previous experiment, first, the percentage of total counts of each material (p_s) were calculated ; then, (p_s) was compared to the material's percentage for the whole lifespan (p_w) to check if $p_s > p_w$ or not. The materials whose contributions were compared are:

- Oilsand: the material containing oil (TS).
- Overburden: the waste above the topsoil (OB).
- Mixture: the mixture of topsoil and overburden (OTS).

The ranges of different feature values for which different material's contributions were analyzed are:

- CP: the relevant segment of *crowd power* presented in the previous experiment.
- HP: the relevant segment of *hoist power* presented in the previous experiment.
- HCP: the relevant segment of *product of hoist and crowd* presented in the previous experiment.
- WHL: the range of the whole cable lifespan.

Predicted Results: It is expected that some material could make more contributions to the *relevant* segments than its contribution to the whole lifespan with obvious differences.

Experimental Results: The percentages of total counts of different materials for each relevant segment and the whole lifespan are shown in Figure4.9 (page 62).

Discussion: Oilsand (TS) has the highest percentages for all relevant segments (CP, HP, and HCP) and the whole lifespan (WHL), and its percentage for each relevant segments is higher than its percentage for the whole cable lifespan with the largest difference occurring at the relevant segment of hoist power (HP). The percentages of mixture (OTS) are almost the same in all relevant segments and the whole lifespan. The percentage of overburden (OB) for each relevant segment is lower than its percentage for the whole cable lifespan. These results indicate that oilsand is more harmful to the cables than the overburden and mixture, although this difference was not always significant for all relevant segments. Specifically, oilsand's contribution to the relevant segment of hoist power (HP) was significantly more than that to the whole lifespan, while its contribution to other two relevant segments (CP and HCP) was not significantly more.

4.2.2 Material-Specific Experiment 2

Question: *For each shovel, is there a specific material (overburden, oilsand or the mixture of them) which have more negative effects on the cable lifespan than the other?* Specifically, what we would like to explore in this experiment is whether the conclusion drawn in the previous experiment is applicable to each shovel separately.

Experimental Outline: The same procedure as in the previous one was used in this experiment. The experiment was done for each of the eight shovels.

Predicted Results: It is expected that the same result in the previous experiment could be obtained for each shovel.

Experimental Results: The percentages of total counts of different materials for each relevant segment and the whole cable lifespan are shown in Figures 4.10 to 4.17 (page 62 – 66) for eight different shovels.

Discussion: With the exception of Shovel82 and Shovel84, oilsand (TS) is more harmful to the cables than other materials, especially for the relevant segment of hoist power (HP).

For Shovel78 (Figure 4.10), TS has the highest percentages for all relevant segments (CP, HP, and HCP) and the whole lifespan (WHL), and its percentages for all relevant segments are higher than its percentage for the whole cable lifespan with the largest difference occurring at the relevant segment of hoist power (HP). For other materials, only OTS's percentage for CP is slightly higher than its percentage for WHL. These results are very similar to those in the previous experiment. Therefore, oilsand (TS) is more harmful to the cables than other materials, especially for the relevant segment of hoist power (HP).

For Shovel79 (Figure 4.11), TS has its percentages for HP and HCP higher than its percentage for WHL, while OB has the higher percentage for CP. TS has the largest difference occurring at HP. Therefore, oilsand (TS) is more harmful to the cables than other materials, especially for the relevant segment of hoist power (HP).

For Shovel80 (Figure 4.12), OTS has its percentages for CP and HCP slightly higher than its percentage for WHL, while TS has its percentages for CP, HP, and HCP higher than its percentage for WHL with the largest difference occurring at HP. Therefore, oilsand (TS) is more harmful to the cables than other materials, especially for the relevant segment of hoist power (HP).

For Shovel81 (Figure 4.13), TS has the largest difference obviously observed at HP, while other differences are more obscure. Therefore, oilsand (TS) is more harmful to the cables than other materials, especially for the relevant segment of hoist power (HP).

For Shovel82 (Figure 4.14), no distinct differences can be observed at CP, HP, and HCP for each material.

For Shovel83 (Figure 4.15), Only TS has its percentages for CP, HP, and HCP higher than its percentage for WHL. Therefore, oilsand (TS) is more harmful to the cables than other materials.

For Shovel84 (Figure 4.16), no distinct differences can be observed at CP, HP, and HCP for each material.

For Shovel85 (Figure 4.17), Only TS has its percentages for CP, HP, and HCP higher than its percentage for WHL. Therefore, oilsand (TS) is more harmful to the cables than other materials.

Although each shovel spent a different ratio of its working time in digging oilsand (lowest ratio for Shovel81 and highest ratio for Shovel85), oilsand's negative effects on the cable lifespan were always more than other materials according to the results. Notably, for all shovels except Shovel82 and Shovel84, oilsand's contributions to the relevant segment of hoist power (HP) were always the highest relative to its contributions to other relevant segments and the whole cable lifespan. The range of HP should be investigated in more detail.

4.3 Operator-Specific Experiments

Question: *Are there any specific teams for each shovel whose operation practices are more harmful to the cables than others?*

Experimental Outline: In this experiment, we used a procedure similar to the ones in the previous experiment. The experiment was done for each individual shovel to compare each team's contribution to a relevant segment with its contribution to the whole cable lifespan.

Predicted Results: Each shovel has some specific team which has more contributions to the relevant segments than its contribution to the whole cable lifespan.

Experimental Results: The percentages of total counts of different teams for each relevant segment and the whole lifespans are shown in Figures4.18 to 4.25 (page 66 – 70) for the eight different shovels.

Discussions: The contributions of different teams to the relevant segments varied for each shovel. While for some shovels all teams perform at the same level overall, for other shovels one or two teams contribute more to the relevant segments than they contribute to the whole cable lifespan.

For Shovel78 (Figure 4.18), team TB has its percentage for CP slightly higher than its percentage for WHL, team TC has its percentages for CP and HCP slightly higher than its percentage for WHL, and team TD has its percentages for HP and HCP higher than its percentage for WHL. TD has the largest difference occurring at HP. Overall, TD contributes the most to the relevant segments, and therefore its operation practices are more harmful than other teams.

For Shovel79 (Figure 4.19), team TA has its percentages for CP and HCP slightly higher than its percentage for WHL, team TC has its percentage for CP slightly higher than its percentage for WHL, and team TD has its percentages for HP and HCP higher than its percentage for WHL. TD has the largest difference occurring at HP. Overall, TD contributes the most to the relevant segments, and therefore its operation practices are more harmful than other teams.

For Shovel80 (Figure 4.20), team TB has its percentages for CP and HP slightly higher than its percentage for WHL, team TC has its percentage for CP slightly higher than its percentage for WHL, and team TD has its percentages for HP and HCP higher than its percentage for WHL. TD has the largest two differences occurring at HP and HCP. Overall, TD contributes the most to the relevant segments, and therefore its operation practices are more harmful than other teams.

For shovel81 (Figure 4.21), team TA has its percentages for CP and HCP higher than its percentage for WHL, team TB has its percentage for HP higher than its percentage for WHL, team TC has its percentages for CP and HP higher than its percentage for WHL, and team TD has its percentage for HCP higher than its percentage for WHL. Overall, we could not find a specific team which contributes the most to the relevant segments.

For shovel82 (Figure 4.22), team TA has its percentage for CP slightly higher than its percentage for WHL, team TB has its percentage for HP slightly higher than its percentage for WHL, team TC has its percentage for HCP slightly higher than its percentage for WHL, and team TD has its percentages for HP and HCP slightly higher than its percentage for WHL. Overall, we would not find a specific team which contributes the most to the relevant segments.

For Shovel83 (Figure 4.23), team TA has its percentage for CP slightly higher than its percentage for WHL, team TB has its percentages for CP, HP and HCP higher than its percentage for WHL, and team TC has its percentage for CP slightly higher than its percentage for WHL. TB has the largest two differences occurring at HP and HCP. Overall, TB contributes the most to the relevant segments, and therefore its operation practices are more harmful than other teams.

For Shovel84 (Figure 4.24), team TA has its percentages for CP, HP and HCP higher than its percentage for WHL, and team TB has its percentage for HCP slightly higher than its percentage for WHL. TA has the largest difference occurring at HP. Overall, TA contributes the most to the relevant segments, and therefore its operation practices are more harmful than other teams.

For Shovel85 (Figure 4.25), team TA has its percentages for CP, HP and HCP slightly higher than its percentage for WHL, and team TD has its percentages for CP, HP and HCP higher than its percentage for WHL. TD has the largest difference occurring at HP. Overall, TD contributes the most to the relevant segments, and therefore its operation practices are more harmful than other teams.TD contributed mostly to the relevant segments, especially on HP.

Some researches on oil shovels operation determined the indicators of shovel performance and then, using the indicators, analyzed the influence of some factors, such as, dipper types and operators, on the performance [27][25][26]. [27] and [25] pointed out that hoist power is the best indicator of shovel performance due to its better representation of diggability of materials than other indicators. In our experiment, the relevant segment of hoist power (HP) is the best to distinguish the different influence on the cable lifespan between different teams or between different materials. Actually, our conclusions specific to the teams or the materials would not be changed if using only HP. The difference between our work and the others is that only a segment of hoist power value was chosen in our work as *relevant*; while the average hoist power consumed in dig cycles was calculated to analyze the shovel performance in [27][25][26].

The operator practices are anticipated to influence the shovel performance. [26] analyzed the operator influence on shovel performance using the average hoist power in dig cycles because the average hoist power is more significantly different for the four teams than other indicators. The tests in [26] show some teams would consume higher hoist power than other teams for the same uniform material, therefore the different practice between teams would influence the shovel performance. It could be anticipated that a team consumed higher hoist power would have more harmful influence on the cables. However, whether each team's influence on the shovel performance is coincident with its influence on the cable lifespan needs further studying.

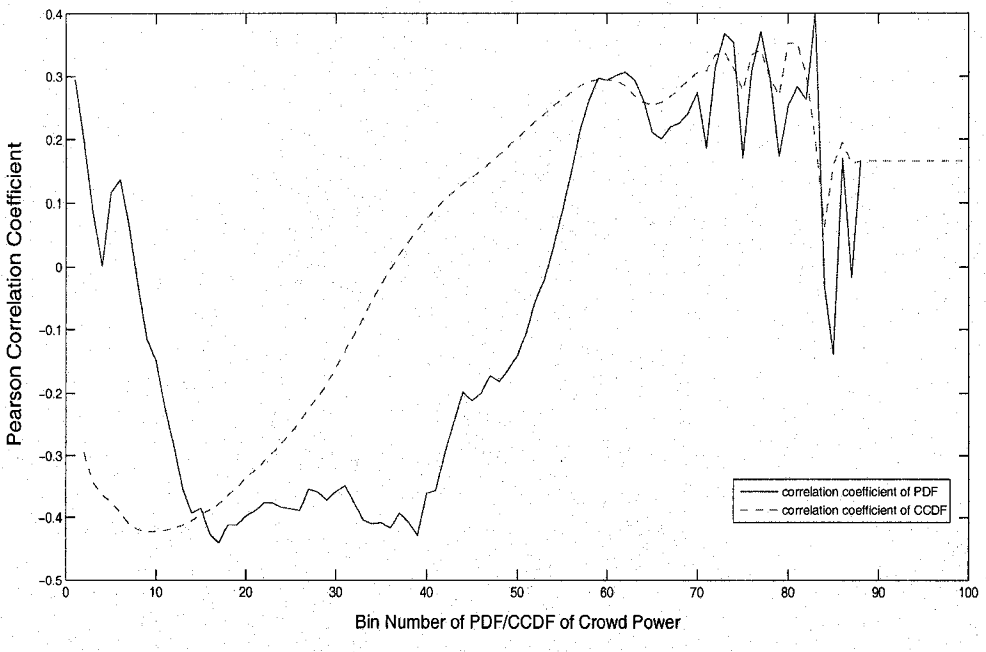


Figure 4.2: Pearson correlation coefficient of crowd power as functions of bin number.

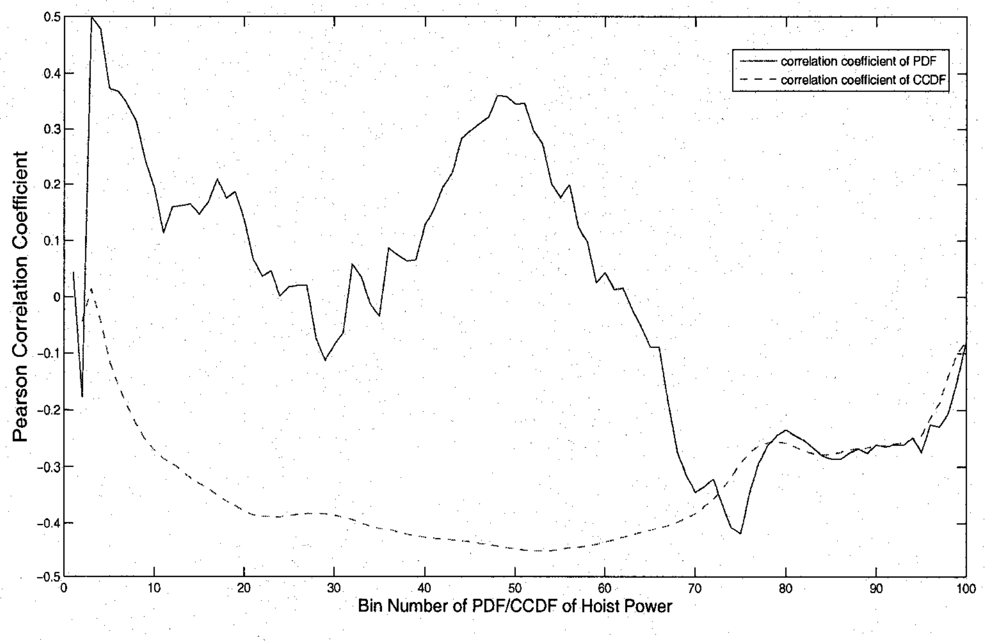


Figure 4.3: Pearson correlation coefficient of hoist power as functions of bin number.

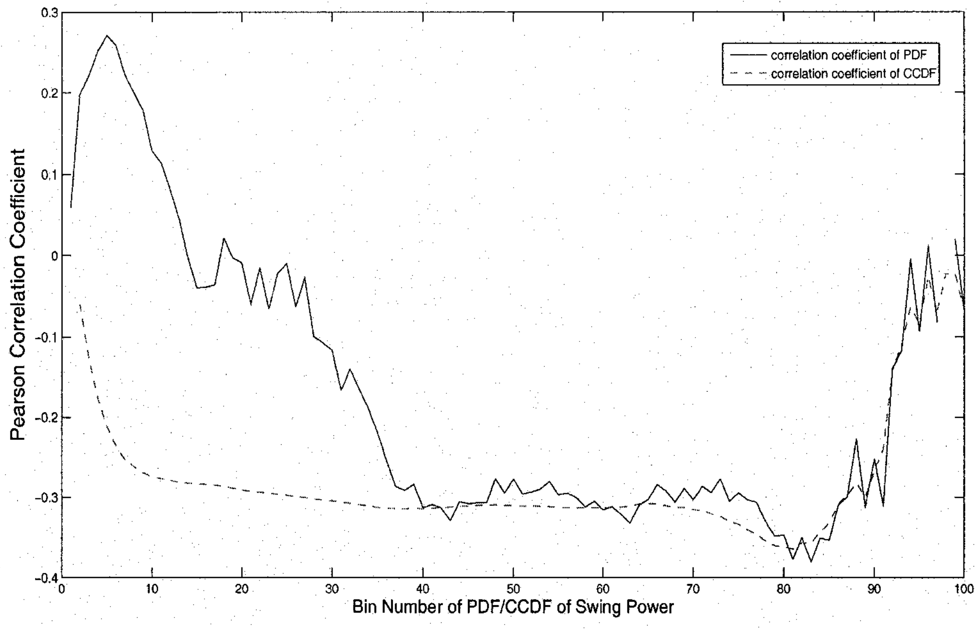


Figure 4.4: Pearson correlation coefficient of swing power as functions of bin number.

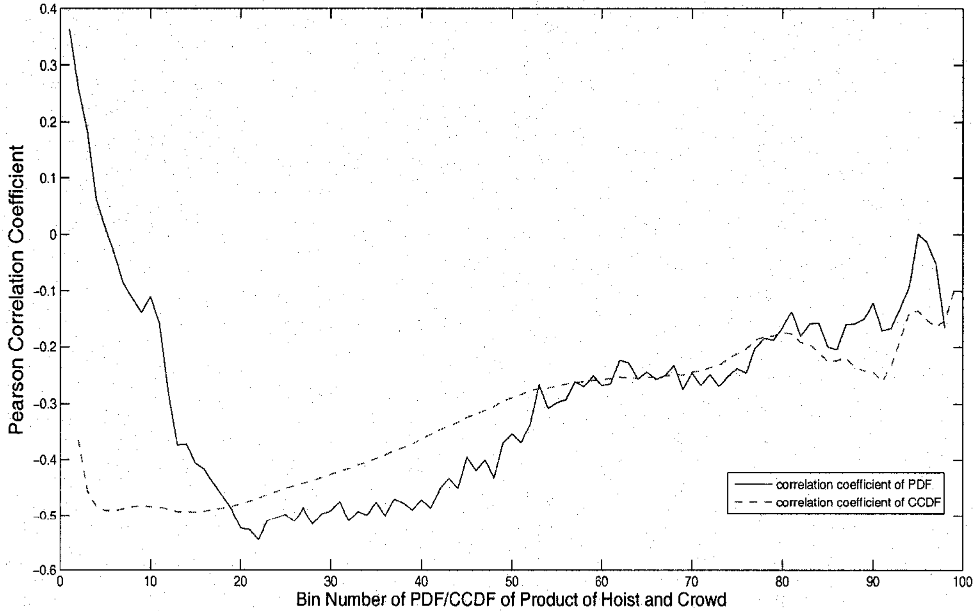


Figure 4.5: Pearson correlation coefficient of product of hoist and crowd as functions of bin number.

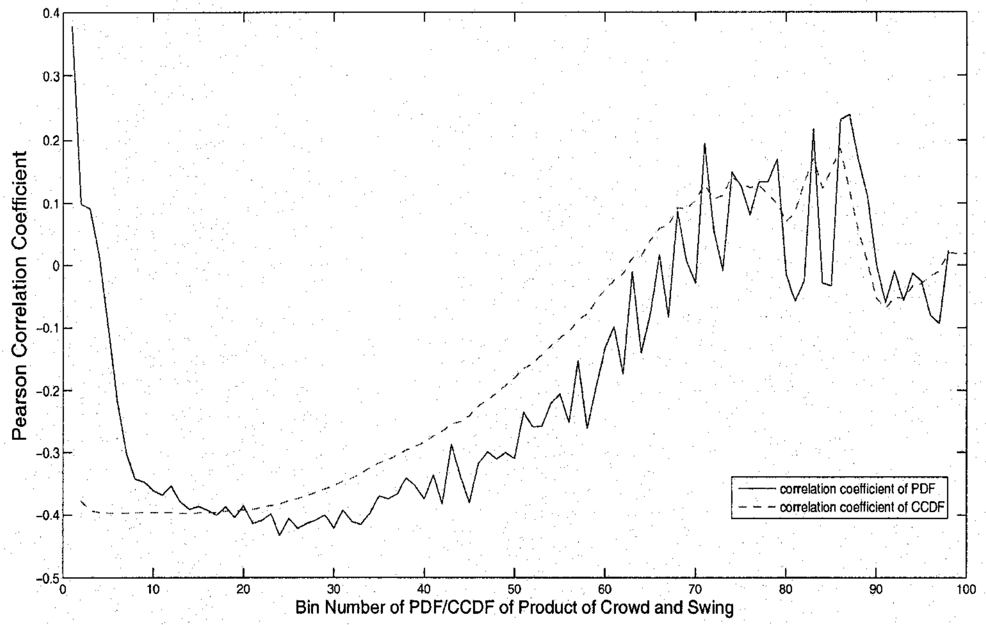


Figure 4.6: Pearson correlation coefficient of product of crowd and swing as functions of bin number.

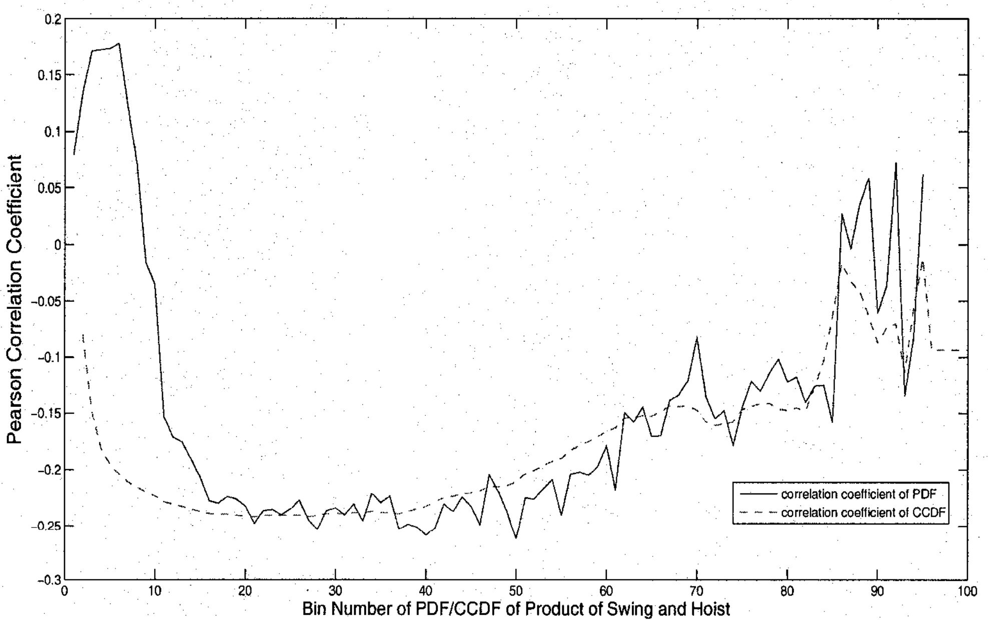


Figure 4.7: Pearson correlation coefficient of product of swing and hoist as functions of bin number.

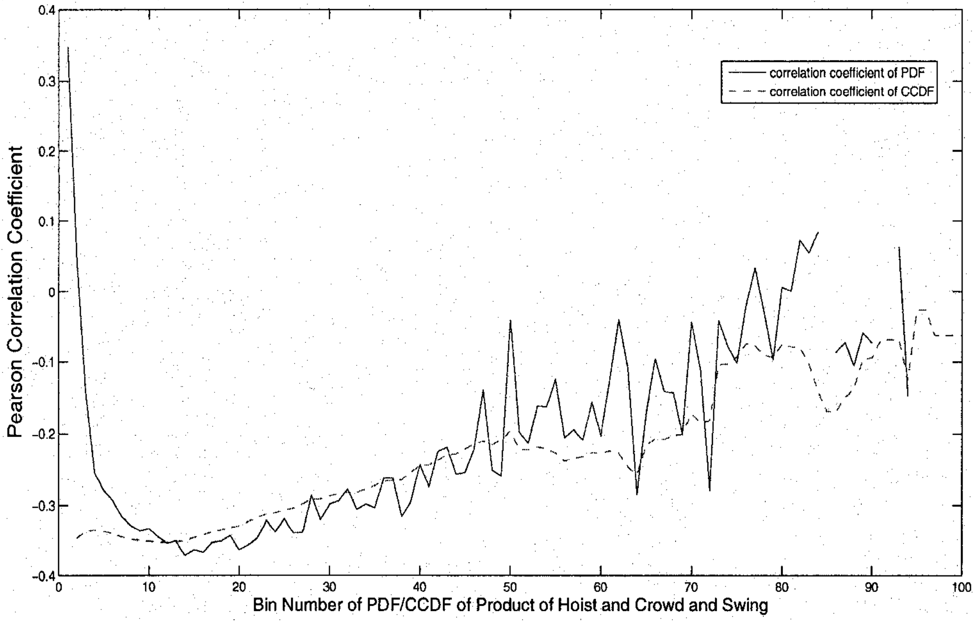
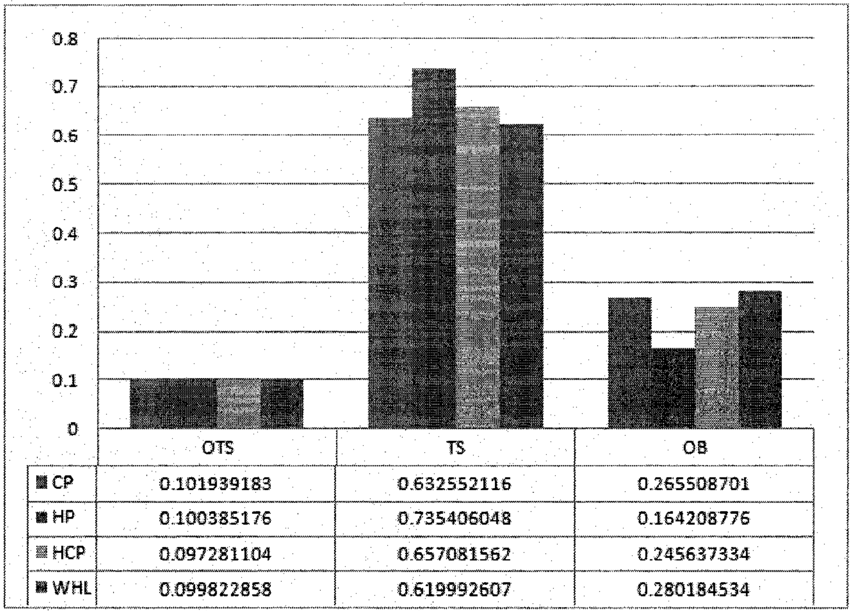
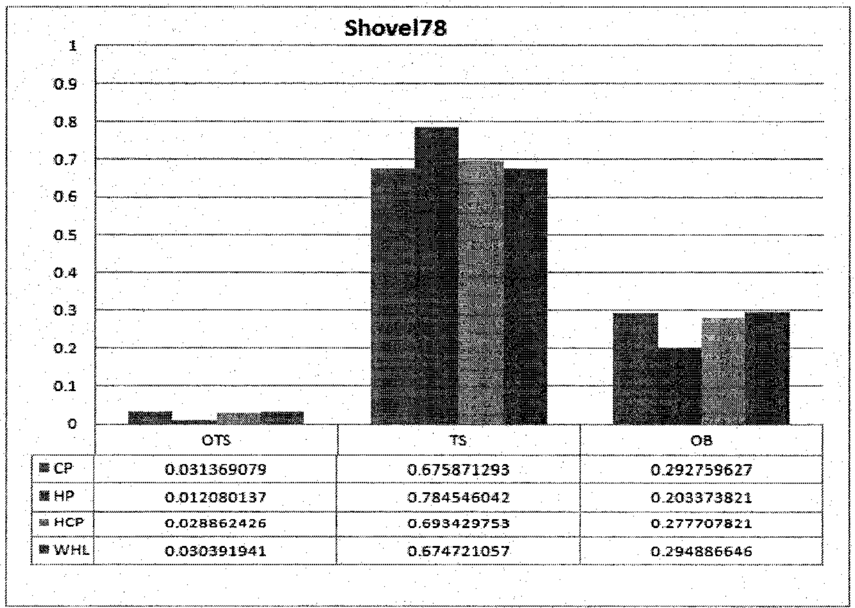


Figure 4.8: Pearson correlation coefficient of product of crowd and hoist and swing as functions of bin number.



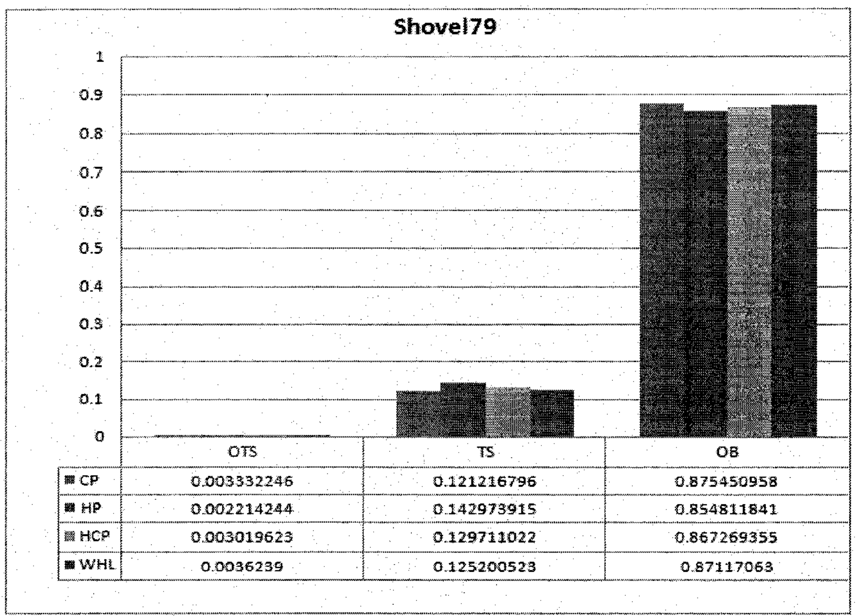
	OTS	TS	OB
■ CP	0.101939183	0.632552116	0.265508701
■ HP	0.100385176	0.735406048	0.164208776
▨ HCP	0.097281104	0.657081562	0.245637334
■ WHL	0.099822858	0.619992607	0.280184534

Figure 4.9: Percentages of total counts for different materials.



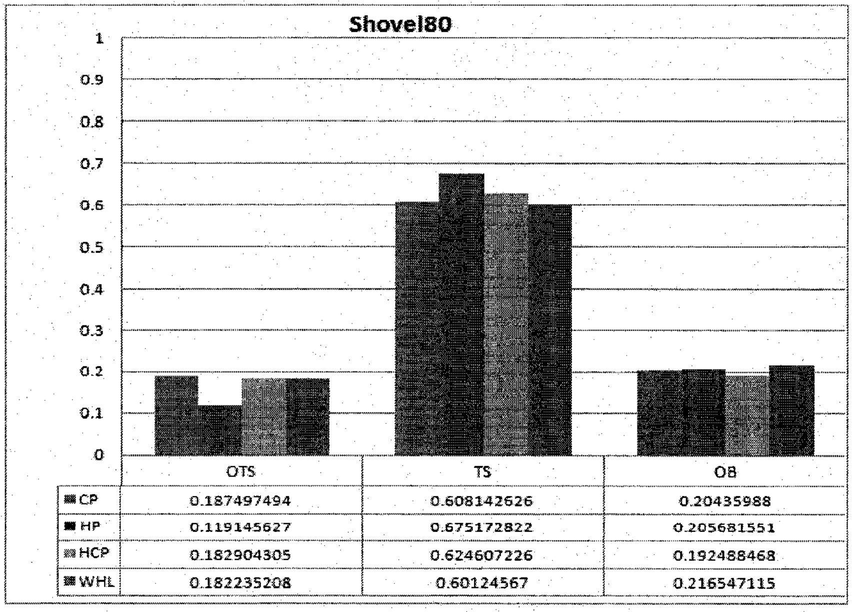
	OTS	TS	OB
■ CP	0.031369079	0.675871293	0.292759627
■ HP	0.012080137	0.784546042	0.203373821
▨ HCP	0.028862426	0.693429753	0.277707821
■ WHL	0.030391941	0.674721057	0.294886646

Figure 4.10: Percentages of total counts for different materials for Shovel78



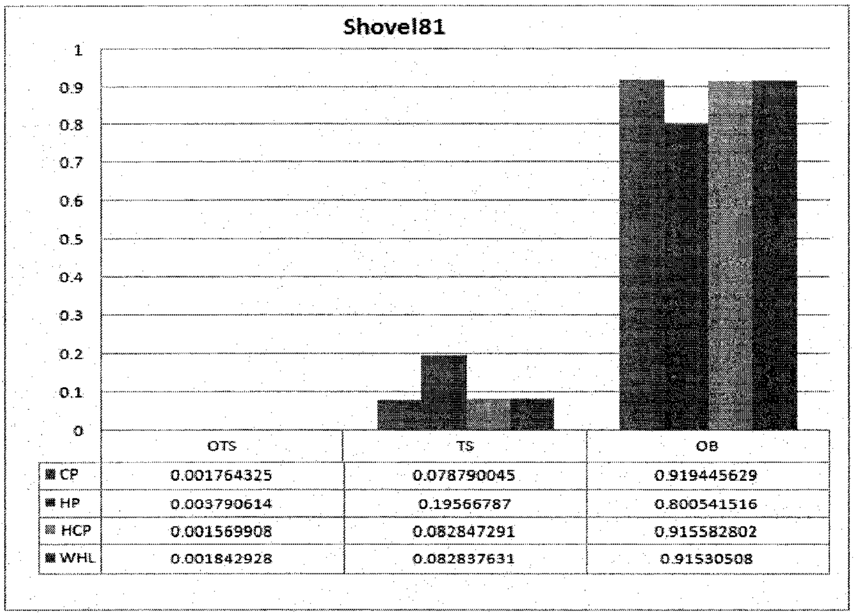
	OTS	TS	OB
▨ CP	0.003332246	0.121216796	0.875450958
■ HP	0.002214244	0.142973915	0.854811841
▨ HCP	0.003019623	0.129711022	0.867269355
■ WHL	0.0036239	0.125200523	0.87117063

Figure 4.11: Percentages of total counts for different materials for Shovel79



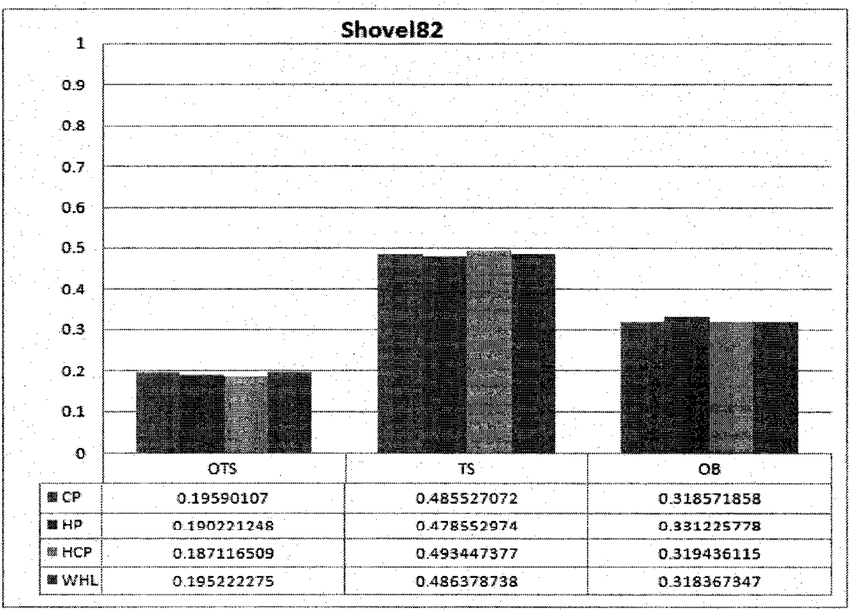
	OTS	TS	OB
▨ CP	0.187497494	0.608142626	0.20435988
■ HP	0.119145627	0.675172822	0.205681551
▨ HCP	0.182904305	0.624607226	0.192488468
■ WHL	0.182235208	0.60124567	0.216547115

Figure 4.12: Percentages of total counts for different materials for Shovel80



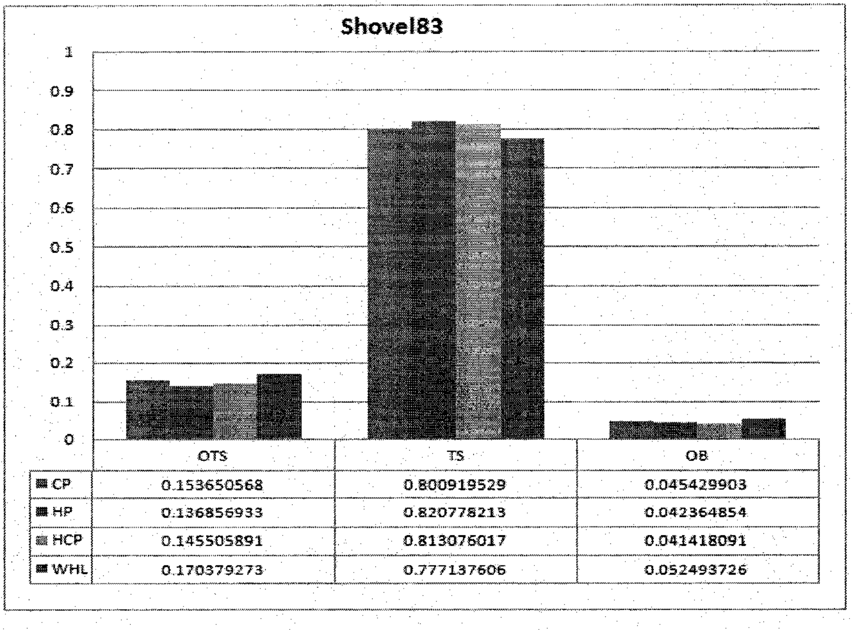
	OTS	TS	OB
▨ CP	0.001764325	0.078790045	0.919445629
▨ HP	0.003790614	0.19566787	0.800541516
▨ HCP	0.001569908	0.082847291	0.915582802
▨ WHL	0.001842928	0.082837631	0.91530508

Figure 4.13: Percentages of total counts for different materials for Shovel81



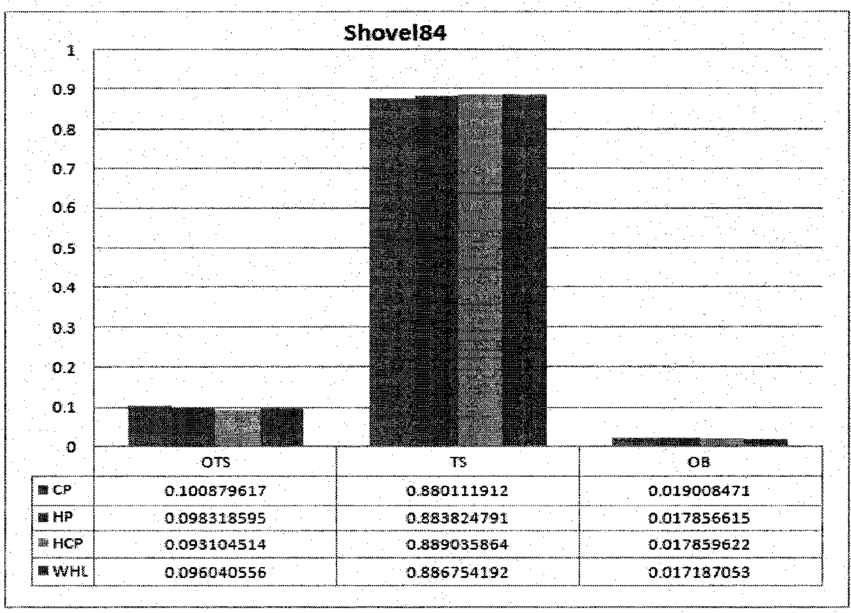
	OTS	TS	OB
▨ CP	0.19590107	0.485527072	0.318571858
▨ HP	0.190221248	0.478552974	0.331225778
▨ HCP	0.187116509	0.493447377	0.319436115
▨ WHL	0.195222275	0.486378738	0.318367347

Figure 4.14: Percentages of total counts for different materials for Shovel82



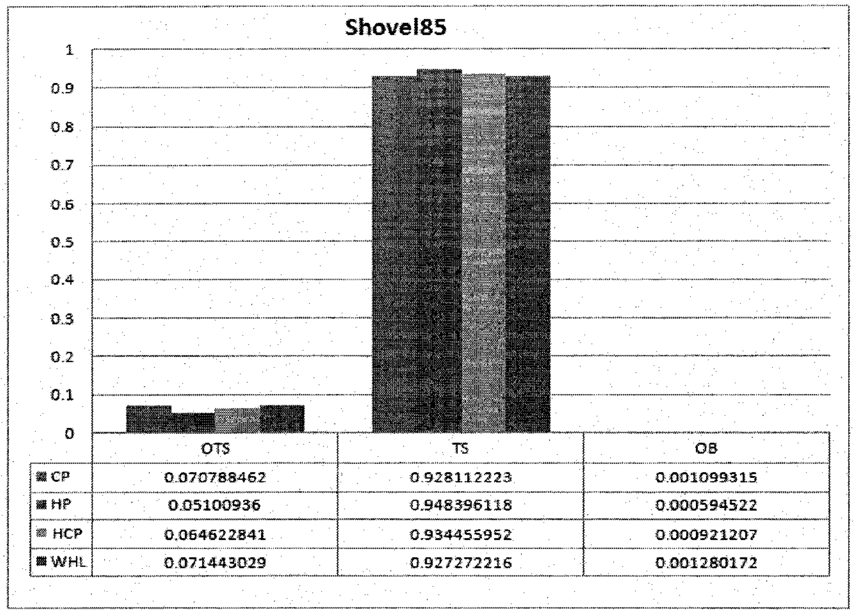
	OTS	TS	OB
■ CP	0.153650568	0.800919529	0.045429903
■ HP	0.136856933	0.820778213	0.042364854
▨ HCP	0.145505891	0.813076017	0.041418091
■ WHL	0.170379273	0.777137606	0.052493726

Figure 4.15: Percentages of total counts for different materials for Shovel83



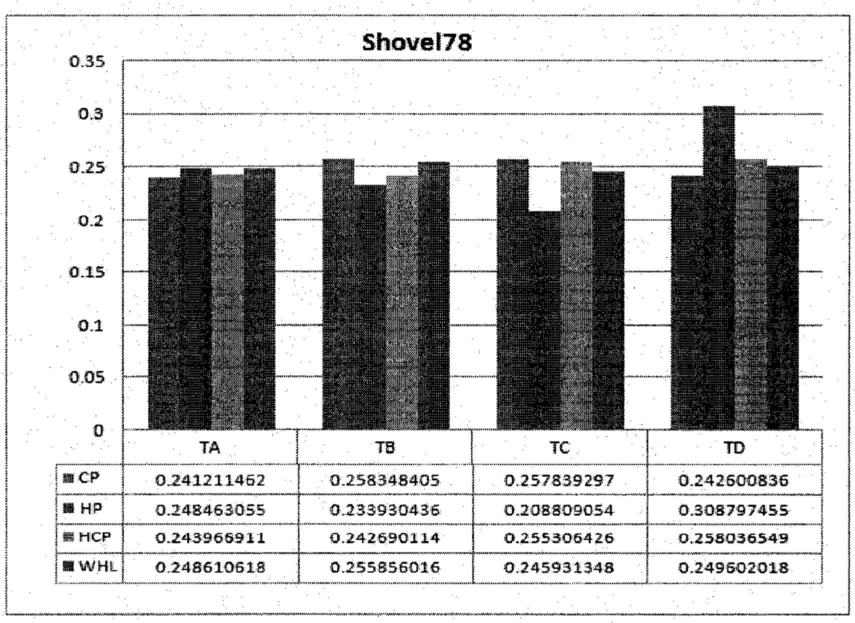
	OTS	TS	OB
■ CP	0.100879617	0.880111912	0.019008471
■ HP	0.098318595	0.883824791	0.017856615
▨ HCP	0.093104514	0.889035864	0.017859622
■ WHL	0.096040556	0.886754192	0.017187053

Figure 4.16: Percentages of total counts for different materials for Shovel84



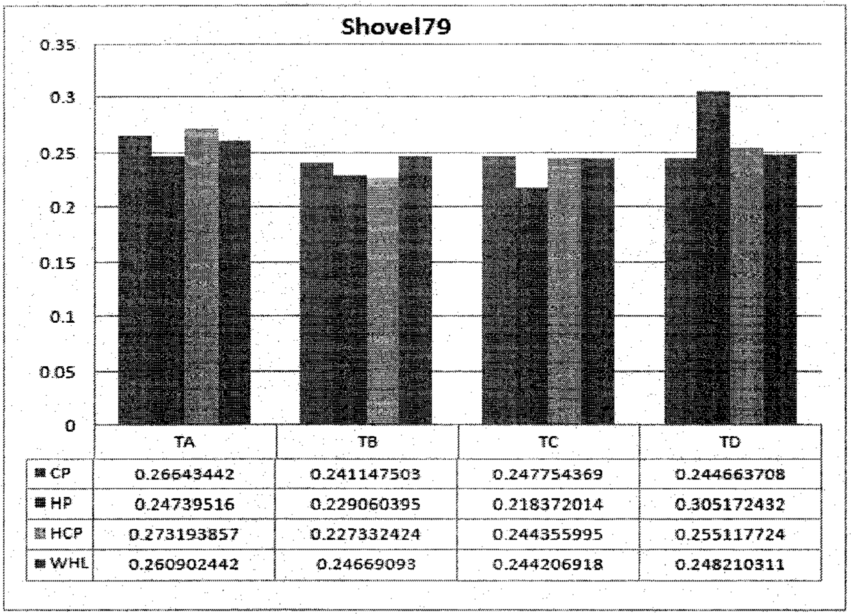
	OTS	TS	OB
▨ CP	0.070788462	0.928112223	0.001099315
■ HP	0.05100936	0.948396118	0.000594522
▨ HCP	0.064622841	0.934455952	0.000921207
■ WHL	0.071443029	0.927272216	0.001280172

Figure 4.17: Percentages of total counts for different materials for Shovel85



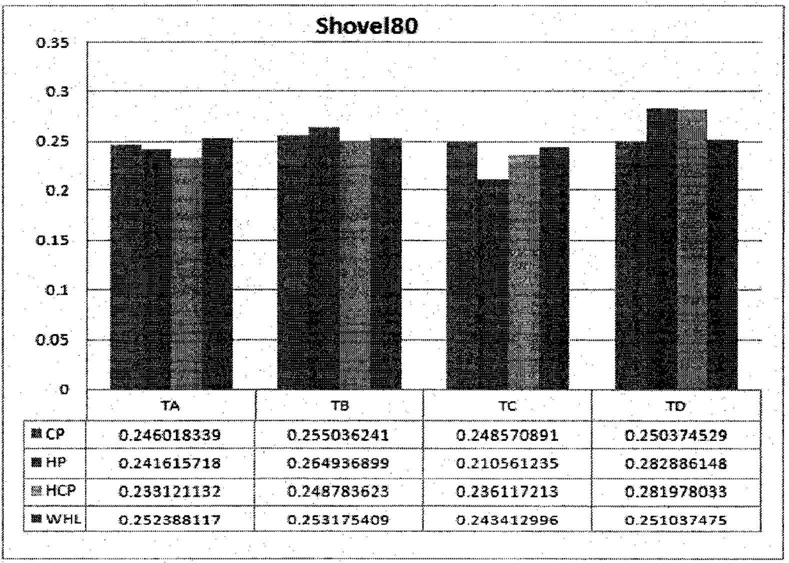
	TA	TB	TC	TD
▨ CP	0.241211462	0.258348405	0.257839297	0.242600836
■ HP	0.248463055	0.233930436	0.208809054	0.308797455
▨ HCP	0.243966911	0.242690114	0.255306426	0.258036549
■ WHL	0.248610618	0.255856016	0.245931348	0.249602018

Figure 4.18: Percentages of total counts for different teams for Shovel78



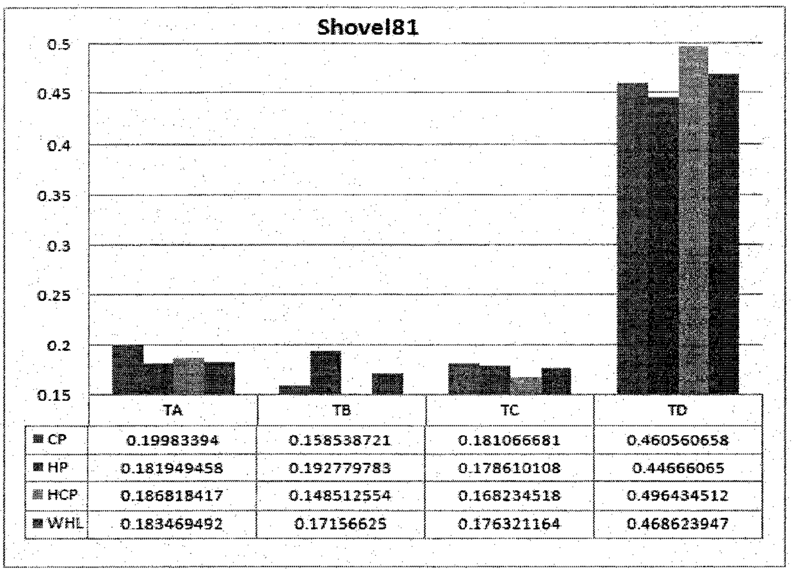
	TA	TB	TC	TD
▨ CP	0.26643442	0.241147503	0.247754369	0.244663708
▨ HP	0.24739516	0.229060395	0.218372014	0.305172432
▨ HCP	0.273193857	0.227332424	0.244355995	0.255117724
▨ WHL	0.260902442	0.24669093	0.244206918	0.248210311

Figure 4.19: Percentages of total counts for different teams for Shovel79



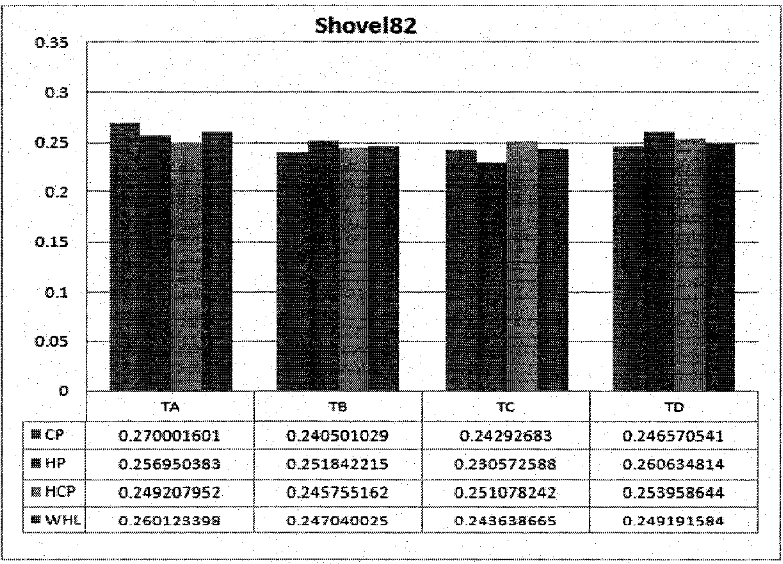
	TA	TB	TC	TD
▨ CP	0.246018339	0.255036241	0.248570891	0.250374529
▨ HP	0.241615718	0.264936899	0.210561235	0.282886148
▨ HCP	0.233121132	0.248783623	0.236117213	0.281978033
▨ WHL	0.252388117	0.253175409	0.243412996	0.251037475

Figure 4.20: Percentages of total counts for different teams for Shovel80



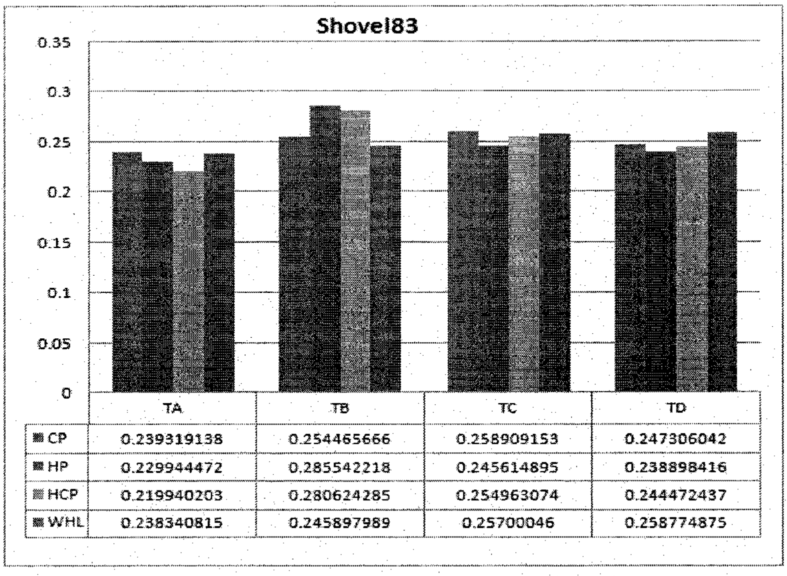
	TA	TB	TC	TD
▩ CP	0.19983394	0.158538721	0.181066681	0.460560658
▩ HP	0.181949458	0.192779783	0.178610108	0.44666065
▩ HCP	0.186818417	0.148512554	0.168234518	0.496434512
▩ WHL	0.183469492	0.17156625	0.176321164	0.468623947

Figure 4.21: Percentages of total counts for different teams for Shovel81



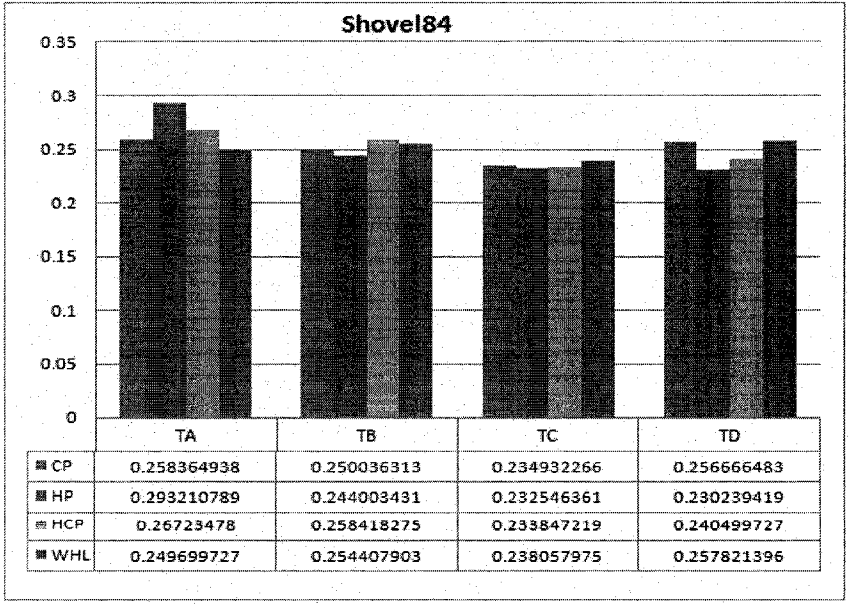
	TA	TB	TC	TD
▩ CP	0.270001601	0.240501029	0.24292683	0.246570541
▩ HP	0.256950383	0.251842215	0.230572588	0.260634814
▩ HCP	0.249207952	0.245755162	0.251078242	0.253958644
▩ WHL	0.260123398	0.247040025	0.243638665	0.249191584

Figure 4.22: Percentages of total counts for different teams for Shovel82



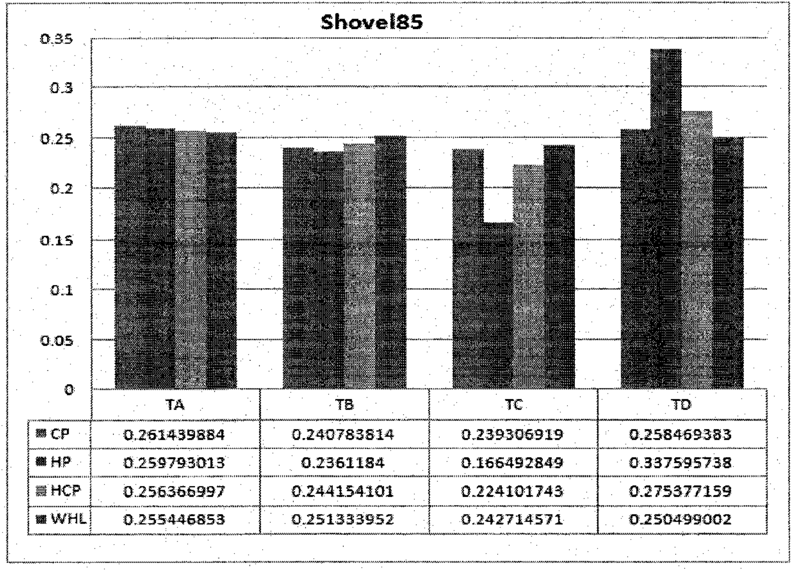
	TA	TB	TC	TD
▨ CP	0.239319138	0.254465666	0.258909153	0.247306042
▨ HP	0.229944472	0.285542218	0.245614895	0.238898416
▨ HCP	0.219940203	0.280624285	0.254963074	0.244472437
▨ WHL	0.238340815	0.245897989	0.25700046	0.258774875

Figure 4.23: Percentages of total counts for different teams for Shovel83



	TA	TB	TC	TD
▨ CP	0.258364938	0.250036313	0.234932266	0.256666483
▨ HP	0.293210789	0.244003431	0.232546361	0.230239419
▨ HCP	0.26723478	0.258418275	0.233847219	0.240499727
▨ WHL	0.249699727	0.254407903	0.238057975	0.257821396

Figure 4.24: Percentages of total counts for different teams for Shovel84



	TA	TB	TC	TD
▨ CP	0.261439884	0.240783814	0.239306919	0.258469383
▨ HP	0.259793013	0.2361184	0.166492849	0.337595738
▨ HCP	0.256366997	0.244154101	0.224101743	0.275377159
▨ WHL	0.255446853	0.251333952	0.242714571	0.250499002

Figure 4.25: Percentages of total counts for different teams for Shovel85

Chapter 5

Conclusion

The previous chapters have introduced the problem of mining the ranges of power-related feature values relevant to the shovel cable lifespan. The study of this problem is practically motivated, but has properties that make it an interesting and challenging time series data mining task. Chapter 3 introduced a framework that combined ideas from the related work into a general method to perform this task, notably expanding on the ideas of correlation-mining in probability distributions. Chapter 4 evaluated this framework with feature-specific, material-specific, and operator-specific experiments. This chapter will begin by a summary of the contributions made in this thesis. This will be followed by a discussion of potential future directions of research that directly follow from this work.

5.1 Contribution

This thesis has presented a framework and an implementation of this framework for the research on oil-shovel cable lifespan. The main components of this framework include Telemetry Data Preprocessing, Energy-Specific Feature Extraction, Dispatch Data Merging, Feature PDF Generation, and finally Correlation Mining. For each of these components, a method as simple as possible was incorporated into the system to perform the task. In addition, a new method was introduced for pattern mining in time series data, which allows robust estimation of non-sequential patterns correlated with events external to the raw time series.

This work showed that some ranges of feature values have higher negative effects on the cable lifespan if such values occur more frequently in the whole cable lifespan. We carried out related experiments using these *relevant* segments, including a wide range in the lower half of *crowd power* values, a wide range from the lower half to the higher end of *product of hoist and crowd* values, and a narrow range in the higher half of *hoist power* values. The frequency of *hoist power* value in this last relevant segment (HP) showed the largest difference to the overall frequency across all materials and teams and therefore needs to be explored further.

In our material-specific experiments where tests were performed on the shares of different materials in the relevant segments and in the whole cable lifespan, we presented that oilsand was more harmful to the cable than overburden. In addition, the material-specific experiments indicate that the conclusion applies to most of eight shovels. Our operator-specific experiments, where tests were performed on the shares of different teams for each shovel, indicate that some teams have more harmful influences on the cable lifespan than others. These conclusions could potentially lead to immediate practical use in the working schedules of the shovels and the training of the shovel operators.

5.2 Future Directions

For the framework described in Chapter 3, potential improvements for each step of the system were suggested in Chapter 3. They include: testing armature voltages and currents in the telemetry data as features, testing with the unused information (*e.g.*, loading information) in the dispatch data set, testing non-parametric correlation measures, and testing different number of bins.

With respect to the correlation pattern mining in this work, there are several obvious future directions to explore. Based on the experimental results, it is likely that significantly *relevant* segment of hoist power values would be expected if further separation was made to the HP segment (the relevant segment of hoist power). This segment always produced in our experiments the clearest discrimination between materials or teams overall, even though the case was not true for few shovels. The experimental results also suggest that mining the relevant segments specific for each shovel might result in more *robust* material-specific or operator-specific conclusions for each shovel, since the relevant segments we found may not fit for some shovels according to the experiment results. It is also likely that a more effective separation method which leads to the more narrow segments could improve results. This is based on visual analysis of the experiment results, where it is clear that the narrower segment (HP) increased the discrimination between materials and teams with respect to the contributions to the relevant segments. Incorporating other factors, such as climate and geology, also represent promising future directions.

In addition to correlation mining, it is worthwhile investigating the underlying mechanism that causes the cable lifespan to be negatively correlated to the relevant segments. In Figure 4.1, only half numbers of cables are around the regression line. These cables need be explored to find what make them different than others.

Our study found the hoist power is more important than other features in our correlation, material-specific, and operator-specific experiments. This coincides with the conclusion in the research on shovel performance. The further study may be done on the relationship between the shovel performance and the cable lifespan. A possible direction is to test whether the *better* shovel performance would cause the longer cable lifespan. It is also possible to test whether the operators with better influence on the performance would have less harmful influence on the cables.

Related to the use of our framework, is the potential to apply this system to different tasks. In Chapter 4, we performed 3 different types of experiments. It is clear that the system could be used in related tasks such as correlation mining for long lifespan cables, short lifespan cables, shovels working at different sites, or shovels with other different features. It is possible that the framework could also be applied for tasks that are not directly related to cable lifespan. For example, it could also be used in systems in which the correlation should be mined between the telemetry data and the lifespan of a component. As a final note, our framework could be adapted for all tasks in which the correlation should be expected between time series data distribution and a group of numerical data.

Bibliography

[1] J. V. Anderson and D. Sornette. Prediction failure using conditioning on damage history: Demonstration on percolation and hierachical fibre bundles. *PHYSICAL REVIEW E*, 72:056124, 2005.

[2] Sergey Brin, Rajeev Motwani, and Craig Silverstein. Beyond market baskets: generalizing association rules to correlations. *ACM SIGMOD Record*, 26(2):265–276, 1997.

[3] Sergey Brin, Rajeev Motwani, Jeffrey D. Ullman, and Shalom Tsur. Dynamic itemset counting and implication rules for market basket data. In *SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data*, pages 255–264, Tucson, Arizona, USA, May 1997.

[4] J. D. Brown. *Understanding research in second language learning: A teacher's guide to statistics and research design.* Cambridge University Press, Cambridge, 1988.

[5] Jake D. Brutlag. Aberrant behavior detection in time series for network monitoring. In *Proceedings of the 14th USENIX conference on System administration*, pages 139–146, New Orleans, Louisiana, 2000. USENIX Association.

[6] Philip K. Chan and Matthew V. Mahoney. Modeling multiple time series for anomaly detection. In *Proceedings of the Fifth IEEE International Conference on Data Mining*, pages 90–97. IEEE Computer Society, 2005.

[7] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley, New York, 1991.

[8] G. Dieter. *Mechanical Metallurgic.* Mc Graw Hill, 1988.

[9] GEOFFREY GRIMMETT. *Percolation.* Springer, second edition, 1999.

[10] Jiawei Han and Micheline Kamber. *Data Mining: Concepts and Techniques.* Data Management Systems. Morgan Kaufmann Publishers, 2nd edition, March 2006.

[11] Evan Hoke, Jimeng Sun, John D. Strunk, Gregory R. Ganger, and Christos Faloutsos. Intemon: continuous mining of sensor data in large-scale self-infrastructures. *Operating Systems Review*, 40(3):38–44, 2006.

[12] Tomoharu Iwata and Kazumi Saito. Visualization of anomalies using mixture models. *Journal of Intelligent Manufacturing*, 16(6):635–643, December 2005.

[13] M. G. Kendall. *Rank Correlation Methods.* Hafner Publishing Co., New York, 1955.

[14] J. F. Kenney and E. S. Keeping. *Mathematics of Statistics.* Van Nostrand, Princeton, NJ, 2nd edition, 1951.

[15] B. Lawn. *Fracture of brittle Solids.* Cambridge University Press, 1993.

[16] Young-Koo Lee, Won-Young Kim, Y. Dora Cai, and Jiawei Han. Comine: Efficient mining of correlated patterns. In *Proceedings of the Third IEEE International Conference on Data Mining*, page 581. IEEE Computer Society, November 19-23 2003.

[17] E. L. Lehmann and H. J. M. D'Abrera. *Nonparametrics: Statistical Methods Based on Ranks.* Prentice-Hall, Englewood Cliffs, NJ, rev. edition, 1998.

[18] M. Mahoney and P. Chan. Trajectory boundary modeling of time series for anomaly detection. In *KDD-05*, Chicago IL, USA, August 2005. ACM.

[19] Dragos Margineantu, Stephen Bay, Philip Chan, and Terran Lane. Data mining methods for anomaly detection kdd-2005 workshop report. *ACM SIGKDD Explorations Newsletter*, 7(2):132–136, December 2005.

[20] W. I. Newman, A. Gabrielov, T. Durand, S. L. Phoenix, and D. Turcotte. An exact renormalization model for earthquakes and material failures. statics and dynamics. *Physica D (Nonlinear Phenomena)*, 77:200–216, 1994.

[21] W. I. Newman, D. L. Turcotte, and A. M. Garbrielov. Log-periodic behavior of a hierarchical failure model with applications to precursory seismic activation. *Physical Review E*, 52(5):4827–4835, November 1995.

[22] Edward R. Omiecinski. Alternative interest measures for mining associations in databases. *IEEE Transactions on Knowledge and Data Engineering*, 15(1):57–69, 2003.

[23] Edward R. Omiecinski. Alternative interest measures for mining associations in databases. *IEEE Transactions on Knowledge and Data Engineering*, 15(1):57–69, Jan/Feb 2003.

[24] Spiros Papadimitriou, Jimeng Sun, and Christos Faloutsos. Streaming pattern discovery in multiple time-series. In *Proceedings of the 31st international conference on Very large data bases*, pages 697–708, Trondheim, Norway, 2005. VLDB Endowment.

[25] S. Patnayak and D. D. Tannant. Performance monitoring of electric cable shovels. *International Journal of Surface Mining, Reclamation and Environment*, 19(4):276–294, December 2005.

[26] S. Patnayak, D. D. Tannant, I. Parsons, V. Del Valle, and J. Wong. Operator and dipper tooth influence on electric shovel performance during oil sands mining. *International Journal of Mining, Reclamation and Environment*, 2007. iFirst article.

[27] Sibabrata Patnayak. *Key Performance Indicators for Electric Mining Shovels and Oil Sands Diggability*. PhD thesis, University of Alberta, 2006.

[28] S. Roux, A. Hansen, H. Herrmann, and E. Guyon. Rupture of heterogeneous media in the limit of infinite disorder. *Journal of Statistical Physics*, 52(1/2):237, 1988.

[29] H. Saleur, C. G. Sammis, and D. Sornette. Discrete scale invariance, complex fractal dimensions, and log-periodic fluctuations in seismicity. *JOURNAL OF GEOPHYSICAL RESEARCH*, 101(B8):17,661–17,677, August 1996.

[30] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[31] Stan Salvador and Philip Chan. Learning states and rules for detecting anomalies in time series. *Applied Intelligence*, 23(3):241–255, December 2005.

[32] S.P. Shah. *Toughening Mechanisms in Quasi Brittle Materials*. Kluwer Accademic Publisher, 1991.

[33] GALIT SHMUELI. Wavelet-based monitoring in modern biosurveillance. Technical Report RHS-06-002, Robert H. Smith School Research Paper, 2005.

[34] D. Sornette and J. V. Andersen. Optimal prediction of time-to-failure from information revealed by damage. *EUROPHYSICS LETTERS*, 74(5):778–784, 2006.

[35] Jimeng Sun, Spiros Papadimitriou, and Philip S. Yu. Window-based tensor analysis on high-dimensional and multi-aspect streams. In *Proceedings of the Sixth International Conference on Data Mining*, pages 1076–1080. IEEE Computer Society, 2006.

[36] Jimeng Sun, Dacheng Tao, and Christos Faloutsos. Beyond streams and graphs: dynamic tensor analysis. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 374–383, Philadelphia, PA, USA, 2006. ACM Press.

[37] Ruey S. Tsay. *Analysis of Financial Time Series*. Wiley-Interscience, 1st edition, October 2001.