UNIVERSITY OF SOUTHAMPTON

FACULTY OF ENGINEERING, SCIENCE & MATHEMATICS

School of Mathematics

Modelling Breakdown Durations in Simulation Models of Engine Assembly Lines

by

Lanting Lu

Thesis for the degree of Doctor of Philosophy

May 2009

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING, SCIENCE & MATHEMATICS

SCHOOL OF MATHEMATICS

Doctor of Philosophy

MODELLING BREAKDOWN DURATIONS IN SIMULATION MODELS OF ENGINE ASSEMBLY LINES

Lanting Lu

Machine failure is often an important source of variability and so it is essential to model breakdowns in manufacturing simulation models accurately. This thesis describes the modelling of machine breakdown durations in simulation models of engine assembly lines. To simplify the inputs to the simulation models for complex machining and assembly lines, the Arrows classification method has been derived to group machines with similar distributions of breakdown durations, where the Two-Sample Cramér-von Mises statistic and bootstrap resampling are used to measure the similarity of two sets of data. We use finite mixture distributions fitted to the breakdown durations data of groups of machines as the input models for the simulation models. We evaluate the complete modelling methodology that involves the use of the Arrows classification method and finite mixture distributions, by analysing the outputs of the simulation models using different input distributions for describing the machine breakdown durations. Details of the methods and results of the grouping processes will be presented, and will be demonstrated using examples.

Contents

List of Figures

List of Tables

Declaration of Authorship

I, Lanting Lu, declare that the thesis entitled Modelling Breakdown Durations in Simulation Models of Engine Assembly Lines and the work presented in it are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of other, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- parts of this work have been published as:
  - Lu, L., Currie, C.S.M., Cheng, R.C.H. and Ladbrook, J. "Classification analysis for simulation of machine breakdowns" in *Proceedings of the Winter Simulation Conference 2007*, S.G. Henderson, B. Biller, M.-H. Hsieh, J. Shortle, J.D. Tew and R.R. Barton as editors, pages 480-487, 2007.
  - Currie, C.S.M. and Lu, L. "Optimal Scheduling Using Length-of-Stay Data for Diverse Routine Procedures". Intelligent Patient Management. S. McClean, P. Millard, E. El-Darzi and C. D. Nugent as editors. Heidelberg, Springer Berlin. 189: 193-205, 2009.
- parts of this work are due to be published as:
  - Currie, C. S. M. and Lu, L. "Comparison of Simulation Output Series Using Bootstrapping". In Proceedings of the 2009 INFORMS Simulation Society Research Workshop. L. H. Lee, M. E. Kuhl, J. W. Fowler, and S. Robinson as editors, 2009.

- parts of this work presently unpublished are:
  - Lu, L., Currie, C.S.M., Cheng, R.C.H. and Ladbrook, J. “Classification Analysis for Simulation of the Duration of Machine Breakdowns”.
  - Currie, C. S. M. and Lu, L. “Evaluation of the Arrows Methods for Classification of Data”.

Signed: .....

Date: .....

Acknowledgements

I would first like to thank my supervisor Dr Christine Currie for the assistance, advice and support she has given freely and generously throughout the course of my PhD study.

I would also like to thank Professor Russell Cheng for his knowledge and guidance throughout my PhD.

I am grateful to the Ford Motor Company who supported me throughout this research and to John Ladbrook who provided warm help and massive data.

Many people at the Ford's Dunton Technical Centre have helped me through the course of my PhD, especially John Ladbrook, Mark, Matthew Loynes and Zou Qi. Thanks also to Phil Smith and Ornella Benedett for their kindhearted support during the times I worked at Dunton.

I would also like to thank Honora Smith, Georgina Mellor, Andrew Drake, Simon Doherty and Marion Penn with whom I have had such a great time sharing offices with.

I would like to give my recognition and thanks to all my family for their continued love and support, without whom my study in Southampton would not be possible.

Last but not least, I would very much like to thank my partner Bo Huang, whose continued encouragement, patience and good humour throughout my time of study has been absolutely invaluable to me.

Chapter 1

Introduction

This thesis describes the methodology of modelling machine breakdown durations in simulation models of engine assembly lines. We derive a classification method termed the Arrows classification method to simplify the inputs to the simulation models for complex machining and assembly lines by grouping machines with similar distributions of breakdown durations. We fit finite mixture distributions to the breakdown duration data of groups of machines that are involved in the engine assembly lines to represent the machine breakdown duration inputs in the corresponding simulation models.

This research is supported by Ford Motor Company and their objective was to find an appropriate mathematical representation of the machine breakdown duration inputs in manufacturing simulation models. We use a pre-existing simulation model of an engine assembly line, to test our methodology. The simulation model is built in WITNESS simulation software (Lanner Group) [102] and is supplied by Ford who also provide the necessary data.

Discrete-event simulation has been widely used in manufacturing industry to model production operations. Ford have used this powerful tool since 1982 to help with the planning of new facilities and the improvement of existing lines in all

of their manufacturing plants. Different scenarios, such as number of resources, length of buffers or layout of the manufacturing lines, can be set in different simulation models. The outputs of these simulation models of machining and assembly lines can and have been used to estimate costs, productivity targets and proper labour requirements and layouts for existing and new engine programs. Therefore, simulation models are required to reflect the real world as accurately as possible.

In manufacturing systems, machine failure is often an important source of variability. Therefore it must be represented correctly in simulation models of the process. Machine and engine repairs and operator stoppages can have a significant effect on the line yield. For example, the total loss due to these repairs and stoppages in the engine assembly line we consider in this thesis, for the last three months of 2007 was 18.7%. However, while Ford have detailed duration data for machine repairs, since the machines are linked to an automatic on-line monitoring system, similar data are not available for engine repairs and operator stoppages because the enormous time and resource requirements for monitoring every single engine repair and operator stoppage are prohibitive. We therefore focus on the development of a methodology to enable the modelling of the machine repair durations.

Currently, historical data are commonly used in Ford as machine breakdown duration inputs to the simulation models while theoretical distributions are only used when there are no historical data available for a machine. However, it is generally preferable to use appropriate theoretical distributions as simulation inputs for several reasons; for example, it is often easier to change a theoretical distribution when performing different experiments on the simulation model. No common statistical distribution has been found to be a reasonable fit for most of the breakdown duration data as each set of data is a mixture of a number of distinct populations, resulting in a multimodal distribution. Therefore, finite mixture distributions have been proposed to fit the breakdown duration data of machines.

There are normally hundreds of different machines involved in each engine assembly line in Ford. A major contribution of this thesis is the simplification of the machine breakdown duration inputs, which is required when modelling such large assembly lines. We have derived a method of grouping machines based on the breakdown duration data available, called the Arrows classification method. The grouping is such that two machines can be placed in the same group only if there is a statistically significant similarity between their breakdown duration data, where the statistical similarity between the breakdown duration data sets of two machines is estimated using the Cramér-von Mises goodness-of-fit statistic [5]. Bootstrapping is used to determine the significance level of the statistic. Finite mixture distributions are fitted to the grouped breakdown duration data so that the fitted finite mixture distributions for each group can be used to represent the breakdown duration inputs for all of the machines in this group. The grouping reduces the number of input distributions that must be estimated and increases the data available for fitting the finite mixture distributions.

1.1 Finite Mixture Models

We use finite mixture models to represent the breakdown duration data for machines in engine assembly lines because the data are generally multimodal. Finite mixture models provide a good description of multimodal data, using parameters that have an intuitive meaning, and their implementation in most standard simulation packages, including the WITNESS software (Lanner Group) [102], which we use to build our simulation models, is simple and convenient.

A continuous finite mixture model is defined by probability density function written as

$$h(x) = \sum_{i=1}^{k} w_i f_i(x|\theta_i), \tag{1.1}$$

where $f_i(x|\theta_i)$ is a component distribution and w_i is its weight and satisfies $w_i > 0$ and $\sum_{i=1}^{k} w_i = 1$.

Parameter θ_i comprises the unknown parameters associated with the ith individual component. Parameters θ_i, weight w_i and number of components k are all unknown. We therefore wish to determine the number of components k and other parameters in the finite mixture model. Since it is possible that the mixture model is composed of components that are not represented in the data, finding k is a statistically non-standard problem. In addition, the likelihood is unusual with certain combinations of parameter values giving rise to an infinite likelihood, and these combinations do not correspond to consistent parameter estimates. Hence making use of standard maximum likelihood methods is impossible in this case.

Instead, a Bayesian framework is used for the fitting process as described in [40]. Using Bayesian statistics, although the posterior distribution may still be multimodal, the prior distribution smooths out the likelihood function. Moreover, the posterior distribution for k is considered to be a more meaningful measure of k in the mixture model than the likelihood function [40]. Importance sampling is used to determine the posterior distribution for the number of components.

1.2 Estimating Similarity

We wish to classify machines involved in the engine assembly lines into groups with similar breakdown duration data, in order to simplify the breakdown inputs for simulation models. To achieve this we first need to estimate the similarities between the machines. As the breakdown duration data sets have uneven numbers of data points, no standard method for measuring similarity is applicable. Thus, we derive a new approach and measure the similarity of two machines by estimating the possibility of the two corresponding breakdown duration data sets having

been drawn from identical distributions. We assume that two samples of breakdown duration data $X = (x_1, x_2, \ldots, x_n)$, and $Y = (y_1, y_2, \ldots, y_m)$ for machines M_x and M_y respectively consist of independent observations. Under the null hypothesis that samples X and Y are drawn from the same distribution, we calculate the Two-Sample Cramér-von Mises goodness-of-fit statistic T, which is a good general purpose goodness-of-fit test method [42] and has an advantage of being a distribution-free method, i.e. there is no need to make any assumptions about the underlying distributions of the data sets being analysed [5]. We reject the null hypothesis if T is too large. Tabulated criterion values for this test are not very extensive and only give standard criterion values for samples with up to 8 data points or with sizes close to infinite [5]; while the number of data points of machine breakdown duration data sets varies from 9 to 1310. Therefore, in order to determine whether T is too large, we need to estimate the p-value of T by estimating $\Phi(T)$, the distribution of the statistics of samples that are drawn from the same distribution. We do this using bootstrapping, which is described further in Chapter 4. The similarities between each and every pair of machines are put together to form the similarity matrix of all of the machines involved.

In Chapter 4 the method for measuring similarity is tested by comparing random samples generated from known distributions. Although this method was originally derived to estimate similarity between the machine breakdown duration data sets, it is widely applicable, and we have also used it to calculate the similarity between medical procedures based on their patients' length-of-stay in a group of private hospitals [41].

1.3 Classification

The machining and engine assembly lines that we are modelling often include hundreds of different machines. Since the breakdown duration data of many machines

follow similar distributions, for the purpose of reducing the number of input distributions, we propose a new classification method for grouping machines based on their breakdown duration data. The fitted mixture distribution for the group can then be used to described the breakdown duration inputs for all of the machines in this group. In the classification process, two machines with significantly different breakdown duration data, as calculated using the bootstrapping method described in Section 1.2, are not allowed to be placed in the same group.

We name this classification method the Arrows method because in this method the strength of connections between objects are defined using arrows. This will be described in Chapter 5. Objects with double-arrow and single-arrow connections are placed in the same groups whenever possible. Objects 1 and 2 are said to have a double-arrow connection if p_{12}, the p-value similarity of the two objects, is the greatest in both row 1 and row 2 of the similarity matrix; but if p_{12} is the greatest in only one of row 1 or row 2, objects 1 and 2 are said to have a single-arrow connection instead. Another major feature of the Arrows method is the setting of a threshold. A similarity threshold, p_0, is set with the assumption that two data sets with a similarity of the threshold value or above are similar enough to be put in the same group. Thus, two objects can be put in the same group only if the p-value for comparing their corresponding data sets is greater than or equal to p_0.

1.4 Evaluation of the Breakdown Inputs

We evaluate the whole process of modelling breakdowns by studying the outputs of a simulation model of an engine assembly line designed by Ford using three different inputs to represent the machine breakdown durations: (1) empirical distributions; (2) fitted finite mixture distributions for individual machines; (3) fitted finite mixture distributions for groups of machines. We assess the simulation outputs of the models with the three machine breakdown duration inputs using three

different methods: graphical comparison, paired-T test and bootstrapping analysis. The bootstrapping analysis uses the same method for calculating the similarity of two sets of simulation output data as is used to measure the similarity of breakdown duration data from pairs of machines. This is another important potential application of the work in this thesis.

We also wish to investigate the impact of the choice of similarity threshold when using the Arrows classification method. Simulation models are built with the breakdown duration inputs represented by different sets of fitted mixture distributions corresponding to the different groups that are generated using the Arrows method with a range of thresholds. The simulation outputs of the same engine assembly line model with different groupings of machines are compared to give some insights.

1.5 Modelling Machine Breakdowns

The models of the manufacturing plants that we consider in this thesis are built in WITNESS simulation software (Lanner Group) [102].

Historical breakdown duration data for machines are available directly from the on-line monitoring system that the engine assembly line is linked to. The collected data need to be validated by deleting unreasonable data points or subtracting some part of durations for some data points; checked for correlations before the data can be used in the subsequent analysis; and transformed for further analysis in the breakdown duration modelling process. We discuss the data preparation further in Section 3.3.

We propose using fitted mixture distributions for groups of machines to represent the machine breakdown durations, i.e. the time to repair machine failures. Fitted mixture distributions cope well with the multimodality present within the data

and can smooth out its irregularities. Our proposed breakdown duration modelling process is shown in Figure 1.1 and comprises three major steps:

1. Data preparation/ transformation:

   Adjustments need to be made to validate the data for the mixture distributions fitting process. We transform the validated breakdown duration data to obtain a better fit of finite mixture distributions.

2. Select component distribution type:

   The type of component distribution is chosen based on the characterisations of the breakdown duration data. A mixture of lognormal distributions is considered to be the most appropriate to represent machine downtimes and is simple to input into the WITNESS models for the engine assembly lines. Section 3.4 describes the rationale behind this choice.

3. Fitting mixture distributions:

   We propose using finite mixture distributions fitted to the amalgamation of the data for all of the machines in a group to represent the machine downtimes for machines in the same group. There are three steps in this part:

   (a) Estimate similarities between the machines

   The similarities between machines are measures by the significance levels of Cramér-von Mises statistics of their corresponding breakdown duration data sets. The method for measuring machines' similarities is described in Chapter 4.

   (b) Machines classification

   Use the Arrows classification method to divide machines into groups based on the similarities between their breakdown duration data. This classification method is described in Chapter 5.

(c) Fitting mixture distributions to the grouped data

This step involves estimating parameters of finite lognormal mixture distributions for representing the breakdown durations for groups of machines. A Bayesian framework is applied to find the posterior distributions of the parameters of the component distributions and that of the number of components in the mixture distribution (see Section 3.2 for details). We fit one mixture distribution to each group of machines. The fitted mixture distribution for one group can be used to represent the breakdown durations for all machines in this group.

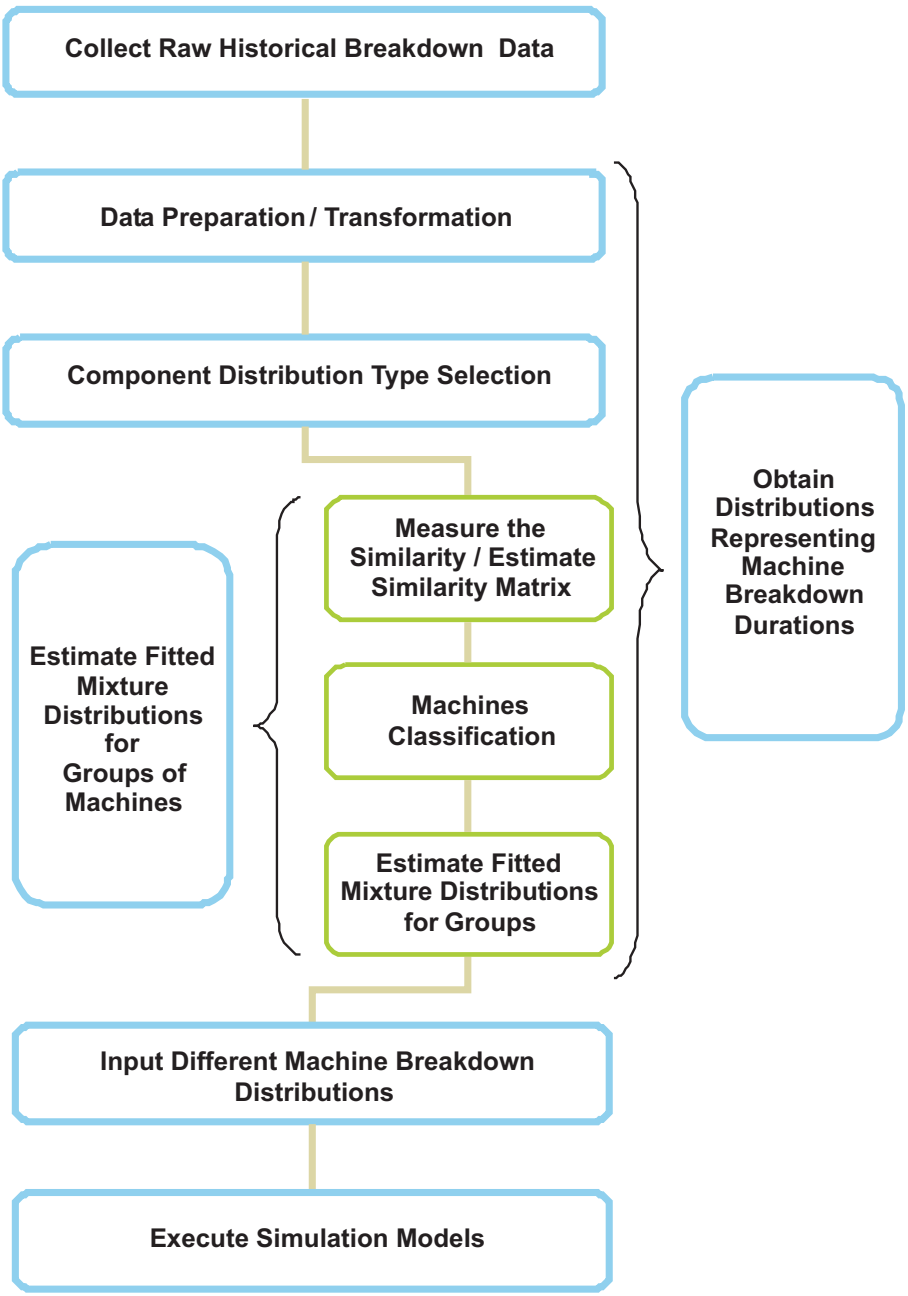

Figure 1.1: Diagram of the proposed machine breakdown duration modelling process.

1.6 Outline of the Thesis

A literature review of machine breakdown input modelling in manufacturing simulation models is given in Chapter 2. We then describe the proposed statistical model for representing the breakdown duration data in Chapter 3. In Chapter 4 we discuss the methodology used to estimate the similarity of data sets of uneven sizes and test the method by implementing it on samples generated from a number of known distributions. The application of this method to the machine breakdown duration data and data on patients' hospital length-of-stay for a set of medical procedures is also given here. The Arrows classification method used for the machines classification process is presented in Chapter 5, including a comparison between the Arrows method and popular cluster analysis methods and examples of the method's application to grouping machines and medical procedures. The machine breakdown modelling process for the simulation model of an existing engine assembly line that is currently in use is described in Chapter 6. In Chapter 7 we describe the methodology for evaluating the proposed machine breakdown duration modelling methodology by constructing experiments on the engine assembly line simulation model. We conclude in Chapter 8.

Chapter 2

Literature Review for Modelling Breakdowns

We aim to develop a new mathematical form to represent the distribution of machine breakdown durations in simulation models of engine assembly lines. As "the most important source of randomness in many manufacturing systems" ([103], P687), machine breakdowns have a very big impact on the system throughput and need to be modelled correctly. While there is a substantial literature on modelling the time between breakdowns ([64], [128], [99], [163], [171] and [68]), there has been relatively little work done on modelling the durations of breakdowns. The lack of literature on this specific subject is indicated in [6], [97] and [103]. It is also suggested that even within the written literature on the topic of modelling breakdowns there is little discussion on the practical implementation [97]. The most practical publication on this subject suggested by [97] is [78]. There are some other good references on modelling breakdowns in manufacturing system models, such as [83], [103], [25] and [6].

This chapter gives a review of the available literature on previous methods for modelling machine breakdowns. We begin by giving the definition of the term breakdown in Section 2.1. A discussion of machine failure rates is then given in

Section 2.2 and the elements of a machine breakdown are described in Section 2.3. As the failure data collection is often problematic, a discussion of data collection methods is given in Section 2.4. Finally, Section 2.5 discusses the approaches that can be used to represent the machine downtimes.

2.1 Definition of a Machine Breakdown

Machine downtimes can be classified into two types:

1. Deterministic downtimes are machine downtimes that can be scheduled: such as shift changes, breaks and planned maintenance [103].

   Modelling this type of machine downtimes can be relatively easy.

2. Random downtimes are unscheduled machine downtimes: such as actual machine failures, broken tool changes, parts being stuck and gauging ([97] and [103]).

   This thesis concentrates on modelling random downtimes.

There are arguments about the randomness of machine breakdowns. Binroth and Haboush [15] believe that breakdowns are time dependent as the occurrence of future events would depend on the random times at which past events happened. Bradford and Martin [21] also consider that machine failures are not entirely random and scheduling the next breakdown in simulation models might be dependent on the machines' previous breakdowns. Some, for example [128], [129], [27] and [37], believe that electronic machine failure rates are related to time and follow the classical Bathtub curve (see Section 2.2). Venton [156] on the other hand states that machine breakdown consists of mechanical failures that often are the result of physical wear, and electronic failures that are invariably concerned with a chance and argues that electronic failures are random while mechanical failures

should really be treated as time dependent events. Although most Ford manufacturing machines are combinations of mechanical and electronic components and the theory of time dependent breakdowns is probably correct, the Productivity Engineers at Ford assume that all breakdowns are random independent events [97]. We do not consider the modelling of the times between breakdowns in this thesis.

A *breakdown* is defined in [97] as "a generalisation for a mechanism failing to perform its required function for an unknown reason when it was capable of doing so". In other words, a breakdown is the event after a mechanism fails and before the machine functions again. The *breakdown duration* includes the amount of time to gather resources to analyse the problems and the length of the actual repair time [103], and this whole period of the breakdown is also referred to as the *repair time* or the *time to repair* (TTR) or the *machine downtime*.

There are many causes that may lead to a breakdown: machine operating times, maintenance conditions, parts replacements, machine weariness, design errors, operator skills and random machine failures [17]. It seems impossible to predict the occurrence of breakdowns ([25] and [16]). Thus, the machine breakdowns are considered to be random downtimes. The main objective of our research is finding accurate statistical distributions for describing machine downtimes.

2.2 Machine Failure Rates

A classical categorisation of failures is based on the time at which they occur, which separates machine failures into three types:

- Early Life: Also referred to as Infant Mortality [37]. In this initial period of time the failure rate gradually decreases with time after time zero ([27] and [37]).

- Useful Life: This long period is also known as the Intrinsic Failure period or Stable Failure period. In this phase the failure rate is roughly constant ([27] and [37]).

- Wearout: In this period of time failures are mainly caused by degradation and the failure rate increases with time ([27] and [37]).

The sum of these three phases is commonly know as the *bathtub curve*, shown in Figure 2.1, which is suggested to be the traditional curve for electronic machines [124]. The basic concept for the bathtub curve was believed to be established in Proschan [128], [129]. There is some discussion, disagreement and development about the true character and the use of the bathtub curve (see, for example, [99], [163], [171], [64] and [68]). Condra [37] states that the argument of correctness of the bathtub curve appears to be very subjective.

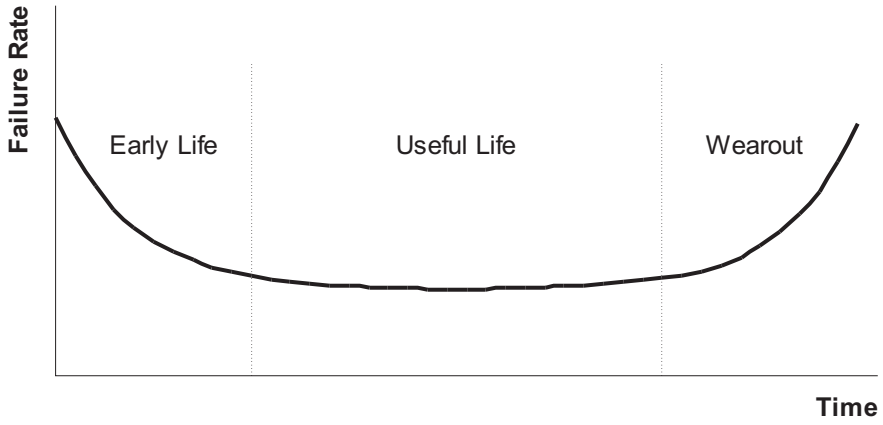

Figure 2.1: Bathtub Curve for machine reliability.

Venton [156] separates machine breakdown into mechanical failures and electronic failures. The former are suggested to be treated as time dependent because they are “often a result of physical removal of material by wear”. The latter can be considered as random events as they are “invariably concerned with a chance excess of applied stress over inherent strength”.

Porter and Finke [127] examined machine breakdowns with forty eight causes on the area of integrated circuits and classified them into four main categories: broken parts, time degradation, mechanical stress and serial effects of time degradation and mechanical abrasion.

Buzacott and Hanifin [23] identified two types:

- Operation dependent cause:

  Cannot happen when the machine is in the idle state; happens after a certain number of operation cycles.

- Time dependent cause:

  Can happen when the machine is idle; is due to some uncertain reason except wear and happens after a certain amount of time.

This categorisation suggests that a breakdown can happen even when the machine is not operating and there is time dependency in the occurrences of breakdowns. However, engineers in Ford assume that a breakdown is a totally random and independent event and cannot happen when a machine is not operating. We make the same assumption in the simulation model and this is discussed further in Chapter 6.

Another categorisation identified by Ibe and Wein [84] is based on the duration of the failures, which is also used by Ford engineers. Law [103] (P320) gives a similar opinion about the types of machine breakdowns. The two types are:

- Permanent failure:

  Commonly classified as inherent failure by machine manufacturers, and "requires the physical repair of a system by the field service crew and usually takes hours to complete" It is referred to as Major failure by Ford and defined as a failure that usually requires highly skilled maintenance staff to fix and normally takes longer than 15 minutes to repair.

- Intermittent failure:

  Commonly classified as operational failure by machine manufacturers, and "can be taken care of by the system operator and usually takes minutes to complete" This is called Minor failure by Ford and defined as a failure that generally needs basic skills to perform the maintenance and usually takes no more than 15 minutes to fix.

2.3 Elements of a Machine Breakdown

Barton et al [10] point out that the time spend on collecting and analysis data is huge, therefore understanding the elements of breakdowns can really help with initial data analysis. It is believed that the time from when the failure occurs until the machine functions again is not only actual repair time. This point is demonstrated in an example given by Feltner and Weiner [54]:"the line stopped at 3:00pm on a Thursday and was not running again until Monday morning, should we use the elapsed time as repair time? Is it possible that the shift finished at 3.30pm and, since part of the press line was not needed for the rest of the week, action was deferred until the No2 shift came on board on Monday". The total time of the failure contains a long period of time in which no repair was carried out. Since the data collected electronically in Ford states only the start and finish time of a failure (see Section 2.4), this is the main reason for requiring data validation (see Section 3.3.1) before the analysis can take place.

Blache and Shrivastava [16] introduced the term of corrective maintenance as corrections have to be undertaken to make a repair. They indicate that there are more actions than just repairing the machine to turn it "from a failed state to an operating or available state". It is stated that the whole period of corrective maintenance can be separated into two main stages:

1. The active stage

   The period needed to change the machine into “a serviceable state”, i.e. actual repair time.

2. The delay stage

   Waiting time caused by the absence of one or more resources, such as tools or maintenance staff.

Law [103] splits the repair time into the same two stages. Human behaviour was cited by Hanifin [77] as an important contribution to uncertainty. Banks et al [7] also blame human behaviour for much of the variability.

A diagram of two major types of machine repair process used at Ford manufacturing plants is given in Ladbrook [97] and is reproduced in Figure 2.2. The repair process has two main types: (a) the left hand side of this diagram, shown as blue arrows, is the process without line side maintenance and (b) the right hand side, shown as purple arrows, is the process with line side maintenance. The rectangles indicate the basic steps of the breakdown process and the blue or purple arrows indicate two different sequences of the basic steps: blue for without line side maintenance process and purple for with line side maintenance process. As shown in this diagram, the biggest difference between the two types is that with line side maintenance, there is no need to “call maintenance operators from a central pool” [97].

Operators can manage to undertake a minor repair and maintenance operators are called if it is identified as a major repair at the initial inspection of the operator. Machine tryouts are test runs carried out by operators or maintenance operators to check whether the machine is fixed properly. If the machine operates successfully during tryouts, the whole maintenance process is considered to be completed.

A sequence of very detailed elements and phases in a maintenance process is identified by Ferrazano in [97], although no explanation of the different phases of

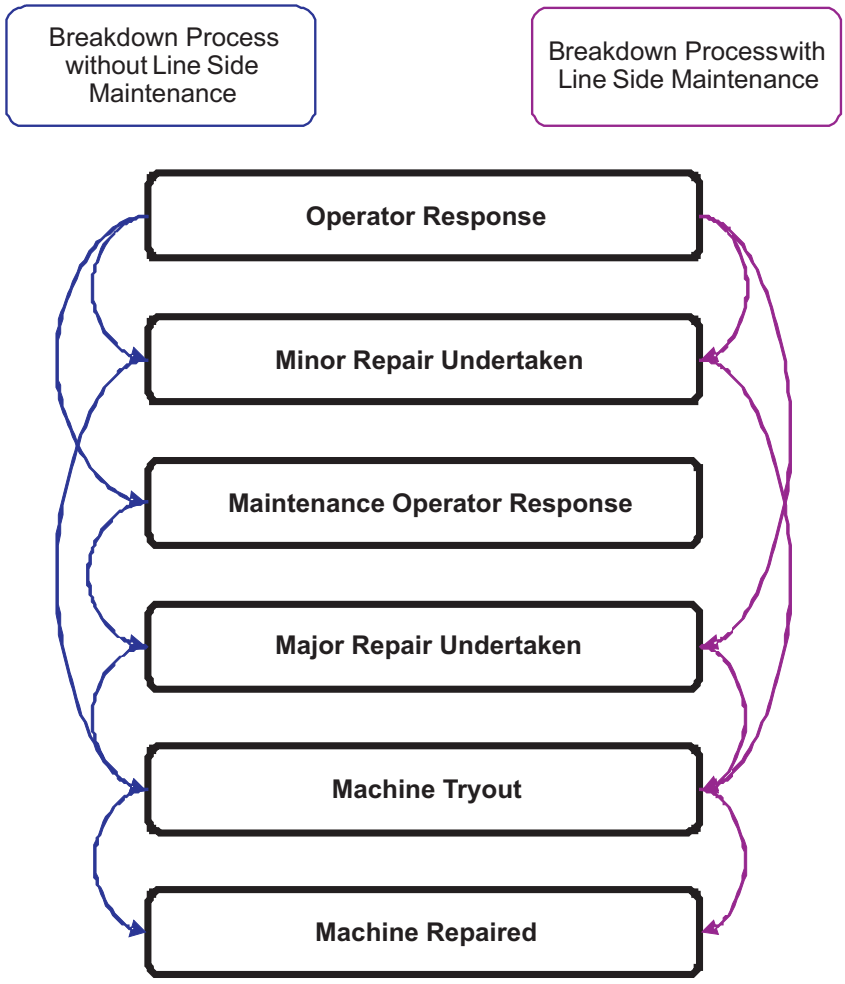

Figure 2.2: Diagram of elements of two types of repair process at Ford.

the whole maintenance route for a breakdown is given. A diagram of the maintenance process is shown in Figure 2.3.

Carrie [25] describes a more straightforward logic for modelling machine breakdowns. After choosing the method to generate the failure times much (Steps 1 and 2), his approach is as follows:

Step 3 “Schedule start of breakdown event at this time.”

Step 4 “When the clock reaches this time take the machine out of service.”

Step 5 “Draw a sample from the repair time distribution and add it to the current

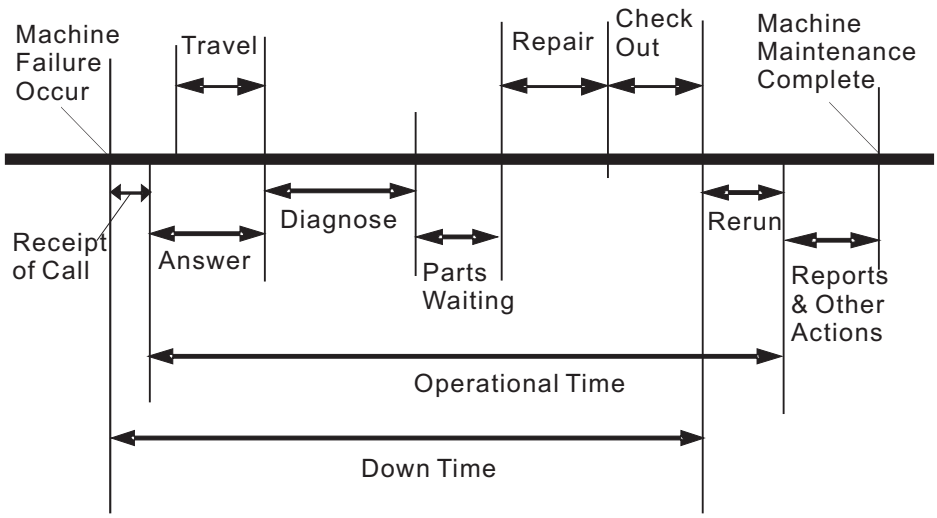

Figure 2.3: Detailed diagram of elements of the maintenance process. Reproduced from [97].

clock time."

Step 6 "Schedule an end of breakdown event at this time."

Step 7 "When the clock reaches this time return the machine to normal service."

Step 8 "Draw a sample from the time between repair distribution and add it to the current clock time."

This logic assumes the time generated for a machine failure is the whole elapsed time of all elements of the breakdown stage. Compared to the detailed model shown by Figure 2.3, the greatest advantage is its simplicity. Ford found it was very time consuming and even unrealistic to collect precise data for each phase shown in Figure 2.3. Besides, experiments have been carried out on simulation models with different detail levels of breakdown durations modelling and no significant differences have been detected [97]. Therefore engineers in Ford make similar assumptions to Carrie's, that all of the elements of breakdowns are included within the generated time to repair.

2.4 Historical Data Collection

There are two main methods of breakdown data collection in Ford: electronic and manual collection. The former data collection method is achieved by using the automatic on-line monitoring system, while the latter requires the work of line foremen, machine operators or productivity engineers. The data we use is all collected automatically. It includes every breakdown of machines that are linked to the on-line monitoring system on the engine assembly line, recorded during a period of three months from January to March 2008. There are 39 machines linked to the monitoring system for this line, and these machines are chosen because they are considered the most important to the running of the line. Each entry of the data has several attributes consisting of the ID of the machine that has broken down, the start time of the breakdown, the finish time of the breakdown and a brief description of the fault that caused the breakdown.

The manual collection in Ford includes two methods: Line Foreman's Records and Productivity Engineers Records. Compared to manual collection, the advantage of electronic collection is that the monitoring system records every failure of machines that have been connected to the system. Manual collection can also be expensive and time-consuming. The disadvantages of electronic collection are described by Ladbrook [97] as the following:

1. The system cannot identify lack of spares, tools collection or tidy up or the shift break times.

2. During a machine breakdown, the maintenance operator sometimes needs to run some 'try outs' to see if the machines is repaired correctly. The system cannot treat the 'try outs' as part of one failure. Hence, one breakdown could be recorded as more.

3. If the machine is powered off during repair, the system may record two stoppages instead of one.

4. A failure occurring on the last production shift of the week could have one of two outcomes. First, the machine is fixed during overtime at the weekend, or second, it is fixed in the first shift of the following week. In either case, the system records the duration of this repair as lasting the whole weekend or lasting until the end of the last shift.

5. The monitoring system may be off during weekend overtime. Thus, it is often not known when the repair is completed during the overtime period.

6. The automatic monitoring system might breakdown. In this case, it is necessary to rely on the engineers responsible for the line to use other methods to collect the data.

The data collected from the on-line monitoring system therefore needs to be validated before subsequence analysis. The cleaning and validation of the raw data was previously carried out manually in Ford, which was a very time consuming process especially when dealing with large data sets that include thousands of breakdown entries. We have derived a program using Visual Basic of Applications in Excel to process the data validation, which has helped the Ford simulation modellers to achieve an enormous saving of time spent on this task.

The data validation may change the raw data significantly. For example, the histogram of the distribution of the raw repair time data for a typical machine in an engine assembly line is shown in Figure 2.4 and the histogram of the validated repair time data for the same machine is given in Figure 2.5. The detail of the data validation process will be discussed in Section 3.3.

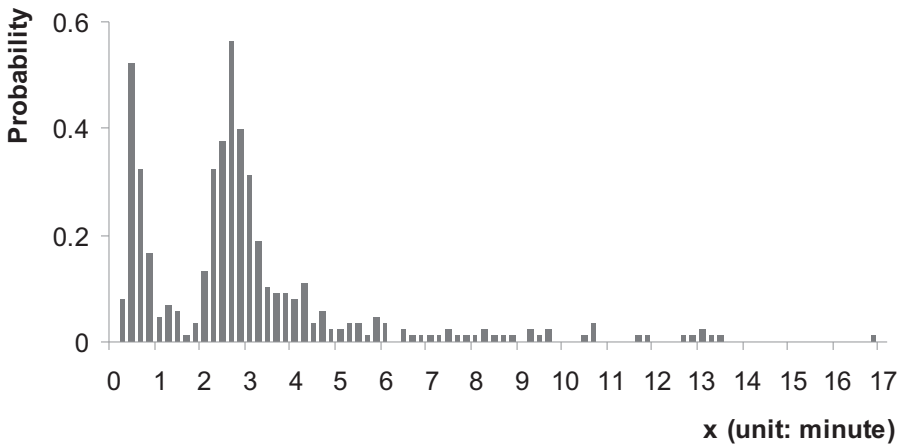

Figure 2.4: Histogram of the distribution of the raw breakdown duration data of a machine involved in engine assembly process.

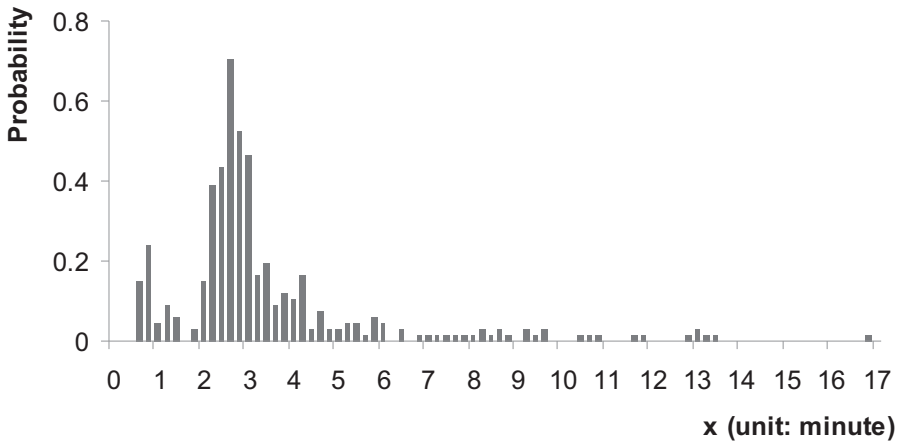

Figure 2.5: Histogram of the distribution of the validated breakdown duration data of the same machine given in Figure 2.4.

2.5 Distributions for Representing Downtimes

Finding a good representation of machine breakdown durations is a crucial part of modelling breakdowns when building a simulation model. Input modelling is used to specify the appropriate form of the distributions representing input random variables to a simulation model. In this section, we will give a discussion of literature reviews in general simulation input modelling and then consider machine breakdown input modelling in more detail.

2.5.1 General Input Modelling

Almost all simulation models of real-world systems require the input random variables that represent the sources of variability to be modelled. For example, in a queuing system, sources of variability include random customer inter-arrival times and customer service times and their probability distributions are required.

There is an extensive discussion of general simulation input modelling. The common recommendation is that if a standard *theoretical distribution* can be found that is a good model for the input data, then this distribution should be used in the simulation model; otherwise, using the *empirical distribution* based on the data is a good option (see, for example, [103], [157] and [12]). While using an empirical distribution seems to be straightforward, an adequately well fitted theoretical distribution is generally preferable for a number of reasons:

- Smooth out the data:

  As the number of data points in the data is finite and sometimes even very small, the empirical distribution may contain irregularities, such as gaps, in which there are no observations in this sample but these values may be possible in other samples ([103] and [12]).

- Represent extreme events:

  Generally, empirical distributions only represent data with limited values because the randomly generated data from an empirical distribution cannot be less than the minimum of the observed data or greater than the maximum of the observed data. Since the chance of extreme events can heavily influence the performance of simulation systems, a fitted theoretical distribution can be a better method of representing the whole process ([103] and [12]).

- Physical reasons:

  Certain physical characteristics of the data, such as nonstationarity or dependence, make it elaborate to obtain the empirical distribution ([103] and [12]).

- Simpler to make changes:

  It is much simpler to make changes to a theoretical distribution. If we want to investigate the system performance in different scenarios with differences in that input data. With theoretical distributions, simply changing the parameters will make all of the changes. But there is no straightforward way for making the changes when using an empirical distribution ([103] and [12]).

- Compact way to represent the data:

  The physical process to input the empirical distribution into the simulation model might be time-consuming especially with a large data set. A theoretical distribution, on the other hand, is a much more compact way to represent the input data [103].

In relevant work using this approach, most authors focus on relatively simple problems where input random variables are independently and identically distributed and follow well-known parametric theoretical distributions, such as gamma, lognormal, normal, Weibull, etc. Since the natures of different kinds of data vary

a lot, the number of choices is correspondingly large. There are a few features of the data that can help narrow down the possible choice to a few that may have a better fit, e.g. the shape of the histogram of data or whether the data consist of negative or positive values([107], [108] and [109]). For example, if the histogram of data skews to the right, the normal distribution can probably be ruled out. Law [103] gives a tutorial on "hypothesizing" distributions that might be a good fit of the data. A good descriptions of the physical features of many standard theoretical distributions can be found in [52] and Chapter 9 of [8].

Law et al. [106] identified that sometimes no standard theoretical distributions can reflect the actual underlying distribution. If no theoretical distribution seems to be a good fit, it is recommended by most text books, such as Law [103], that an empirical distribution should be used. Biller and Barry [12] also suggest that an empirical distribution can be a good option "when an adequate sample is available, the data are thought to be representative and there is no compelling reason to use a probability model (including the case that nothing appears to fit well)". Barton et al. [10] express their concerns on the common approach of using fitted theoretical distributions as simulation input and advocate the use of empirical distribution for its simplicity and "transparent" meanings.

There is a growing recognition of problems where input random variables are multivariate or correlated. Some recent work, such as Nelson and Yamnitsky [123], Deler and Nelson [47], Ghosh and Henderson [65], Biller and Nelson [13] and [14], Lada et al. [96] and Kuhl et al. [94], have studied these two situations.

There are also cases where no standard theoretical distribution can be a reasonable fit for the data: "the data are a mixture of two or more heterogeneous populations" [103]. Cheng and Currie [35] indicate that many of these cases can be generalised to the situation where input random variables are drawn from finite mixture distributions. Most of Ford's machine breakdown duration data are multimodal and so can be described by finite mixture distributions. The term finite

mixture distribution and the methodology and process of fitting mixture distribution to Ford's machine downtimes will be introduced in Chapter 3.

2.5.2 Input Modelling of Machine Downtimes

When considering the more specific case of input modelling for manufacturing systems most of the literature recommends modelling machine breakdown durations by assuming the time between failures (TBF) and the time to repair (TTR) are independently and identically distributed and follow a well-known theoretical distribution, such as Weibull, Erlang or exponential.

In the very early stage of breakdown modelling, the exponential distribution was suggested to be a plausible distribution for all data sets of breakdown durations ([43] and [51]). Then, more researchers and modellers became aware that exponential distribution may not be a good model for machine breakdown durations as many real-life random variables cannot be well described by the exponential distribution ([128] and [129]). The normal distribution is another distribution that has been widely assumed to be an appropriate distribution for modelling breakdown durations. However, this is disputed by Law et al [106]. Other distributions have been studied on representing breakdowns in later work. Kay [92] believes that "life to failure distribution" can be demonstrated by the Weibull distribution. Some other authors like [104], [158], [159] and [105] believe that machine downtimes can be correctly represented by theoretical distributions provided that adequately well fitted theoretical distributions can be found.

Nevertheless, there are researchers who advocate the use of empirical distributions, such as [78], [54] and [142]. Carson [142] suggests that the use of an empirical distribution is probably the simplest way to use the data. Feltner and Wiener [54] also prefer the use of empirical distributions as the process for estimating a fitted theoretical distribution is very complex. Hanifin and Liberty [78] consider

that modelling machine breakdowns with theoretical distributions has risks and indicate that first, there is no actual theoretical proof that the assumed theoretical distribution fits data from a real transfer line and second, important variables in the data are "disregarded, assumed constant or forced to fit". In their work, they generated machine breakdown durations in the simulation that were exactly the same as the data they collected. The input was fixed and set as the sequence of actual start time and finish time of machine failures collected in a certain period. Therefore, under their approach, every run has exactly the same sequence of breakdown durations. However, this means that the length of the simulation run time can not be more than the amount of time over which the breakdown data has been collected. Hence, if a particular event has low frequency and a relatively short length of breakdown input is used, the simulation run length may not be sufficient to reflect the true impact of the rare events.

Some of the research on breakdown modelling of manufacturing simulation supports the use of theoretical distributions. Bradford and Martin [21] studied 10 transfer line machines' breakdown behaviour and compared the performance of average throughput of two simulation models consisting of these 10 machines. One of the two models uses actual historical data to model machine breakdowns and the other uses a negative exponential distribution to model machine up durations and uses a Erlang-2 to represent machine downtimes. The conclusion is that the averaged line yield produced with the use of standard theoretical distributions was "as accurate as using historical data". However, it is also indicated that no one distribution used (negative exponential, Weibull, Poisson and Erlang-2) could represent the time between failures and the breakdown durations accurately for all of the machines, and the breakdown durations were modelled especially badly. Some other authors like [104], [158], [159] and [105] believe that simulation models using theoretical distributions to represent machine downtimes produce accurate performance, but only when adequately well fitted theoretical distributions can be

found.

Some projects on breakdown modelling that have been undertaken in Ford preferred the use of historical data (empirical/user-defined distribution). Crosby and Murton [39] conclude that the theoretical distribution could not truly reflect the underlying distribution as the outputs were very different. Ikonen [85] states that an empirical distribution was believed to be the more accurate way to represent the actual data. Ladbrook [97] expresses his concerns that no theoretical distribution seems to be an appropriate representation of the breakdown data.

It is also indicated that much of the relevant mathematical and statistical knowledge of theoretical distribution selection and estimation of parameters are very complex [32] and "beyond the understanding of many manufacturing engineers" who happen to be the simulation modellers. Correspondingly, it takes much longer for the engineers to learn and build simulation models if applying theoretical distributions.

The factor of time limitations has been emphasized in a number of manufacturing simulation studies, such as [111], [119], [97] and [98]. Therefore, as Ma and Kochhar [111] state, it is ideal to obtain accurate repair times representation with simple and intuitively meaningful mathematical formulations that can be easily implemented in simulation software, which our proposed method aims to provide.

Chapter 3

Statistical Models of Breakdown Duration Data

The machine breakdown duration data is a collection of machine breakdown durations over a period of manufacturing time. Currently, empirical distributions are used for representing the time to repair by Ford since no common distribution appears to fit the data of breakdown durations well. The empirical distributions are input into the WITNESS simulation models in the form of a histogram.

In the case when there are no historical data available or a new machine is being modelled, Ford usually use the Erlang-2 or exponential distributions to describe the distributions of machine breakdown durations. Only the mean breakdown duration is needed to fit the Erlang-2 and exponential distributions, and this is normally provided by the machine manufacturer.

If we plot a single histogram of the entire collection of breakdown durations for each machine, we see two or more distinct peaks for most of the histograms, i.e. the breakdown duration data is *multimodal*. Figure 3.1 shows the distribution of breakdown durations for a typical machine and is clearly multimodal. Therefore, the more common statistical distributions, such as Erlang-2 and exponential,

will produce poor fits to these data. Instead, we use finite mixture distributions, allowing us to describe the multimodality.

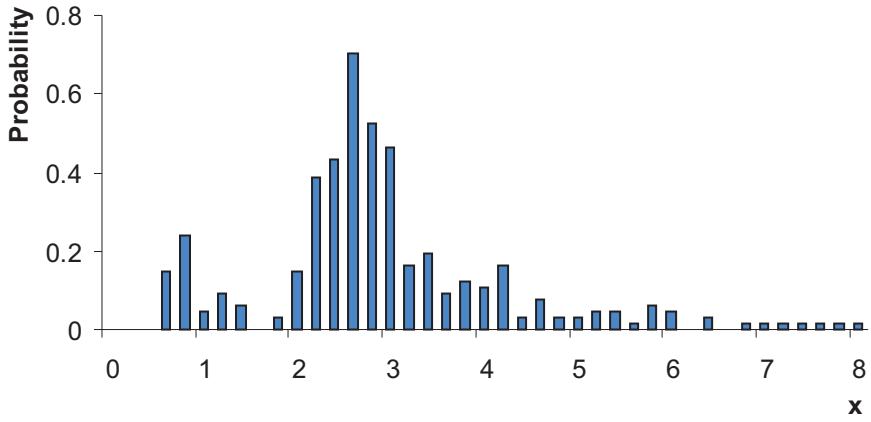

Figure 3.1: Histogram showing the distribution of the machine breakdown duration data of a machine involved in engine assembly process.

We begin with a description of finite mixture distributions, our proposed method, in Section 3.1, stating the estimation problem of fitted finite mixture models. We use a Bayesian approach for the fitting methodology and this is discussed in Section 3.2, including a brief description of the implementation of the importance sampling used to fit the finite mixture models. Section 3.3 addresses some of the issues in the raw data before carrying out the actual fitting process for the machine breakdown duration data. Section 3.4 discusses the selection of the distribution for the individual components. We investigate the relations between the components of the fitted mixture distributions for the breakdown durations of a machine and the different types of faults that cause failures of the machine in Section 3.5.

3.1 Multimodal Distributions

In statistics, a multimodal distribution is a continuous probability distribution that has multiple *modes*, i.e. whose density function has two or more distinct peaks; as illustrated in Figure 3.2. Sharing the same physical features, multimodal distributions can be used to fit a dataset that is composed of a number of distinct modes.

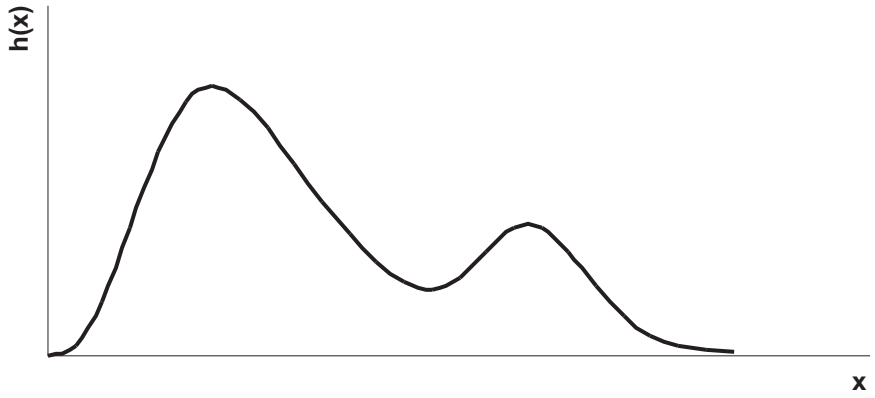

Figure 3.2: Histogram corresponding to a probability density function of a multimodal distribution with two local modes.

Mixture models are a common form of multimodal distributions. A *finite mixture model* is defined as having probability density function [115]:

$$h(x) = \sum_{i=1}^{k} w_i f_i(x|\theta_i), \tag{3.1}$$

where

$$0 < w_i \leq 1 \text{ for } i = 1, \ldots, k \tag{3.2}$$

and

$$\sum_{i=1}^{k} w_i = 1 \tag{3.3}$$

are the weights of the components whose individual densities are $f_i(x|\theta_i)$ for $i = 1, \ldots, k$. The parameter k is the number of components in the finite mixture model.

Being a particularly flexible and useful method of modelling, finite mixture models have been receiving more attention recently [76] and have been successfully applied in both practical and theoretical fields (e.g. [136], [11], [131], [35], [110] and [1]).

Other multimodal distributions exist for fitting data that are not distributed according to common stochastic models. These are generally based on using flexible families of distributions, such as the Bézier distribution ([161], [160], [123], [95] and [103]) or the Johnson family (see Chapter 12 of [90], or page 297 of [103]). The Bézier distribution exploits the properties of Bézier curves and allows the modeller to fit the cumulative distribution function $F(x)$ to a wide range of distributions of data, its flexibility being due in part to the fact that the number of parameters to be used is not fixed. Johnson distributions are based on transformations of normal variables and, although they offer a wide range of shapes of distributions, do not cope as well with multimodality.

The advantage of the use of finite mixture models is that they provide a good description of multimodal data, using parameters that have an intuitive meaning, which will make it more understandable for engineers with little expertise. They are also easy to implement in most standard simulation packages using a two-stage approach, where the component is sampled in the first step and then the input value is sampled from the component density.

We use software developed by Cheng and Currie [40] to estimate the best fitted mixture models for breakdown duration data sets. The assumption made in [40] is that all of the component densities take the same form. If we allow the component densities to take different forms in the mixture, the time spent on the fitting process will increase massively, especially with a large selection of different distribution

types and a high number of components. Therefore the probability density function 3.1 can be written as

$$h(x) = \sum_{i=1}^{k} w_i f(x|\theta_i), \tag{3.4}$$

where $0 < w_i \leq 1$ for $i = 1, \ldots, k$ and $\sum_{i=1}^{k} w_i = 1$.

In this work, we have assumed that the components follow a lognormal distribution, and so

$$f(x|\theta_i) = \frac{1}{x\tau_i\sqrt{2\pi}}\, e^{-\frac{(\ln x - \mu_i)^2}{2\tau_i^2}} \tag{3.5}$$

where

$$\theta_i = (\mu_i, \tau_i)^T \tag{3.6}$$

The choice of distributions for the component densities should be dependent on the characterisations of the data being modelled, for example the shape of the corresponding histogram and the range of the data, and the selection is further discussed in Section 3.4.

It is assumed that none of the θ_i nor the number of components k are known in the model. It is possible for components to be present in the mixture that are not represented within the data. Fitting such models is therefore a non-standard statistical problem. The main issue of the estimating problem is that standard asymptotic theory does not hold when the number of components is not known. Thus, suitable statistical tests are difficult to be constructed to identify the correct number of components. We adopt a Bayesian framework that makes use of importance sampling ([35] and [40]). This is discussed further in the next section.

3.2 Bayesian Fitting Process

We first give a brief introduction of Bayesian statistics. In Bayesian statistics, the parameters of a model are treated as random variables, such that the parameter θ is the realised value of a random variable Θ. We define the prior distribution initially, which represents the prior information about the parameter θ before the data D that the model is describing are obtained. We combine the prior information about Θ encapsulated in the prior distribution $\pi(\theta)$, with the likelihood function $P(D|\theta)$ to obtain the posterior distribution $P(\theta|D)$, such that

$$P(\theta|D) = \frac{\pi(\theta)P(D|\theta)}{P(D)}. \tag{3.7}$$

The posterior distribution represents the information about θ given the knowledge of the data and the prior information. The function $P(D)$ is a normalising factor, which is required to ensure that the posterior distribution integrates to one [19].

Formula 3.7 states that the posterior distribution is proportional to the product of the likelihood and the prior distribution, and so only the product of the likelihood and the prior distribution at any point in Θ, the parameter space of θ, need be evaluated to describe the shape of the posterior probability distribution. However, to obtain a proper probability distribution, we need to evaluate the constant of proportionality $P(D)$. The calculation of $P(D)$ is given by

$$P(D) = \int_{\Theta} \pi(\theta)P(D|\theta)\, d\theta, \tag{3.8}$$

the product of the prior probability distribution and the likelihood integrated over parameter space. In the finite mixture distribution fitting problem that we consider here, the integral cannot be computed analytically, and we use importance sampling to evaluate it. We describe the process briefly here and refer the reader to [35] for more details.

In the following we let k^* denote the unknown true number of components and let θ_i^* denote the unknown true values of the parameters of component distribution $i = 1, 2, \ldots, k^*$. For simplicity, we also assume that we can specify a maximum number of components, K, where, $0 < k^* < K$.

We use a prior distribution for the unknown parameters of the mixture model

$$\pi(\psi^k|k)\pi(k),\ k = 1, 2, \ldots, K \tag{3.9}$$

where

$$\pi(k),\ k = 1, 2, \ldots, K \tag{3.10}$$

is the prior distribution for k, and $\pi(\psi^k|k)$ for given k, is the conditional prior density of the component parameters $\psi^k = (\theta_1, \theta_2, \ldots, \theta_k,\ w_1, w_2, \ldots, w_k)$.

Suppose we fit the finite mixture model to a *sample* of breakdown duration data $x = (x_1, x_2, \ldots, x_n)$, then the posterior distribution is given as

$$p(\psi^k, k|x) = \frac{p(x|\psi^k, k)\pi(\psi^k|k)\pi(k)}{\sum_{k=1}^{K} \pi(k) \int p(x|\psi^k, k)\pi(\psi^k|k)d\psi^k},\ k = 1, 2, \ldots, K \tag{3.11}$$

where $p(x|\psi^k, k)$ is the likelihood corresponding to the mixture model with k components.

In order to determine $p(\psi^k, k|x)$, the main problem is in evaluating the denominator in Equation 3.11. The most popular sampling method used to find the posterior distribution without evaluating the denominator explicitly is Markov chain Monte Carlo (MCMC), which is described in [66]. However, in our case, as MCMC requires random moves between different k values and the form of these moves is not easy to identify, it is difficult to implement. Other authors have proposed several methods for doing this: [75] and [130] describes the reversible jump methods; [63] described an approach using indicator variables and [33] proposed a simpler approach without using the indicator variables.

We use importance sampling as the sampling method to determine the denominator in Equation 3.11 and thus the posterior distribution. *Importance sampling* is a method to evaluate a general integral $I = \int_{\Theta} m(\theta)\, d\theta$ numerically. In importance sampling, the integral can be estimated by sampling from a candidate distribution $w(\theta, \beta)$ and calculating the ratio of the integrand $m(\theta)$ at each sample point θ_i to the value of the candidate distribution at that point. By taking n samples, the integral I can be estimated by

$$\hat{I}_w = \frac{1}{n} \sum_{i=1}^{n} \frac{m(\theta_i)}{w(\theta_i, \beta)}.$$

The integral of interest here is the normalisation of the posterior probability distribution in Bayesian statistics involved in Equation 3.11, i.e. $m(\theta)$ is the product of the prior, $\pi(\psi^k|k)$ and likelihood distributions, $p(x|\psi^k, k)$.

In importance sampling, sample points are chosen from a distribution which concentrates the points where the function being integrated is large, instead of sampling them from a uniform distribution. This means that it is important to know something about the function being sampled prior to sampling. Therefore, we find the modes and covariance matrices for the posterior distribution before setting the candidate distribution. The requirement to have some knowledge of the function means that when dealing with problems in which the form of the posterior is not clear in advance, importance sampling is generally considered to be less robust than MCMC, but it is simpler to implement in the case of mixture models [35].

3.2.1 Implementation

A more detailed discussion of the implementation of the methodology we use for importance sampling can be found in [40]. We describe it here briefly.

The Nelder Mead optimization method [122] is chosen as the optimization

routine for finding the mode of the posterior distribution. The optimization routine starts by fitting a model with one component. The starting parameters for the model with k components, $1 < k < K$, are decided by the best estimates for the model with $k - 1$ components by determining the greatest discrepancy between the model and the data.

Defining

$$\widetilde{\psi}^k = arg\ max[p(x|\psi^k, k)\pi(\psi^k|k)] \tag{3.12}$$

conditional on each $k = 1, 2, \ldots, K$ as the modes of the posterior distribution, the candidate distribution for the importance sampling of a model with k components is

$$q(\psi_k, k) = \Phi(\psi^k|\widetilde{\psi}^k, \Xi^k), \tag{3.13}$$

where $\Phi(\psi^k|\widetilde{\psi}^k, \Xi^k)$ is the degenerate multivariate normal density with mean $\widetilde{\psi}^k$ and covariance matrix Ξ^k, equal to the generalised inverse of the information matrix at the mode. The reason it is degenerate is that the weights must sum to 1 (Equation 3.3).

The candidate distribution for the number of components is a uniform distribution such that

$$q(k) = K^{-1}, \quad k = 1, 2, \ldots, K. \tag{3.14}$$

Thus, the complete candidate distribution for the importance sampling procedure is

$$q(\psi, k) = q(k)q(\psi^k|k) = K^{-1}\Phi(\psi^k|\widetilde{\psi}^k, \Xi^k). \tag{3.15}$$

The implementation of the importance sampling is quite straightforward. Draw a sample of m values of $(k_j,\ \psi_j^{k_j}),\ j = 1, 2, \ldots, m$, from the candidate distribution $q(k)q(\psi^k|k)$, then the posterior distribution sample is

$$p(\psi^{k_j}|x) = \frac{p(x|\psi^{k_j}, k_j)r(\psi^{k_j}, k_j)}{\sum_{j=1}^{m} p(x|\psi^{k_j}, k_j)r(\psi^{k_j}, k_j)},\ j = 1, 2, \ldots, m, \tag{3.16}$$

where

$$r(\psi^{k_j}, k_j) = \frac{\pi(\psi^{k_j}|k_j)\pi(k_j)}{q(\psi^{k_j}|k_j)q(k_j)}. \tag{3.17}$$

The posterior distribution for k, the number of components is then equal to

$$p(k|x) = \sum_{j=1}^{m} p(\psi^{k_j}|x)\ \delta_{kk_j},\ k = 1, 2, \ldots, K, \tag{3.18}$$

where δ_{kk_j} is the Kronecker delta, such that

$$\delta_{kk_j} = \begin{cases} 1, & \text{if } k_j = k \\ 0, & \text{otherwise} \end{cases} \tag{3.19}$$

We use the value of k for which $p(k|x)$ is maximised as our final estimate for the number of components.

In addition to the advantage of easy implementation, another feature of this method is that the posterior distribution sample given in Equation 3.16 is a random sample of independent variables. Also, if the shape and location of the candidate distribution are similar to those of the posterior distribution, then the values of the posterior distribution sample will tend to be reasonably constant, and thus the integration over the posterior distribution can be performed quite accurately even with a relatively small sample size.

3.3 Data Preparation

The machine breakdown data are collected using the automatic on-line monitoring system that is connected with the machines. Due to the system setting and human errors as discussed in Section 2.4, the data contain some inaccuracies. We make some initial adjustments to the raw data, followed by a check for autocorrelation. As the ranges of most of the data sets are large, we need to find a method to

transform the data in order to obtain a good fitted model. A discussion of the data transformation is given with an example.

3.3.1 Data Validation

As we discussed in Section 2.4, there are issues concerning accuracy or availability using the existing two data collecting methods. Ford currently use electronic collection as the main method for breakdown data collection. The data collected directly from the monitoring system are known as the *raw data*. It is important to analyse and validate the raw data and make modifications if necessary, before fitting input distributions.

Carson [142] emphasised that caution needs to be taken when validating raw data. For example, Feltner and Weiner [54] studied Ford's systems and pointed out that the time difference between a failure starting and finishing was the total repair time, however this is not always real as there is a possibility of shift breaks or other activities happening within that period, as discussed in Section 2.3.

It is reasonable to model the breakdown duration data of all the elements as a whole. Therefore, we only need to extract the period of shift breaks out of the breakdown duration. We ignore any stoppage that starts inside a shift and finishes outside a shift and delete any stoppage that occurs during breaks or subtract any part of that stoppage that is overlapping with break(s).

The raw data often contain data points with very small values that are less than 30 seconds. These extremely small values appear to be suspicious. The engine assembly line that the raw data are collected from has a cycle time of 24 seconds, so it is not possible that the duration of a machine failure is smaller than half a minute due to the limitation of the response time. Three potential reasons of the recorded stoppages being less than 30 seconds were identified [97]:

1. Actual machine failures.

2. Extended cycle time but mistakenly recorded as failures.

3. Not stoppages but recorded due to errors in the setting up of the monitoring system.

However, a comprehensive investigation is required to find out the exact reason. The current assumption made by Ford engineers is that these short periods are extended cycle times and thus should be removed from the data set of machine failure times. We decided to make the same assumption when processing the raw breakdown data.

3.3.2 Data Correlation

We wish to check whether the sequence of breakdown durations demonstrates any autocorrelation. This may occur for individual machines if, for example, the machine is wearing out. In this case, breakdown durations may get longer and longer as the machine gets harder to fix. Alternatively, it may happen for the whole line if the maintenance team reacts to a lengthy period spent fixing one machine by working slowly on the next or it takes longer for a machine to be fixed because a long time is spent waiting for resources during an extremely busy period for the maintenance team. We thus wish to check whether there are any correlations within the valid breakdown duration data for all machines as well as for individual machines.

We denote a sequence of observations of machine breakdown durations, a *time-series*, as $x_1, x_2, \ldots, x_n$. The interval j unit(s) (in this case, j breakdowns) between two observations x_i and x_{i+j} is referred to as the *lag*; and for a sequence of n observations, there are $n - 1$ possible lags. The lag j autocorrelation is defined as the correlation between $x_1, x_2, \ldots, x_{n-j}$ and $x_{j+1}, x_2, \ldots, x_n$. Correlation between x_i and x_{i+j} would indicate that the time to repair a machine is possibly dependent on previous repair time data and the breakdown duration data cannot be considered as independent random variables.

For the breakdown duration data of all of the 39 machines in the assembly line, there are 7493 observations. Figure 3.3 is the plot of autocorrelations of all possible lags $1, 2, \ldots, 7492$ of this data set with approximate $\alpha = 0.05$ critical bands for the hypothesis that the correlations are equal to zero, generated by Minitab. It is seen from this plot that the autocorrelations of some lags exceed the approximate $\alpha = 0.05$ critical bands, which suggests that the absolute value of autocorrelations of these lags are statistically significantly greater than zero. However, the largest of all, lag 1211 autocorrelation, is 0.0958, which is a quite small value. Since there are 7493 observations included in the data set, we wish to check whether it is the influence of outliers that causes the lag 1211 autocorrelation to be relatively high.

We thus examine the lag 1211 autocorrelation more closely by making a scatter plot of $X_1, X_2, \ldots, X_{5981}$ against $X_{1212}, X_{1213}, \ldots, X_{7493}$, which is given in Figure 3.4. As shown in this scatter plot, there is no obvious correlation between the majorities of points in the two time series. It is possible that the one outlier circled in Figure 3.4 might be the reason that lag 1211 autocorrelation is high. Thus, we delete that one outlier. Carrying out a Pearson correlation statistic test for the two time-series of lag 1211 after deleting the outlier, the p-value is 0.826, which suggests there is no significant correlation between the two time-series. Based on this more detailed analysis, we believe we may still assume that the breakdown duration data for all of the 39 machines is made up of independent observations, i.e. the repair time of the current failure of any machine does not have influence on the repair time of the $1211st$ failure later of any machine in the assembly line.

The relationship between the current repair time and the next repair time is of most interest. If there are other factors that might affect the breakdown durations, such as the availability of maintenance operators or the age of a machine, the lag 1 autocorrelation should be able to indicate this by having a very large value. In other words, in this case it is whether the lag 1 autocorrelation is zero that is of most interest rather than any other autocorrelation with a greater lag. Therefore,

we focus on the calculation and analysis of the lag 1 autocorrelation for the whole data set of the 39 machines as well as for the data sets of individual machines.

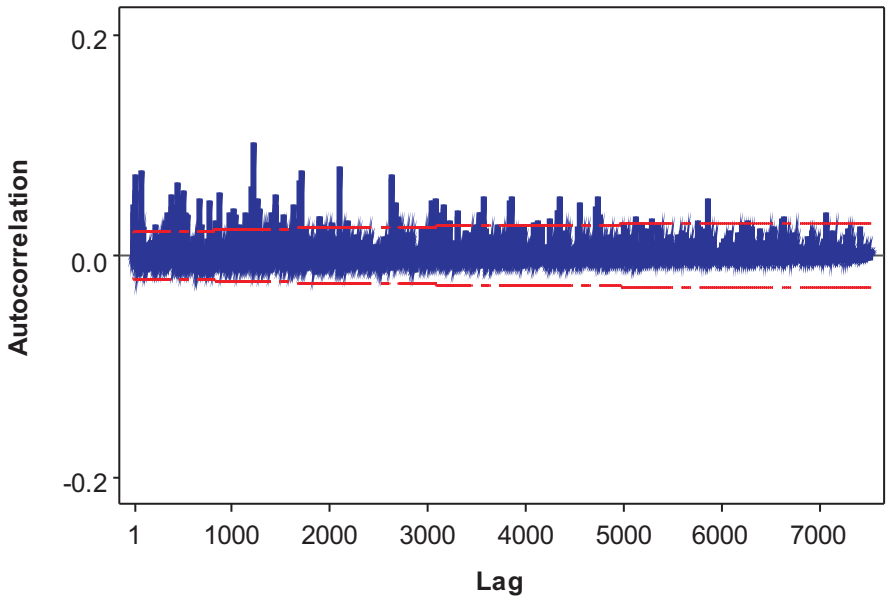

Figure 3.3: Autocorrelation of lags $1, 2, \ldots, 7492$ within the data set of breakdown durations for all 39 machines in the assembly line. Red curve indicates the 5% significance limits for the autocorrelations.

The lag 1 autocorrelation for the whole data set of the 39 machines is 0.0448, which is an extremely small value. Although the 5% significance limits shown in Figure 3.3 suggests that 0.0448 is statistically significantly greater than zero, it is possibly because the whole data set for all machines contains such a large number of observations (7493) that the statistical test rejects the hypothesis that the correlations are equal to zero. Thus, we assume that there is no influence on the next repair time of any machine from the duration of the current repair.

For the individual machines, 36 out of 39 have lag 1 autocorrelations that are not significantly different from zero. For example, Figure 3.5 gives the autocorrelations of lag 1 and all other possible lags for the breakdown duration data of machine ML08, in which we can see that the values are all fairly small and can

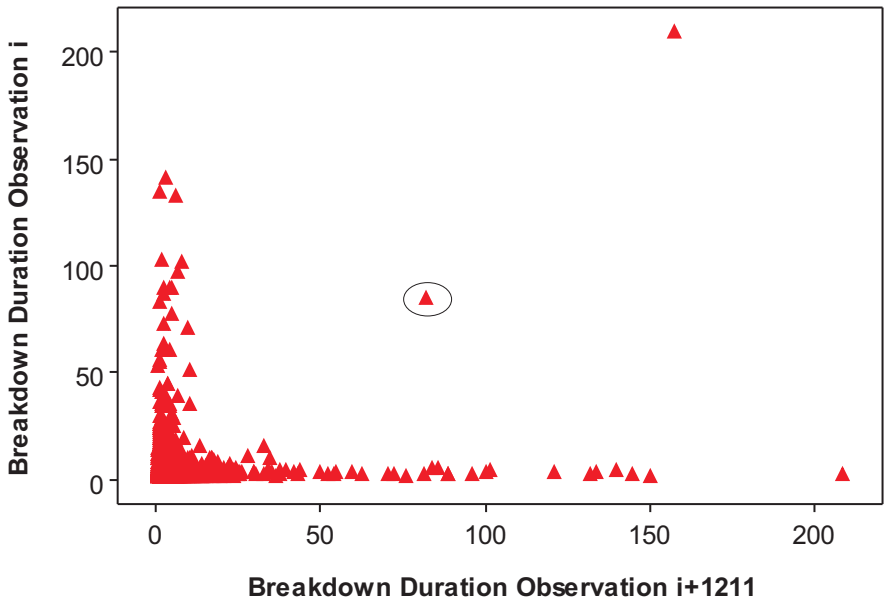

Figure 3.4: Scatter plot of observation i vs. observation $i + 1211$ in the breakdown duration data set for all 39 machines. The circled point indicates an outlier.

be considered as zero according to the 5% significance limits. However, there are 3 machines: ML17, ML07 and ML36, which have lag 1 autocorrelations that are significantly different from zero. Therefore, we examine the data sets for these three machines more closely to decide whether we can assume there is no apparent autocorrelation within the breakdown duration data for these three machines.

For machine ML17, the lag 1 autocorrelation is 0.104, which is still fairly close to zero. Since the breakdown duration data set for ML17 has 1310 observations, it is possible that the statistical test rejects the hypothesis that the correlations are equal to zero because of the size of the data set. As this data set has a large number of data points, with the majority falling into a very small range, the test can pick up spurious correlations. Thus, we believe that for machine ML17, there is no apparent correlation between the repair time for previous failure and that for the current failure.

For machines ML07 and ML36, we believe the relatively high lag 1 autocorre-

lations are probably due to the effect of an extremely small number of outliers. Within these two machines, machine ML07 appears to be more problematic as ML36's lag 1 autocorrelation is less than 0.20 while ML07's is greater than 0.30. Thus, we use the investigation of the data of machine ML07 as an illustration to demonstrate the impact of outliers.

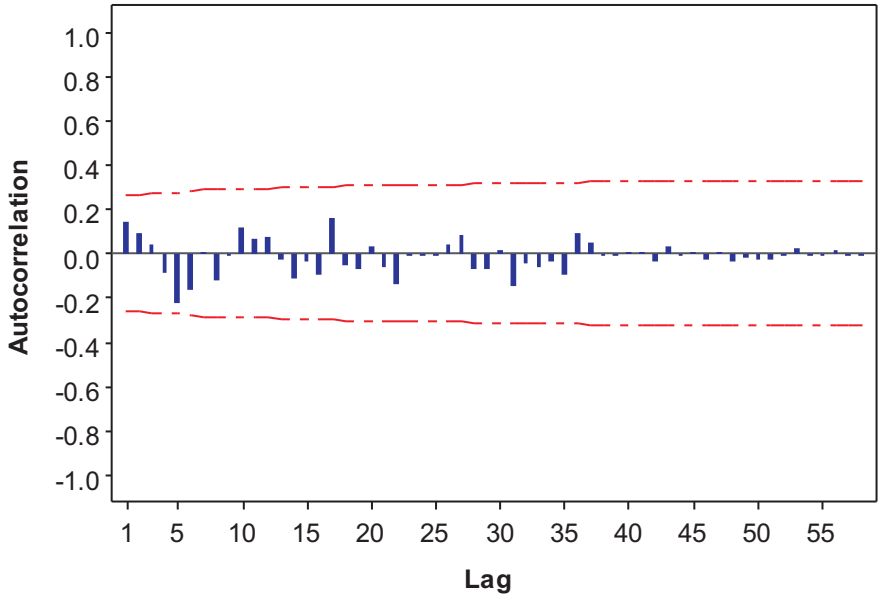

Figure 3.5: Autocorrelation of lags $1, 2, \ldots, 58$ within the data set of breakdown duration for machine ML08. Red curve indicates the 5% significance limits for the autocorrelations.

Figure 3.6 gives the autocorrelation of lags $1, 2, \ldots, 60$ for the breakdown duration data set of machine ML07, and it can be seen that only the lag 1 value is suggested to be significantly higher than zero. We make the scatter plot of the two lag 1 stochastic process given in Figure 3.7. There is no obvious correlation between the majority of points in the two time series that can be seen in this scatter plot. It is possible that the two outliers circled in Figure 3.7 might be the reason that the lag 1 autocorrelation of ML07 is relatively big. Thus, we delete those two outliers and get the new scatter plot in Figure 3.8, in which there seems to be no obvious correlation. After deleting the two outliers, the p-value result of the

Pearson correlation statistic test for the two stochastic processes of lag 1 is 0.826, which suggests that there is no significant correlation between the two stochastic processes.

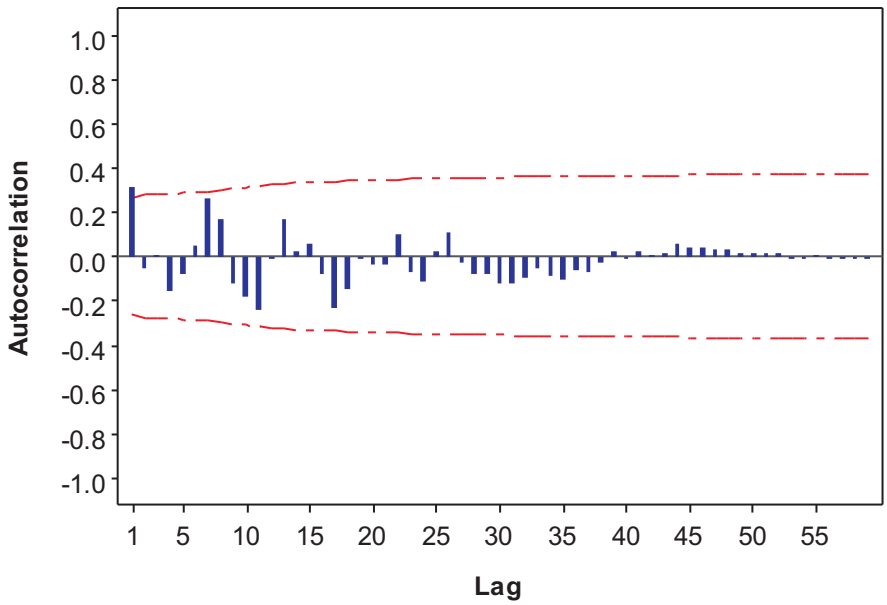

Figure 3.6: Autocorrelation of lags $1, 2, \ldots, 60$ within the data set of breakdown duration for machine ML07. Red curve indicates the 5% significance limits for the autocorrelations.

Since it is illustrated that the lag 1 autocorrelation for machine ML07 is relatively high because of the two outliers, we believe that we can still assume that the breakdown duration data for machine ML07 are independent observations, i.e. the time to repair the current failure of machine ML07 does not have any effect on the time to repair the next failure of ML07. We also believe that it is due to the impact of only one outlier in the data set for machine ML36 that the autocorrelations are statistically non-zero, as after deleting that outlier, the lag 1 autocorrelation drops dramatically from 0.193 to 0.028.

Therefore, from the analysis of the autocorrelation values and testing results, we are able to assume that there is neither obvious correlation between the failure

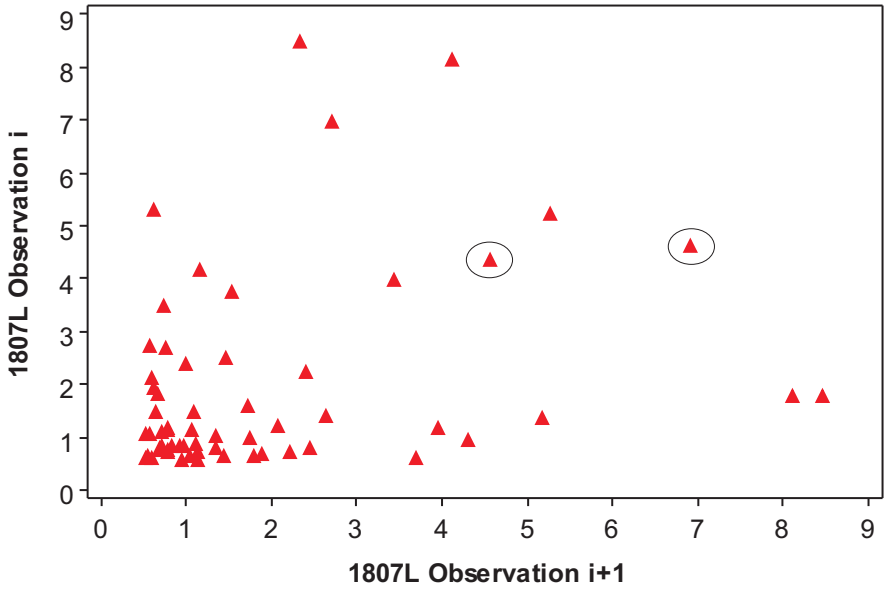

Figure 3.7: Scatter plot of observation i vs. observation $i+1$ in the breakdown duration data set for machine ML07. The circled points are identified as outliers.

durations of one machine and that of any other machine in the assembly line nor apparent correlation between the current repair duration and the next repair duration for the same machine; i.e. the breakdown durations are independent of each other.

We also wish to check whether there is any correlation between the breakdown durations of a machine failure and the time this failure occurred, e.g. durations may be longer at the end of a week. The time series plot for the whole breakdown duration data set of 39 machines shown in Figure 3.9 shows no apparent correlation between the two. Similar results can be drawn from the time series plots for individual machines. Thus, it is believed that the time a failure happens does not have any impact on the time that it takes to repair it.

Based on the above analysis of correlations for the breakdown duration data, we may assume that the breakdown durations are independent random variables and furthermore have no obvious correlation with the time the failures occur.

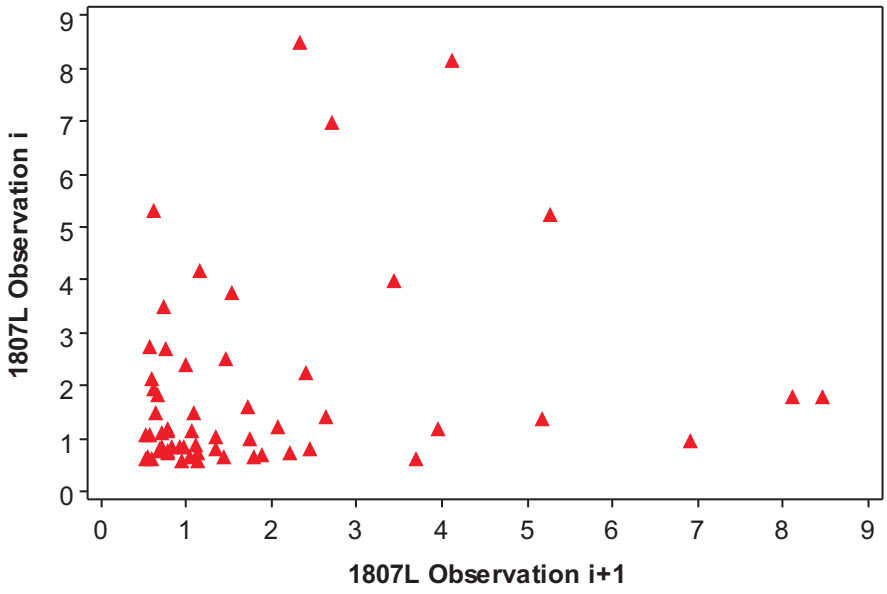

Figure 3.8: Scatter plot of observation i vs. observation $i + 1$ in the breakdown duration data set for machine ML07, after deleting the two outliers circled in the previous scatter plot in Figure 3.7.

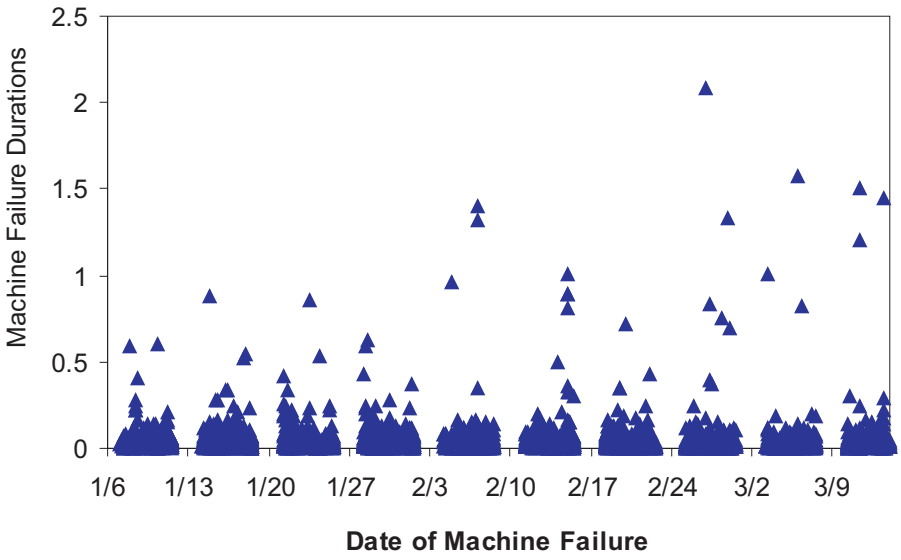

Figure 3.9: Time series plot of the breakdown duration data set for all 39 machines in the engine assemble line collected in the period between 07 January 2008 and 14 March 2008.

3.3.3 Data Transformation

After removing the invalid data points, the data has a wide range of values. We transform the data in order to reduce its range so as to improve the accuracy of the fitting process. We considered two transformations: (1) taking logs; and (2) taking the square root.

When taking logs, durations of less than one minute are transformed to negative values. This limits the choice of component distribution that can be used.

Taking the square root of the original data shrinks the data's range and ensures all of the transformed data are positive.

Here we show the advantage of the data transformation using an example. We fit mixture distributions for a sample of valid breakdown duration data and also for the transformed data of the same sample and then compare the two fittings. We here assume the components of the mixture model are lognormal distributions. We obtain the best-fit lognormal mixture distribution for the valid untransformed data first, which has 3 components. The histogram of the original data and the plot of the fitted model's Probability Density Function (PDF) are given in Figure 3.10. Plots of the original data's Empirical Distribution Function (EDF) and the fitted mixture model's Cumulative Distribution Function (CDF) on four different scales are given in Figure 3.11 (a, b, c, d).

Both Figure 3.10 and Figure 3.11 show that the fitted mixture model is not very accurate. In Figure 3.11, (a) and (b) show that the fitted model fits the part where data are greater than 10 minutes quite well; (c) and (d) suggest that the distribution is a poor fit to the data that are smaller than 8 minutes. More than 87% of the data in this example is smaller than 8 minutes, which means that the fitted mixture model appears to fail to fit the majority of the sample well.

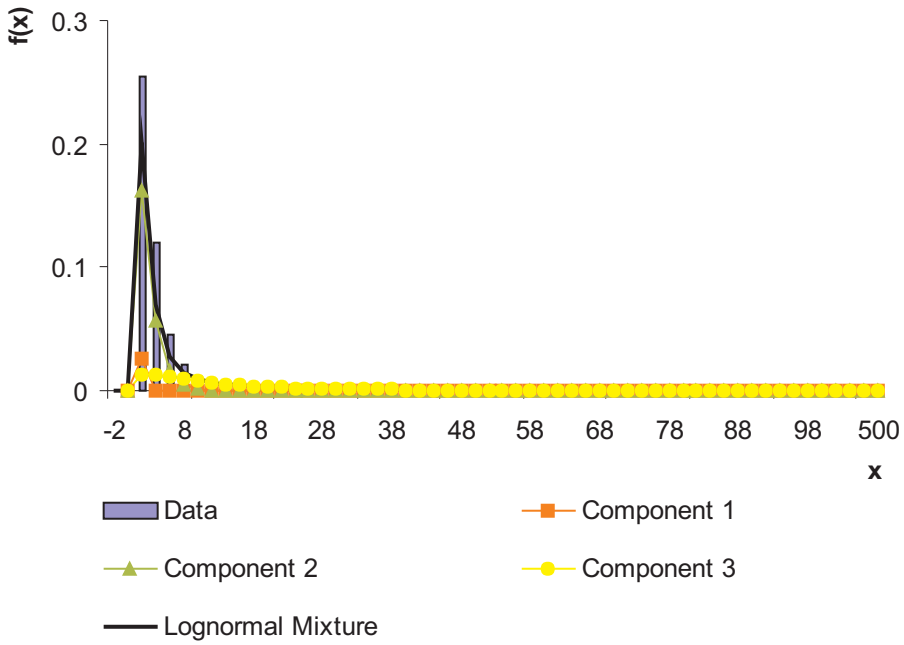

Figure 3.10: Histogram of the valid untransformed data and plot of the PDF of the fitted 3-component lognormal mixture model.

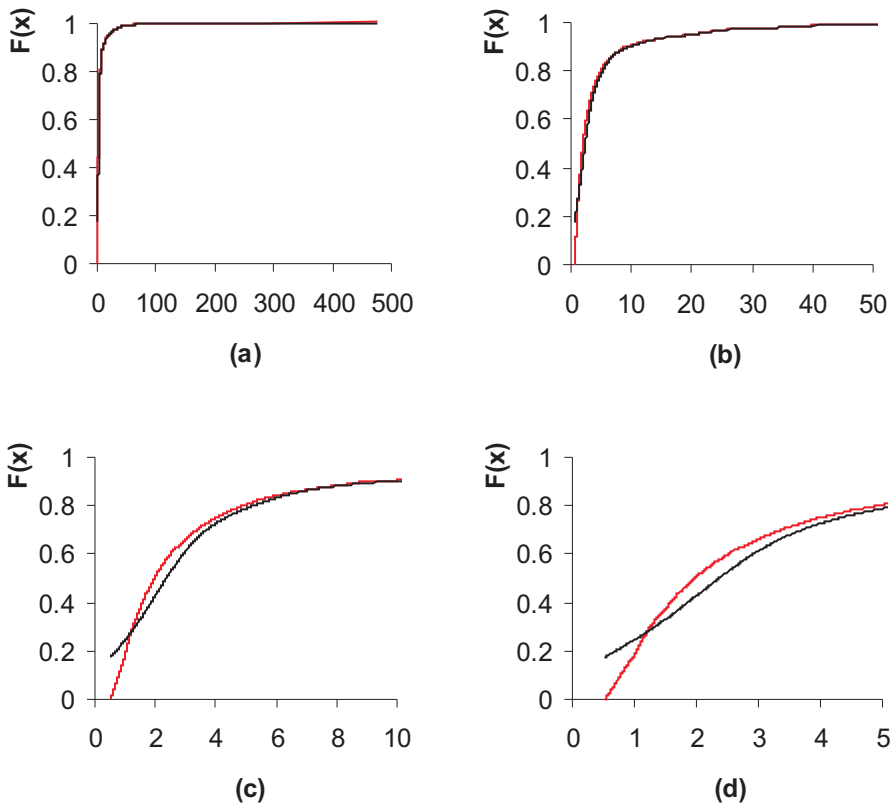

Figure 3.11: Plots of the EDF and the best-fit CDF of the untransformed data on four different scales. Red line for EDF and black line for CDF in all four plots.

We then obtain the best-fit lognormal mixture model for the transformed data (square roots of the same data). This fitted mixture model has 4 components. The histogram of the transformed data and the plot of the best-fit mixture model's PDF, and plots of the transformed data's EDF and the fitted mixture model's CDF are given in Figure 3.12.

Both of the charts in Figure 3.12 show that the best-fit distribution is a reasonably good fit to the transformed data, which means that our fitting method deals with the transformed data set better than with the untransformed one. Also for the implementation in simulation models, the transforming is straightforward; simulations first generate the transformed data from the fitted models and then transform back to the breakdown duration data by taking their squares.

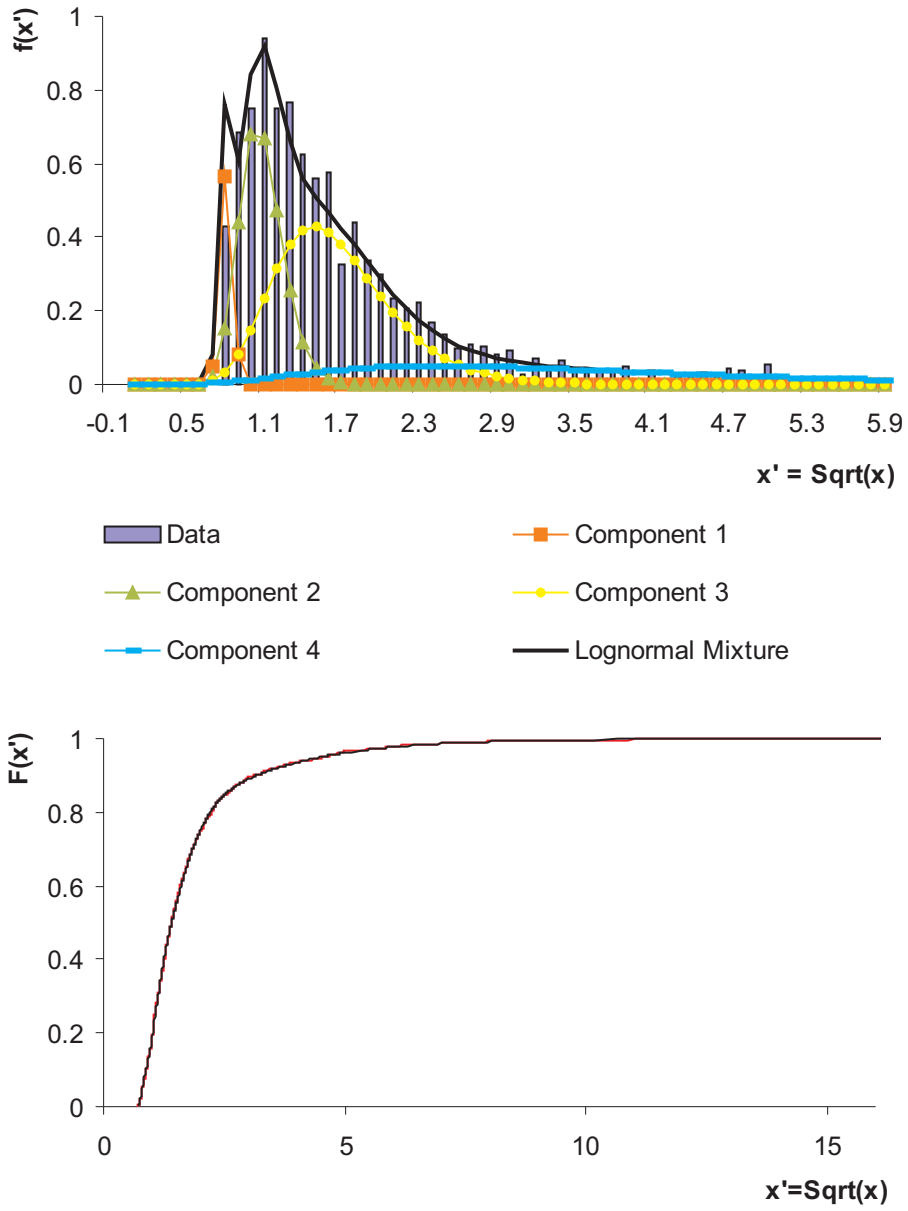

Figure 3.12: The first chart includes the histogram of the transformed data and the PDF of the fitted 4-component lognormal mixture model. The second chart includes the EDF of the transformed data and the CDF of the fitted lognormal mixture distribution; red line for EDF and black line for CDF.

3.4 Component Distribution Selection

In order to get an adequately fitted mixture distribution, it is important to choose an appropriate component distribution. In our program for estimating fitted mixture distribution, there are seven choices for component distribution: extreme, negative extreme, Weibull, normal, lognormal, gamma and inverse Gaussian distributions. The most appropriate distribution for representing the component distributions is selected from within these seven types.

The histogram of the transformed breakdown duration data generally skews to the right and has a long tail, therefore normal or negative extreme are considered to be inappropriate distributions as the PDF curve of the former is symmetric and that of the latter distribution skews to the left. The remaining distributions: extreme, Weibull, lognormal, gamma and inverse Gaussian, seem to be reasonable choices, as the PDF curves of these five distributions all have a similar shape to the breakdown duration data. To find the best distributions for components out of the remaining five choices, we fit mixture distributions using the five different component distributions for the same sample of transformed breakdown duration data used in Section 3.3.3 and then compare the five fitted distributions.

The histogram and EDF plot of the data and the plots of the fitted mixture models' probability density functions and cumulative density functions using the five different component distributions: lognormal, Weibull, gamma, extreme and inverse Gaussian are shown in Figures 3.12, 3.13, 3.14, 3.15 and 3.16, respectively.

Comparing the five fitted distributions, it appears that three distributions: the extreme, inverse Gaussian and lognormal mixture distributions, are the most robust as their best-fit distributions contain only 4 components each and fit the data very well. The Weibull mixture distribution contains 8 components and gamma mixture distribution contains 6 and both still seem to fail to fit the highest peak in the data. Furthermore, as the mixture distributions are ultimately required to be input into

the simulation models built in the WITNESS software, it is essential to choose a distribution that is convenient and simple to code in the software language. Thus, the lognormal distribution is selected to be the component distribution to analyse the breakdown duration data as it is the only one of the three remaining types of distributions that can be easily input into the WITNESS models.

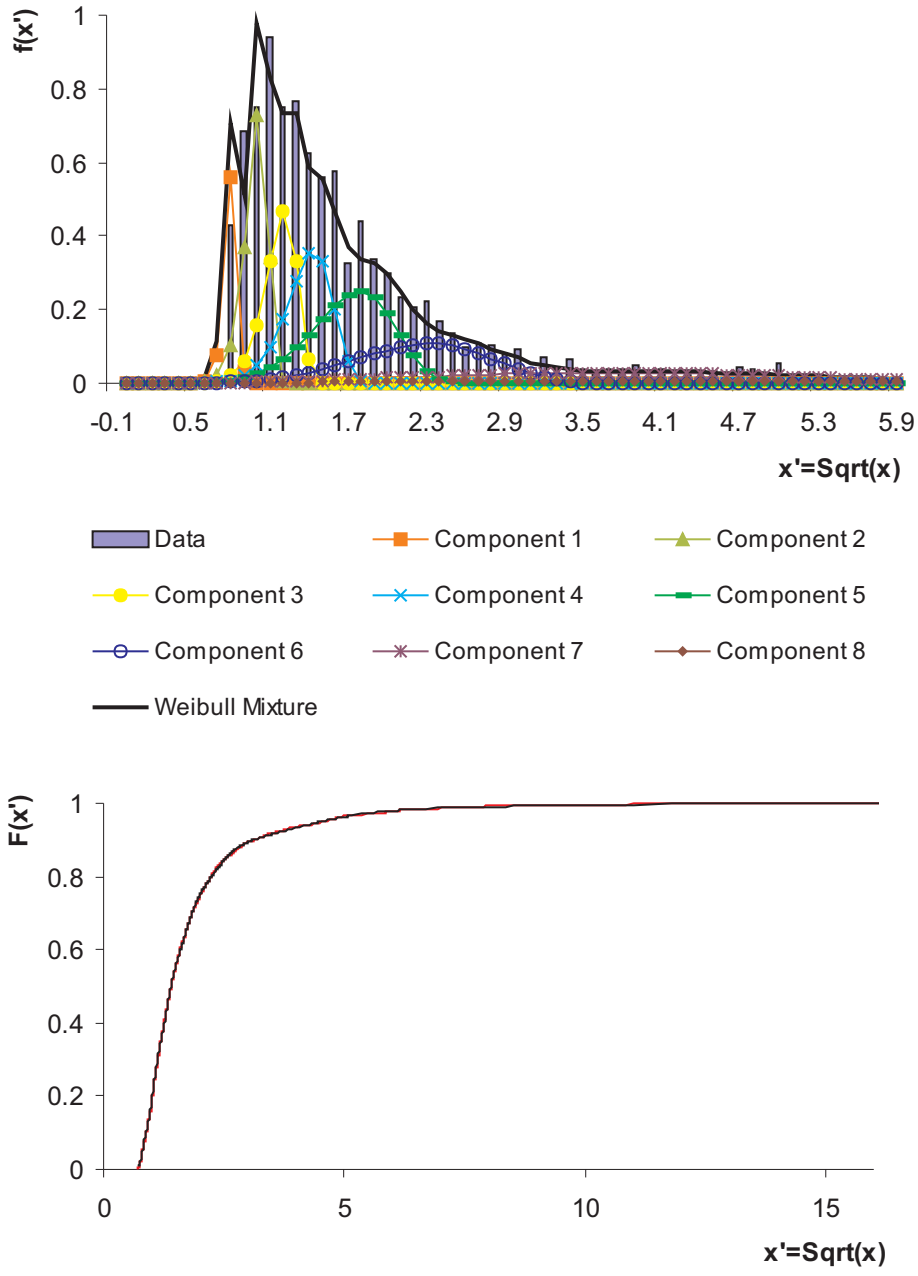

Figure 3.13: The first chart includes the histogram of the same sample of transformed data shown in Figure 3.12 and the PDF of the fitted 8-component Weibull mixture model. The second chart includes the EDF of the transformed data and the CDF of the fitted Weibull mixture distribution; red line for EDF and black line for CDF.

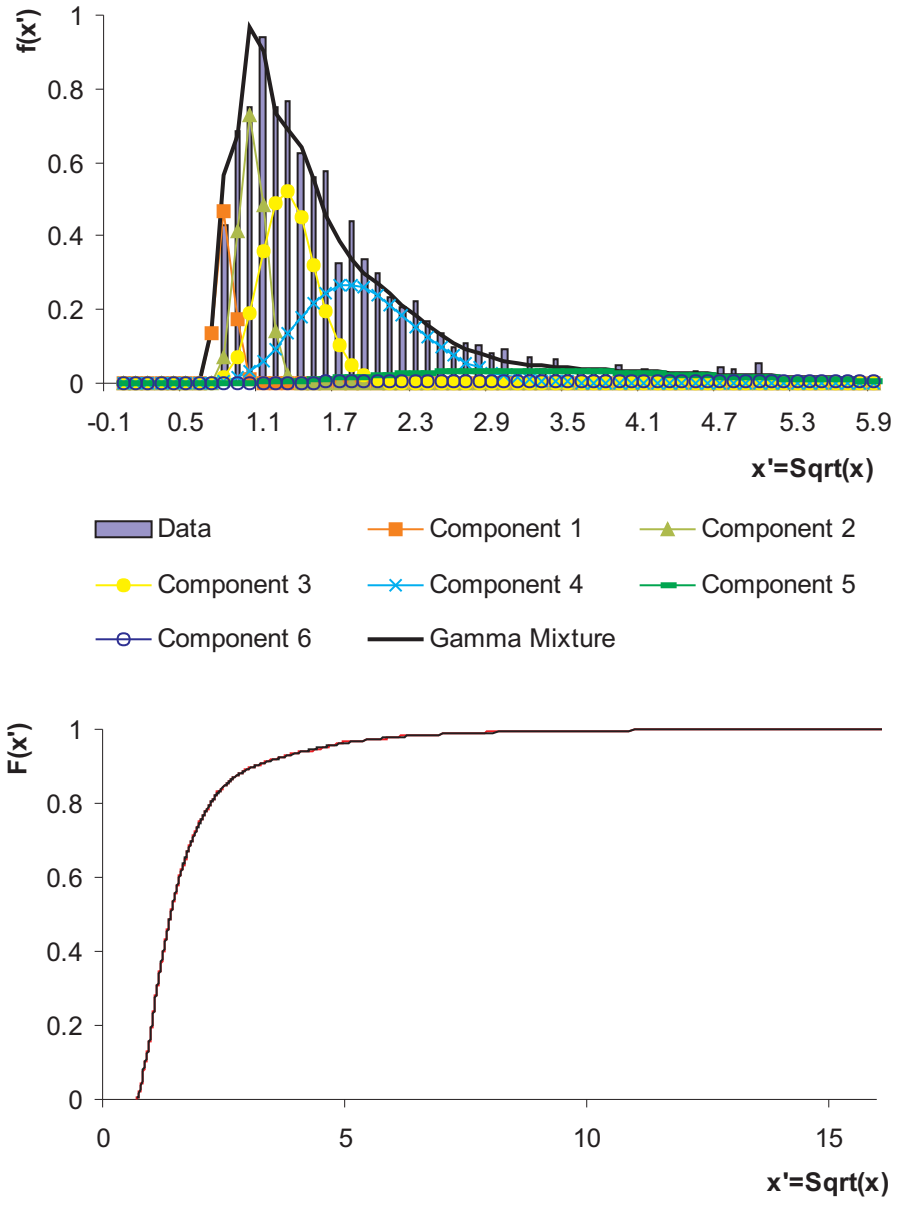

Figure 3.14: The first chart includes the histogram of the same sample of transformed data shown in Figure 3.12 and the PDF of the fitted 6-component gamma mixture model. The second chart includes the EDF of the transformed data and the CDF of the fitted gamma mixture distribution; red line for EDF and black line for CDF.

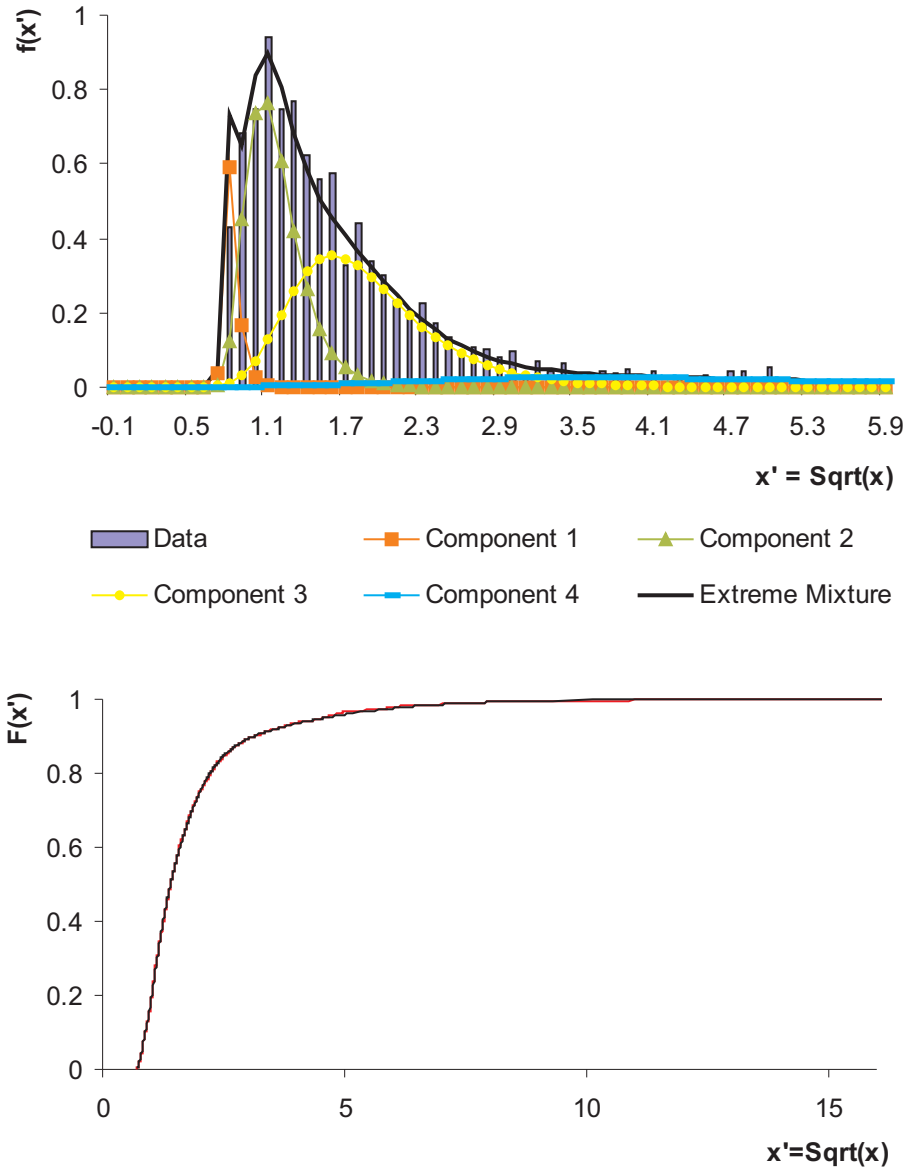

Figure 3.15: The first chart includes the histogram of the same sample of transformed data shown in Figure 3.12 and the PDF of the fitted 4-component extreme mixture model. The second chart includes the EDF of the transformed data and the CDF of the fitted extreme mixture distribution; red line for EDF and black line for CDF.

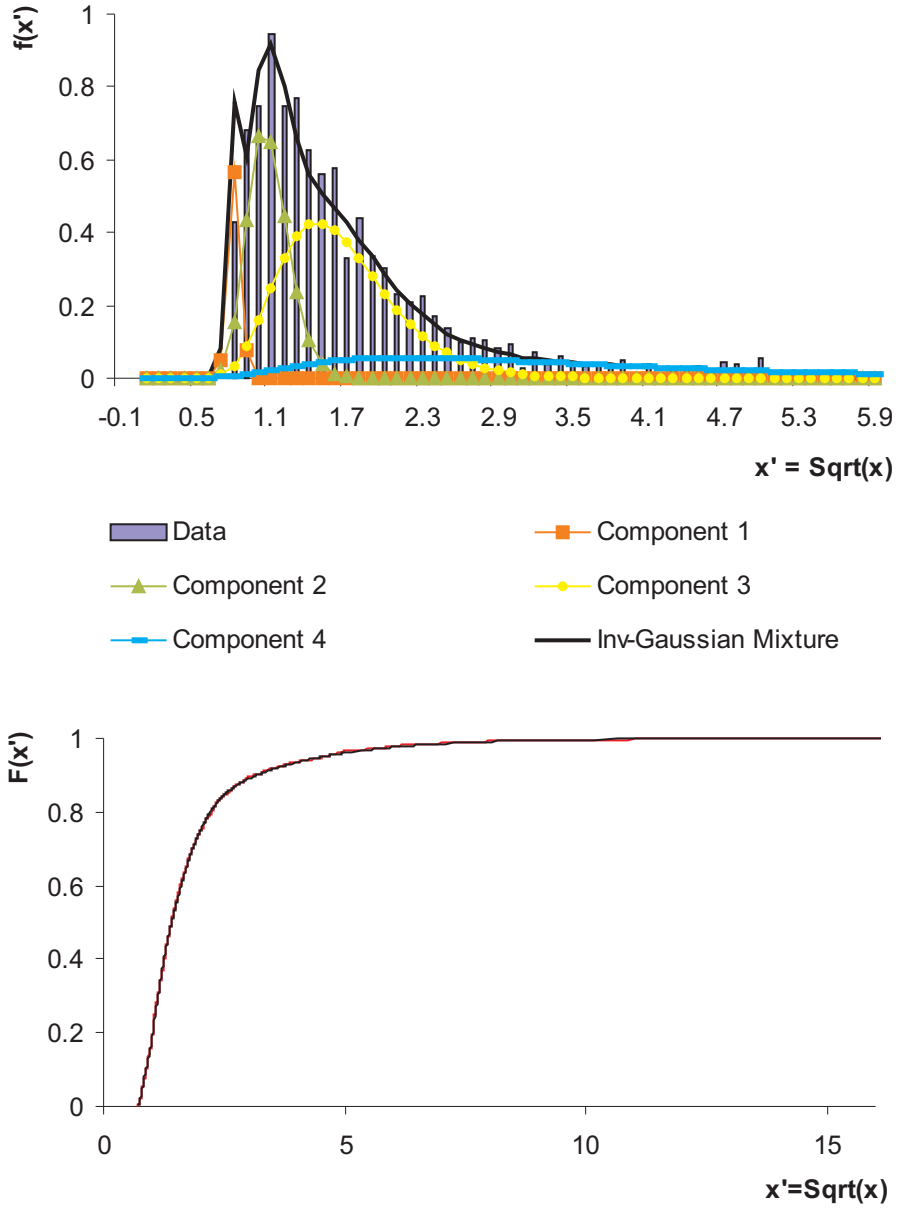

Figure 3.16: The first chart includes the histogram of the same sample of transformed data shown in Figure 3.12 and the PDF of the fitted 4-component inverse Gaussian mixture model. The second chart includes the EDF of the transformed data and the CDF of the fitted inverse Gaussian mixture distribution; red line for EDF and black line for CDF.

3.5 Relating Components with Faults

The motivation of this section is to investigate whether each component in the fitted mixture distribution for the breakdown duration data of all faults reflects one particular fault or one particular group of similar faults. We here use a sample of breakdown durations data collected within a period of three months for machine ML01 as an example to show the relations between the groups of faults and components in the fitted mixture distribution for the data. The data set includes 170 failures that are caused by the occurrence of 12 different faults. In this data set, repair duration varies from 52 seconds up to a maximum of 59 minutes for all failures. For failures caused by the same fault, the durations for two different repairs can differ by more than 10 minutes.

We obtain the best-fit lognormal mixture distribution for the breakdown duration data set of ML01, which has 3 components: 2 distinct components with means at 0.93 and 1.81, and one with a fairly flat shape spread out over the whole data range. The probability histograms for the repair durations of failures that are caused by each of the 12 different faults and the PDF plot of the fitted lognormal mixture distribution are given in Figure 3.17, where the faults are distinguished by different colours. As shown in this figure, the repair times data for the 12 different faults are fairly spread out. Nevertheless, it can be seen that the histograms of some faults have only one peak corresponding to either component 1 or component 2; while the histograms of some other faults, such as 18997, 29685, 29621, have two peaks corresponding to both components 1 and 2.

On the whole, it is reasonable to say that there are no remarkable relations between the components in the fitted mixture distribution for the data and the individual faults that cause the failures recorded in the data.

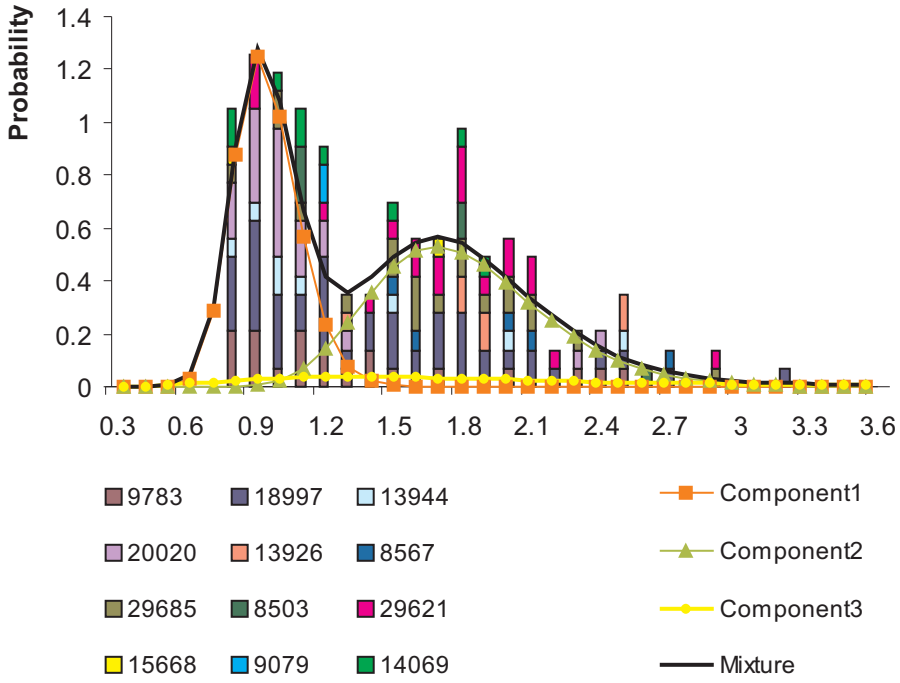

Figure 3.17: Histogram of breakdown duration data for machine ML01; the different colours represent different groups of faults, and the plot of the PDF of the fitted 3-component lognormal mixture distribution.

Chapter 4

Estimating the Similarity Matrix

Before classifying the machines, we measure their similarity by calculating the goodness of fit statistic between the two sets of breakdown duration data. The breakdown duration data sets have uneven numbers of data points and we do not wish to assume distributions for the data at this stage. The Two-Sample Cramér-von Mises goodness of fit statistic [5] can cope with these characteristics of the data, although p-values are only tabulated for a few examples. Bootstrap resampling allows estimation of the sample distribution of almost any statistic using only very simple methods. We therefore use bootstrap resampling [50] to estimate the p-values for each comparison.

We first give a brief literature review in Section 4.1. Section 4.2 gives an introduction of the Cramér-von Mises statistic as well as some other goodness of fit statistics. The basic process of bootstrapping and its common applications are introduced in Section 4.3. We then describe the methodology that we have used to generate the similarity matrix in Section 4.4. An explicit study of the method is given in Section 4.5 by testing on random samples generated from known distributions. The method is applicable in a wide range of situations, not strictly for analysing breakdown duration data. Two real-life examples are given in Section 4.6: (1) assessing similarities between six machines using their real breakdown

duration data; (2) analysing the similarities between hospital procedures based on patients' lengths-of-stay in a group of private hospitals [41].

4.1 Index of Similarity

The raw data matrix is an $n \times p$ matrix, X, that consist of observations x_{ik}, where x_{ik} denotes the value of the kth variable observed for the ith object. The raw data matrix is required to be transformed into an $n \times n$ matrix of pairwise dissimilarities or pairwise similarities for many classification methods. The dissimilarity or similarity matrix consists of d_{ij}, where d_{ij} denotes the dissimilarity or similarity between the ith and jth objects. Twelve similarity structures, S, are listed in [79]. A large number of empirical studies have proposed different methods of proceeding from X to S ([48], [30], [22], [24], [148], [121], [59], [117], [31], [100], [72], [20] and [91]).

One of the most commonly used similarity structures is the Euclidean distance. When all variables are quantitative, it can be measured by calculating the sum of the Euclidean distances between the data points from object i and those from object j. Other similarity structures tend to work on a similar principle but different distance measures are used. Exceptions are where the raw data matrix is not an $n \times p$ matrix, where the number of data points of object i is not necessarily the same as the number of data points of object j, such as in the data we have.

We measure the similarity of the breakdown duration data of any two machines using the Two-Sample Cramér-von Mises goodness of fit statistic [5]. Bootstrapping is used to determine the p-value, i.e. the significance level, of the statistic of the pair of machines, which gives the probability that the breakdown duration data for these two machines are drawn from the same distribution. The similarity matrix is then made up of the p-values of every pair of machines and thus is symmetric and real-valued.

4.2 Goodness of Fit Statistics

Generally, the goodness of fit problem is to test the null hypothesis that a sample comes from a population defined by a distribution function, given a random sample and a distribution function. The goodness of fit statistic is compared with tabulated criterion values to describe how well the distribution fits the given sample. For most commonly used tests for this problem, such as the χ^2 tests, information about the underlying distribution is required before constructing the test [151]. We use the Cramér-von Mises statistic to test whether two samples of breakdown duration data from two machines come from the same unspecified distribution. The advantage of the Cramér-von Mises statistic is that it is distribution-free and therefore there is no need to make any assumptions about the distributions of the data sets being analysed [5]. It also allows for the data sets having uneven sizes.

The most obvious contenders to the Cramér-von Mises statistic are three non-parametric statistics: the Kolmogorov-Smirnov [151], Somer's D concordance statistic [153] and Mann-Whitney tests [114]. The Mann-Whitney test aims to determine whether the data points in one set of data are greater than those in the other, whereas we wish to establish whether the data coming from two objects could have been drawn from the same distribution; the Mann-Whitney test is therefore less appropriate here. In the general situations we consider here, the data sets may have different number of data points; thus, the Somer's D concordance statistic, which describes the strength of concordant relations between pairs of variables and deal with data sets with identical size, is less applicable here. The Kolmogorov-Smirnov statistic is the closest in form and objective to the Cramér-von Mises statistic but has been shown in simulation studies to have a lower power ([151] and [42]).

Given two samples of breakdown duration data $X = (x_1, x_2, \ldots, x_n)$, and $Y = (y_1, y_2, \ldots, y_m)$ for machines M_x and M_y respectively, the Cramér-von Mises

T criterion for testing that the two samples, X and Y, come from the same unspecified continuous distribution is

$$T = [nm/(n+m)] \int_{-\infty}^{\infty} [F_n(v) - G_m(v)]^2 dH_{n+m}(v), \tag{4.1}$$

where $F_n(v)$ is the EDF of the first sample; that is, $F_n(v) = (\text{no. of } x_i \leq v)/n$; $G_m(v)$ is the EDF of the second sample and $H_{n+m}(v)$ is the EDF of the two samples together; that is, $(n+m)H_{n+m}(v) = nF_n(v) + mG_m(v)$.

As $H_{n+m}(v)$ gives each observation in the pooled sample a weight of $1/(n+m)$, Equation 4.1 can be calculated by

$$T = [nm/(n+m)^2] \left\{ \sum_{i=1}^{n} [F_n(x_i) - G_m(x_i)]^2 + \sum_{j=1}^{m} [F_n(y_i) - G_m(y_i)]^2 \right\}, \tag{4.2}$$

Let r_i and s_j be the ranks in the pooled sample of the ordered observations of the two samples X and Y, respectively, where $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m$. Then

$$F_n(v) - G_m(v) = i/n - (r_i - i)/m, \tag{4.3}$$

where $v = x_i$, the ith x-observation and

$$F_n(v) - G_m(v) = (s_j - j)/n - j/m, \tag{4.4}$$

where $v = y_j$, the jth y-observation. Thus we can write the criterion T as

$$T = \frac{U}{nm(n+m)} - \frac{4nm-1}{6(n+m)}, \tag{4.5}$$

where

$$U = n \sum_{i=1}^{n} (r_i - i)^2 + m \sum_{j=1}^{m} (s_j - j)^2. \tag{4.6}$$

To test the null hypothesis that the two samples are drawn from the same distribution, all of the observations are ordered, the ranks $r_1 < r_2 < \ldots < r_n$ of the n observations from the first sample and the ranks $s_1 < s_2 < \ldots < s_m$ of the m observations from the second sample are then determined and T is computed. If T is too large, we reject the null hypothesis, that the samples are drawn from the same distribution.

Generally, tabulated criterion values are used to decide the significance level of the goodness of fit statistic. However, for the Two-Sample Cramér-von Mises goodness-of-fit test, tabulated criterion values are not very extensive and do not cover the samples that we are dealing with: for example, only standard criterion values for samples with up to 8 data points and that for samples both with infinite number of data points are given in Anderson [5], while the number of data points of breakdown duration data sets for machines varies from 9 to 1310. Therefore, bootstrapping is used to determine the p-values of the Cramér-von Mises statistics for the breakdown duration data sets of each possible pair of machines.

4.3 Basic Bootstrapping

Bootstrapping is a practical and effective method for estimating the standard error, the confidence intervals or the distribution of statistical estimates of variables by resampling ([44] and [34]). Efron and Tibshirani [50] state that bootstrap is a computer-based implementation of basic statistical concepts. Suppose we have a random sample that is generated from a unknown probability distribution. We have calculated a statistic of interest such as the mean from the observed data and we wish to know the statistic's behaviour, for example its distribution. A number of bootstrap samples can be drawn from the empirical distribution of the observed data and thus the same number of replications of the statistic can be calculated to form a distribution of the statistic, the process of which is described in the

following.

Let $s(\mathbf{Y})$ denote the statistic calculated from samples $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$. Assume that the Y_i are mutually independent samples, i.e. Y is a random vector. Bootstrapping is a numerical method for finding $G(s)$, the distribution of the statistic $s(Y)$. Generally, $F(y)$, the distribution of Y, is unknown, but $F_n(y|\mathbf{y})$, the EDF of the observed data $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ is available. We generate a sample from $F_n(y|\mathbf{y})$ instead of $F(y)$, which is equivalent to drawing a sample of the same size n from the original set of y's with replacement, as **y** is a set of observations that can be assumed to be independent and identically distributed. We call such a sample a *bootstrap sample*, and write it as $\mathbf{y}^* = (y^{1*}, y^{2*}, \ldots, y^{n*})$. As in the basic process above, B numbers of such bootstrap samples are drawn, and the statistic $s^{j*} = s(y^{j*})$ is calculated from each bootstrap sample. Then the *empirical distribution function* (EDF) of the bootstrap statistics $\mathbf{s}^* = (s^{1*}, s^{2*}, ..., s^{B*})$ given by

$$G_B(s|\mathbf{s}^*) = \frac{(\text{no. of } s^{j*} \leq s)}{B} \tag{4.7}$$

is our estimate of $G(s)$, as it will converge to $G(s)$ with probability one as B tends to infinity ([50] and [34]).

The Bootstrap Sampling Process is then:

Given a random sample $y = (y_1, y_2, \ldots, y_n)$ from $F(y)$

Form the EDF $F_n(y|\mathbf{y})$

For $j = 1$ to B

  For $i = 1$ to n

    Draw y_i^{j*} from $F_n(y|\mathbf{y})$

  Next i

  Calculate $s^{j*} = s(\mathbf{y}^{j*})$

Next j

Form $G_B(s|\mathbf{s}^*)$

Bootstrapping may also be used for constructing hypothesis tests [34]. Efron and Tibshirani [50] describe the application of the bootstrap to hypothesis testing on a two-sample problem, where there are two random samples from two probability distributions and we wish to test the null hypothesis that the two distributions are identical. Bootstrapping can be used to estimate the distribution of the test statistic θ and hence the significance level of the test. The value of the test statistic is initially calculated for the two samples of observations. Bootstrap samples are then drawn from the two empirical distributions for the two observed random samples, and for each pair of bootstrap samples, the test statistic is calculated. We can draw as many bootstrap samples as we want and hence we can calculate as many bootstrap replications of the statistic of interest as we want. Thus, the distribution of the statistic can be determined in a direct and intuitive way. Having observed θ and the distribution of θ, the significance level of the test can be computed straightforwardly. This problem is similar to the problem we consider in this thesis and we use a similar bootstrapping method to estimate the significance level of the Cramér-von Mises statistics.

4.4 Bootstrapping for Estimating the Similarity Matrix

We wish to measure the similarity of the two samples of breakdown duration data $X = (x_1, x_2, \ldots, x_n)$, and $Y = (y_1, y_2, \ldots, y_m)$ for machines M_x and M_y respectively by estimating the significance level of the Two-Sample Cramér-von Mises goodness of fit statistic T. As we mentioned earlier, tabulated criterion values are not very extensive and do not cover the samples that we are dealing with. Thus, in order to assess whether the Cramér-von Mises goodness of fit statistic T is too large, we need to estimate its p-value by using bootstrapping to determine the

distribution of T, $\Phi(T)$. The p-value gives the probability that the breakdown duration data for the two machines are drawn from the same distribution and therefore is considered to indicate the similarity between these two machines. The p-values of any two of the machines are stored in the *similarity matrix*. This is then input into the Arrows classification method to group the machines, which is discussed in the next chapter.

For each pair of machines M_x and M_y, we combine the breakdown data, X and Y, in order to form the pooled sample of the breakdown duration data, $Z = (z_1, z_2, \ldots, z_{n+m})$. The EDF of Z is denoted by $H_{n+m}(z)$. In each iteration of the bootstrapping, we generate two samples out of the original pooled set of observations, Z, with replacement: one of size n, written as $X^* = (x_1^*, x_2^*, \ldots, x_n^*)$, and the other of size m, written as $Y^* = (y_1^*, y_2^*, \ldots, y_m^*)$; this is called one pair of *bootstrap samples*. We calculate the Cramér-von Mises statistic, T^*, for each pair of bootstrap samples, X^* and Y^*. In order to estimate $\Phi(T)$, we generate B pairs of bootstrap samples from $Z : (X^{*1}, Y^{*1}), (X^{*2}, Y^{*2}), \ldots, (X^{*B}, Y^{*B})$ and calculate the statistic T^{*j} for each pair of these samples. The EDF of the sample $T^* = (T^{*1}, T^{*2}, \ldots, T^{*B})$ is then written as

$$\Phi_B(T) = \frac{(\text{no. of } T^{*j} \leq T)}{B} \tag{4.8}$$

Since the bootstrap distribution $\Phi_B(T)$ will converge to the true distribution $\Phi(T)$ with probability one as B tends to infinity ([50] and [34]), we can use $\Phi_B(T)$ as our estimate of $\Phi(T)$.

The Bootstrapping Process can be briefly described as:

For $j = 1$ to B

  For $i = 1$ to n

    Draw x_i^{*j} from Z (with replacement)

  Next i

For $i = 1$ to m

Draw y_i^{*j} from Z (with replacement)

Next i

Calculate T^{*j} by comparing X^{*j} with Y^{*j}

Next j

Form the EDF of T^*, $\Phi_B(T)$.

The p-value describing the fit of data from machine M_x to data from machine M_y is then obtained by checking the calculated T with $\Phi_B(T)$. The whole process of estimating the p-value is illustrated in Figure 4.1. This procedure is carried out for all pairs of machines to form the similarity matrix.

As a measure of the similarity between machine M_x and machine M_y, the higher the p-value, the greater the possibility that the breakdown duration data of the two machines have been drawn from the same distribution and thus the more similar the two machines. For example, Figure 4.2 shows that the p-value corresponding to T is under 0.10, which means that the data from the two machines being compared are significantly different at a similarity threshold level of 0.10 and have not been drawn from the same distribution. In contrast, Figure 4.3 shows that the p-value of T is over 0.90, which means that the data from the two machines being compared can be assumed to have been drawn from the same distribution, with a probability of more than 0.90.

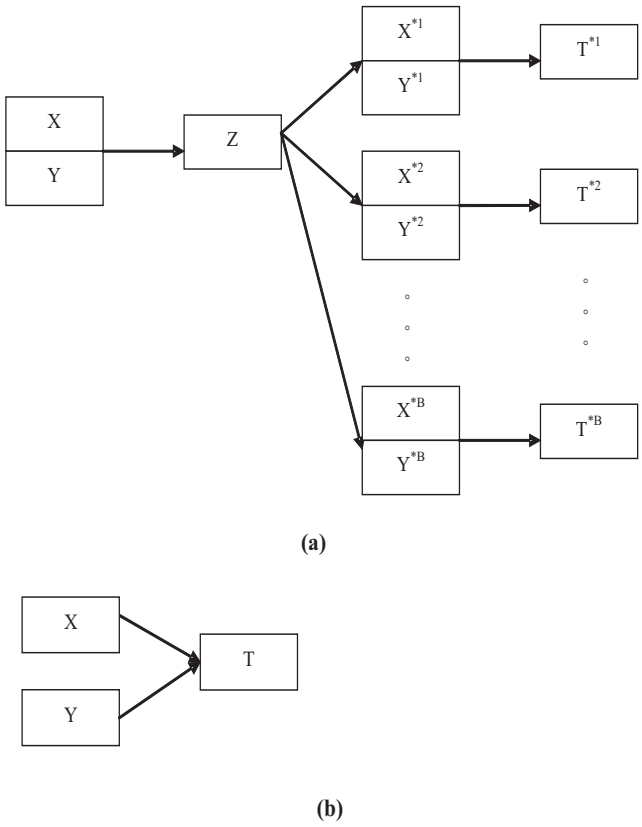

Figure 4.1: (a) The bootstrapping process used to determine the null distribution of T, $\Phi(T)$, and (b) the evaluation of the Cramér-von Mises statistic for the original samples, which is compared with $\Phi(T)$ to determine the p-value for the similarity of the two machines.

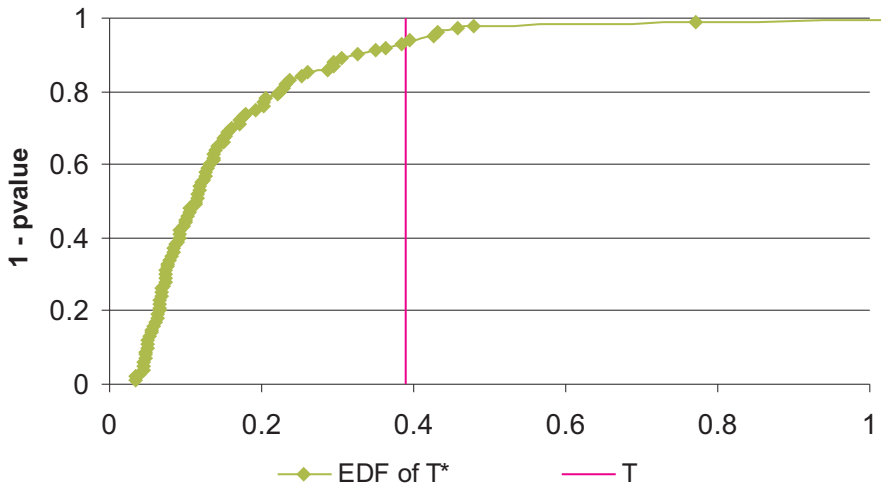

Figure 4.2: M_1 vs. M_2, $p_{12} < 0.10$

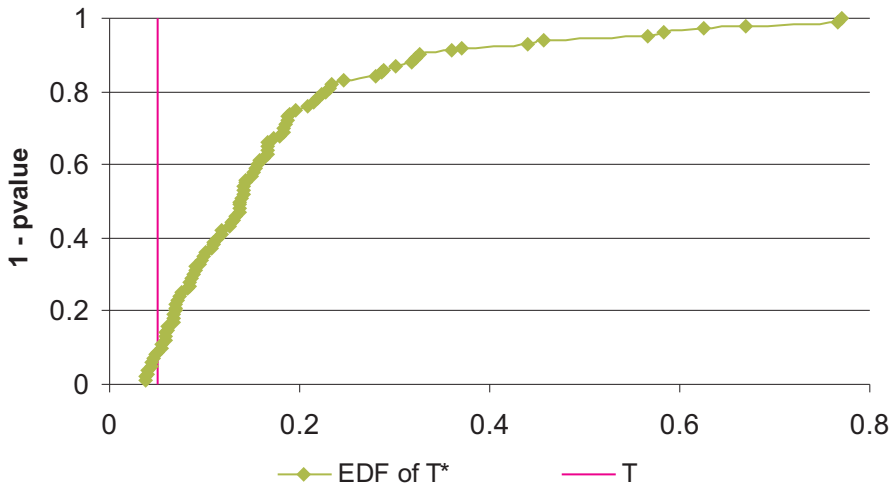

Figure 4.3: M_1 vs. M_3, $p_{13} > 0.90$

4.5 Testing the Estimation of Similarity

The testing procedure consists of five phases:

- Phase 1. Assess the impact of the number of bootstrapping iterations on the p-value results and find an appropriate number of bootstrap samples to run.
- Phase 2. Check the influence of the sample size, i.e. the number of data points in the sample, on the resultant p-value.
- Phase 3. Examine the performance of the method when dealing with samples that are drawn from the same type of distribution with the same variance but different mean.
- Phase 4. Investigate the method using samples that are generated from the same type of distribution with equal means but different variances.
- Phase 5. Test the method with samples generated from different types of distributions.

4.5.1 Phase 1: the impact of the number of bootstrap iterations

We can use the EDF of the bootstrap samples of T, $\Phi_B(T)$ as an estimateof the true distribution $\Phi(T)$ when B is big enough. We here investigate how large B should be for $\Phi_B(T)$ to be a good approximation to $\Phi(T)$. In general, there are three types of data sets in terms of their similarities: (a) very similar samples, (b) neither very different nor very similar and (c) distinctly different. We randomly generate four samples of size 100 from the 3 different distributions given in Table 4.1 below: two samples from distribution N1 and one each from N2 and N3. We choose 100 as the sample size as it is of a similar order to the machine duration data sets we analyse in the assembly lines. An investigation of the influence of

the sample size on the resultant p-value is given in the next phase of this testing process.

We examine three pairs of samples corresponding to the three types of data sets listed above: (a) very similar - N1S100a and N1S100b, two samples both from N1; (b) neither very similar nor very different - N1S100a and N2S100, one sample from N1 and the other from N2; and (c) distinctly different - N1S100a and N3S100, one sample from N1 and one from N3. With each pair of random samples, we use seven different and widely spread number of bootstrapping numbers: $B = 50$, 100, 200, 300, 500, 1000, 2000. For each of the three pairs of samples, we run the comparison seven times, once for each B; for each of these seven comparisons, we repeat the method 100 times, which gives 7 sets of p-values for the comparison of each pair of samples. The inter-quartile ranges of the total 21 sets of p-values are given in Table 4.2.

Code	Distribution	Notation	Mean	Variance	Sample ID
N1	Normal	$N(5.0, 1.0)$	5.0	1.0	N1S100a & N1S100b
N2	Normal	$N(5.1, 1.0)$	5.1	1.0	N2S100
N3	Normal	$N(7.0, 1.0)$	7.0	1.0	N3S100

Table 4.1: The 3 different distributions from which 4 random samples in total are generated.

We then study the influence of the choice of B by comparing the inter-quartile ranges of the p-value results using all the different B. As shown in Table 4.2, as B increases, the results for the comparison of the two pairs of samples become more stable, and the variability decreases, as the inter-quartile range tends to shrink as B gets larger.

As the two samples, N1S100a and N1S100b, are generated from the same distribution N1, in theory, they should be very similar to each other and thus the comparison should give very high p-value results. The inter-quartile ranges of the

	N1S100a vs.		
B	**N1S100b**	**N2S100**	**N3S100**
50	$(0.940, 0.975)$	$(0.000, 0.159)$	$(0, 0)$
100	$(0.935, 0.964)$	$(0.007, 0.113)$	$(0, 0)$
200	$(0.947, 0.976)$	$(0.029, 0.119)$	$(0, 0)$
300	$(0.953, 0.972)$	$(0.043, 0.114)$	$(0, 0)$
500	$(0.955, 0.969)$	$(0.063, 0.105)$	$(0, 0)$
1000	$(0.958, 0.971)$	$(0.062, 0.105)$	$(0, 0)$
2000	$(0.958, 0.969)$	$(0.070, 0.105)$	$(0, 0)$

Table 4.2: The inter-quartile ranges of each set of the 100 p-values resulting from 100 random runs with each different number of iterations of bootstrapping when comparing each of the 3 pairs of random samples.

p-values using the 7 choices of B for this pair of samples are within the range of $(0.935, 0.976)$, which shows that the two samples are very similar, as expected.

The two distributions $Normal(5.0, 1.0)$ and $Normal(5.1, 1.0)$ are not identical, but are very close, therefore, in theory, the two samples, N1S100a and N2S100, should be neither very similar nor very different and thus the p-values for the comparison should be neither very high nor very low. The inter-quartile ranges for this pair of samples given in Table 4.2 show that the majority of the p-values are within the range of $(0, 0.159)$, which is as expected. Samples of this type are neither very similar nor very different and thus tend to be on the edge of groups in classification analysis, i.e. they are fairly similar to a large number of other data sets but not very similar to any. Assuming we use 0.10 as the threshold significance level in classification analysis, such that two data sets with p-value smaller than 0.10 can not be placed in the same group, samples such as N1S100a and N2S100 might be put in two different groups with one run of bootstrapping process and then be placed in the same group with a subsequent run, as the p-value might be smaller than 0.10 with one run and then might become larger than 0.10 with a subsequent run. Although N1S100a and N2S100 are drawn from two different distributions, N1 and N2 are so close that it would not be unreasonable to

place them in the same group. Nevertheless, the difference would still be distinguished as the p-value would be relatively low. Moreover, if a higher similarity level within the final groups is required, a threshold significance level higher than 0.10 can be set for the classification analysis process to achieve that.

For the third comparison, since the two distributions $Normal(5.0, 1.0)$ and $Normal(7.0, 1.0)$ are very different, the two generated samples N1S100a and N3S100 should be very different and thus the p-values for the comparison should be very low. The inter-quartile ranges for this pair of samples given in Table 4.2 are all equal to zero, which shows that these two samples are extremely different, as expected.

To conclude, the method provides sensible p-values for all three types of samples even for small values of B. The p-values do, however, become more stable when more bootstrap samples are run. The bootstrapping process with $B = 2000$ will take much longer than that with $B = 50$, especially when a large number of data points are involved. Nevertheless, while running more bootstrap samples may improve the stability of the p-values, the resultant p-values when using a bootstrapping number as small as 50 are quite reasonable. Therefore, to reduce the computational cost, we use $B = 100$ to estimate the p-values in this work.

4.5.2 Phase 2: the influence of the sample size

As the method has been derived to estimate the similarity between data sets with uneven numbers of data points, we design this phase to test this ability. Since our comparison method is a distribution-free approach, the data sets in question may be drawn from any distribution. Hence, we use samples generated from four different types of distributions. A set of six samples, two of size 20, two of size 100 and two of size 200, is randomly generated from each of the four distributions listed in Table 4.3 below. In this phase, we only compare like distributions.

Code	Distribution	Notation	Mean	Variance
N1	Normal	$N(5.0, 1.0)$	5.0	1.0
Ga1	Gamma	$Ga(10.0, 0.5)$	5.0	2.5
E1	Exponential	$E(0.2)$	5.0	25.0
LN1	LogNormal	$LN(1.109, 1.0)$	5.0	43.0

Table 4.3: The 4 different distributions from which 24 random samples in total are generated.

	$N1S20a$	$N1S20b$	$N1S100a$	$N1S100b$	$N1S200a$	$N1S200b$
$N1S20a$	–	0.69	0.91	0.95	0.84	0.78
$N1S20b$	0.69	–	0.76	0.79	0.75	0.70
$N1S100a$	0.91	0.76	–	0.97	0.73	0.70
$N1S100b$	0.95	0.79	0.97	–	0.87	0.56
$N1S200a$	0.84	0.75	0.73	0.87	–	0.25
$N1S200b$	0.78	0.70	0.70	0.56	0.25	–

Table 4.4: Similarity Matrix for the six generated samples from distribution $N(5.0, 1.0)$.

For each set of the three pairs of samples drawn from distributions $N(5.0, 1.0)$, $Ga(10.0, 0.5)$, $E(0.2)$ or $LN(1.109, 1.0)$, we run 100 bootstraps for each pair of samples to determine the p-values. The p-values in each of the four resultant similarity matrices are fairly high and are all greater than 0.10, which is what we would expect as the samples for each matrix are indeed drawn from the same distribution. All four p-value matrices show a similar tendency, that the p-values between the samples with 200 data points are much smaller than the other p-values. For example, in the p-value matrix for the samples generated from N1 given in Table 4.4, the p-value between N1S200a and N1S200b, the two samples with the largest size, is 0.25, while the smallest of the rest of the p-values is 0.56. The reason for this range is likely to be that the more data points, the more possibilities that a statistical test will find dissimilarities between the data sets. Nevertheless, the four smallest p-values from the four matrices are all still higher than 0.10. Overall, the method manages to provide reasonable p-value results for data sets with different sizes.

4.5.3 Phase 3: distinguishing samples with different means

In this phase, we check the performance of the method in distinguishing samples that are drawn from the same type of distribution with the same variance but different means. We also test it on four different distribution types. The 9 distributions we generate samples from are listed in Table 4.5. As the exponential distribution has only one parameter, the two exponential distributions we test on have different means and variances.

Code	Distribution	Mean	Variance	Sample ID
N1	Normal	5.0	1.0	N1S100
N2	Normal	5.1	1.0	N2S100
N3	Normal	7.0	1.0	N3S100
Ga1	Gamma	5.0	2.5	Ga1S100
Ga2	Gamma	7.0	2.5	Ga2S100
E1	Exponential	5.0	25.0	E1S100
E2	Exponential	7.0	49.0	E2S100
LN1	LogNormal	5.0	43.0	LN2S100
LN2	LogNormal	7.0	43.0	LN2S100

Table 4.5: The 9 different distributions with the same variance but different means, from which 9 random samples are generated.

We generate one random sample of size 100 out of each of the 9 distributions. Then, we run 100 bootstraps for each pair of samples that are drawn from the same type of distribution to get the p-values shown in Table 4.6. All of the p-values except $p(N1S100, N2S100)$ are extremely small, which is sensible as the distributions that the pairs of samples come from are clearly distinct. The p-value for the comparison between N1S100 and N2S100 is 0.118, greater than our suggested threshold of 0.10. Therefore, we would assume that these two samples had been drawn from the same distribution. Although this is not the case, the distributions are so close that it would not be an unreasonable assumption. This comparison was included to test the method and the fact that the p-value is so close to the threshold

is encouraging.

Distribution	Samples	p-value
Normal	N1S100 vs. N2S100	0.118
	N1S100 vs. N3S100	0.000
	N2S100 vs. N3S100	0.000
Gamma	Ga1S100 vs. Ga2S100	0.000
Exponential	E1S100 vs. E2S100	0.041
LogNormal	LN1S100 vs. LN2S100	0.000

Table 4.6: The 6 p-values comparing the 6 pairs of random samples.

4.5.4 Phase 4: distinguishing samples with different variances

In this phase, we check the performance of the method in distinguishing samples that are drawn from the same type of distribution with the same mean but different variances. As we have considered the exponential distribution in Section 4.5.3, we do not include it in this test. The 6 distributions we generate samples from are listed in Table 4.7.

Code	Distribution	Mean	Variance	Sample ID
N1	Normal	5.0	1.0	N1S100
N4	Normal	5.0	4.0	N4S100
Ga1	Gamma	5.0	2.5	Ga1S100
Ga3	Gamma	5.0	5.0	Ga3S100
LN1	LogNormal	5.0	43.0	LN2S100
LN3	LogNormal	5.0	415.9	LN2S100

Table 4.7: The 6 different distributions with the same mean but different variances, from which 6 random samples are generated.

We generate one random sample of size 100 from each of the 6 distributions. We run 100 bootstraps for each pair of samples that come from the same type of distribution to get the p-values given in Table 4.8. As shown in this table,

$p(N1S100, N4S100)$ and $p(LN1S100, LN3S100)$ are both very small, which is sensible as the distributions that the two pairs of samples come from are different. However, $p(Ga1S100, Ga3S100)$ is 0.210, which indicates the two samples appear to be quite similar. The PDF curves for these two distributions and the histograms for these two samples are given in Figure 4.4, and show that distributions Ga1 and Ga3 are not that different. Therefore, it is unsurprising that the p-value for samples Ga1S100 and Ga3S100 is greater than the threshold of 0.10.

Distribution	**Samples**	**p-value**
Normal	N1S100 vs. N4S100	0.007
Gamma	Ga1S100 vs. Ga3S100	0.210
LogNormal	LN1S100 vs. LN3S100	0.000

Table 4.8: The 3 p-values comparing the 3 pairs of random samples.

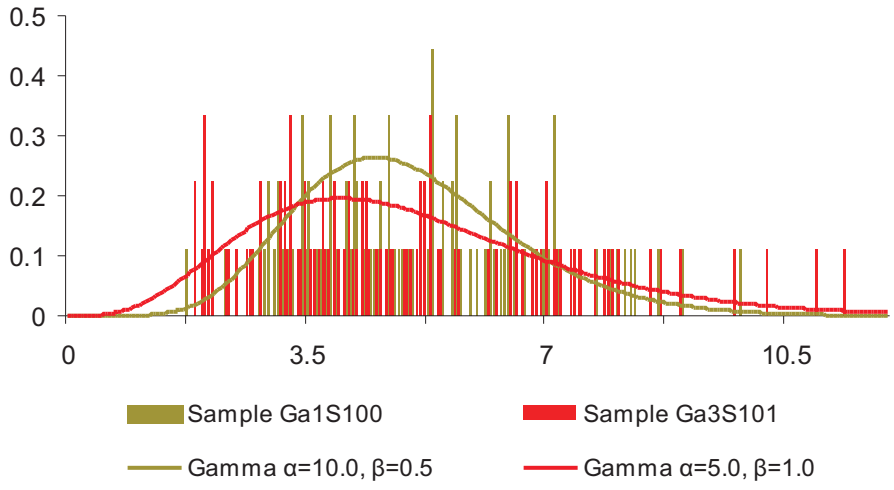

Figure 4.4: Plots of the PDF of $Gamma(10, 2.5)$ and $Gamma(5.0, 1.0)$ and the histograms of the two random samples, Ga1S100 and Ga3S100, generated from each of the two distributions respectively.

4.5.5 Phase 5: distinguishing samples generated from different types of distributions

We wish to assess the method's ability to find the similarities between samples with similar mean but different distribution shapes. We randomly generate one sample of size 100 from each of the 4 different distributions listed in Table 4.3 above, which gives a collection of 4 random samples in total. The plots of probability density functions for these distributions are given in Figure 4.5. We run 100 bootstraps for each pair of samples to determine the p-values and hence the similarity matrix given in Table 4.9.

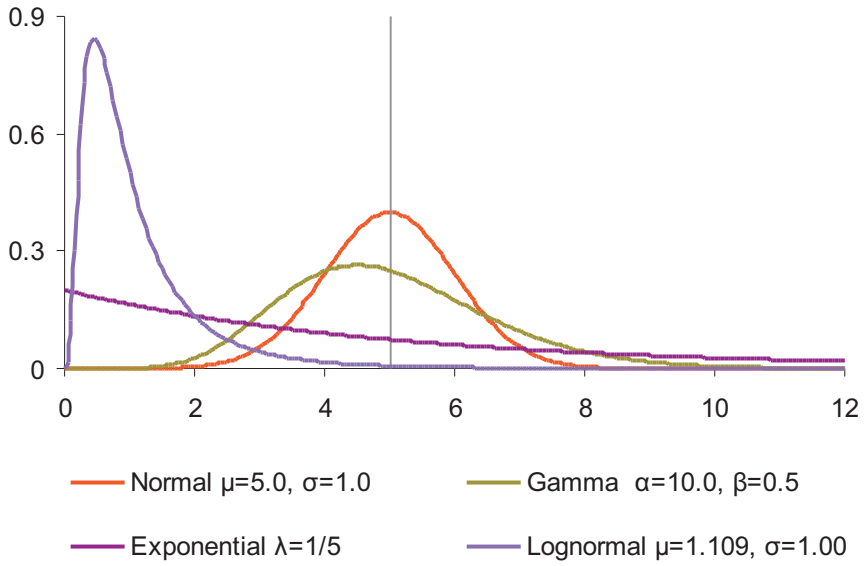

Figure 4.5: Plots of the PDF curves of the 4 different distributions listed in Table 4.3.

The p-values given in Table 4.9 are all extremely low, as expected, except that between the exponential and the lognormal. Both have their modes close to or at zero and then decline, and so although the lognormal has less weight in the tails of the distribution, the general shapes are similar. Furthermore, it is seen from the histograms shown in Figure 4.6 that the two particular random samples are

	$N1S100$	$Ga1S100$	$E1S100$	$LN1S100$
$N1S100$	–	0.00	0.00	0.00
$Ga1S100$	0.00	–	0.00	0.00
$E1S100$	0.00	0.00	–	0.12
$LN1S100$	0.00	0.00	0.12	–

Table 4.9: Similarity Matrix for the four random samples generated from distributions $Normal$(5.0, 1.0), $Gamma$(10.0, 0.5), $Exponential$(0.2) and $LogNormal$(1.109, 1.0) respectively.

fairly close and so it is not unreasonable to have a p-value slightly higher than the threshold 0.10.

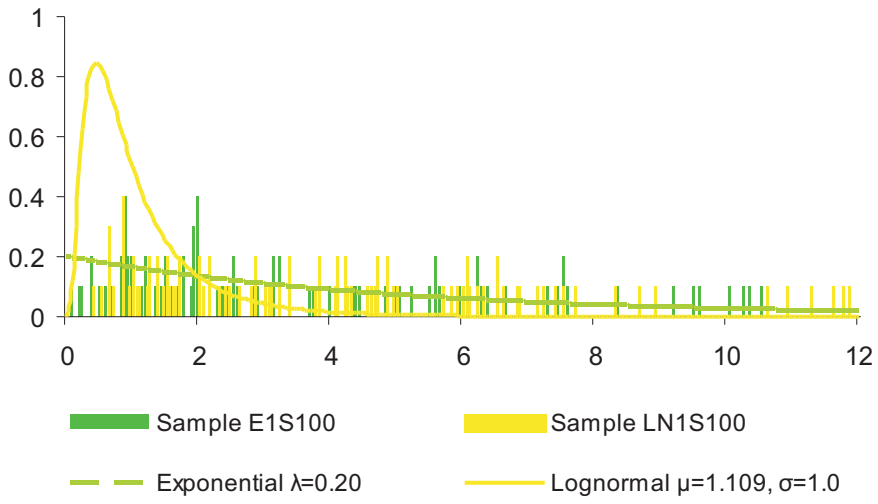

Figure 4.6: Plots of the PDF of $Exponential(0.20)$ and $Lognormal(1.109, 1.0)$ and the histograms of the two random samples, E1S100 and LN1S100, generated from each of the two distributions respectively.

4.6 Examples

Although this method of measuring similarity was originally derived to analyse machine breakdown duration data, it is widely applicable. In this section, we consider two examples:

- breakdown duration data collected over a period of 3 months for six machines.
- hospital length-of-stay data for patients recovering from medical procedures.

4.6.1 Breakdown Duration Data

We here consider real breakdown duration data sets for six machines: ML01, ML02, ML03, ML04, ML05, ML06, in an engine assembly line at one of Ford's plants. The size of these six data sets are 170, 319, 112, 113, 60 and 460 data points, respectively. They come from machines with different functionalities in different stations. The histograms of the breakdown duration data for the six machines are given in Figure 4.7. We wish to group the machines based on their breakdown duration data and so we need to produce the similarity matrix for the six machines.

We run 100 bootstraps for each pair of machines to determine the p-values matrix given in Table 4.10. The p-value between machine ML05 and ML06 is the highest value in the matrix, the p-values between any one of these two machines and any machine of the other four machines are extremely small, which tells that these two machines are very similar to each other and not similar to any other machines in terms of their breakdown behaviour. Within the other four machines: machine ML01 has a 0.21 similarity to ML04 and a 0.11 similarity to ML02 but a similarity of less than 0.10 to ML03; Machine ML02 seems to be significantly similar to ML01 and ML04, especially similar to ML04 as they have a much higher p-value, but has a nearly zero similarity to ML03; machine ML04 has high p-values for the comparison with ML02, ML01 and ML03; and ML03 only has a significant p-value (greater than 0.10) for its comparison with ML04. Referring to the histograms in Figure 4.7, both of the histograms for machines ML05 and ML06 have high peaks around 1.60 while the histograms for the other four machines

	$ML01$	$ML02$	$ML03$	$ML04$	$ML05$	$ML06$
$ML01$	–	0.11	0.08	0.21	0.01	0.00
$ML02$	0.11	–	0.03	0.51	0.00	0.00
$ML03$	0.08	0.03	–	0.20	0.00	0.00
$ML04$	0.21	0.51	0.20	–	0.00	0.00
$ML05$	0.01	0.00	0.00	0.00	–	0.89
$ML06$	0.00	0.00	0.00	0.00	0.89	–

Table 4.10: Similarity Matrix for six machines in a Ford engine assembly line, based on their breakdown duration data.

have high peaks within the range of $(0.60, 1.30)$; and both histograms have more symmetric shapes than the other four. These features of the histograms confirm the reliability of the p-value results.

4.6.2 Length-of-Stay Data

The method of estimating similarity presented here has been implemented to calculate the similarity between medical procedures based on the hospital length-of-stay data of the corresponding patients, where the data comes from a group of private hospitals [41]. More information about the data and the grouping process will be described in Section 5.5. We here use a small example of the length-of-stay data of five procedures to illustrate our method.

The five procedures are coded as Q13.1, Q20.2, Q38.3, W37.1 and W42.1 (the codes are called OPCS-4 codes and are used by NHS). The first three procedures are procedures on the uterus and fallopian tubes: Q13.1 is implantation of fertilised egg into uterus, Q20.2 is a biopsy of lesion of uterus, and Q38.3 is therapeutic endoscopic operations on fallopian tube; and the last two procedures are joint replacements: W37.1 is hip joint replacement and W42.1 is knee replacement.

We run 100 bootstraps for each pair of procedures to determine the p-values. The similarity matrix is given in Table 4.11. The p-values between any pair of

	$Q13.1$	$Q20.2$	$Q38.3$	$W37.1$	$W42.1$
$Q13.1$		0.91	0.46	0.00	0.00
$Q20.2$	0.91		0.60	0.00	0.00
$Q38.3$	0.46	0.60		0.00	0.00
$W37.1$	0.00	0.00	0.00		0.31
$W42.1$	0.00	0.00	0.00	0.31	

Table 4.11: Similarity Matrix for five procedures based on their patients' length-of-stay data.

the procedures Q13.1, Q20.2 and Q38.3 are all larger than 0.40, which suggests the length-of-stay data of these three procedures have been drawn from the same distribution with a high probability. This seems to reflect the real situation because these three procedures are similar operations on similar organs. In particular, the similarity between Q13.1 and Q20.2, 0.91, is much higher than the similarity between Q13.1 and Q38.3 and that between Q20.2 and Q38.8, which makes sense intuitively as both Q13.1 and Q20.2 are operations on the uterus and Q38.3 is a procedure on the fallopian tubes. The matrix also shows that there are significant differences between the set of procedures Q13.1, Q20.2 and Q38.3 and the set of procedures W37.1 and W42.1 as the p-value between any procedure from the former set and any from the latter set is zero, which is sensible because the former set of procedures are very distinct from the joint replacements. That the p-value between procedures W37.1 and W42.1 is larger than 0.30 is also reasonable as there are definite similarities between the recovery time from a hip joint replacement and a knee replacement. Overall, the resultant similarity matrix of the five procedures appears to be reflecting the real situation quite well.

Since the p-value demonstrates the probability of the data sets having been drawn from the same distribution, it shows the similarity of the corresponding distributions of the data sets. The histograms of the length-of-stay data for the five procedures given in Figure 4.8 add more confidences to the similarity results. For instance, the way the histograms for procedures Q13.1, Q20.2 and Q38.3 distinct

from those for W37.1 and W42.1 clearly supports the extremely small similarity between the two sets of procedures.

These two examples show that the method we have described in this chapter is an appropriate distribution-free method for estimating the similarity between data sets that may be of different sizes. Although this method has been derived to estimate the similarity index between breakdown duration data sets, it is also applicable to other data sets, such as the hospital length-of-stay data in Example 2.

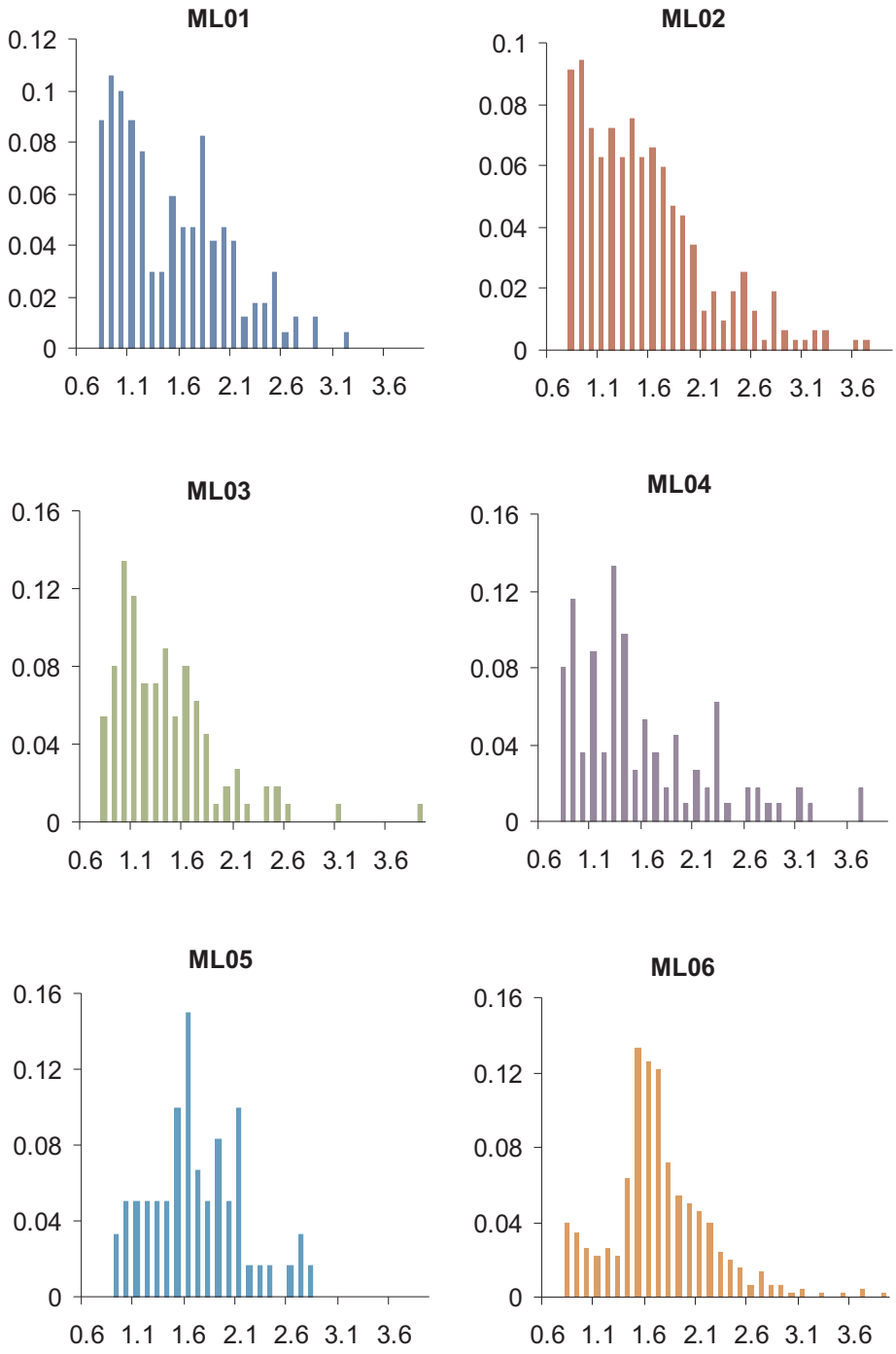

Figure 4.7: Histograms of the breakdown duration data for the six machines.

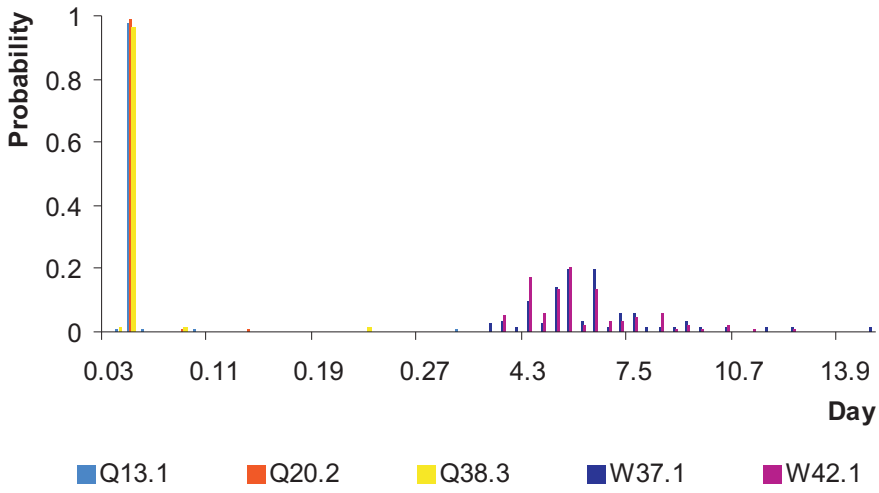

Figure 4.8: Histograms of the patients' hospital length-of-stay data for the five procedures.

Chapter 5

Classification of Machines

Having found the similarity matrix of the machines, we discuss the classification method we propose to use for grouping the machines in this chapter. The grouping is such that two machines with statistically significantly different breakdown duration data cannot be placed in the same group. We can later fit finite mixture distributions to the grouped breakdown duration data. The aim is to use the fitted finite mixture models for groups to represent the breakdown duration inputs for all of the machines in the same group.

A review of the literature on classification methods is given in Section 5.1. Then, in Section 5.2 we present a description of the Arrows classification method. Section 5.3 gives an example of 20 machines involved in an engine assembly line to demonstrate the Arrows classification process.

The Arrows method has similarities with cluster analysis and a comparison with the cluster analysis method is given in Section 5.4 using some standard data as examples and using the same example considered previously of the 20 machines. A study of the features of the Arrows method is also included in this section. The classification process described in this chapter could be applied to classify data from a wide range of applications, in addition to manufacturing. We present an

example of the classification of hospital procedures by their patients' length-of-stay data in Section 5.5. A short conclusion drawing together the main ideas of this chapter is given in Section 5.6.

5.1 Classification

Classification is normally understood as the activity of allocating objects into a smaller number of classes so that objects in one class are similar to one another. It is also called *identification* or *assignment* [38].

There are a large number of classification methods, Section 5.1.1 gives suggested categorisations for these methods. We introduce two main targets of classification methods in Section 5.1.2. We then go on to describe the sorting strategies and algorithms of the procedure generally used for finding clusters in Section 5.1.3. A brief comparison of different sorting strategies is given in Section 5.1.4.

5.1.1 Types of Classification

Two general types of classification methods can be specified based on distinction of the classification process (see for example, Grabmeier and Rudolph [74] and Fielding [55]):

(i) Hierarchical Classification:

Generally known as being able to transform a raw data matrix, similarity matrix or dissimilarity matrix into a dendrogram.

(ii) Partitioning:

The result is a partition of the set of objects.

In addition to these two types, Cormack [38] indicates that there is another major type: clumping, where the resultant classes can overlap. There are other types such as model-based, density-based, factor analysis variants and graph theoretic methods (see for example, Aldenderfer and Blashfield [2], Anderberg [3], Everitt [53] and Fielding [55]).

Another categorisation of classification made based on the distinction of the process is given by Kendall [93]:

(a) *Classification*:

Objects in one class are needed to be isolated from objects in another class.

(b) *Dissection*:

Objects in one class are not necessarily isolated from objects in another class.

Gengerelli [62] gives an example to demonstrate the differences: "If there are two dense clusters of buildings separated by much empty space, we have no difficulty in perceiving the existence of two villages; whereas if a village by one name coalesces with a village by another name, we feel that the separation is artificial and that there exist not two entities, but one". It seems to be understandable that all sets of objects can be dissected but not all can be classified.

It is emphasized by Cormack [38] that different methods of classification can be achieved by one algorithm; for example, a sorting strategy with a particular algorithm gives a hierarchical classification but produces a partition or clump when a stopping rule is applied. The Arrows classification method, which we introduce in Section 5.2, can be described as a combination of an hierarchical classification method and a partitioning method. A dendrogram is formed by clusters merging at different similarity levels but a threshold, whose value is chosen by the user, is

used as a stopping rule: any amalgamation of two clusters with a similarity smaller than the chosen similarity threshold is not allowed.

5.1.2 Method Targets

Methods strive to maximize either internal cohesion or external isolation or some combination of the two, where internal cohesion can be defined such that an object should be added into an existing cluster if its smallest similarity with any member in the cluster is larger than some chosen threshold [29] and external isolation focuses on the isolation between clusters such that there should be a clear distinction between clusters, and similar objects shall not be divided into different clusters [137]. Székely and Rizzo [152] state that many standard clustering procedures aim only at within cluster distance minimization, i.e. internal cohesion maximization, or at between cluster distance maximization, i.e. external isolation maximization. Cormack [38] states that often both are included in one classification method. For example, Gengerelli [62] discusses a method satisfying the requirement that the distance between any two objects in one group is less than the distance between any object in the group and any not in it. Needham [121] describes a method in which the sum of the similarities of any object to the other objects in one group should exceed the sum of its similarities to objects in other groups and vice versa for objects in other groups. Our method also is a combination of the two ideas.

5.1.3 Obtaining Classes

Cormack [38] indicates that there are three types of procedure generally used for finding clusters:

- Agglomerative: merging n objects into classes.

- Divisive: dividing one initial class containing n objects into a larger number of classes.
- Clustering: reallocating objects between sets of some initial classes.

The first two types: agglomerative and divisive, are suggested by a number of authors to be the two major types of hierarchical classification, which is generally known as the procedure for transforming the raw data matrix, similarity or dissimilarity matrix into a dendrogram. There are other types of algorithms of hierarchical classification methods; for example, Gordon [70] identifies two additional types: constructive and direct optimization algorithms. The former progress by "successively adding new objects to a hierarchical classification of a smaller data set" and algorithms have been introduced by Sibson [146] and Defays [46] to update single linkage and complete linkage dendrograms. The latter has been advocated by Hartigan [79], Carroll [26] and De Soete [45].

Grabmeier and Rudolph [74] give a diagram of a taxonomy of classification methods and clustering algorithms and a simpler version of this diagram is shown in Figure 5.1, reproduced from Fielding [55]. Hierarchical classification methods are considered to be the most popular classification methods. The agglomerative and divisive algorithms will be further described below.

Agglomerative

There are n single-object classes initially, and the most similar pair of classes is merged at each stage. Different sorting strategies are distinguished by their way of determining the similarity between two classes of objects. There is a general agglomerative algorithm proposed by Lance and Williams [100, 101], in which the measures of dissimilarity between class C_k and a new class $C_{(ij)}$ that is formed by

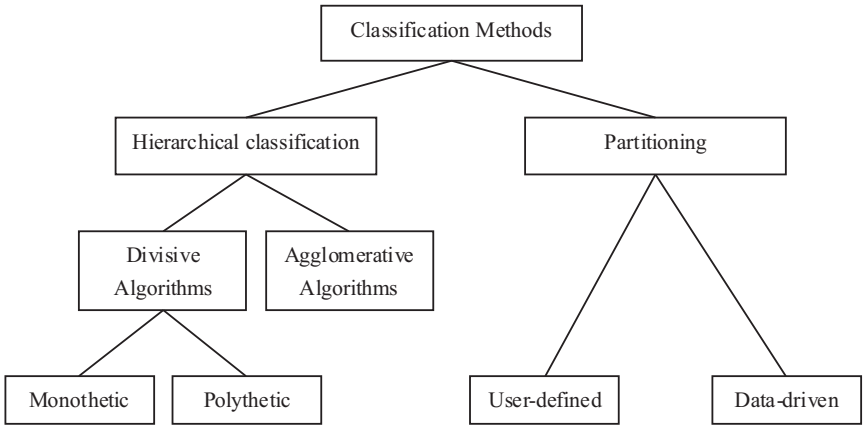

Figure 5.1: A Taxonomy of classification methods and sorting algorithms. Reproduced from [55].

combining class C_i and class C_j can be defined as:

$$d_{k(ij)} = \alpha_i d_{ki} + \alpha_j d_{kj} + \beta d_{ij} + \gamma \left| d_{ki} - d_{kj} \right| \tag{5.1}$$

A similar but more general form for Equation 5.1 was proposed by Jambu [87], with three new parameters introduced,

$$d_{k(ij)} = \alpha_i d_{ki} + \alpha_j d_{kj} + \beta d_{ij} + \gamma \left| d_{ki} - d_{kj} \right| + \delta_i h_i + \delta_j h_j + \epsilon h_k \tag{5.2}$$

where h_i is the height of class C_i in the dendrogram representing the clustering process.

Gordon [70] states “an advantage of the general formulation is that the initial matrix of pairwise dissimilarities need not be retained, but can be overwritten as the amalgamation proceeds”.

The values of the parameters for a number of well-known clustering strategies are given in Table 5.1, reproduced from Gordon [70]. In this table w_i is the weight of class C_i, and is set equal to the number of objects in C_i, i.e. $w_i = n_i$.

The single linkage clustering strategy is also referred to as the *nearest neigh-*

Method & References	α_i	β	γ	γ_i	ϵ
Single linkage (Sneath [147]; Sokal and Sneath [149])	$\frac{1}{2}$	0	$-\frac{1}{2}$	0	0
Complete linkage (McQuitty [116]; Sokal and Sneath [149])	$\frac{1}{2}$	0	$\frac{1}{2}$	0	0
Group average linkage (Sokal and Michener [148]; McQuitty [118])	$\frac{w_i}{w_i+w_j}$	0	0	0	0
Weighted average linkage (McQuitty [118], [118])	$\frac{1}{2}$	0	0	0	0

Table 5.1: Values of the parameters for clustering strategies. Reproduced from Gordon [70].

bour method. For *single linkage* clusters, the distance between two clusters is defined as the distance between the two most similar objects in the two clusters [53]. It is said to be "the simplest agglomerative sorting procedure" [38]. An advantage of this strategy is that consecutive merging always occurs at lower levels of inter-cluster similarity.

One drawback of single linkage is that clusters may be forced to be merged due to only one object from one cluster being similar to another object from the other cluster, even if many other objects in each cluster are very distant from each other: a situation described as the *chaining phenomenon*. Another pitfall identified by Hodson in [82] is that when there are "transitional" objects between distinct clusters, single linkage cannot provide reasonable results. Such transitional objects were referred to as "intermediates" and suggested to be treated as noise in Cormack [38]. Wishart [170] and Baron and Fraser [9] propose methods to eliminate noise from objectives and from variables respectively. Shepherd and Willmott [145] suggest an extra constraint that an object is allowed to join a cluster only if its

similarities to a certain number of members of that cluster are all larger than some chosen threshold and this can be used to break the chaining phenomenon.

Complete linkage is also known as the *furthest neighbour* sorting method. It is the opposite of single linkage; the distance between two clusters here is specified as the distance between the two farthest objects in the two clusters [53]. Thus, one feature of the *complete linkage* method is that it gives compact clusters [3]. However, the outliers greatly affect the merging process. This strategy is not appropriate if random noise is present in the data, but is useful if the expected clusters are very distinct in the multi-dimensional space. Similar to single linkage, merging occurs "monotonically with inter-cluster similarity" [38].

Group average linkage and *Weighted average linkage* methods define the similarity between two clusters as the unweighted and weighted averaged similarity between the objects from one cluster and those from the other [53] and both therefore need numerical calculations. Accordingly, their clustering effect is in-between the single linkage and complete linkage. Both methods produce monotonic cluster trees. The two methods are almost identical, the only difference is that with the weighted average linkage, the numbers of objects contained in the two clusters are used as weight [53]. Sokal and Sneath [149] formulate the similarity between clusters C_i and C_j as:

$$S_{ij} = \frac{\sum_{a \in C_i} \sum_{b \in C_j} (s_{ab} w_a w_b)}{\sum \sum w_i w_j} \tag{5.3}$$

where s_{ab} is the similarity index between objects a from C_i and b from C_j. So for the group average linkage strategy w_a is equal to 1 and for the weighted average linkage strategy $w_a = n_i$. Therefore, the weighted average linkage method is suggested to be applied if the cluster weights are expected to be significantly uneven.

For the agglomerative hierarchical part, our Arrows method has similarities to

the complete linkage and average linkage methods, with a major additional constraint that the similarities of every pair of objects in every cluster should all be greater than or equal to some specified threshold.

Divisive

Divisive algorithms start with one big class including all n objects. At each stage of the algorithm, the current class is divided into two smaller classes. The divisive hierarchical method can be thought of as the opposite of the agglomerative hierarchical method. It is stated by Fielding in [55] that it is not widely used as it appears to have computational difficulties. This method can be divided into two types: monothetic and polythetic [55]. The former divides the class on the basis of the possession of only a single variable and often leads it to "misclassify" [167], while the latter uses the values taken by more than one variable ([125] and [80]). Chipman and Tibshirani [36] have proposed a hybrid method that combines the solutions of agglomerative hierarchical clustering and divisive hierarchical clustering.

5.1.4 Strategy Comparison

Jardine and Sibson [88] and many other authors have identified that methods and algorithms can have distinct meanings. For example, Rohlf [139] has proved that the single linkage method can be achieved by a number of different algorithms.

Gower [73] believes that if there is a huge distinction between objects and clear distinct clusters any useful clustering strategy would classify the objects correctly. However, different clustering methods can and do generate different classification solutions to the same data set when the distinction is less clear cut ([53] and [70]). The single linkage method has been proposed by Jardine and Sibson in [88] to be the method that satisfies a number of desirable properties, but, there is no single

method believed to be uniquely suitable for any data set. Therefore, it is important to choose the appropriate clustering methods for different data sets and a number of approaches have been proposed to do so.

There are simulation studies investigating the behaviour of clustering methods. Milligan [120] gives a review of this type of study. Although detailed information about the clustering procedures can be accumulated in these simulation studies, they provide little guidance on the most appropriate method for a particular data set without knowing its characteristics. A second approach is to obtain a number of requirements that it is desirable to see in the analysis of a data set and examine various sorting strategies to ascertain whether the requirements can be satisfied. Fisher and Van Ness [56] and Van Ness [154] have proposed this approach and provided a list of properties. An example is given in Gordon [70]: if a clustering method is required to be *monotone admissible* (that is, if a monotone transformation is made on the entire similarity or dissimilarity matrix, the clustering solution stays the same.) Single linkage and complete linkage methods are the only two strategies in Table 5.1 that satisfy this requirement. Another approach is to use more than one clustering method to classify the data set and synthesize the obtained results so that the combined solution may "represent genuine structure in the data" [70]. Rohlf [138] has proposed an adaptive agglomerative sorting algorithm to adapt the index of dissimilarity corresponding to the data structure. Diday and Moreau [49] have used the information obtained from a training set whose clusters are given by the analyser to choose suitable values of the parameters in formulation 5.2 for analysing a new data set of a larger size. It is suggested by Gordon [70] that the adaptive agglomerative clustering algorithm and the training set strategies can be applied to specify the structure in the data set and thus to help in selecting suitable clustering methods for it.

5.2 Arrows Classification Method

We aim to classify the machines involved in the assembly line into a smaller number of groups, based on their breakdown duration data, such that no pair of machines in a group has sampled breakdown duration data with significantly different distributions. The similarity matrix that is used to classify the machines into groups is made up of the p-values describing the probabilities of pairs of samples having been drawn from the same distribution as described in Section 4.4.

First, we define two terms that we associate with the name of the method. Machines M_i and M_j have a *double-arrow connection* if p_{ij}, the p-value comparing their corresponding sets of data, is the highest in both row i and row j of the similarity matrix and p_{ij} is greater than the specified threshold p_0. Machines M_i and M_k have a *single-arrow connection* if p_{ik} is the highest in only one of the rows i or k and p_{ik} is greater than the specified threshold p_0.

We follow the steps below to determine the groups.

1. Choose the threshold p-value, p_0, for assuming that two sets of data are similar enough to be grouped together. If the p-value for the fit between the breakdown duration data of a pair of machines is greater than or equal to p_0 then they can be put in the same group; otherwise, the data are assumed to be significantly different. We currently use 0.10 as a threshold p-value. Increasing the p-value threshold to, e.g. 0.20, may increase the average similarities within groups but may also increase the number of groups.

2. Search the similarity matrix,

   (a) If M_i and M_j are not grouped and they have the greatest double-arrow connection in the pool of ungrouped machines, put machine i and machine j into one group, say group C_a.

(b) Place all ungrouped machines that have double-arrow connections or single-arrow connections to machine i or machine j in group C_a.

(c) Place in C_a all ungrouped machines that have double-arrow connections or single-arrow connections to any of the machines added to group C_a in step 2(b). Continue until there are no more possible machines to be grouped together.

(d) Start to find a new group from step 2(a). Search the whole similarity matrix, until no more new machine groupings can be made. It is possible that some groups are made up of only one machine.

3. Check the p-values of all pairs of machines in each group, e.g. for group C_a: if the values are all greater than or equal to p_0, the threshold we choose, keep C_a; otherwise, for pairs with p-values less than p_0, use the following decision process to determine which machine in the pair to keep and which machine should be deleted from group.

(a) If M_i and M_j have a double-arrow connection keep both of the machines in C_a as machines with double-arrow connections form the core of the groups. This also reduces the number of machines that we need to search over in the following step of the algorithm.

(b) Take out the machine with the weakest connection with the others in the group and repeat this until there are no pairs of machines with p-values under p_0 in C_a, where the strength of a connection of an machine M_i to its group C_a is measured by its *inside connection* defined as

$$\overline{p}_{(i,a)} = \frac{\sum p_{ij}}{N_a - 1}, M_j \in C_a \text{ and } j \neq i, \tag{5.4}$$

where N_a is the number of machines in group C_a. This is effectively the average of the p-values between machine M_i and the other ma-

chines in C_a.

4. For each possible pair of groups, check the p-values between the machines in the first group and those in the second group. If all pairs of machines in group C_a and group C_b have p-values greater than or equal to p_0, these two groups can be combined into a new group. If we can combine group C_a with group C_b or with group C_d, combine the groups which have the greater *average connection* between them, where the average connection between groups C_a and C_b is defined to be

$$\overline{p}_{(a,b)} = \frac{\sum_{M_i \in C_a} \sum_{M_k \in C_b} p_{ik}}{N_a N_b}, \tag{5.5}$$

where p_{ik} is the similarity between machine i from group C_a and machine k from group C_b, and N_a and N_b are the numbers of machines in group C_a and group C_b respectively. This is effectively the average of the p-values for the comparisons between the machines in group C_a and those in group C_b. Search until all of the groups have been processed and combined where possible, including groups formed during step 4.

The above classification procedure has been implemented in Visual Basic for Applications. Although this method has been devised to classify machines, it is widely applicable. We next consider its application to a number of example data sets.

5.3 An Example of Machine Classification

We illustrate the classification method using an example of twenty machines involved in one of Ford's engine assembly lines. We currently use 0.10 as the threshold p-value for assuming two sets of breakdown duration data are similar enough

to be grouped together. Increasing the threshold may improve the homogeneity of the groups but also would increase the number of groups. It is therefore necessary to set the threshold for p-values to achieve a balance between the two conflicting aims of homogeneity and a small number of groups. A study of the influence of the threshold on grouping results using the Arrows method will be given in Section 5.4. Using the groups found by the Arrows method we then fit a different mixture distribution for each group, and in the simulation use this as the breakdown duration distribution for all of the machines in the group. The influence of the choice of threshold for machines grouping on the resultant output of simulation models using fitted mixture distributions for different groups will be investigated in Section 7.3.

	$M01$	$M02$	$M03$	$M04$	$M05$	$M06$	$M07$	$M08$	$M09$	$M10$	$M11$	$M12$	$M13$	$M14$	$M15$	$M16$	$M17$	$M18$	$M19$	$M20$
$M01$	–	0.02	0.00	0.00	0.11	0.00	0.03	0.00	0.09	0.14	0.50	0.50	0.06	0.00	0.04	0.00	0.06	0.00	0.13	0.13
$M02$	0.02	–	0.62	0.00	0.03	0.00	0.00	0.00	0.08	0.22	0.02	0.03	0.00	0.00	0.17	0.00	0.00	0.01	0.01	0.07
$M03$	0.00	0.62	–	0.00	0.03	0.00	0.00	0.00	0.01	0.05	0.00	0.00	0.00	0.00	0.12	0.00	0.00	0.00	0.02	0.08
$M04$	0.00	0.00	0.00	–	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$M05$	0.11	0.03	0.03	0.00	–	0.00	0.00	0.03	0.00	0.19	0.15	0.16	0.57	0.00	0.12	0.00	0.14	0.00	0.85	0.93
$M06$	0.00	0.00	0.00	0.00	0.00	–	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
$M07$	0.03	0.00	0.00	0.00	0.00	0.00	–	0.17	0.00	0.36	0.00	0.00	0.31	0.01	0.00	0.00	0.82	0.00	0.18	0.01
$M08$	0.00	0.00	0.00	0.00	0.03	0.00	0.17	–	0.00	0.29	0.00	0.00	0.20	0.00	0.00	0.00	0.63	0.00	0.32	0.01
$M09$	0.09	0.08	0.01	0.00	0.00	0.00	0.00	0.00	–	0.02	0.05	0.08	0.00	0.00	0.01	0.00	0.00	0.00	0.02	0.01
$M10$	0.14	0.22	0.05	0.01	0.19	0.00	0.36	0.29	0.02	–	0.27	0.26	0.82	0.03	0.14	0.02	0.23	0.07	0.53	0.36
$M11$	0.50	0.02	0.00	0.00	0.15	0.00	0.00	0.00	0.05	0.27	–	0.38	0.29	0.00	0.06	0.00	0.01	0.00	0.38	0.25
$M12$	0.50	0.03	0.00	0.00	0.16	0.00	0.00	0.00	0.08	0.26	0.38	–	0.26	0.00	0.09	0.00	0.01	0.00	0.30	0.30
$M13$	0.06	0.00	0.00	0.00	0.57	0.00	0.31	0.20	0.00	0.82	0.29	0.26	–	0.02	0.05	0.00	0.62	0.02	0.57	0.48
$M14$	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.03	0.00	0.00	0.02	–	0.00	0.00	0.37	0.00	0.06	0.00
$M15$	0.04	0.17	0.12	0.00	0.12	0.00	0.00	0.00	0.01	0.14	0.06	0.09	0.05	0.00	–	0.00	0.00	0.10	0.38	0.45
$M16$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	–	0.07	0.00	0.01	0.00
$M17$	0.06	0.00	0.00	0.00	0.14	0.00	0.82	0.63	0.00	0.23	0.01	0.01	0.62	0.37	0.00	0.07	–	0.00	0.30	0.11
$M18$	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.00	0.02	0.00	0.10	0.00	0.00	–	0.51	0.15
$M19$	0.13	0.01	0.02	0.00	0.85	0.01	0.18	0.32	0.02	0.53	0.38	0.30	0.57	0.06	0.38	0.01	0.30	0.51	–	0.71
$M20$	0.13	0.07	0.08	0.00	0.93	0.00	0.01	0.01	0.01	0.36	0.25	0.30	0.48	0.00	0.45	0.00	0.11	0.15	0.71	–

Table 5.2: Similarity Matrix for the 20 machines based on their breakdown duration data.

For these 20 machines the Arrows classification process proceeds as follows:

1. Step 1

   Choose the p-value threshold $p_0 = 0.10$.

2. Step 2 (see Figure 5.2)

   Form 8 groups based on identifying the single-arrow and double-arrow connections, which are displayed in Figure 5.2 as black arrows with heads at either one end (single-arrow connections) or both ends (double-arrow connections). For example, machines M01 and M11 have a double-arrow connection as the p-value for the comparison between these two machines is the greatest in row 1 and row 11 of the similarity matrix and is greater than p_0.

3. Step 3 (see Figure 5.2)

   Identify 6 pairs of machines in 3 groups that are formed in step 2 that have significantly different breakdown duration data. The connections between these pairs are coloured red in Figure 5.2. Decide which machine or machines to remove from the corresponding groups to ensure that there are no groups containing pairs of machines with p-values less than p_0, i.e. no red connections. The three groups with red connections are groups 2, 3 and 4. We consider each of the three groups in turn:

   (a) Group 2: The priority is to keep pairs of machines with double-arrow connections in the same group; therefore, M09 is removed from the group to eliminate the red connection.

   (b) Group 3: M05 and M20 have a double-arrow connection and should be kept in the same group. M18 has the weakest inside connection and is discarded. The resultant group has no red connections.

   (c) Group 4: M07 and M17 have a double-arrow connection and should be kept in the same group. Of the remaining machines, M14 has the

weakest inside connection and is deleted. The resulting group has no red connections.

4. Step 4 (see Figure 5.3)

   Combine groups 4 and 5 after step 3 as no pairs of members are significantly different, i.e. there are no red connections after the amalgamation. This is the only merging that can take place without creating red connections.

   Finally 10 groups are obtained, as shown in Figure 5.3, the largest group contains 5 machines and there are 6 groups that contain only one machine.

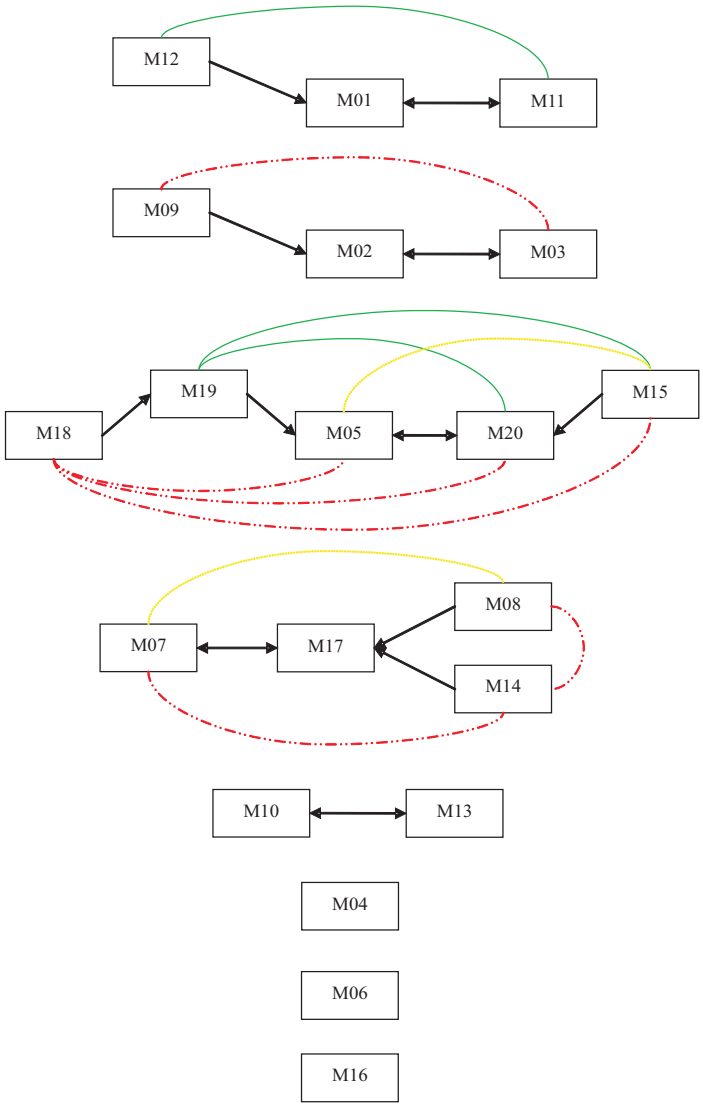

Figure 5.2: Steps 1 and 2 of the example of 20 machines, showing groups with double-arrow and single-arrow connections and the strength of the connections within each group. Red curve (— - - — -): p-value of the two connected machines is significantly different; yellow curve (- - - - - - - -): p-value of the two connected machines is on the borderline; green curve (————): p-value of the two connected machines is not significantly different.

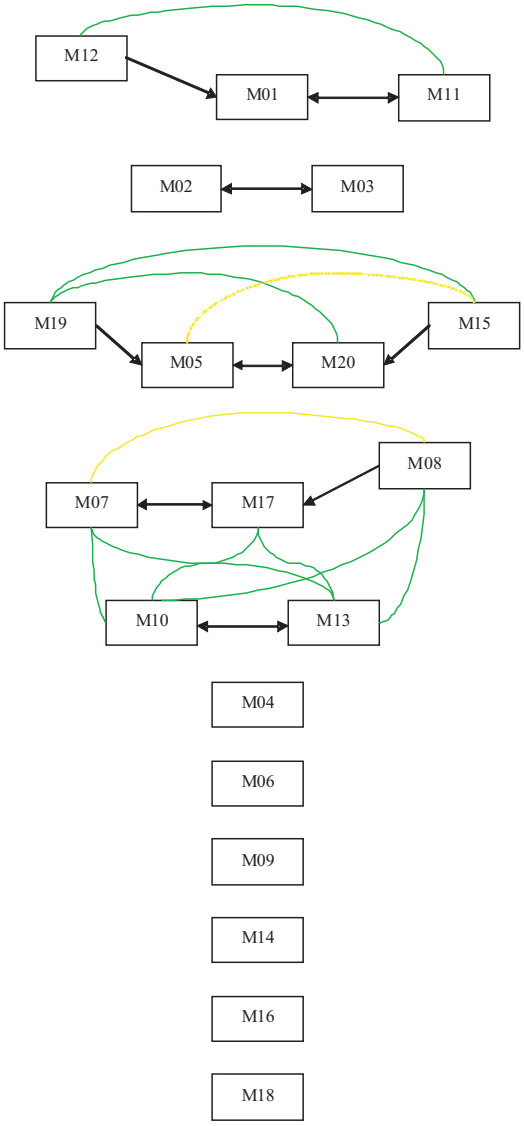

Figure 5.3: Step 4 of the example of 20 machines in which we try to combine the primary groups without red connections

5.4 Comparison with Cluster Analysis

In this section, we compare the Arrows method with cluster analysis. The Arrows classification method has similarities with complete linkage clustering and average linkage clustering methods. The complete linkage cluster analysis algorithm proceeds iteratively, combining the two most similar machines or groups of machines at each iteration, where the distance between any two groups is defined to be the greatest distance (in this case, the smallest p-value) from any member of one group to any member of the other group. The average linkage cluster analysis algorithm is the same as the complete linkage except the distance between any two groups is now defined to be the average of the distances from any member of one group to any member of the other group.

The Arrows classification method uses a threshold distance or similarity to ensure that all of the objects in a group have significant similarities. It is thus very easy to control the similarity level in the final groups when using the Arrows method. Where the two methods differ is that the clustering method searches the whole matrix to find the most similar groups to merge while the Arrows method aims to keep together objects that have what we term an double-arrow connection. Two objects have a double-arrow connection if one object has the greatest similarity to the other object and vice versa for the other object and thus keeping these objects together is a way to enhance the internal cohesion of groups resulting from the Arrows method.

The following gives a comparison between the complete and average linkage cluster analysis methods and the Arrows method by first using an example distance matrix from a text book and then extending this example to better highlight the features of the Arrows method. Finally, we show how the Arrows method works in practice, using the 20 machines example that has been described in Section 5.3.

	1	2	3	4	5
1	0.0	2.0	6.0	10.0	9.0
2	2.0	0.0	5.0	9.0	8.0
3	6.0	5.0	0.0	4.0	5.0
4	10.0	9.0	4.0	0.0	3.0
5	9.0	8.0	5.0	3.0	0.0

Table 5.3: Distance Matrix of Example 1 from Everitt [53] P9.

5.4.1 Example 1

We use a distance matrix obtained from Everitt [53] (P9) as an example; the distance matrix is given in Table 5.3. We apply complete linkage and average linkage clustering methods, and the Arrows method. The grouping results for the cluster analyses are presented in the two dendrograms given in Figure 5.4. For the Arrows method, we set a distance threshold of 10.00 with the purpose of getting a complete dendrogram, as shown in Figure 5.5. (In all of the dendrograms shown in Section 5.4, the first column of numbers is the corresponding distance or similarity level at each amalgamation, and the second column of numbers denote the order of each amalgamation only.) As the opposite of setting a similarity/p-value threshold, a distance threshold is set so that a pair of objects can be put in the same group only when the distance between them is less than or equal to this distance threshold. So, in this case, a distance threshold of 10.00 is equivalent to a similarity threshold of zero.

The dendrograms of this example resulting from the three methods are all seen to be similar in shape. Such is not always the case, as will be seen in Sections 5.4.2 and 5.4.3. Moreover, it may not be possible for the Arrows method to show the grouping results of different similarity levels by a continuous and complete dendrogram such as Figure 5.5, since the merging of some objects or groups might change when the threshold is set to a different value, which will also be illustrated in Sections 5.4.2 and 5.4.3.

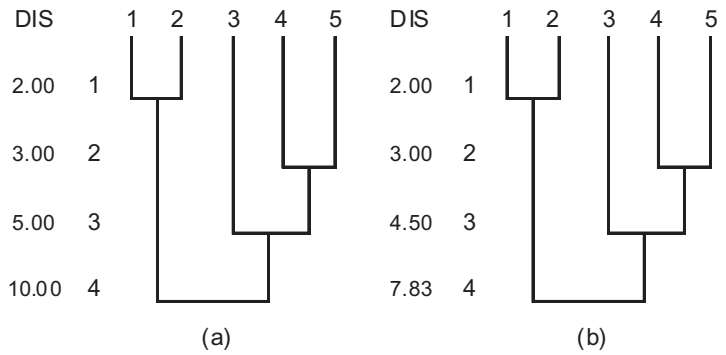

Figure 5.4: Dendrograms of the grouping results for objects with the distance matrix given in Table 5.3: (a) from the complete linkage cluster analysis; (b) from the average linkage cluster analysis. The first column of numbers is the corresponding distance between the objects or groups at each amalgamation.

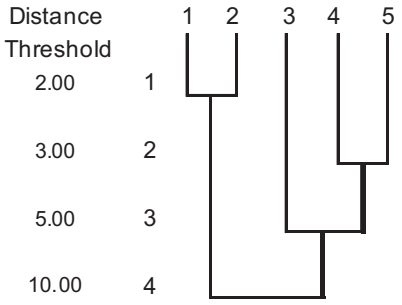

Figure 5.5: Dendrogram of the grouping results from the Arrows method for objects with distance matrix given in Table 5.3. The first column of numbers is the distance threshold.

In this example, objects 1 and 2, and objects 4 and 5 have double-arrow connections and are also the closest and second closest pairings and therefore the merging of these two pairs of objects will occur first using all of the three methods. Although at the dissimilarity level of 5.00, object 3 is in the same group with (4, 5) rather than with (1, 2) using all of the three methods, the criteria and process of getting the group (3, 4, 5) differs between the three methods. For complete linkage clustering and average linkage clustering, the only difference is the way of calculating the distance from object 3 to the existing two groups (1, 2) and (4,

5). For the former method, the distance between object 3 and group (1, 2) is 6.00, which is larger than the distance between object 3 and group (4, 5), so with complete linkage object 3 will be combined with group (4, 5). For average linkage, the distance between object 3 and group (4, 5) is 4.50, which is smaller than the distance between object 3 and group (1, 2), 5.50, so the next merging is again between object 3 and group (4, 5).

The Arrows method gives objects with single-arrow connections some priority by combining all objects with single-arrow and double-arrow connections at the beginning of the grouping process, right after the threshold has been set. In this case, objects 3 and 4 have an single-arrow connection, as the distance between 3 and 4 is the smallest in column 3 and row 3 of the distance matrix, and the distances between objects 3 and 4 and objects 4 and 5 are both smaller than the chosen distance threshold; thus the Arrows method combines object 3 with group (4, 5) rather than with group (1, 2) at the second step of the classification process described in Section 5.2, when the chosen distance threshold is 5.00.

5.4.2 Example 2

We extend Example 1 by changing the distances between objects 1 and 3 and objects 2 and 3, and adding two new objects. The new set up is designed to highlight the features of the Arrows method and the distance matrix is given in Table 5.4. The new dendrograms of grouping results from the complete linkage clustering and average linkage clustering are given in Figure 5.6. The dendrogram of groups resulting from the Arrows method using a distance threshold of less than 5.00 is given in Figure 5.7; it is not possible to show the grouping results of similarity levels that are greater than or equal to 5.00 properly in the same dendrogram, since the merging of object 3 changes when the threshold is set to 5.00 or greater, which will be illustrated later in this section.

	1	2	3	4	5	6	7
1	0.0	2.0	5.5	10.0	9.0	11.0	11.0
2	2.0	0.0	3.1	9.0	8.0	11.0	11.0
3	5.5	3.1	0.0	4.0	5.0	4.6	4.6
4	10.0	9.0	4.0	0.0	3.0	11.0	11.0
5	9.0	8.0	5.0	3.0	0.0	11.0	11.0
6	11.0	11.0	4.6	11.0	11.0	0.0	3.5
7	11.0	11.0	4.6	11.0	11.0	3.5	0.0

Table 5.4: Distance Matrix of Example 2.

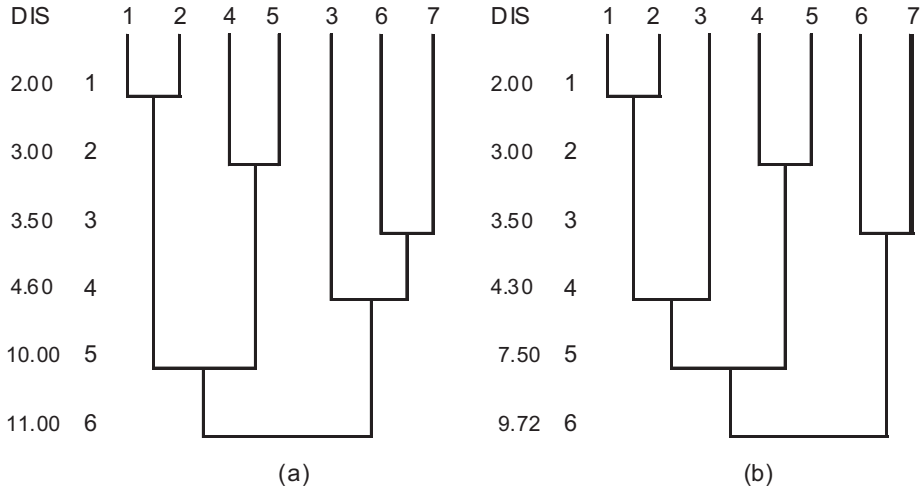

Figure 5.6: Dendrograms of the grouping results for objects with distance matrix given in Table 5.4: (a) from the complete linkage cluster analysis; (b) from the average linkage cluster analysis. The first column of numbers is the corresponding distance between the objects or groups at each amalgamation.

Using the new distance matrix, objects 1 and 2, objects 4 and 5, and objects 6 and 7 have double-arrow connections and are the closest pairs of objects and therefore the merging of these three pairs of objects make up the first three amalgamations. At the dissimilarity level of 4.60, the complete linkage clustering and the Arrow method differ from the average linkage clustering over where they place object 3. For the complete linkage clustering, the distance between object 3 and group (6, 7) is 4.60, which is smaller than the distances between object 3 and group (1, 2) or group (4, 5); and so object 3 is combined with group (6, 7). For

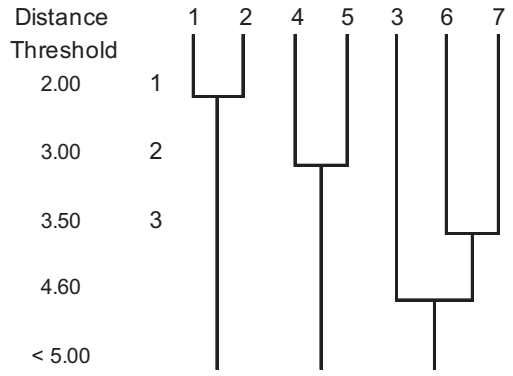

Figure 5.7: Dendrogram of the grouping results from the Arrows method using distance threshold lower than 5.00 for objects with distance matrix given in Table 5.4. The first column of numbers is the distance threshold.

the Arrows, object 3 is amalgamated with group (6, 7) rather than with group (1, 2) or group (4, 5) because the distances between object 3 and objects 1 or 5 are both higher than the specified distance threshold. While for the average linkage clustering, the distance between object 3 and group (1, 2) is 4.30, which is smaller than the distances between object 3 and group (4, 5) or group (6, 7); the next merging is therefore object 3 and group (1, 2).

For the Arrows method, multiple criteria are used to decide the next merging. First it ensures that no objects that are further apart than the threshold distance can be placed within the same group, then it ensures that objects with double-arrow connection are placed in the same group. The Arrows method prefers to keep objects with single-arrow connections together, if all relevant distances are below the threshold distance, even when there are other potential amalgamations satisfying the first criterion. If there are no objects with single-arrow connections involved, it allows the merging of objects or groups with lower or the lowest average distance (i.e. higher or the highest average connection).

It is possible that one object or group may be combined with different groups or objects when the distance threshold changes. This might occur as a result of

the method's intention of keeping objects with single-arrow connections in the same group, while satisfying the condition that every pair of objects in the same group should have a distance that is below the threshold distance. For example, the grouping results for object 3 are different when the distance threshold is changed from 5.50 to 5.00, as shown in Table 5.5. When there is no relevant influence from single-arrow connections, one object may also be grouped with different groups or objects when a different distance threshold is selected due to the method's aim to merge objects or groups with higher average connections, while satisfying the condition that every pair of objects in the same group should have a distance that is below the selected threshold distance. An illustration of this situation is also shown in Table 5.5: the different merging for object 3 when the distance threshold is changed from 5.00 to 4.60.

Distance Threshold	**Group**	**Objects**
5.50	1	1, 2, **3**
	2	4, 5
	3	6, 7
5.00	1	1, 2
	2	**3**, 4, 5
	3	6, 7
4.60	1	1, 2
	2	4, 5
	3	**3**, 6, 7

Table 5.5: Grouping results of Example 2 using the Arrows method with a distance threshold of 4.60, 5.00 or 5.50.

Selecting a distance threshold of 5.50, object 3 is placed in the same group as objects (1, 2) in step 2 of the Arrows classification process described in Section 5.2, because object 3 has a single-arrow connection with object 2 and the distance between objects 3 and 1 is no greater than 5.50, the distance threshold. However, when the distance threshold is set to be 5.00, object 3 can no longer be put in the

same group with (1, 2) because the distance between objects 3 and 1 is now larger than the distance threshold; thus, object 3 is grouped with group (4, 5) in step 4 of the classification process described in Section 5.2, as the distances between object 3 and object 4 or object 5 are both no greater than the current distance threshold and the average distance between object 3 and group (4, 5) is smaller than the average distance between object 3 and group (6, 7). Moreover, when the threshold is changed to be 4.60, the grouping result for object 3 is different again; object 3 is amalgamated with group (6, 7) rather than with group (4, 5) because the distance between objects 3 and 5 is now higher than the specified distance threshold and hence object 3 cannot be merged with group (4, 5) even though the average distance between object 3 and group (4, 5) is smaller than the average distance between object 3 and group (6, 7).

Since using different thresholds means the grouping results for object 3 may be different, it is not possible for the Arrows method to show the grouping results of different similarity levels by a continuous and complete dendrogram; only the incomplete dendrogram of using a distance threshold of less than 5.00 shown in Figure 5.7 can be drawn, from which the grouping results can be read straightforwardly when a distance threshold is set to be any value less than 5.00.

It is seen that the three methods give similar results; for instance, the core of the groups, (1, 2), (4, 5) and (6, 7), stay the same. From the groupings resulting from the Arrows method using different distance thresholds, it seems that when a lower similarity level is required within the groups, the Arrows method appears to be more similar to the average linkage clustering, however, when a higher similarity level needs to be achieved, the Arrows method tends to be closer to the complete linkage clustering.

5.4.3 Example 3

It is seen from the previous two examples that the Arrows classification method has similarities with the complete linkage and average linkage hierarchical cluster analysis [3]. We here use a more complicated example to study the differences between the methods as well as the influence of the threshold on the results of the Arrows method; the similarity matrix is given in Table 5.2. The dendrograms of the grouping process of the 20 machines using the complete linkage and average linkage clustering are given in Figures 5.8 and 5.9, respectively. The dendrogram resulting from the Arrows classification method for p-value thresholds $p_0 > 0.046$ is given in Figure 5.10. When the threshold is set to be less than or equal to 0.046, group (M10, M13) may be combined with different machines or groups of machines and thus the corresponding grouping results cannot be properly displayed in the same dendrogram.

For the machines data, we generally assume that two machines with a p-value smaller than 0.10 are considered to be significantly different and therefore cannot be combined in to one group. Thus, it is reasonable to ignore the grouping results obtained at a similarity level below 0.046.

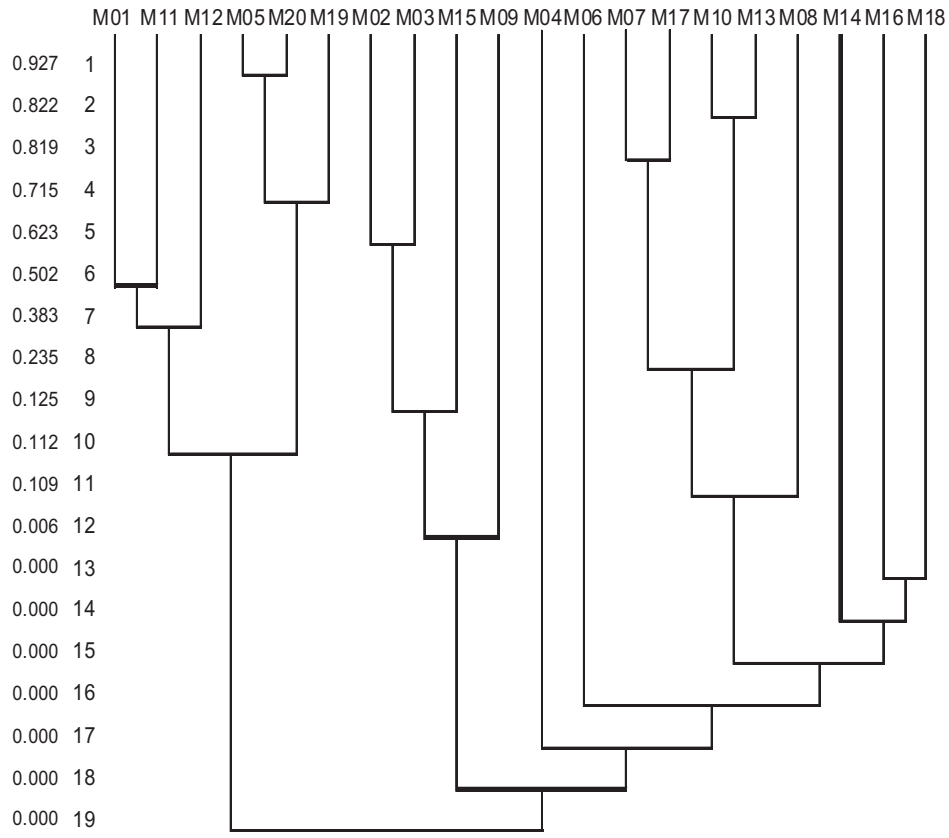

Figure 5.8: Dendrogram from the complete linkage cluster analysis for the example of 20 machines. The first column of numbers is the corresponding similarity level at each amalgamation.

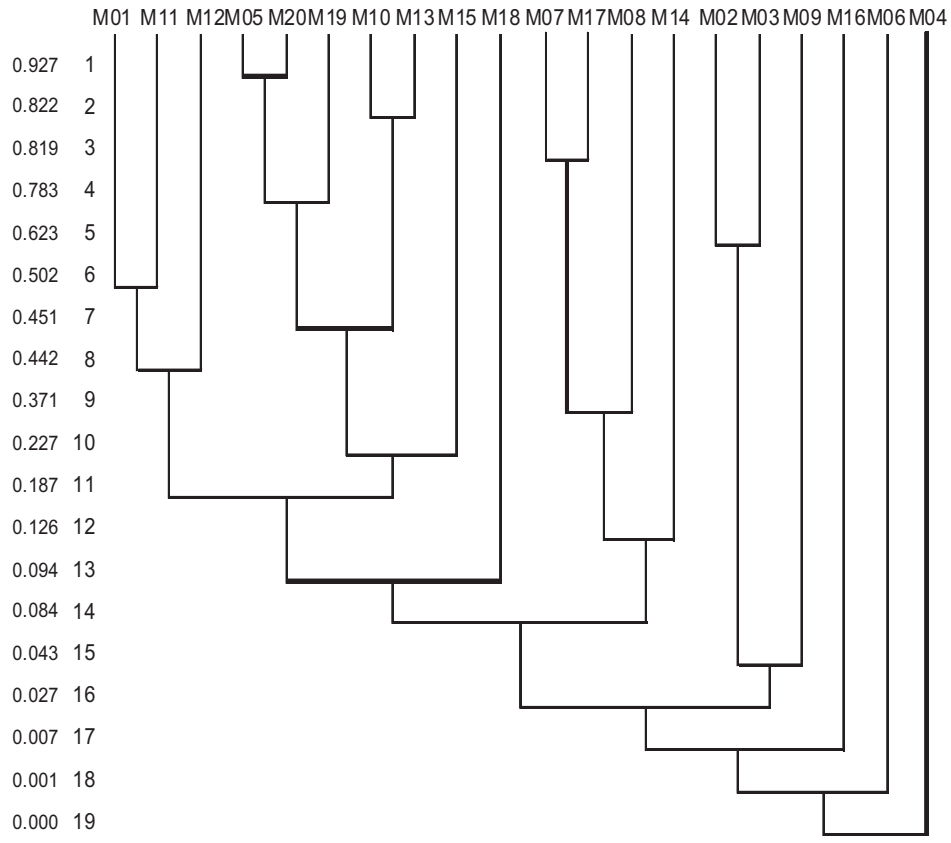

Figure 5.9: Dendrogram from the average linkage cluster analysis for the example of 20 machines. The first column of numbers is the corresponding similarity level at each amalgamation.

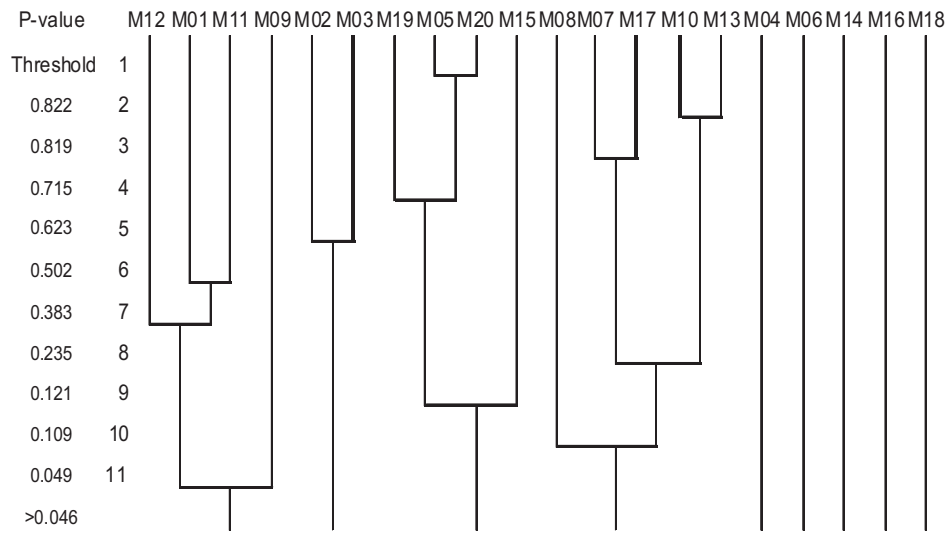

Figure 5.10: Dendrogram from the Arrows clustering method using similarity threshold $p_0 > 0.046$ for the example of 20 machines. The first column of numbers is the corresponding p-value/similarity threshold.

We compare the classification methods by examining the grouping results at a similarity level of 0.10 as we assume that two sets of data are similar enough to be grouped together when their p-value is above 0.10.

We consider the average linkage clustering initially. The dendrogram is given in Figure 5.9, and this shows that the 12th merging happens at similarity level of 0.126 while the 13th amalgamation happens at 0.094. Thus, at a similarity level of 0.10, it gives 8 groups, as listed in Table 5.6. Since the average linkage clustering uses the average similarity between groups as the only measure to decide groups, there are pairs of machines with very low p-values that are included in the same groups. For example, in the first group in Table 5.6, the similarity between M15 and M13 is 0.046; in the third group, M07 and M14 have an extremely small similarity p-value of 0.009, while M08 and M14 has a even lower p-value of zero, which statistically means there is zero possibility that the breakdown duration data of M08 and M14 are drawn from the same distribution. Thus, although the final number of groups at the similarity level of 0.10 is fewer than the number of groups resulting from the other two methods, the homogeneity of the groups is not sufficiently high. Hence, in the following we focus on comparing the Arrows method with the complete linkage clustering.

Group	**Machines**
AL1	M01, M05, M10, M11, M12, M13, M15, M19, M20
AL2	M02, M03
AL3	M07, M08, M14, M17
AL4-AL8	(Single machine groups) M04, M06, M09, M16, M18

Table 5.6: Grouping results of the 20 machines at a similarity level of 0.10 using the average linkage clustering method.

In the complete linkage clustering dendrogram shown in Figure 5.8, the 11th merging happens at a similarity level of 0.109, while the 12th amalgamation happens at a p-value equal to 0.006. Therefore, there are 9 groups, listed in Table

5.7, at a similarity level of 0.10. Figure 5.3 and the dendrogram in Figure 5.10 show that using the Arrows method with a threshold of 0.10 produces 10 groups, as listed in Table 5.8.

Group	Machines	Average Similarity within Group
CL1	M01, M05, M11, M12, M19, M20	0.385
CL2	M02, M03, M15	0.303
CL3	M07, M08, M10, M13, M17	0.420
CL4-CL9	(Single machine groups) M04, M06, M09, M14, M16, M18	-

Table 5.7: Grouping results of the 20 machines at a similarity level of 0.10 using the complete linkage clustering method.

Group	Machines	Average Similarity within Group
AR1	M01, M11, M12	0.462
AR2	M02, M03	0.623
AR3	M05, M15, M19, M20	0.574
AR4	M07, M08, M10, M13, M17	0.420
AR5-AR10	(Single machine groups) M04, M06, M09, M14, M16, M18	-

Table 5.8: Grouping results of the 20 machines at a similarity level of 0.10 using the Arrows classification method.

It can be seen from Tables 5.7 and 5.8 that the results are similar, for example, the single machine groups are exactly the same; the reason for this similarity between the two classification methods appears as both methods ensure that every pair of objects within the same group has a similarity that is above the similarity level, 0.10: for complete linkage clustering, it is achieved by using the smallest p-value within one group as the similarity level of that group; while for Arrows method, it is achieved by setting a p-value threshold as one of the main features of this classification method.

The differences between the two grouping results come from the different groupings of machine M15 only and this demonstrates one of the major features of the Arrows method, that is, it aims to keep together objects with single-arrow connections when possible. Using the Arrows method, M15 has a single-arrow connection with machine M20 and has above-threshold ($p_0 = 0.10$) similarities with M05 and M19. Therefore it is amalgamated with group (M05, M19, M20) during the second step of the process described in Section 5.2. The complete linkage clustering method uses the furthest distance as the only index for grouping, in this case, the smallest p-value. Using complete linkage, M15 is merged with (M02, M03) instead of (M05, M19, M20) because the smallest p-value between M15 and (M02, M03) is higher than the smallest between M15 and (M05, M19, M20). The differences between the grouping results coming from the complete linkage and the Arrows method can be seen in Tables 5.7 and 5.8: CL1 and CL2 vs. AR1, AR2 and AR3. The average similarities within the three groups resulting from the Arrows method are all higher than those within the two groups resulting from the complete linkage clustering. Thus, it is believed that the Arrows classification method achieves more homogeneity within the resultant groups than the complete linkage clustering at the similarity level of 0.10, although the latter method gives a slightly smaller number of groups.

At similarity levels of 0.20, 0.30, $\ldots$, 0.90, all of the grouping results of the Arrows method and complete linkage clustering are the same, despite their different methods for merging groups. The results are shown in the dendrograms in Figures 5.8 and 5.10 and are listed in Appendix A. It is seen that by increasing the threshold p-value the homogeneity of the groups is improved but the number of groups needed to describe the data increases.

On the whole, it seems that the proposed Arrows classification method produces similar results to the hierarchical cluster analysis. The major difference between the two is that the clustering method searches the whole matrix to find the

most similar groups to merge while the Arrows method prefers to keep together objects with double-arrow and single-arrow connections. The use of a threshold distance or similarity is also a characteristic of the Arrows method, which ensures that any two objects whose similarity is less than the selected similarity threshold will not be allowed to be put in the same group. The Arrows classification method therefore allows us to control the similarity level in the resultant groups more easily than cluster analysis.

5.5 Classification of Hospital Length-of-Stay Data

The Arrows classification method is a general method and could be applied to classify data from a wide range of applications, in addition to manufacturing. We here include an example involving a health care application where it has also been applied. This example comes from [41], where the ultimate purpose was to use Gallivan and Utley's linear programming approach for setting up optimal schedules for hospital procedures [61]. As we mentioned in 4.6.2, we wish to group procedures based on the similarity of their patients' length-of-stay data.

This classification of procedures into groups before the optimising process has three benefits [41]. First, the schedules output by the optimisation program have more flexibility. Instead of insisting that a set number of procedures of a particular type X need to be performed on a certain day, the schedules output are able to suggest that a set number of procedures of Group G_X need to be performed on a certain day, where Group G_X may include more than one type of procedure. Therefore, if a cancellation or a last minute request for a procedure occurs, substitution is relatively easy. Second, the number of variables in the optimisation program can be reduced by the grouping and the subsequent computation time required to find the optimal schedule can be decreased. This saving can be significant when setting up a schedule of a large number of procedures for several weeks.

Third, the demand for a group of procedures will be more accurately forecast than the demand for individual procedures.

We have length-of-stay data for $10,929$ different episodes recorded over a period of 7 months coming from 655 different procedures. There are a large number of rare procedures for which we have little data. After the primary analysis of the data (see [41] for details), there are 147 procedures or procedure groups that we wish to classify into a smaller number of groups. The aim of this example is to group these 147 different procedures or procedure groups based on their length-of-stay data; which means that two procedures or procedure groups can be put in the same group if there is no statistically significant difference between the distributions of their length-of-stay data. Beforehand, we need to obtain the similarity matrix of the procedures using the method we introduced in Section 4.4. We run 100 bootstraps for each pair of procedures or procedure groups to determine the p-values. Here, we again set 0.10 to be the p-value threshold for the Arrows classification procedure.

The results of the Arrows method suggest that there should be 48 groups, and these are given in Table 5.9 (the codes are called OPCS-4 codes and are used by NHS; www.hesonline.nhs.uk provides a facility for decoding these codes). The largest group contains 8 procedures and there are four groups of 7 procedures; 14 groups contain only one procedure. Overall, the groups make sense intuitively. For example, group 19 is mainly made up of rare inpatients procedures; group 23 includes only endoscopic procedures on the fallopian tubes and uterus; and group 28 contains hip and knee replacements.

Group	Procedures
1	25120, A52.1, F09.5, SO8.2, Ear, nose and throat Outpatients
2	A57.3, E35.2, Q18.1, W92.4, Anaesthetics Mixed, Paediatrics Outpatients, Gynaecology Outpatients

Group	Procedures
3	25012, A57.6, A57.7, C71.2, S64
4	A65.1, A65.8, S70.1
5	B27.8, C13.4, F34.4, F34.8, L85.8, T20, T21, W08.1
6	B28, C13.2, W78, Urology Mixed
7	B28.2, B28.8, H51, M79.4
8	B31, J18.3, S01
9	B31.3, W79.1
10	C13.3, N18.1, Q17, W85
11	C17, C18.1, N30.3, W82
12	D03.3, S06.4, S25, W90.4, ultrasound guided biopsy
13	B31.2, E02.6, W86
14	E03.6, E14.3
15	F34, L85.2, L85.3, T27, W81.9
16	H55.1, Q38, S62.2, T24, W87, Orthopaedics Mixed, Ear, Nose and Throat Mixed
17	J18.8, M11.1, W03, W08.6, Plastic Surgery Inpatients
18	L85.1, W77.1, Paediatrics Mixed
19	M42.1, Orthopaedics Inpatients, General Surgery Inpatients, Urology Inpatients
20	M42.3, W08.5, W28.3, W82.8, General Surgery Mixed, Plastic Surgery Mixed, Ophthalmology Mixed, General Medicine Mixed
21	N13.4, Q48.1
22	D15.1, F09.1, H20, P27.3, T80.5, Anaesthetics Outpatients, Urology Outpatients
23	Q13.1, Q20.2, Q38.3

Group	Procedures
24	B31.2, E02.3, S01.4, T79.1, Ophthalmology Inpatients, Ear, Nose and Throat Inpatients
25	C22.2, S06, Ophthalmology Outpatients, Plastic Surgery Outpatients, Oral Surgery Outpatients
26	T59, T72.3
27	B27.4, T85.2, V33.6
28	W37.1, W42.1
29	Medical admission, Non-procedure related admission
30	G65, H25
31	Q07.4, T41.3, W37.15
32	M45.1, Orthopaedics Outpatients
33	T20.1, T21
34	W90.3, General Medicine Outpatients
35 to 48	(Single procedure groups) C12.3, K65.1, M14, M65.3, Q39, S02.1, S06.3 , S60.4, V25.4, W74.2, W82.3, W86, Gynaecology Inpatients, Gynaecology Mixed

Table 5.9: Grouping results of the hospital procedures.

5.6 Conclusion

The Arrows classification method has been demonstrated using a simple distance matrix from a text book as well as practical and more complicated similarity matrices. The method is widely applicable and we have described its use in the

classification of medical procedures [41], as well as the classification of machines. When a larger similarity threshold is set, the homogeneity of the groups improves but the number of groups generally increases. The balance between the competing requirements of homogeneity within groups and having a small number of groups therefore can be achieved by selecting an appropriate threshold p-value.

The Arrows method gives groupings similar to those resulting from complete linkage and average linkage hierarchical cluster analysis. In general, when a lower similarity level is required within the groups the Arrows method tends to be more similar to the average linkage clustering, while when a higher similarity level needs to be achieved the Arrows method performs more similarly to the complete linkage clustering. This flexibility in the Arrows method allows the same algorithm to be used to satisfy different aims by simply changing the similarity threshold, whereas with cluster analysis it can be necessary to switch to a different algorithm. Moreover, the Arrows classification method has been implemented in Visual Basic for Applications in Excel, allowing it to be used by a non-expert; for example, the engineers at Ford.

In the case of classifying machines based on their breakdown duration data, the target might be to use fewer groups to gain a greater saving on the time spent estimating fitted mixture models. Using the Arrows method we can set a lower threshold and using cluster analysis, we may choose to use the average linkage clustering. If it is necessary to be cautious with the classification, and only to group machines with fairly high similarities we can use a higher threshold to achieve this in the Arrows method, but using cluster analysis, we might need to switch to the complete linkage clustering.

Chapter 6

Simulation

Discrete-event simulation is commonly used in the manufacturing industry to investigate the design and operation of different production lines ([89] and [140]). Ford has been using discrete-event simulation modelling to evaluate new designs for assembly and machining lines and to improve the efficiency of existing lines since 1982. Engine Assembly lines produce saleable engines by assembling components together, most of which are manufactured on automatic transfer lines. We focus on the study of the machine breakdown modelling process for simulating an existing engine assembly line. The line will be referred to as 'DuntonL01'. In this case, the simulation is used to perfect the design of the layout in line DuntonL01. For example, the layout design department may introduce a new design for a particular part of this line. If the new design is launched, the buffer sizes, conveyor length or number of machines may need to be adjusted. Thus, corresponding changes are made in the simulation model to generate new simulation outputs, e.g. line yield and costs, which are used to verify the feasibility of the new design and to estimate its effectiveness.

In this chapter we first briefly introduce the engine assembly lines and transfer lines in Ford manufacturing plants in Section 6.1 and then describe the construction process of simulation models in Section 6.2. In Section 6.3, the modelling for

machine breakdowns, engine repairs and operator stoppages are introduced briefly and the maintenance settings are described. We then focus on the machine breakdown modelling process in Section 6.4.

As the simulation models are built by Ford using the WITNESS simulation software (Lanner Group) [102], one of the major steps of the machine breakdown modelling process is to select a breakdown mode in the software to decide the method for modelling the machine breakdown behaviour, and this is discussed in Section 6.5. As the whole cycle of a machine during manufacturing consists of a sequence of cycles of two segments [103]: up segment where the machine is busy, blocked or idle and down segment where the machine is broken down. Estimating distributions for machine breakdown data thus contains two parts as well: deciding the distributions for representing the time between failures as the machine up segment modelling and estimating the distributions for representing the breakdown durations as the machine down segment modelling. We do not focus on modelling the up segment in this work. A brief description of the modelling method for the machine up segment is given in Section 6.6. Finally, there are a number of issues concerned with the simulation settings, these are described in Section 6.7.

6.1 Manufacturing and Engine Assembly Lines

The engine assembly process generally involves automatic, semi-automatic and manual machines, material handling and machine linking systems, human services including operators, engineers and maintenance operators and other facilities including electrical and coolant materials, tool and parts stores and computerised support and monitoring systems ([97] and [135]). The major standardized engine components that are required for the assembly process are manufactured in automatic transfer lines [21]. A transfer line consists of machining facilities including different machines for various tasks, material handling systems that connect the

machines such as powered conveyors, gantries or palletised loops, manual services and the other facilities mentioned above for assembly lines. Rough parts are processed and machined into completed components in a transfer line. The appropriate engine components are later transported to engine assembly lines, where they are assembled together in a defined sequence and finished as a saleable engine.

The number of machines in an engine assembly line varies based on the type of engines being assembled and the quantity required. Figures 6.1 and 6.2 are layout diagrams of the DuntonL01 engine assembly line simulation model built in WITNESS that we are working on. The former shows the whole view of the assembly line but no details are legible as there are 192 main operations and over 200 machines involved in this line; the latter shows the details of a small part of this line where the yellow blocks with the print of "OP" on indicate machines and the other yellow blocks with small image of conveyor on indicate the conveyors that link the machines together.

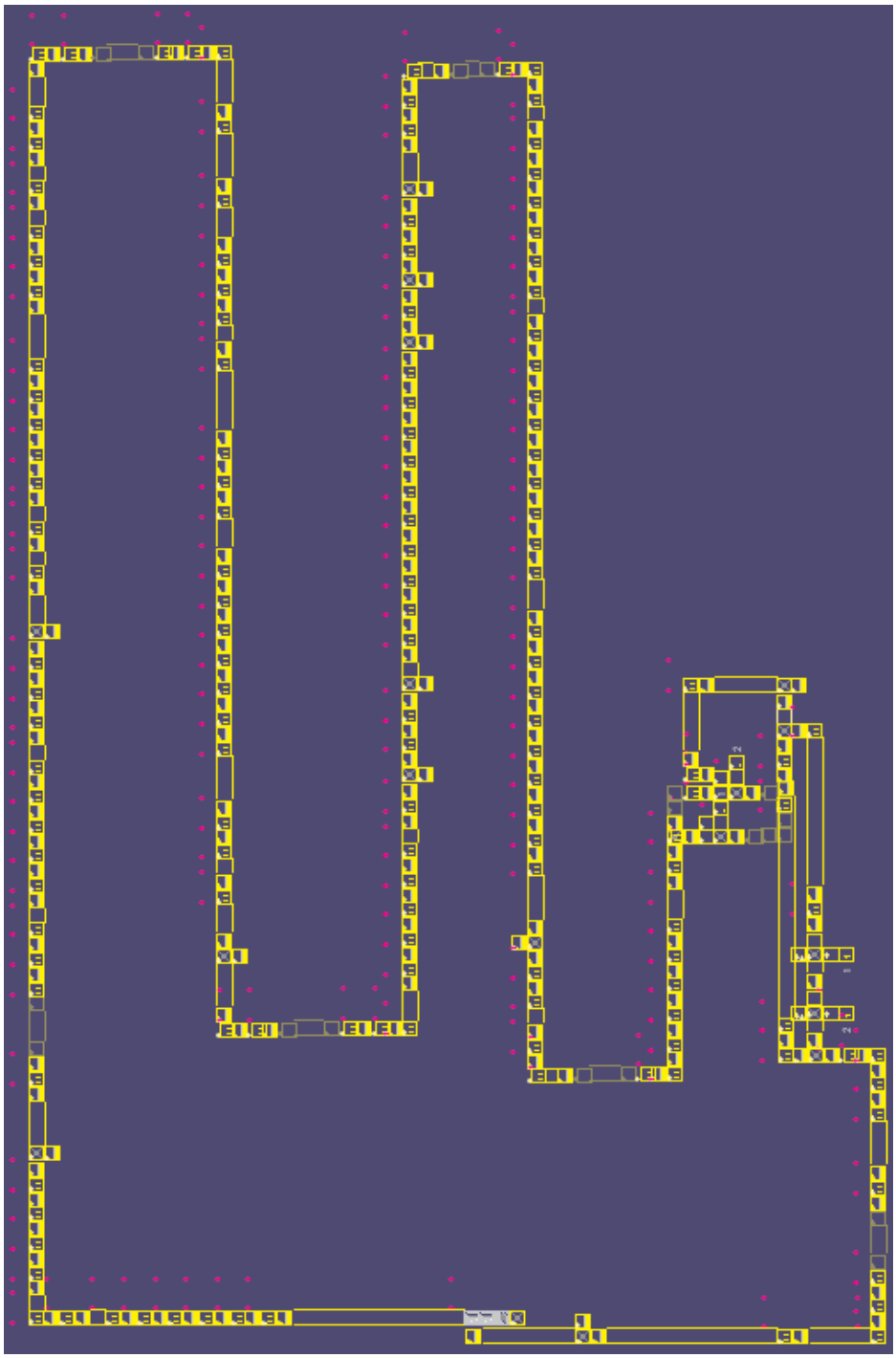

Figure 6.1: Layout diagram of the whole view of the DuntonL01 engine assembly line built in the WITNESS 2008 version software.

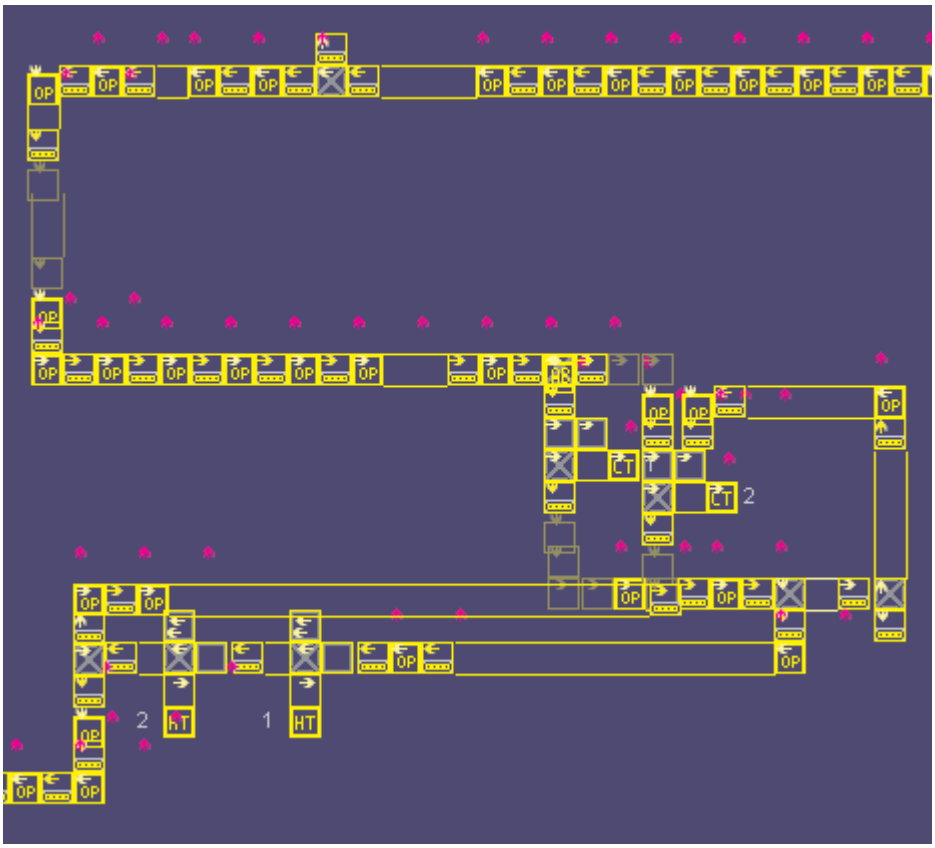

Figure 6.2: Layout diagram of a part of the DuntonL01 engine assembly line built in WITNESS 2008 version software.

6.2 Construction of Simulation Systems

Ford Motor Company have developed several interfaces using Microsoft Excel so that simulation models can be created automatically once the engineers have filled all required entries in the spreadsheets [162]. The initial interface was called FIRST standing for Fast Interactive, Replacement Simulation Tool [98]; created for easier and quicker use by Ford manufacturing engineers.

The interfaces have being used as replacements for detailed simulation construction. These tools enable manufacturing engineers to construct a simulation model by simply inputting required data that is marked and explained clearly in the spreadsheets. Generally, operation numbers, machine identification names, cycle times, setup rates, breakdown settings for machines, shift patterns and a lot of other data are required to be added into these spreadsheets. Using Visual Basic macros inside these spreadsheets, all the data can then be saved directly into the WITNESS system and simulation models with the specified design will be automatically created.

The simulation models constructed through these spreadsheet interfaces mostly have 2D schematics of the whole production line layouts such as that shown in Figure 6.1. Every entry into the spreadsheets by engineers corresponds to their design for the model. For example, positioning data of facilities can be specified in the spreadsheets so that the next facility in the production line is automatically placed in a position relative to the current facility in the built model [162].

A model built through the Excel interface is no different to a simulation model that is built directly on the WITNESS simulation system interface. Figure 6.3 shows a simple conveyor system and a machine details setting dialog. All details included in the dialog are built in when the simulation model is automatically created by the spreadsheet in which the engineers have already input all details that are required to define the machines and other elements in the model.

Figure 6.3: A sample WITNESS layout diagram from a Ford simulation model showing a typical simulation dialog which contains control rules and timings for the each operation and facility within the plant using the WITNESS software, given in [162].

The DuntonL01 engine assembly line model whose layout diagram is shown in Figure 6.1 is built through an Excel interface called FAST. There are 192 main operations and over 200 machines involved in this line. Building a simulation model as complicated as this, the use of spreadsheet tools obviously appear to be a much simpler way and saves considerable time. In addition, small changes to the created simulation models can be made on the WITNESS system interface as well as through the spreadsheets interface.

6.3 Breakdown and Maintenance Logic

The simulation model of the complete engine assembly line is developed in the WITNESS simulation system. As we mentioned in Chapter 1, there are three major causes of production loss: the machine repairs, engine repairs and operator stoppages. This work focuses on the modelling of machine breakdown durations and we propose to use finite mixture distributions fitted to grouped breakdown

duration data as the simulation inputs for machines involved. Engine repairs are simply modelled using the percentage of engines with production quality issues, since there are no available data for more detailed analysis. An operator who is attending the machine may suddenly fail to perform the job on rare occasions. Human breakdown modelling for machines is included to model these rare cases, where generally an Erlang distribution is used to represent the time of operator stoppages, and an extremely low percentage is used to model the frequency of occurrences.

The maintenance logic for machine repairs in the model assumes that an immediate repair will be made when a machine fault occurs and an operator or maintenance operator is available [135]. The failure's duration, which is generated from the machine breakdown duration input distribution, is used to determine the skill level of the maintenance staff required to complete the repair. For example, when the time to repair a failure is generated to be longer than 15 minutes, the highly skilled maintenance operator will be called; otherwise the operator attending the machine will carry out the repair.

The assumption made is that the generated repair time includes the time to wait for maintenance to become available and also the maintenance operator's travel time. In the design of the simulation model, the waiting time for maintenance to become available is generated separately in situations where all of the maintenance staff are busy. In order to meet the assumption of the wait for maintenance being included in the repair time, while still using standard resources settings, a bypass designed in the model is to set a large number of resources so that there are always maintenance staff available for attending a repair when a machine failure happens.

6.4 Machine Breakdown Modelling Process

The machine breakdown modelling process for engine assembly line simulation models is considered to be the typical breakdown modelling methodology at Ford and is the one to be used in all production models such as transfer line models. The general process used to model breakdowns includes six main steps as shown in Figure 6.4, and is described below.

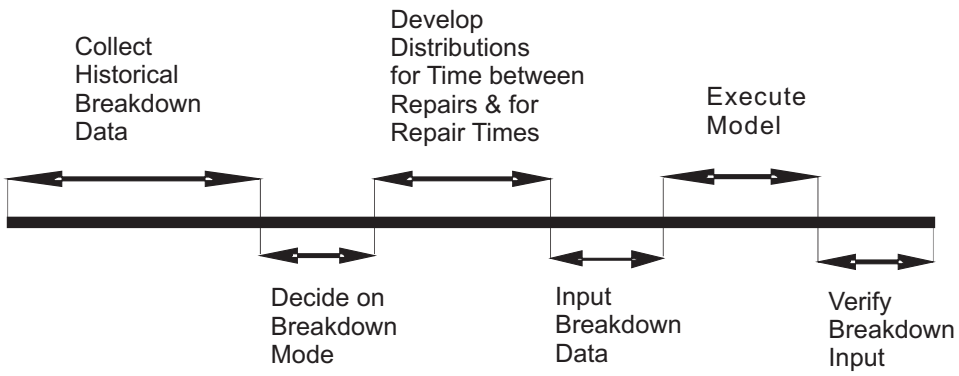

Figure 6.4: Diagram of the machine breakdown modelling methodology.

1. Collect historical data:

   When building a simulation model for an existing line, collect raw breakdown duration data for all machines from that line; when building a simulation model for an in-planning new line, collect raw breakdown duration data from existing machines which will be involved in the new line or are similar to the machines that will be placed in the new line. We use an automatic on-line monitoring system to collect breakdown data (see Section 2.4 for details). The raw data collected directly from the monitoring system need to be validated and checked (see Sections 3.3.1 and 3.3.2), and then can be used in the subsequent analysis.

2. Decide on breakdown mode:

An appropriate breakdown mode in the WITNESS simulation software must be chosen as the method for modelling the machine breakdown behaviour during simulation runs, i.e. the way that the WITNESS software calculates the time between successive repairs. This step will be discussed in Section 6.5.

3. Develop distributions for representing time between machine failures and repair times:

   In order to use the models to reflect the real-world situation, information relating to the breakdowns of the machines must be entered. This is normally in the form of a downtime distribution and a time between failures distribution [21]. An exponential distribution has been used to represent the time between failures, to parameterise which only the value of mean time between failures (MTBF) needs calculation; MTBF is calculated using formulations that have been established by Ford engineers and will be introduced in Section 6.6. To represent the breakdown durations, Ford usually use empirical distributions; we propose to use the finite mixture models for groups of machines, for which the fitting process has been discussed in Chapter 3.

4. Input breakdown data:

   Input the empirical distributions or finite mixture distributions that represent the machine failure durations and negative exponential distributions that represent the time between failures.

5. Execute model:

   It is usually executed for a warm up of one day and a length of 10 days in Ford due to time limitation. An investigation of choosing appropriate warm up period will be explained in Section 6.7.1.

6. Verify breakdown input:

Initially, this is to check whether the target machine downtime levels were met. The simulation evaluation process will be introduced in the next chapter, Chapter 7.

6.5 Using WITNESS to Model Breakdowns

There are three methods for modelling machine breakdown behaviour in the WITNESS simulation software:

1. Available Time Mode:

   This method is also referred to as Calendar Time Mode [103]. In this mode, machines can break down whether they are operating or not. Failures can occur when a machine is idle, busy, blocked, being setup, being repaired or waiting for labour. The time between failures refers to the total elapsed time that the machine has spent in any of the above listed states ([141], [150], [97], [103] and [102]).

   Two drawbacks of the available time method have been identified by Law [103]. One is that it may not be realistic for machines to break down when they are in the idle state. The other is the problem that when a specified machine is in two different systems with a number of other machines. Since there is the same distribution of time between failures for this machine in both systems, the generated time between failures will be the same in both systems. Due to different operating times and conditions in the two simulations, this particular machine may have significantly less breakdowns for one system than for the other. Thus, this approach may not be very realistic.

2. Busy Time Mode:

   Using this option, machines can only break down while they are operating. In other words, a failure can only happen when the machine is processing at

least one part. The time between failures here refers to the elapsed time the machine has spent only in the busy state ([141], [150], [97], [103] and [102]). It is generally believed to be a more natural approach than the available time mode [103].

3. Number of Operations Mode:

   This method is referred to as number of completed parts in [103]. Selecting this mode, a machine will break down after a certain number of operations ([141], [150], [97], [103] and [102]). The time between failures is expressed as the number of operations that a machine has completed since the last failure. Many manufacturing machines do not follow this kind of breakdown pattern; therefore this method is not as well-known as the other two.

Ladbrook [97] also suggested that care should be taken when using the Available Time mode. It was noticed that some scheduled breakdowns were delayed since both the time to the next failure and the repair time of this failure are generated from the input distributions at the start of a breakdown. We use the Busy Time mode in this work.

6.6 Time Between Failures

We do not focus on modelling the time between failures in this work and use the standard method employed by Ford's engineers. This assumes that the time between failures follows an negative exponential distribution, with mean equal to the mean time between failures (MTBF), which is the way the time between failures are currently modelled in the simulation model for line DuntonL01. The WITNESS simulation model will then generate the time of the next failure on a machine from the negative exponential distribution at the start time of a breakdown.

This method for modelling the time between failures is verified in some of the research undertaken in Ford's Engineering Department [97]. It is believed that the averaged line yield produced by the simulation model with the use of negative exponential distribution as the inputs of the time between failures is "as accurate as using historical data" [21]. Nonetheless, it is also indicated that in [21] the negative exponential distribution can not represent the time between failures accurately for all of the machines. Without available data and further research, this is believed to be the best representation of the time between failures. But, we believe there may be a better representation and further suggestions are described in Section 8.5.

We calculate the MTBF for a machine to be

$$MTBF = \frac{TT - TTR}{No.ofFailures}, \tag{6.1}$$

where TT is the time period over which the raw breakdown duration data are collected, and TTR is the total time a machine is broken down during the data collection period. To calculate TTR, we split the breakdown duration data into n bins with thresholds $b_1, b_2, \ldots, b_n$, where the bins do not necessarily need to have the same width. Thus,

$$TTR = \sum_{i=1}^{n-1} \frac{(b_{i+1} - b_i)}{2} F_i, \tag{6.2}$$

where $F_i, i = 1, \ldots, n$ is the number of observations in bin i. The MTBF is then used as the parameter in the exponential distribution.

6.7 Issues with Model Execution

In order to carry out our analysis of the simulation output data, we need to be able to assume that we have a set of *independent and identically distributed* (IID) observations. For this to be true, the stochastic process must be covariance-stationary

and demonstrate no autocorrelations. An output stochastic process beginning at time zero in a simulation is unlikely to be covariance-stationary and is likely to present autocorrelations [103]. We therefore wish to estimate the appropriate warm-up period when executing the simulation to ensure that the output process of the engine assembly line simulation is in a steady state when we start collecting results. We then need to check that the steady-state output exhibits no autocorrelations.

We let $y_1, y_2, \ldots, y_n$ denote a sequence of throughput observations of a simulation run of the assembly line model, which is known as a discrete-time *stochastic process*. It is said to be *covariance-stationary* if

$$\mu_i = \mu, \text{ for } i = 1, 2, \ldots n \text{ and } -\infty < \mu < \infty \tag{6.3}$$

$$\sigma_i^2 = \sigma^2, \text{ for } i = 1, 2, \ldots n \text{ and } \sigma^2 < \infty \tag{6.4}$$

where μ_i and σ_i^2 denote the mean and variance of y_i, respectively; and $C_{i,i+j} = Cov(y_i, y_{i+j})$ is independent of i for $j = 1, 2, \ldots, n - i$.

The definition of covariance-stationary means that the covariance between two observations y_i and y_{i+j} depends only on the time interval, lag j. Therefore, the lag j autocorrelation of stochastic process $y_1, y_2, \ldots, y_n$ is

$$\rho_j = \frac{C_{i,i+j}}{\sqrt{\sigma_i^2 \sigma_{i+j}^2}} = \frac{C_j}{\sigma^2} = \frac{C_j}{C_0}, \text{ for } j = 0, 1, 2, \ldots, n$$

where C_j and ρ_j denote the covariance and correlation between y_i and y_{i+j}, respectively. With autocorrelated simulation output data the sample mean $\overline{x}(n)$ remains the unbiased estimator of the distribution mean μ, but the sample variance $S^2(n)$ is a biased estimator of σ^2 ([103] and [4]):

$$E[S^2(n)] = \sigma^2 \left[1 - 2\frac{\sum_{j=1}^{n-1}(1 - j/n)\rho_j}{n-1} \right]. \quad (6.5)$$

Hence, if $y_1, y_2, \ldots, y_n$ are autocorrelated, i.e. $\rho_j > 0$, then $E\left[S^2(n)\right]$ will be smaller than σ^2. However, as shown in Equation 6.5, when $\rho_j \to 0$, $E[S^2(n)] \to \sigma^2$. Thus, we can assume that a covariance-stationary stochastic process is a set of IID random variables if ρ_j is significantly small.

If $y_1, y_2, \ldots, y_n$ is an output stochastic process of jobs completed per hour (JPH) of a simulation run beginning at time zero, it is unlikely to be covariance-stationary. However, $y_{k+1}, y_{k+2}, \ldots, y_n$ could reach a *steady-state distribution* ([103] P488) and can be assumed to be covariance-stationary if k is large enough. The length k is the warm-up period and its estimation will be described in the following Section 6.7.1, using two different methods.

Before we can assume the covariance-stationary output stochastic process y_{k+1}, $y_{k+2}, \ldots, y_n$ is composed of IID observations, we need to estimate the autocorrelations. The calculation of the autocorrelations is discussed in Section 6.7.2. Only if the autocorrelation is small enough can we assume that $y_{k+1}, y_{k+2}, \ldots, y_n$ are a set of IID random variables and perform our analysis of the simulation output data later in the next chapter.

6.7.1 The Influence of the Initial Transient

In order to remove any initialisation bias in the simulation output, we only wish to collect results when it has reached a more stable state. There is an elaborate discussion of initial transient and steady-state distributions in [164], and a list of relevant papers and books can be found in [67]. If the selected warm-up period is too short, the output stochastic process has not reached a steady-state, which can cause misleading data to be presented in the collected output. On the other hand, if we select a very long warm-up period, it is a waste of time and resources.

Therefore, we need to estimate an appropriate warm-up period. Over the last 40 years of research into estimating warm-up periods for discrete-event simulation models, various methods have been proposed. There are five main types of warm-up estimating methods ([132], [133], [134] and [81]):

1. Graphical Methods:

   A visual inspection of time-series of the simulation output. This set of methods can be implemented simply but relies on the expertise of the analyst for a proper decision ([71], [164], [7], [133] and [103]). The simplest and most popular methods are simple Time-series Inspection ([71] and [133]) and Welch's method ([164] and [103]).

2. Heuristic Approaches:

   Rules for determining the length of the stabilising process. These methods have the advantage of easy implemention. Compared to the graphical methods, the use of rules reduces the risk factor of human judgement ([57], [58], [60], [126] and [165]).

3. Statistical Methods:

   Statistical principles are applied. These methods are more complicated and require more specific knowledge ([103] and [172]).

4. Initialisation Bias Tests:

   These tests are strictly speaking tests for determining whether initialisation bias exists in a series of data. Therefore, these methods can be combined with the above methods to verify whether the selected warm-up period is long enough. These tests can lack accuracy for certain kinds of data ([143], [144], [155] and [69]).

5. Hybrid Methods:

A combination of initialisation bias tests with any of the first three methods ([126] and [86]).

A table of 42 warm-up methods found in currently published literature is given in [81].

There is little research evaluating the performance of the various methods. The only few papers we found were: [168], [169], [60], [126], [28], [112], [166] and [113]. Although the advantages and disadvantages of the tested methods are observed and some warm-up estimation methods are recommended for use on some types of simulation models, no single method is found to work well for all types of models. It is suggested that we apply several methods in order to achieve an accurate estimate of the warm-up length.

We applied two widely used methods, simple time-series inspection method and Welch's method, to determine the warm-up period of our simulation models.

6.7.1.1 Simple Time-Series Inspection

Only one replication is required to carry out this graphical method. Thus, we made a replication of the assembly line model of 200 hours. We plot the hourly throughputs of the engine assembly line model for hour $1, 2, \ldots, 200$. The time-series appears to be quite randomly distributed after 48 hours, as shown in Figure 6.5.

6.7.1.2 Welch's Method

This graphical technique requires multiple replications. Welch's method [164] is carried out in the following four steps as given in [103] (P509):

1. Make 15 replications of the simulation of equal length, $l = 200$ hours. Let

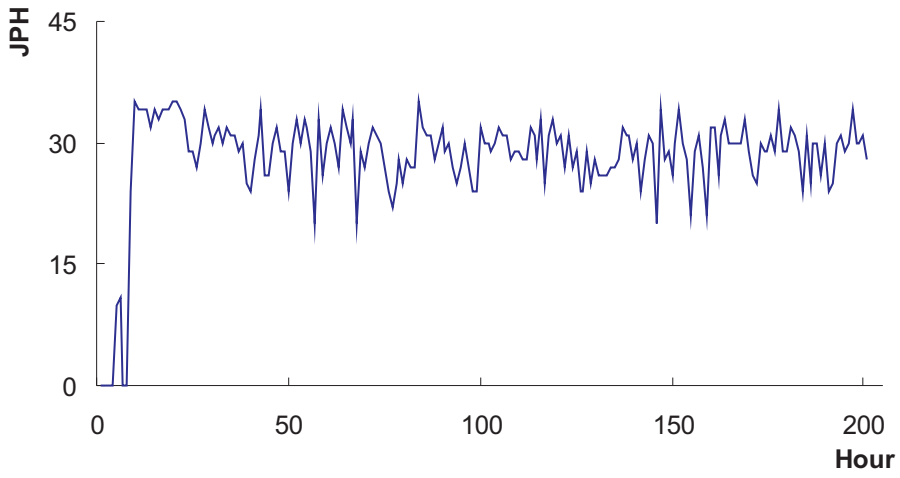

Figure 6.5: Hourly throughputs (Jobs completed per hour), DuntonL01 model.

N_{ij} denote JPH in the jth hour from the ith replication, $i = 1, 2, \ldots, 15$ and $j = 1, 2, \ldots, 200$.

2. Let $\overline{N}_j = \sum_{i=1}^{15} N_{ij}/15$ for $j = 1, 2, \ldots, 200$. The averaged process $\overline{N}_1, \overline{N}_2, \ldots, \overline{N}_{200}$ has the same mean as the original but only $1/15$th of the variance. The plot of the averaged process is shown in Figure 6.6.

3. To highlight the long-run trend of interest, we smooth out the high-frequency oscillations in the averaged process by using the moving average,

$$\overline{N}_r(w) = \begin{cases} \frac{\sum_{s=-(r-1)}^{r-1} \overline{N}_{r+s}}{2r-1}, & \text{if } r = 1, \ldots, w \\ \frac{\sum_{s=-w}^{w} \overline{N}_{r+s}}{2w+1}, & \text{if } r = w+1, \ldots, 200-w \end{cases}$$

where w is termed the *window* and is an integer satisfying $1 \leq w \leq 50$. We calculate the moving averages for $w = 5$ and $w = 10$.

4. Plot $\overline{N}_r(w)$ for $r = 1, 2, \ldots, 200-w$ for both $w = 5$ and $w = 10$ and choose the warm up length k to be that value of r beyond which the plot seems to have converged. The plots are shown in Figures 6.7 and 6.8. We choose a warm-up period of $k = 48$ hours from the smoother plot for $w = 10$.

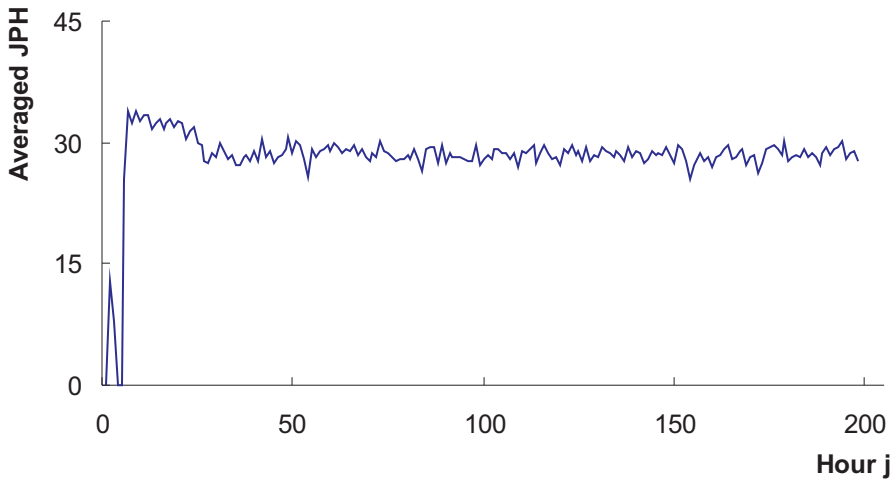

Figure 6.6: Averaged process for hourly throughputs (Jobs completed per hour), DuntonL01 model.

Both methods suggested a warm-up period of 48 hours, i.e. 2880 minutes.

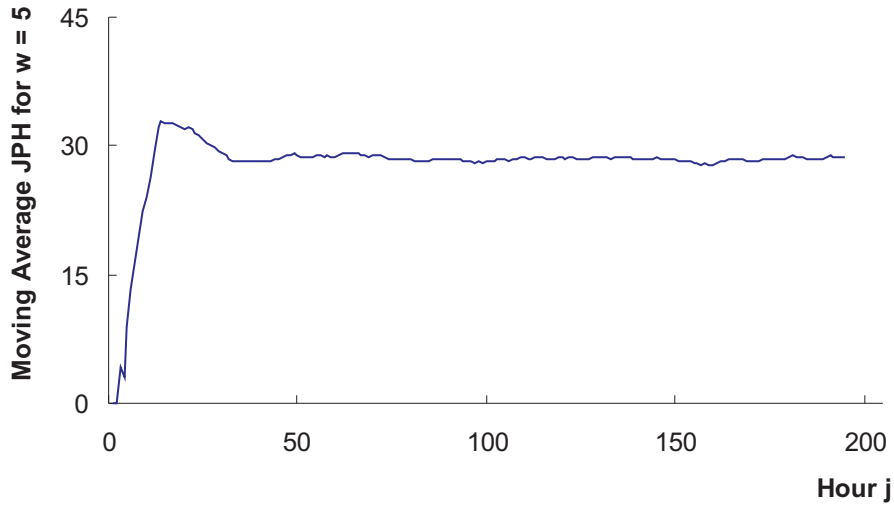

Figure 6.7: Moving averages ($w = 5$) for hourly throughputs (Jobs completed per hour), DuntonL01 model.

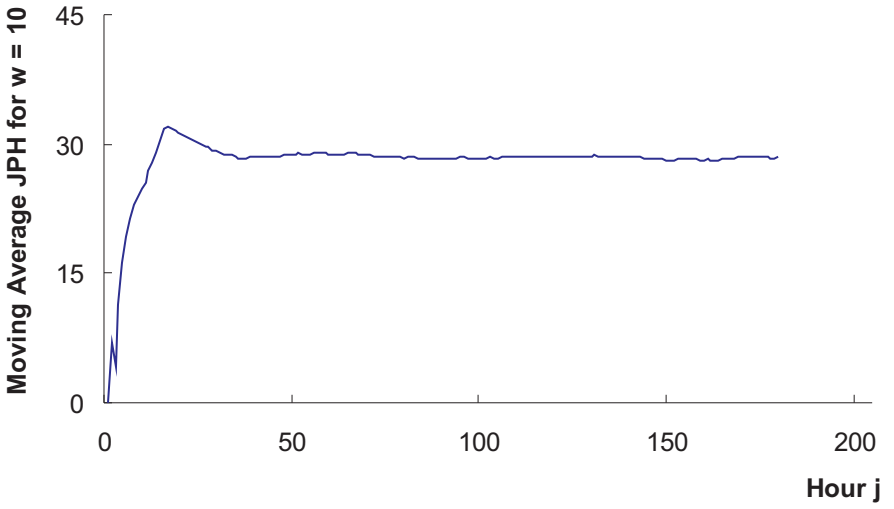

Figure 6.8: Moving averages ($w = 10$) for hourly throughputs (Jobs completed per hour), DuntonL01 model.

6.7.2 Checking for Dependence

We can only apply standard statistical analysis methods to IID data. Therefore, we need to calculate the autocorrelation, ρ_j, of the simulation output data to determine whether the data are independent. Consider random variables $x_1, x_2, \ldots, x_n$ as a covariance-stationary stochastic process. The autocorrelation ρ_j can be estimated by Equation 6.6 ([103] P231):

$$\hat{\rho}_j = \frac{\hat{C}_j}{S^2(n)}, \quad \hat{C}_j = \frac{\sum_{i=1}^{n-j}[X_i - \overline{X}(n)][X_{i+j} - \overline{X}(n)]}{n-j} \; . \tag{6.6}$$

When n is very large, we can use $n-1$ instead of $n-j$ in Equation 6.6 and use the autocorrelation function in Equation 6.7 [18]:

$$\hat{\rho}_j = \frac{\sum_{i=1}^{n-j}[X_i - \overline{X}(n)][X_{i+j} - \overline{X}(n)]}{\sum_{i=1}^{n}[X_i - \overline{X}(n)]^2} \; . \tag{6.7}$$

We make one replication of the engine assembly simulation model (the LION model) of length $m = 259,200$ minutes (180 working days, excluding the warm-up period) and collect the averaged JPH of every 5 days as one observation of the output, which gives 36 observations. We then calculate ρ_j for all possible lags of the output stochastic process of theses 36 observations, $X_1, X_2, \ldots, X_{36}$. The plot of the autocorrelation function generated by Minitab 15, is given in Figure 6.9. Approximate 0.05 critical bands for the hypothesis that the correlations are equal to zero are included on the plot. As shown in this figure, the autocorrelations for all lags $1, 2, \ldots, 35$ of the simulation output are small and can be considered as zero according to the 5% significance limits.

According to the plot of the autocorrelations, there appear to be no significant inter correlations within the output of the engine assemble line model. Therefore, as the simulation output of 36 JPH observations is obtained when the simulation model has reached a steady-state, it can be assumed to be a set of IID random

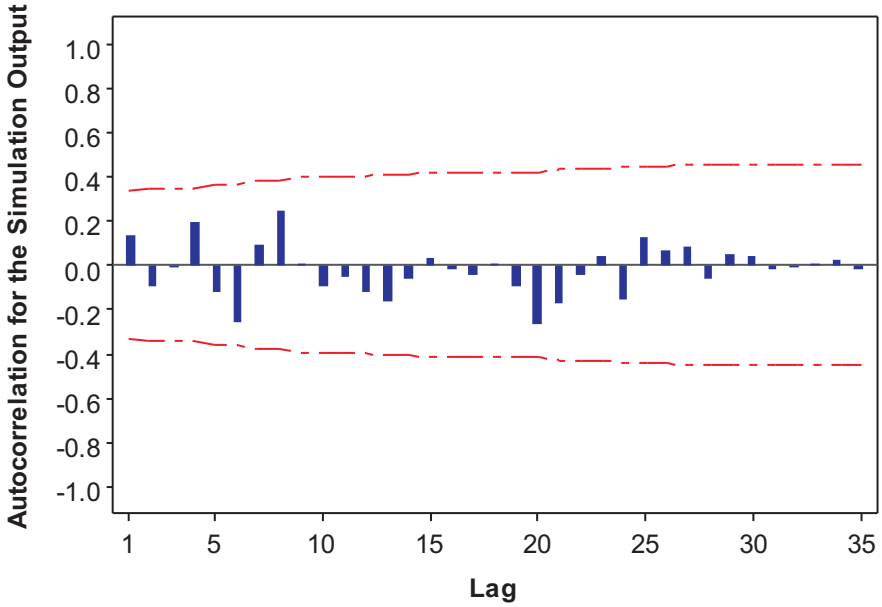

Figure 6.9: Autocorrelation of all possible lags within the JPH output of the simulation run. Red curve indicates the 5% significance limits for the autocorrelations.

variables. Thus, we may carry on the simulation evaluation assuming that, which is given in Chapter 7.

Chapter 7

Simulation Evaluation

In this chapter, we evaluate our method for modelling breakdown durations using a simulation model of an engine assembly line. We compare three different representations of breakdown duration inputs:

(i) empirical distributions for individual machines;

(ii) fitted mixture distributions for individual machines;

(iii) fitted mixture distributions for groups of machines derived using the Arrows classification method.

The methodology for evaluating the inputs compares the simulation outputs of the simulation models with the different breakdown duration inputs using several different methods: graphical comparison, paired-T test and bootstrapping.

The results of the evaluation process suggest that the throughput of the simulation model is not particularly sensitive to the machine breakdown durations, which is confirmed by further investigation of the causes of the total production loss. The engine repairs and operator stoppages seem to be responsible for a larger portion of the line loss and their impact on the simulation model overpowers the effect of the machine breakdowns and effectively masks any differences in output resulting

from different breakdown duration inputs. However, although in this case the simulation output shows only a light dependence on the breakdown duration inputs, the methodology for the evaluation that we introduce in this chapter is still of interest and could be applied in other simulations to evaluate the effect of different inputs on simulation output statistics.

We also further investigate the impact of the choice of p-value threshold for the Arrows grouping by evaluating the simulation outputs of the models with different collections of fitted mixture distributions of different sets of groups obtained using the Arrows classification method with different p-value thresholds.

We describes the three types of input in Section 7.1. The methodology of assessing the three different representations of breakdown durations are described in Section 7.2, including an investigation of the sources causing the line production loss. In Section 7.3 we investigate the impact of the threshold for grouping the machines on the simulation performance. A discussion of the results of the study and a conclusion is given in Section 7.4.

7.1 Breakdown Input for Simulation Model

The simulation model we use describes the DuntonL01 engine assembly line, one of Ford's lines used for the assembly of engines, which is made up of over 200 machines, but for the modelling of breakdown durations, we consider only 39 of these. Among the other machines, some are small pieces of equipment and thus are not linked to the on-line monitoring system, therefore breakdown duration data for these machines are not available. For these small machines, the reliability data including the frequency of failures and the average breakdown duration provided by the machine manufacturers are used to model their breakdown behaviours within the simulation model. The rest of the machines such as buffers and conveyors,

are thought to be very reliable machines, and Ford make the assumption that they rarely break down.

We use WITNESS (Lanner Group 2008) [102] to simulate the assembly plant using Busy Time breakdown mode, assuming that machines can only break down when they are working. An exponential distribution is used to simulate the time between failures, to parameterise which we calculate the mean time between failures (MTBF) for a machine using the method described in Section 6.6.

Meanwhile, the modelling of the other two factors that cause production loss is described in Section 6.3. Engine repairs are simply modelled using the percentage of engines with quality issues. The modelling of operator stoppages is also included in the simulation model, where generally an Erlang distribution is used to represent the time of operator stoppages, and an extremely low percentage is used to model the frequency of occurrences.

The three different methods for generating breakdown durations we compare initially are: (1) sampling from historical data, i.e. using empirical distribution functions (EDF); (2) sampling from the fitted mixture distributions (FMD) for individual machines; (3) sampling from the fitted mixture distributions for the groups of machines obtained using the Arrows classification method with a specified threshold (we here use $p_0 = 0.10$); the similarity matrix for the 39 machines being modelled and the grouping results are given in Appendix B.

7.2 Output Evaluation

We set the warm-up period to be 2880 minutes as discussed in Section 6.7.1 and make 10 independent runs, where the length of each run is 36 weeks, for each of the three different models. We make 36 observations in each run, each observation being the averaged number of jobs shipped per hour (JPH) in each of the 36 weeks.

Thus, we obtain 360 averaged JPH observations for each model.

We can compare these JPH outputs with the real line yield, which includes a set of 36 weekly averaged JPH observations. This set of real line yield observations has a mean of 28.306 with a 95% Confidence Interval (CI) of $(27.828, 28.783)$ and a median of 28.000 with a 95% CI of $(28.000, 29.000)$, though the actual values of the observations are not provided for reasons of confidentiality. All of the 95% CIs in this chapter are calculated using the standard formula by assuming normality within the data.

7.2.1 Graphic Comparison

We first use a graphical method to compare the outputs visually and statistically. The boxplot and 95% confidence interval plot of the three sets of JPH outputs for the engine assembly line simulation model using the three different methods for sampling breakdown durations, together with the real JPH data, are given in Figures 7.1 and 7.2 respectively.

As we can see from the two plots, the inter-quartile ranges and 95% confidence intervals of the three JPH data sets overlap, showing a high degree of similarity between the outputs. The medians of the three sets of JPH are 28.608, 28.604 and 28.608, which are all within the 95% CI for the median estimated from the real JPH data set. Moreover, 95% confidence intervals for the means of the three sets of JPH output: $(28.578, 28.636)$, $(28.581, 28.638)$ and $(28.578, 28.635)$ all fall within the 95% CI for the mean in the real JPH data set. It is also noticeable that the spread of the real JPH data is much wider than the simulated JPH data, indicating that the observations obtained from the simulations are less variable than the real data. The reason for this, suggested by Ford, is that there are other sources of variability in real world that are not modelled (or rather are too complicated to be modelled) in the simulation model. For example, in real world, there are situations

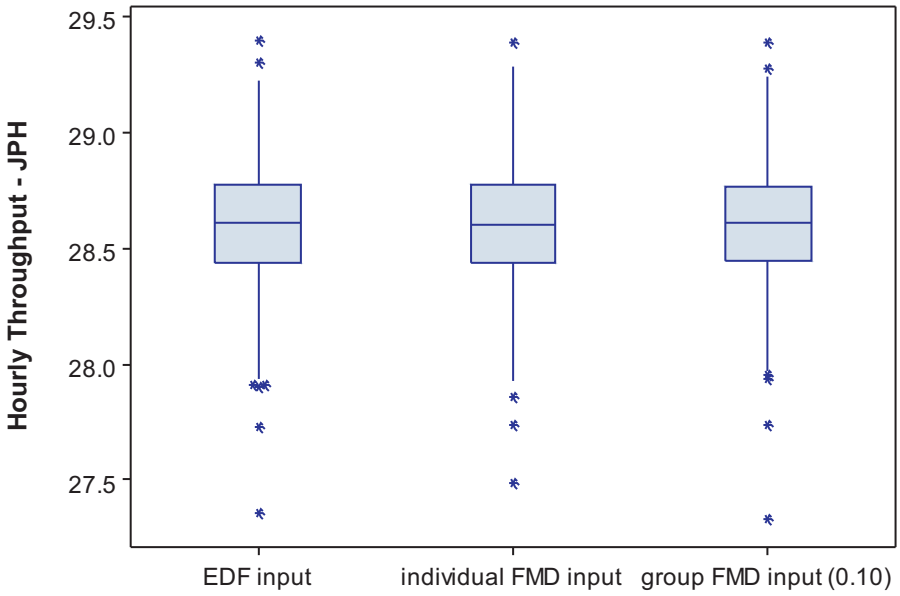

Figure 7.1: Boxplot of simulation output JPH using the three methods for sampling breakdown durations. The central line shows the median and the box spans the inter-quartile range.

where operators have team meetings during shifts, or have early lunches or late start or are absent; which would give lower averaged JPH. It is also possible for the operators to accumulate overtime work to give the next week a higher averaged JPH.

7.2.2 Paired T-Test

We use a paired t-test for testing the mean difference between paired observations of the JPH outputs of simulation models using the different breakdown duration input methods. The null hypothesis is

$$H_0 : \mu_d = \mu_0,$$

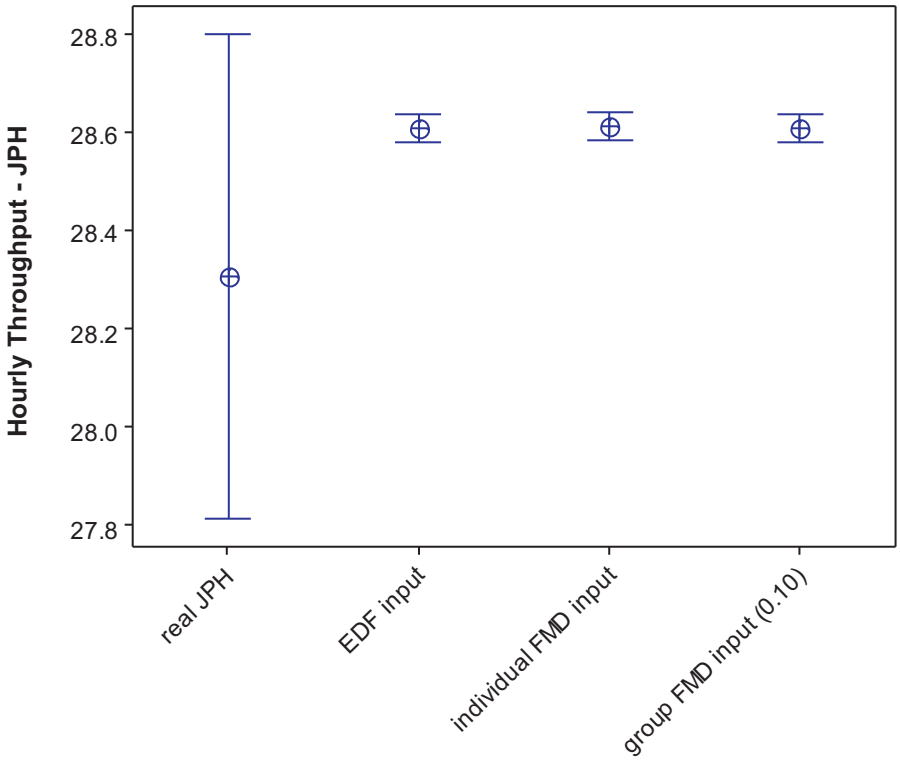

Figure 7.2: Interval plot of the set of real JPH observations and simulation output JPH using the three methods for sampling breakdown durations. The central circle shows the mean and the interval describes the 95% confidence interval for the mean.

where μ_d is the population mean of the differences and μ_0 is the hypothesized mean of the differences. Since this test is comparing the difference between paired observations of the outputs, it is applied to evaluate the simulation performance at approximately the same time while using three different breakdown duration inputs.

The results of the paired t-tests are given in Table 7.1. The confidence intervals for the mean difference between any two output process of the model using any two breakdown duration inputs all include zero, which suggests there is no obvious difference between any two of the simulation outputs. The high p-values further suggest that the data are consistent with $H_0 : \mu_d = \mu_0 = 0$, that is, any two outputs

do perform equally.

Paired T-Test	95% **CI for Mean Differences**	**T-Value**	**P-Value**
empirical data input vs. individual FMD input	$(-0.00868, 0.00497)$	-0.53	0.594
empirical data input vs. group FMD input(0.10)	$(-0.00451, 0.00618)$	0.31	0.759
individual FMD input vs. group FMD input(0.10)	$(-0.00405, 0.00942)$	0.78	0.433

Table 7.1: The results of the paired t-tests between the outputs of models using the three breakdown duration inputs.

7.2.3 Bootstrapping Analysis

We have investigated the differences between the medians and means of the JPH outputs and the differences between paired JPH observations using the graphical method and the statistical test. We here wish to study the distributional properties of the simulation JPH outputs, i.e. to examine the similarities between the underlying distributions of the JPH outputs of the models using the three different breakdown duration inputs, where the similarities are measured by the possibilities that any two sets of the JPH observations have been drawn from the same distribution. The larger the possibility, the more similar the two sets of JPH outputs and thus the more similar the two breakdown duration inputs. We use the method described in Chapter 4 to calculate the p-value similarity between the distributions of the JPH outputs of simulation models using the three different breakdown duration inputs.

The resultant p-values are given in Table 7.2. As shown in this table, the p-values are all quite high, which indicates that the distributions of the JPH outputs of the three simulation models using different breakdown inputs are all very similar

to each other and thus suggests the three representations of the breakdown durations as simulation input have a similar effect on the whole system's production performance.

Bootstrapping Process of Comparison	P-Value
empirical data input vs. individual FMD input	0.371
empirical data input vs. group FMD input(0.10)	0.536
individual FMD input vs. group FMD input(0.10)	0.736

Table 7.2: The p-values obtained from the bootstrapping process of comparison between the outputs of models using the three breakdown duration inputs.

7.2.4 Further Investigation

As discussed in the previous three sections, the evaluation results all suggest that the JPH outputs of models using the three breakdown duration inputs are very similar. The outputs are so close that it appears that the machine breakdowns may have only a small impact on the throughput. We therefore check this inference by comparing the output of the model using the group FMD input with two other possible input distributions: (1) using one lognormal distribution and (2) using one FMD for the whole data set of all machines. The differences between these three breakdown duration inputs are statistically significant, and so we would expect there to be significant differences in the outputs. The boxplot and 95% confidence interval plot of the three sets of JPH outputs are given in Figure 7.3, and suggest that the JPH again appears to be insensitive to the changes made to the machine breakdown duration inputs, which confirms our inference.

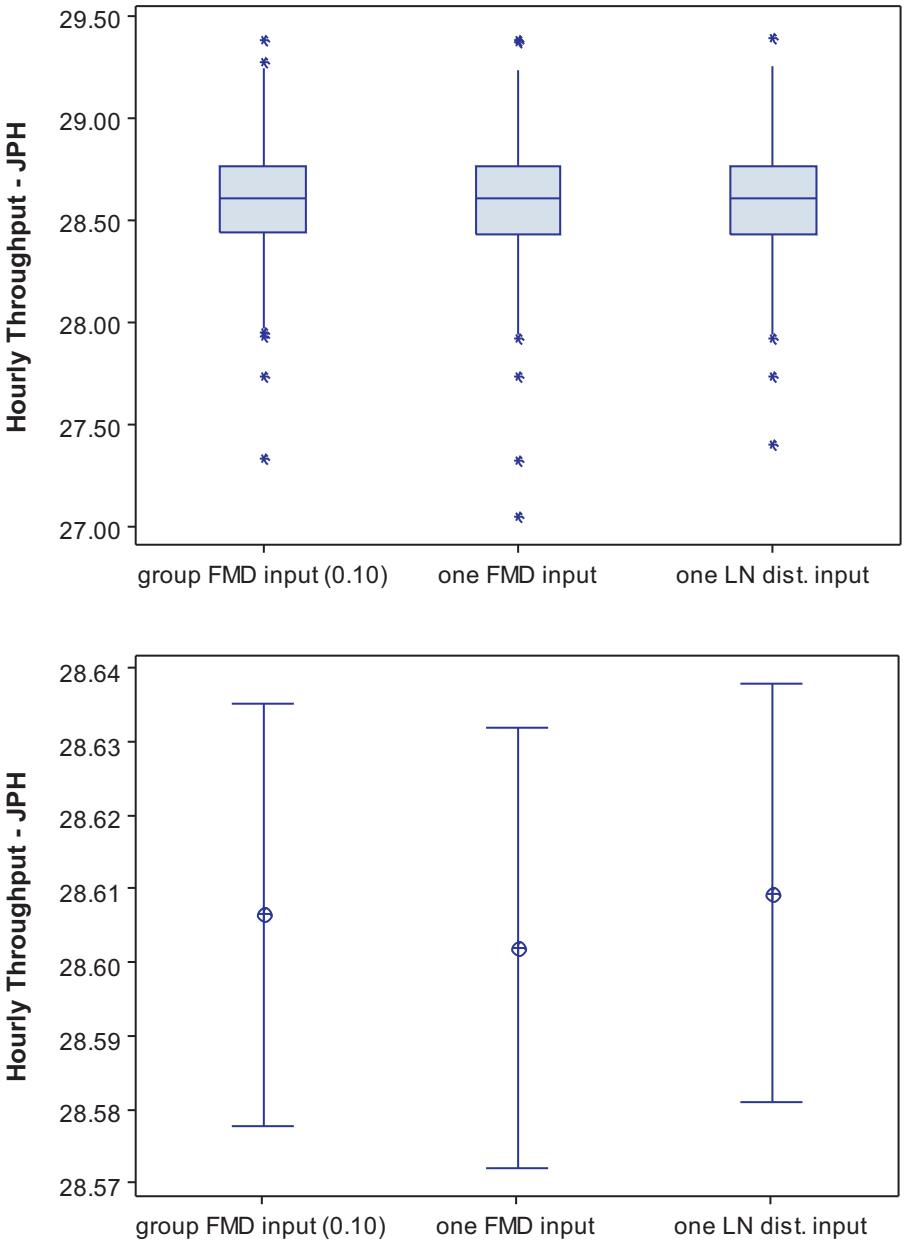

Figure 7.3: Boxplot and Interval plot of simulation output JPH using three methods for sampling breakdown durations: group FMD ($p_0 = 0.10$), one FMD for all 39 machines and one lognormal distribution for all 39 machines. The central line shows the median and the box spans the inter-quartile range. The central circle shows the mean and the interval describes the 95% confidence interval for the mean.

Therefore we investigate the causes of production loss further. Since there are three main causes of production loss: the machine breakdowns, engine repairs and operator stoppages, we shut down the engine repairs and operator breakdowns, and run the simulation models with only the factor of machine breakdowns on to gain a better and clearer picture of the solo impact of the machine breakdowns on the system throughput.

The boxplot and interval plot of the JPH outputs of models with the factors of engine repairs and human breakdowns taken out and using four methods for describing the machine breakdown durations: historical data, individual FMD, group FMD with a threshold of 0.10 in the grouping process and one FMD for all machines, are shown in Figures 7.4 and 7.5. It is seen that the machine repairs are only responsible for a small portion of the loss, as the JPH outputs are much higher than the outputs when all of the three factors: machine breakdowns, engine repairs and operator stoppages, are included in the simulation model. Thus, it seems that the engine repairs and operator stoppages are responsible for a larger portion of the production loss and when all three factors are functioning, their impact on the simulation model overpowers the effect of the machine breakdowns and effectively masks any differences in output resulting from different breakdown duration inputs.

Although the simulation model with the engine repairs and operator stoppages turned off is not a complete model, the outputs show the true impact of the machine breakdowns on the line throughput, without the interaction with other factors that are also affecting the total loss in real world. From Figures 7.4 and 7.5, it is seen that the inter-quartile ranges and 95% confidence intervals of the four outputs all overlap, which suggests that there are similarities between the four breakdown duration inputs. Another interesting observation to be made is that as we move to more general models, i.e. from individually fitted models, to fitted models for groups of machines, to a model for all of the machines, the 95% confidence interval

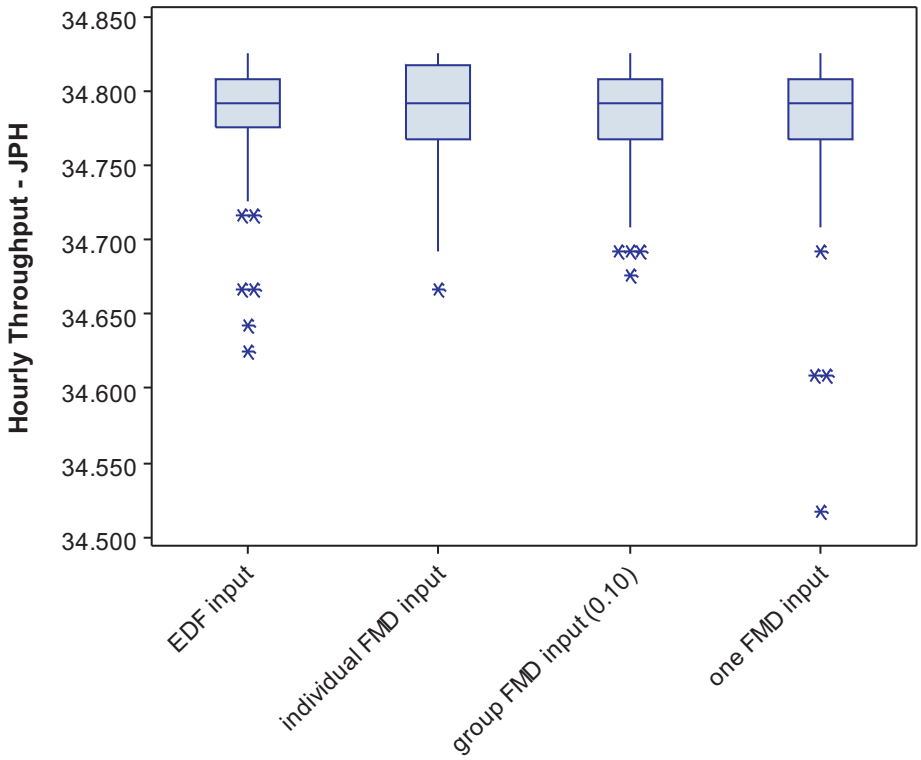

Figure 7.4: Boxplot of simulation output JPH using four different methods for sampling breakdown durations: EDF, individual FMD, group FMD ($p_0 = 0.10$) and one FMD for all 39 machines; while the engine repairs and operator stoppages are set to be turned off. The central line shows the median and the box spans the inter-quartile range.

for the output increases.

We focus on the models using the first three methods for representing the machine breakdown durations: historical data, individual FMD, group FMD with a threshold of 0.10. It can also be seen in the interval plot given in Figure 7.5 that using empirical distributions results in a slightly lower JPH than the output using FMD inputs. We use the breakdown duration data of machine ML06 as an example to study a possible reason of these differences. Figure 7.6 shows the histogram of the breakdown duration data for ML06, the fitted mixture model for machine ML06 only and the fitted mixture model for G03, the group of machines including ML06 (see Appendix B for more details). It can be seen in the histogram that there

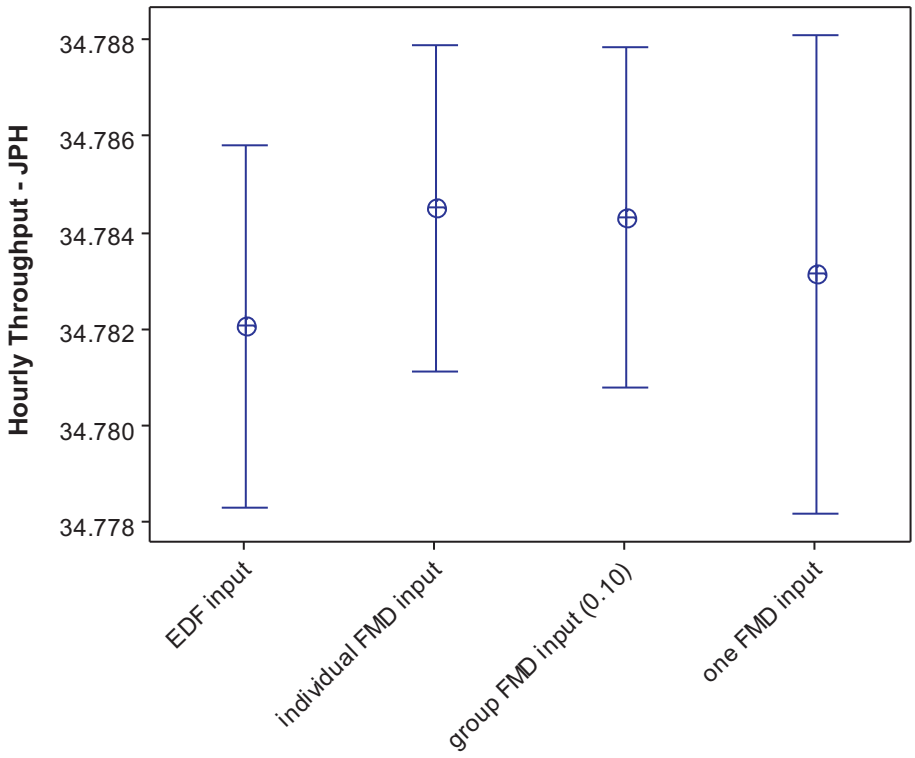

Figure 7.5: Interval plot of simulation output JPH using four different methods for sampling breakdown durations: EDF, individual FMD, group FMD ($p_0 = 0.10$) and one FMD for all 39 machines; while the engine repairs and operator stoppages are set to be turned off. The central circle shows the mean and the interval describes the 95% confidence interval for the mean.

is one extreme outlier for which the breakdown duration is around 133 minutes (i.e. near 11.5 in the X-axis, as the data shown in the plot is the transformed data of the real breakdown durations), resulting in the whole assembly line being down for a relatively long period. The fitted mixture model for ML06 and the fitted mixture model for G03 are both much smoother than the empirical distribution for ML06, and by using a continuous curve are unlikely to sample durations of 133 minutes or greater as often as when using the empirical distributions. Hence, the JPH with the EDF inputs could be lower than that with the mixture distribution inputs.

Since the cycle time of the assembly is 103 seconds, if a repair for any machine needs a long time to be fixed all machines need to stop after a while; therefore long

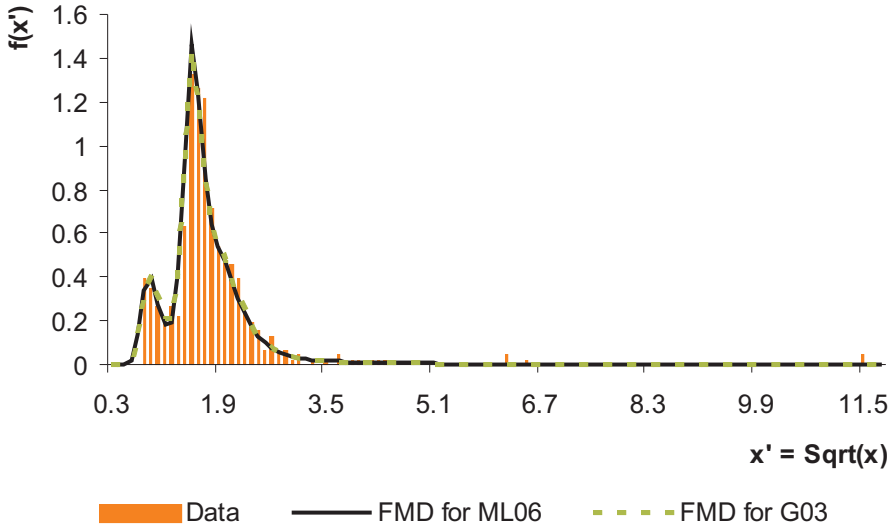

Figure 7.6: Histogram of the transformed breakdown duration data of machine ML06 and plots of its fitted mixture distribution's PDF and its group fitted mixture distribution's PDF.

repair durations have a greater effect on the line production. As the high value outliers of breakdown durations have a significant impact on the resultant JPH output, we calculated the frequency of generating long breakdown durations (greater than 50 minutes) in the WITNESS models using the three different representations of breakdown durations for machine ML06. The results are given in Table 7.3. The frequency of long breakdowns is the highest when using the empirical distribution as the breakdown duration input. Moreover, the three models are using the same distribution to simulate time between two successive failures, so the fact that when running for the same amount of time, the model using the empirical distribution as its input has the lowest efficiency is quite reasonable.

TTR Input	$P(TTR > 50mins)$
Empirical distribution	0.002632
FMD for ML06	0.000609
FMD for G03	0.000611

Table 7.3: Frequency of generating long breakdown durations (greater than 50 minutes) for machine ML06 using the three different distributions. TTR is short for time to repair.

7.3 Impact of the Threshold

The grouping results of the Arrows classification method vary for different thresholds, and so we here study the influence of the choice of threshold on the output of simulation models using fitted mixture distributions for different groups. We consider the following thresholds: 0.05, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90. Most of the corresponding grouping results are different with two exceptions: the groupings with $p_0 = 0.50$ match those with $p_0 = 0.60$; and the groupings with $p_0 = 0.70$ match those with $p_0 = 0.80$; therefore, we have 8 different sets of fitted mixture distributions for 8 different sets of groups. We use these 8 sets of group fitted mixture distributions as the breakdown duration inputs of the same engine assembly line simulation model and make 10 independent runs of 36 weeks for the models to get 8 sets of JPH observations.

The boxplot and interval plot of the sets of JPH output for the engine assembly line simulation model using the individual FMD breakdown input together with the groups FMD breakdown input at different threshold levels are given in Figures 7.7 and 7.8 respectively. As shown in both plots, there are no significant differences between the JPH outputs of models using FMD for groups that are obtained at different threshold levels.

The similarities can be further confirmed by the paired t-test results and bootstrapping analysis, as described in Sections 7.2.2 and 7.2.3, which are shown in

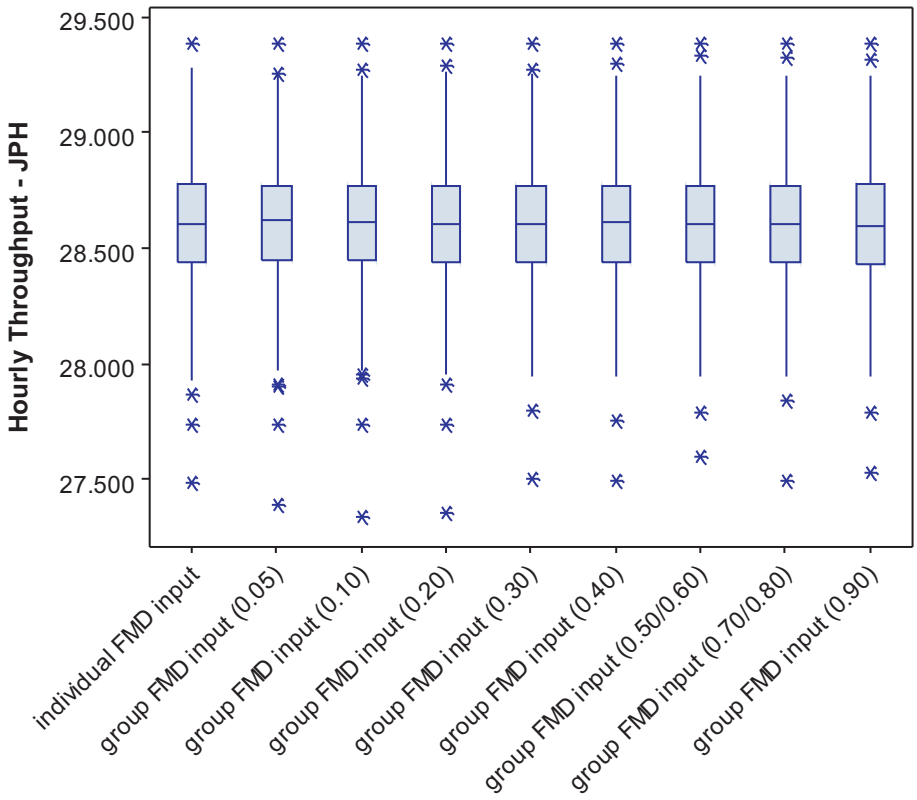

Figure 7.7: Boxplot of simulation output JPH using the FMD for individual machines together with the FMD for groups classified at different threshold levels using the Arrows method for sampling breakdown durations. The central line shows the median and the box spans the inter-quartile range.

Tables 7.4 and 7.5, respectively. All of the confidence intervals in Table 7.4 include zero, and the p-values are all quite high; both of which suggest that there is no apparent difference between any pair of the simulation outputs and thus all of the 8 simulation outputs perform equally. All of the p-values in Table 7.5 are all quite high, which indicates that the distributions of the JPH outputs of the 8 simulation models are all very similar to each other and thus consistently suggests the 8 representations of the breakdown duration inputs have a similar effect on the system production performance. Therefore, it is believed that the choice of threshold in finding the groups of machines does not have a significant impact on the simulation performance when using group FMD as breakdown duration inputs. This

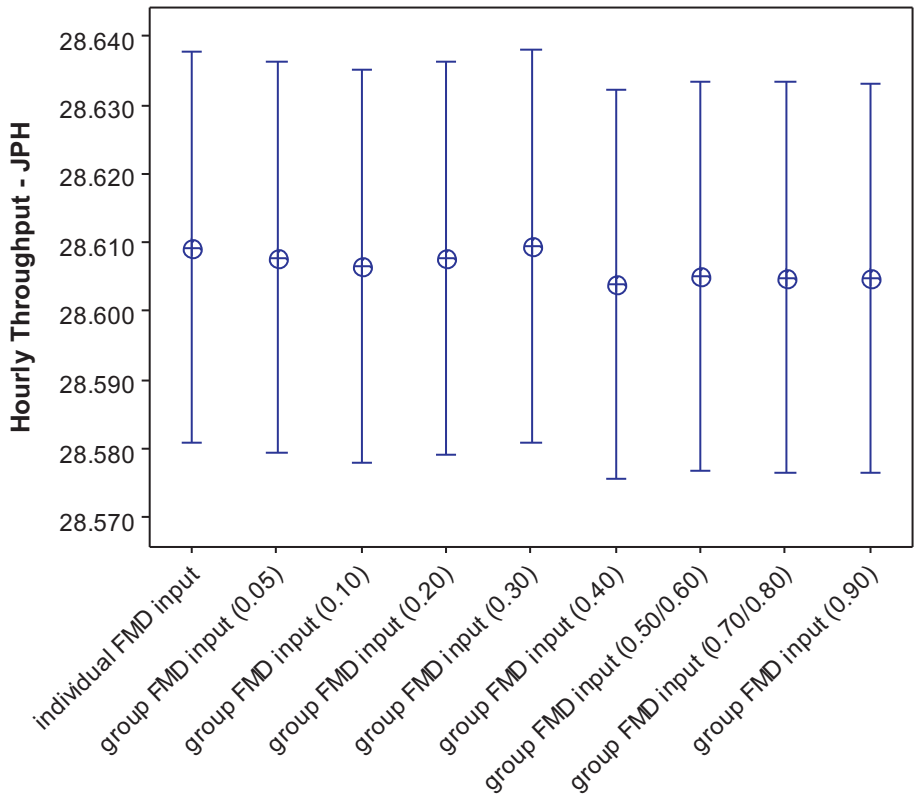

Figure 7.8: Interval plot of simulation output JPH using the FMD for individual machines together with the FMD for groups classified at different threshold levels using the Arrows method for sampling breakdown durations. The central circle shows the mean and the interval describes the 95% confidence interval for the mean.

suggests that $p_0 = 0.05$ may be chosen as it is the smallest value of the thresholds and thus provides the smallest number of groups; and hence decreases the time spent estimating the fitted distributions for all machines and also reduces the time spent inputting the breakdown settings.

We next investigate the impact of the threshold using the simulation models with the engine repairs and operator stoppages turned off. The plots of outputs are shown in Figures 7.9 and 7.10, and these also suggest that there is little difference in the outputs for the different thresholds.

Paired T-Test individual FMD input vs.	95% **CI for Mean Differences**	**T-Value**	**P-Value**
group FMD input(0.10)	$(-0.00405, 0.00942)$	0.78	0.433
group FMD input(0.05)	$(-0.00520, 0.00802)$	0.42	0.675
group FMD input(0.20)	$(-0.00509, 0.00819)$	0.46	0.646
group FMD input(0.30)	$(-0.00650, 0.00603)$	-0.07	0.941
group FMD input(0.40)	$(-0.00170, 0.01223)$	1.49	0.138
group FMD input $(0.50/0.60)$	$(-0.00295, 0.01125)$	1.15	0.251
group FMD input $(0.70/0.80)$	$(-0.00265, 0.01151)$	1.23	0.219
group FMD input(0.90)	$(-0.00246, 0.01124)$	1.26	0.209

Table 7.4: The results of the paired t-tests comparing the simulation output of the model using individual FMD and those of models using FMD for different groups of machines resulting from the Arrows method using different thresholds.

individual FMD input vs.	**P-Value**
group FMD input(0.10)	0.736
group FMD input(0.05)	0.695
group FMD input(0.20)	0.637
group FMD input(0.30)	0.691
group FMD input(0.40)	0.896
group FMD input$(0.50/0.60)$	0.779
group FMD input$(0.70/0.80)$	0.603
group FMD input(0.90)	0.803

Table 7.5: The p-value results obtained from the bootstrapping process comparing the simulation output of the model using the individual FMD and those of models using FMD for different groups of machines resulting from the Arrows method using different thresholds.

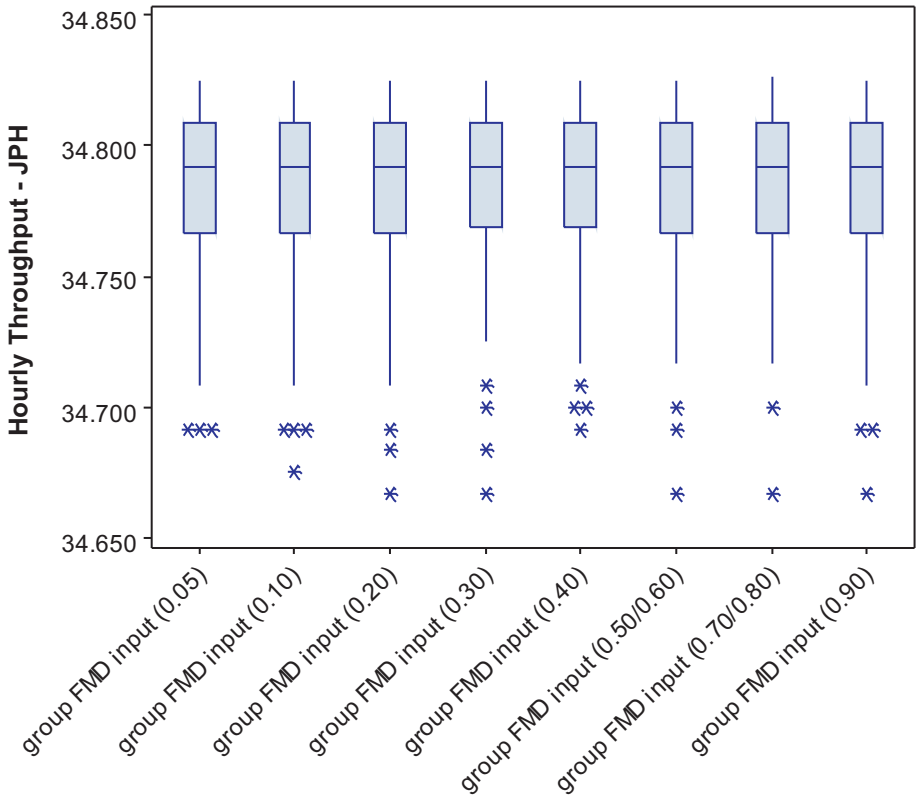

Figure 7.9: Boxplot of simulation output JPH using the FMD for groups classified at different threshold levels using the Arrows method for sampling breakdown durations in the model with the engine repairs and operator stoppages turned off. The central line shows the median and the box spans the inter-quartile range.

7.4 Discussion

The first observation to be made is that the machine breakdowns have only a small impact on the JPH, and the engine repairs and operator stoppages are responsible for a much greater portion of the total loss than the machine breakdowns. Therefore, it is reasonable that the JPH outputs of the model using the three different machine breakdown inputs appear to be similar, which may indicate why this topic has not been discussed much before. The evaluation process was carried out to investigate the influence of the machine breakdown inputs on the simulation throughput and the fact that the outputs are so similar even when the engine repairs

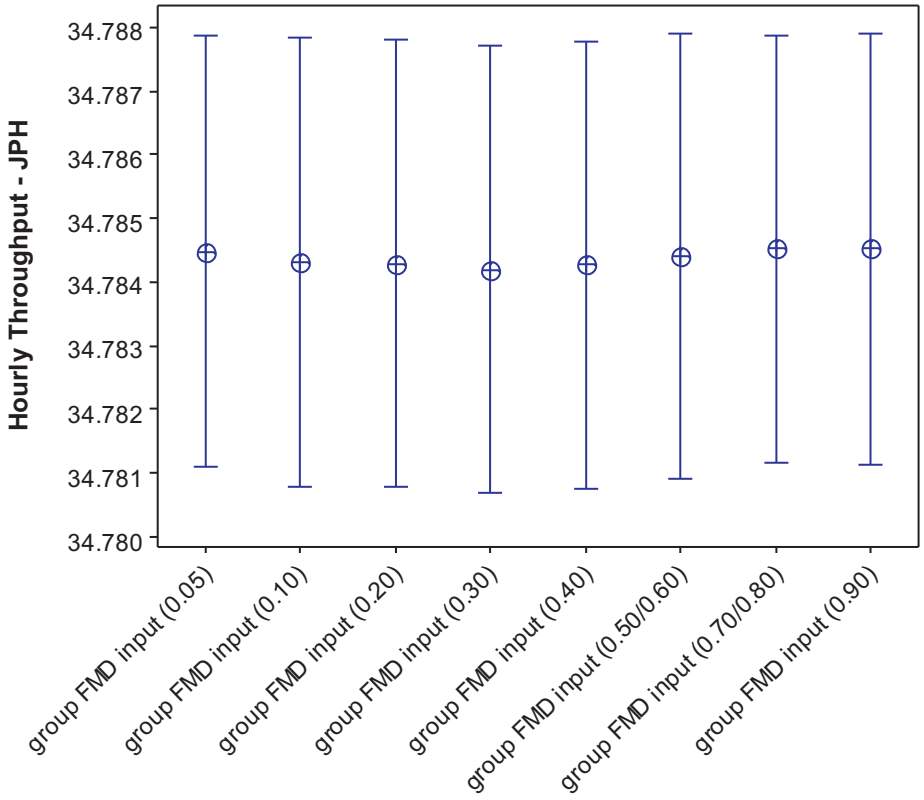

Figure 7.10: Interval plot of simulation output JPH using the FMD for groups classified at different threshold levels using the Arrows method for sampling breakdown durations in the model with the engine repairs and operator stoppages turned off. The central circle shows the mean and the interval describes the 95% confidence interval for the mean.

and operator stoppages are turned off, is encouraging.

The results of the comparison of the simulation outputs of the simulation models that have only the machine breakdowns functioning, i.e. with the engine repairs and operator stoppages turned off, show that when FMD inputs are used, the JPH output by the model is higher than the output when empirical distribution inputs are used. It is believed that the reason for this is that the possibility of getting extremely high breakdown durations in the WITNESS models using historical data is greater than that of the model using individual FMD or group FMD inputs; and the very long breakdown durations have a significant impact on the JPH of the line.

Nevertheless, as suggested by the analysis, the simulation outputs using the three different breakdown duration representations are all quite similar.

The use of mixture distributions for representing simulation inputs has advantages over using EDF inputs. Since the EDF is estimated from a random sample, it may contain irregularities and have a limitation that data generated from it can only be within a certain range. A mixture distribution is a continuous distribution that copes well with the multimodality present within the data, thus can smooth out the irregularities in the data. It is a compact way to represent the duration data and also makes it simpler to make changes for experimental reasons.

On the whole, the similar simulation performance using FMD and group FMD strongly suggests that the classification of machines based on their breakdown duration data is good enough for this purpose. Moreover, there are a number of advantages of using grouped FMD instead of individual FMD. First, less fitting processes need to be carried out; and the number of data sets and variables in the simulation can be reduced by the grouping and thus the subsequent input time required for all machines can be decreased. The total saving of time is significant, even when taking into account the time spent implementing the Arrows method. Second, in the situations that a machine needs to be modelled while there is no available data for it or it is a new machine, an experienced engineer could probably help with identifying which groups of machines the no-data/new machine belongs to and so the FMD for that group could be used to represent the breakdown duration input of this new machine. The accuracy of identifying the machine as being similar to a group of machines should be higher that that of identifying a similarity with one particular machine whose breakdown duration data are available.

While the different thresholds suggest different groupings, it appears that the simulation outputs of models using different collections of group fitted mixture distributions for different sets of groups are not significantly different, with or without the factors of engine repairs and operator stoppages. The paired t-tests

and bootstrapping analysis that compares the simulation output of the model using individual machine FMD and that of the models using FMD for groups obtained using different thresholds confirm this. Thus, it is concluded that the choice of threshold for the Arrows grouping process does not have a significant influence on the simulation throughput of models using corresponding group fitted distributions in this example. Therefore, we could use a relatively low threshold for the classification analysis to gain a greater saving on the time for the fitting and inputting processes.

Chapter 8

Conclusions and Future Research

Simulation modelling is used widely within manufacturing industry to evaluate new designs for production lines and to improve the efficiency of existing lines. As an important source of variability in many manufacturing systems [103], machine breakdowns need to be modelled correctly in manufacturing simulation models. Our work has focused on an existing engine assembly line within a Ford manufacturing plant, where over two hundred different machines are involved in the assembly process. Although many authors have considered the machine failure rates occurring on a production line ([64], [128], [99], [163], [171] and [68]), we have found little work in the literature on modelling the duration of machine breakdowns ([97] and [103]). A review of the literature on machine breakdown modelling in manufacturing simulation models has been given in Chapter 2.

In this thesis we have described a modelling process to represent machine breakdown durations in engine assembly line simulation models. We use finite mixture distributions to model machine breakdown durations, allowing us to describe the multimodality present within the data. Since the simulation models generally contain a large number of machines and can be very complex, we have derived the Arrows classification method to group machines with similar distributions of breakdown durations, where the Two-Sample Cramér-von Mises statistic

is used to measure the similarity of two sets of breakdown duration data for two machines, with bootstrapping being used to assess the significance of the similarity. The grouping is such that two machines with statistically significantly different breakdown duration data cannot be placed in the same group. Finite mixture distributions are fitted to the grouped breakdown duration data sets and one fitted mixture distribution for the group is used in the simulation model to model the breakdown durations for all of the machines in the group.

We have implemented the breakdown duration modelling methodology with the simulation model of the engine assembly line and have evaluated the classification and mixture distributions fitting procedure by comparing the throughput of the simulation model when running with three different machine breakdown duration inputs: mixture models fitted to individual machine breakdown durations; mixture models fitted to group breakdown durations; and historical data. Three different methods have been used for the outputs comparison and the results suggest that the modelling methodology successfully produced an appropriate representation of machine breakdown duration inputs for the simulation model.

8.1 Finite Mixture Models

Finite mixture models are multimodal and have been found to be an appropriate statistical model of the breakdown durations of machines in engine assembly lines. Their use has advantages over the historical data and common theoretical distributions for modelling the breakdown durations. Historical data may contain irregularities and have strict upper and lower boundaries. Commonly used theoretical distributions may be worse representations of breakdown durations as most breakdown duration data sets are not unimodal, while common theoretical distributions are. In comparison, finite mixture distributions are particularly appropriate as they can cope with the multimodality present in most of the breakdown duration data

sets and adequately fitted mixture distributions can smooth out the irregularities within the historical data, represent extreme events and make it simpler to make changes for experimental reasons. From a practical point of view, fitted mixture distributions are found suitable as well as they contain parameters with intuitive meanings and can be input into simulation models in a simple way.

Since the original data has a wide range of values, we found that using a data transformation could help in getting more accurately fitted mixture distributions by reducing the range of the data so that the fitting process coped better. By taking the square root of the original data, the range of the transformed data shrinks and all of the transformed data stay positive. We found that the lognormal mixture distribution was most robust for representing the machine breakdown duration data sets.

8.2 Method for Estimating Similarity

A new method for estimating the similarities between machines based on the breakdown duration data sets was described in Chapter 4. The method uses the Two-Sample Cramér-von Mises goodness of fit to compute a statistic, T, of two data sets by testing the null hypothesis that the two samples are drawn from the same distribution, and then applies bootstrap resampling to estimate the significance level of the statistic by determining the distribution of T, $\Phi(T)$. The Cramér-von Mises goodness of fit statistic was used as it has advantages when dealing with the machine breakdown duration data sets, compared with other goodness of fit statistics. For other goodness of fit statistics, such as the χ^2 statistic and the Anderson-Darling statistic, information about the underlying distribution of the data is required before constructing the goodness of fit tests [151]. In comparison, computing the Cramér-von Mises statistic is relatively straightforward, as it is distribution-free and therefore there is no need to make any assumptions about the

distributions of the data sets being analysed [5]. In addition, the Cramér-von Mises goodness of fit statistic copes well with the fact that the data sets contain very uneven numbers of data points. The tabulated criterion values for the Cramér-von Mises test are not very extensive and do not cover the samples that we are dealing with and so we use bootstrap resampling to produce the distribution of Cramér-von Mises goodness of fit statistics. The similarity of the two samples of breakdown duration data is then measured by the significance level, i.e. the p-value, of T, which is obtained by simply comparing T with $\Phi(T)$.

We tested the new method on samples drawn from (a) identical distributions; (b) distributions with the same variance but different mean; (c) distributions with equal means but different variances; and (d) different types of distributions. The method is especially successful when applied to cases where the samples are clearly distinct or are identical to each other, where extremely small or high p-values were obtained as expected. It is more difficult to calculate p-values for samples drawn from close although not identical distributions. In these cases the method gives p-values that are not extremely low but are close to our suggested threshold. Given how close the distributions used are for some examples, it is not unreasonable to sometimes obtain a result suggesting that the samples are generated from the same distribution.

We applied the method to estimate the similarity matrix for all machines involved in the engine assembly line we focused on. An example of six machines was given in Section 4.6.1 and the reliability of the p-values was confirmed by the check of the features of the breakdown duration data sets. The method is widely applicable and we have demonstrated its application to estimating the similarity between medical procedures based on the patients' hospital length-of-stay data [41]; an example of five procedures was given in Section 4.6.2, where the similarity results made sense intuitively. This method has also been used to evaluate the similarities between simulation outputs of models using the current and

the proposed breakdown duration modelling. Overall, this method appears to be an appropriate distribution-free method for estimating the similarity between data sets that may contain different numbers of data points.

8.3 Arrows Classification Method

The Arrows classification method was derived to group machines based on the similarity matrix, consisting of the similarities between machines. The similarity between two machines is assumed to be the p-value for the Cramér-von Mises goodness of fit test for the comparison between their breakdown duration data sets, as described in the previous section. We found that this classification method performed well for a simple distance matrix from a text book, as well as for practical and more complicated similarity matrices such as the machine breakdown duration data. The method could be applied to classify data from a wide range of applications and it also gave sensible results when we applied it to grouping medical procedures based on the similarities between their patients' hospital length-of-stay data.

There are three main features of the Arrows method: (1) it ensures that objects with similarities below a specified threshold are not placed in the same group; (2) it ensures that objects with double-arrow connections are put in the same group; and (3) it prefers to keep objects with single-arrow connections in the same group when possible. Two machines have a double-arrow connections only if their similarity is greater than the specified threshold and is the highest among the similarities between the two machines and all of the other machines; two machines have a single-arrow connections if their similarity is greater than the specified threshold and is the highest among the similarities between either one of the two machines and all of the other machines. One characteristic of the Arrows method resulting from the multiple criteria is that it is possible that one object or group may be

combined with different groups or objects when the threshold changes. This can occur as a result of the method's intention of keeping objects with single-arrow connections in the same group, while satisfying the condition that every pair of objects in the same group should have a p-value that is above the threshold. When there is no relevant influence from single-arrow connections, this can also happen as a result of the method's intention of merging objects or groups with higher average connections, while satisfying the condition that every pair of objects in the same group should have a p-value that is above the selected threshold.

The method has similarities with complete linkage and average linkage hierarchical cluster analysis. The Arrows method places objects with double-arrow connections in the same group and prefers to keep together objects with single-arrow connections, which is different from cluster analysis in which the clustering method searches the whole similarity matrix to find the most similar groups to amalgamate. The results from the three methods suggest that the Arrows method seems to give more similar results to average linkage clustering when a lower similarity level is required, but when a higher similarity level is required the Arrows method tends to be more similar to complete linkage clustering. An advantage of the Arrows method over the two forms of cluster analysis considered here is that it allows us to control the similarity level in the resultant groups more easily through the use of a threshold, such that any two objects whose similarity is less than the threshold will not be placed in the same group.

8.4 Evaluate Breakdown Duration Input Modelling

In Chapter 7 we described the methodology used to evaluate the modelling of the machine breakdown durations, by comparing the system throughput of the same engine assembly model using three different breakdown duration inputs. The methodology could be useful for comparing system configurations, by evaluat-

ing the similarities between the stochastic outputs coming from the corresponding simulation models.

The evaluation process revealed that the machine breakdown duration settings did not affect the system throughput significantly. Further work on investigating the causes of production loss was carried out and it was found that the main sources of variability in the line yield are the engine repair process and operator stoppages, and these mask the effect of changes in machine breakdown durations on the system throughput. The three representations of the machine breakdown durations considered here (empirical distributions, fitted mixture distributions for individual machines, and fitted mixture distributions for the groups of machines obtained using the Arrows classification method) generated simulation outputs that were all within the 95% confidence interval of the real line yield data, suggesting any of them could be used as input models. The mixture distribution fitted to groups of machines is likely to be the most appropriate representation of the breakdown duration inputs for several reasons. First, it overcomes some shortcomings of the use of empirical distributions as simulation inputs as discussed in Section 8.1. Furthermore, comparing the use of group fitted mixture distributions to using individual fitted mixture distributions, the former has a couple of advantages over the latter: (a) the total saving of time for the fitting processes and the inputting of breakdown setting is considerable, even when taking into account the time spent implementing the Arrows method for the grouping; (b) for situations where a machine without available data or a new machine is being modelled, an experienced engineer could probably help with identifying which group of machines the no-data/new machine belongs to and so the fitted mixture distribution for that group could be used to represent the breakdown duration input of this machine; and the accuracy of identifying the machine as being similar to a group of machines should be higher that that of identifying one particular machine whose breakdown duration data are available as a similar machine. In addition, the similar simulation performance using inputs

of individual FMD and group FMD strongly justified the use of the classification method.

The choice of threshold for the Arrows grouping process did not appear to have a significant influence on the simulation throughput of models using corresponding group fitted distributions in this model. The impact of using different values for the threshold in the Arrows classification method on the system throughput was studied. In this case, we adjusted the simulation model so that machine breakdown was the only major source of variability in the system throughput (the engine repairs and operator stoppages were turned off) and the results showed that the simulation outputs of models using different group fitted mixture distributions are not significantly different. Therefore, a relatively low threshold, producing a low number of groups, can be chosen for the purpose of using group fitted mixture distribution for representing the machine breakdown duration input of simulation models.

8.5 Future Work

We have considered only a small part of the total breakdown process in this thesis and we would like to develop a complete model of breakdowns. While machine breakdown durations are important, the current method of modelling the time between failures may also be influencing the model output. Improving the representation of the time between failures could use the basic methodology with most of the additional work probably being the collection of data of time between failures.

The breakdown duration data provided by Ford included not only the actual repair time but also some waiting time for some resources, e.g. maintenance team or parts. In this work we focused on developing a statistical model of the total breakdown duration. Splitting the breakdown duration up into its constituent parts and modelling them separately would allow a better description of breakdowns in

the simulation model. The methodology would not need to change substantially and most of the work to make this extension would be involved in distinguishing and recording the data for the actual repair stage and for the waiting stage.

Engine repairs and operator stoppages, which are essentially product quality issues and human behaviour breakdowns, are responsible for a great part of the total loss of the line productions in the engine assembly plant. Therefore, it is important that they are modelled accurately. As MODAPTS, a technology involved in recording all motions required for a person to complete a task and analysis for methods improvement, has been introduced and used in more manufacturing companies, human behaviour can also be recorded more accurately. Accordingly it should be possible to extend the methodology to incorporate modelling of human breakdowns and response times. This would allow a complete and integrated model of machine breakdown behaviour to be developed including the modelling of time to repair failures, waiting time for resources, time between failures, human response times and human breakdowns. In the future we should also consider implementing the methodology described in this thesis to model the engine repairs process. Together with the extensions of modelling machine breakdowns discussed above, this would result in a complete system for modelling the total loss in manufacturing processes due to machine breakdowns, operator performances and product quality issues.

Simulation input modelling is an important part of simulation construction. The methodology for modelling breakdown durations presented in this thesis could be extended to model variable inputs in other simulation applications, where the inputs are multimodal, outside of the manufacturing area.

The Arrows classification of machines has been examined using the collected historical breakdown duration data and we would like to be able to validate the classification using the machines' future performance. New breakdown duration data may provide more confidence in the methods or may lead to the groups being

updated. It would be useful to devise an update procedure that does not involve a complete recalculation of the similarity matrix and rerun of the Arrows method. We have suggested that when modelling the breakdown of new machines, experienced engineers may decide they may be similar to a group of machines. It may also be useful to collect their real breakdown duration data during a period when they are used in the actual production, which can then be analysed to assess the engineers' decision.

The Arrows method could be extended to classify objects in other applications. The distribution-free method for estimating the similarity between data sets that may be of different sizes has the scope to be useful in fields other than manufacturing. For example we have shown their applications to the grouping of medical procedures in this thesis.

8.6 Discussion

We have demonstrated the modelling of machine breakdown durations in an engine assembly line simulation model. We found that fitted finite mixture distributions for groups of machines were suitable for representing machine breakdown durations as simulation inputs, and used parameters with an intuitive meaning. Grouping like machines serves to decrease the total time spent on fitting the input models considerably, as well as simplifying the breakdown duration inputs required for the simulation model. The Arrows classification of the machines based on the similarities between their breakdown duration data sets serves this purpose well.

The method for estimating similarity that we have introduced can be used to calculate the similarity between data sets with uneven numbers of data points and being a distribution-free method, its application is relatively simple and widely applicable.

We introduce the Arrows Classification procedure in Chapter 5 and tested it on a textbook example as well as two different sets of data coming from widely different applications (manufacturing and health care). The results suggest that it produces similar results to cluster analysis, while making it much easier to control the similarity level in each resultant group in order to achieve different classification targets.

Ford have been using the program we developed for the data validation, which has achieved a huge saving on the data process time. Meanwhile, they have showed interest in using the proposed method for modelling machine breakdown durations. However, as Ford use Excel interfaces to generate simulation models, these interfaces need to be upgraded, in order to allow the engineers and simulation modellers to use fitted mixture distributions to model the machine breakdown durations.

In conclusion, if there is multimodality present in a data set, the machine breakdown duration modelling process described in this thesis can be used to obtain a representation of the random inputs for simulation models. We have demonstrated its use on machine breakdown duration modelling in the manufacturing simulation model of an engine assembly line. The calculation of similarity and the Arrows Classification method introduced in this thesis would be applicable in a wide range of situations, not simply for analysing machine breakdown duration data. We have demonstrated their use on grouping machines and medical procedures. The methodology of simulation evaluation has been successfully used for evaluating the machine breakdown duration inputs and could also be applied to evaluate other sources of variability in simulation models.

Glossary

Arrows Classification Procedure: A classification method we have derived. It has a setting of similarity threshold that can be specified by the user, which allows the user to easily control the similarity level in the resultant groups.

Available Time Mode: In this mode, machines can break down whether they are operating or not.

Breakdown Duration: The whole period of a machine breakdown, which is also generally referred to as the *repair time* or the *time to repair* (TTR) or the *machine downtime*.

Busy Time Mode: In this mode, machines can only break down while they are operating.

CDF: Cumulative Distribution Function.

Double-Arrow Connection: A definition of similarity between objects that associate with the Arrows Classification method. Objects O_i and O_j have a *double-arrow connection* if p_{ij}, the p-value comparing their corresponding sets of data, is the biggest in both row i and row j of the similarity matrix and p_{ij} is greater than the specified threshold p_0.

EDF: Empirical Distribution Function.

FMD: Fitted finite Mixture Distribution.

Forman: A generic title of a supervisory person in a manufacturing plant and can be a male or female.

JPH: Jobs completed Per Hour for a machining or engine assembly line.

Maintenance Operator: A worker who has been trained to obtain the required skills to identify and rectify the faults of equipment that fails to function.

Major Repair: A machine failure that takes longer than 15 minutes to repair and generally requires a highly skilled maintenance operator to fix.

Minor Repair: A machine failure that takes less than 15 minutes to repair and generally only require a basic level of skill to fix.

Monitoring system: An automatic data record system that keeps track of all stoppages that occur on machines that are connected to the system.

MTBF: An acronym stands for Mean Time Between Failure.

MTTR: An acronym stands for Mean Time to Repair.

Number of Operations Mode: In this mode, machine breaks down after a certain number of operations.

Operator: A worker who is responsible for ensuring the efficient functioning of equipment in the assigned department.

PDF: Probability Density Function.

Productivity Engineering Department: A department usually known as Industrial Engineering department which used the skills of Time and Method Study. But due to changes in operating philosophy the name was changed.

Single-Arrow Connection: A definition of similarity between objects that associate with the Arrows Classification method. Objects O_i and O_k have a *single-arrow connection* if p_{ik}, the p-value comparing their corresponding sets of data, is the biggest in only one of row i or row k of the similarity matrix and p_{ik} is greater than the specified threshold p_0.

TTR: An acronym stands for Time to Repair.

Appendix A

Grouping Results of the 20 Machines

For the 20 machines with Similarity Matrix given in Table 5.2, the Arrows Classification method and complete linkage clustering give the same grouping results at similarity levels of $0.20, 0.30, \ldots, 0.90$. These grouping results are given in Table A.1.

P-value Threshold	Group	Arrows Method or Complete Linkage Clustering Method
0.20	1	M01, M11, M12
	2	M02, M03
	3	M05, M19, M20
	4	M07, M10, M13, M17
	5-12	(Single machine groups) M04, M06, M08, M09, M14, M15, M16, M18
0.30	1	M01, M11, M12
	2	M02, M03
	3	M05, M19, M20
	4	M07, M17
	5	M10, M13
	6-13	(Single machine groups) M04, M06, M08, M09, M14, M15, M16, M18
0.40/0.50	1	M01, M11
	2	M02, M03
	3	M05, M19, M20
	4	M07, M17
	5	M10, M13
	6-14	(Single machine groups) M04, M06, M08, M09, M12, M14, M15, M16, M18
0.60	1	M02, M03
	2	M05, M19, M20
	3	M07, M17
	4	M10, M13

P-value Threshold	Group	Arrows Method or Complete Linkage Clustering Method
	5-15	(Single machine groups) M01, M04, M06, M08, M09, M11, M12, M14, M15, M16, M18
0.70	1	M05, M19, M20
	2	M07, M17
	3	M10, M13
	4-16	(Single machine groups) M01, M02, M03, M04, M06, M08, M09, M11, M12, M14, M15, M16, M18
0.80	1	M05, M20
	2	M07, M17
	3	M10, M13
	4-17	(Single machine groups) M01, M02, M03, M04, M06, M08, M09, M11, M12, M14, M15, M16, M18, M19
0.90	1	M05, M20
	2-19	(Single machine groups) M01, M02, M03, M04, M06, M07, M08, M09, M10, M11, M12, M13, M14, M15, M16, M17, M18, M19

Table A.1: Grouping results of the 20 machines with Similarity Matrix given in Table 5.2, using the Arrows Classification method and complete linkage clustering.

Appendix B

Similarity Matrix and Grouping Results of the 39 Machines in DuntonL01 Engine Assembly Line

The estimated Similarity Matrix of the 39 machines involved in the engine assembly line, DuntonL01, is given in Tables B.1 and B.2. The similarities are estimated using the method descrined in Chapter 4. The matrix has been split across the two tables for presentation purposes.

The grouping results of the 39 machines using the Arrows Classification method with a specified threshold of 0.10 are given in Table B.3.

	$ML01$	$ML02$	$ML03$	$ML04$	$ML05$	$ML06$	$ML07$	$ML08$	$ML09$	$ML10$	$ML11$	$ML12$	$ML13$	$ML14$	$ML15$	$ML16$	$ML17$	$ML18$	$ML19$	$ML20$
$ML01$	–	0.12	0.07	0.21	0.00	0.00	0.02	0.00	0.00	0.00	0.93	0.00	0.00	0.00	0.00	0.07	0.00	0.00	0.07	0.00
$ML02$	0.12	–	0.03	0.48	0.00	0.00	0.00	0.00	0.00	0.00	0.87	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.01	0.00
$ML03$	0.07	0.03	–	0.20	0.00	0.00	0.01	0.00	0.00	0.00	0.36	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.18	0.00
$ML04$	0.21	0.48	0.20	–	0.00	0.00	0.00	0.00	0.00	0.02	0.89	0.02	0.00	0.00	0.00	0.15	0.00	0.00	0.03	0.00
$ML05$	0.00	0.00	0.00	0.00	–	0.89	0.00	0.00	0.01	0.02	0.21	0.00	0.00	0.30	0.00	0.23	0.62	0.06	0.00	0.00
$ML06$	0.00	0.00	0.00	0.00	0.89	–	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.18	0.01	0.00	0.00	0.00
$ML07$	0.02	0.00	0.01	0.00	0.00	0.00	–	0.00	0.00	0.00	0.09	0.00	0.16	0.00	0.00	0.00	0.00	0.00	0.49	0.06
$ML08$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	–	0.00	0.00	0.01	0.00	0.00	0.01	0.23	0.33	0.00	0.00	0.00	0.00
$ML09$	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	–	0.00	0.07	0.00	0.00	0.00	0.00	0.12	0.00	0.00	0.00	0.00
$ML10$	0.00	0.00	0.00	0.02	0.02	0.00	0.00	0.00	0.00	–	0.94	0.01	0.00	0.00	0.00	0.16	0.00	0.00	0.00	0.00
$ML11$	0.93	0.87	0.36	0.89	0.21	0.10	0.09	0.01	0.07	0.94	–	0.73	0.00	0.12	0.00	0.29	0.15	0.02	0.16	0.00
$ML12$	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.01	0.73	–	0.00	0.00	0.00	0.14	0.00	0.00	0.00	0.00
$ML13$	0.00	0.00	0.00	0.00	0.00	0.00	0.16	0.00	0.00	0.00	0.00	0.00	–	0.00	0.00	0.00	0.00	0.00	0.02	0.26
$ML14$	0.00	0.00	0.00	0.00	0.30	0.00	0.00	0.01	0.00	0.00	0.12	0.00	0.00	–	0.00	0.37	0.00	0.00	0.00	0.00
$ML15$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.23	0.00	0.00	0.00	0.00	0.00	0.00	–	0.06	0.00	0.00	0.00	0.00
$ML16$	0.07	0.10	0.04	0.15	0.23	0.18	0.00	0.33	0.12	0.16	0.29	0.14	0.00	0.37	0.06	–	0.20	0.10	0.02	0.00
$ML17$	0.00	0.00	0.00	0.00	0.62	0.01	0.00	0.00	0.00	0.00	0.15	0.00	0.00	0.00	0.00	0.20	–	0.00	0.00	0.00
$ML18$	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.10	0.00	–	0.00	0.00
$ML19$	0.07	0.01	0.18	0.03	0.00	0.00	0.49	0.00	0.00	0.00	0.16	0.00	0.02	0.00	0.00	0.02	0.00	0.00	–	0.01
$ML20$	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.26	0.00	0.00	0.00	0.00	0.00	0.01	–
$ML21$	0.00	0.00	0.20	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.13	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.04	0.00
$ML22$	0.24	0.05	0.42	0.11	0.00	0.00	0.37	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.87	0.00
$ML23$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.00
$ML24$	0.04	0.00	0.29	0.03	0.00	0.00	0.21	0.00	0.00	0.00	0.32	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.84	0.00
$ML25$	0.00	0.00	0.00	0.00	0.00	0.00	0.19	0.00	0.00	0.00	0.01	0.00	0.61	0.00	0.00	0.00	0.00	0.00	0.03	0.47
$ML26$	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.18	0.00	0.06	0.00	0.00
$ML27$	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.64	0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.00
$ML28$	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.72	0.00
$ML29$	0.39	0.46	0.04	0.38	0.00	0.00	0.00	0.00	0.00	0.00	0.91	0.00	0.00	0.00	0.00	0.09	0.00	0.00	0.01	0.00
$ML30$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.20	0.00	0.00	0.01	0.00	0.00	0.01	0.04	0.68	0.00	0.00	0.00	0.00
$ML31$	0.28	0.14	0.86	0.23	0.00	0.00	0.20	0.00	0.00	0.00	0.36	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.50	0.00
$ML32$	0.36	0.12	0.51	0.29	0.00	0.00	0.14	0.00	0.00	0.00	0.49	0.00	0.00	0.00	0.00	0.07	0.00	0.00	0.28	0.00
$ML33$	0.00	0.00	0.04	0.01	0.00	0.00	0.02	0.00	0.00	0.00	0.47	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.16	0.00
$ML34$	0.05	0.11	0.06	0.23	0.02	0.00	0.00	0.00	0.00	0.12	0.53	0.27	0.00	0.00	0.00	0.25	0.00	0.00	0.00	0.00
$ML35$	0.00	0.00	0.00	0.00	0.00	0.00	0.23	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.88	0.01
$ML36$	0.00	0.00	0.00	0.00	0.09	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.17	0.00	0.22	0.00	0.00
$ML37$	0.13	0.09	0.05	0.29	0.06	0.03	0.00	0.06	0.00	0.18	0.58	0.12	0.00	0.05	0.00	0.68	0.02	0.01	0.02	0.00
$ML38$	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.04	0.00	0.00	0.19	0.00	0.00	0.00	0.00	0.62	0.00	0.00	0.00	0.00
$ML39$	0.06	0.18	0.03	0.41	0.02	0.00	0.00	0.00	0.00	0.35	0.94	0.13	0.00	0.00	0.00	0.19	0.00	0.00	0.00	0.00

Table B.1: Part a of the Similarity Matrix of the breakdown duration data for the 39 machines involved in DuntonL01 engine assembly line, estimated using the method described in Chapter 4.

	$ML21$	$ML22$	$ML23$	$ML24$	$ML25$	$ML26$	$ML27$	$ML28$	$ML29$	$ML30$	$ML31$	$ML32$	$ML33$	$ML34$	$ML35$	$ML36$	$ML37$	$ML38$	$ML39$
$ML01$	0.00	0.24	0.00	0.04	0.00	0.00	0.00	0.00	0.39	0.00	0.28	0.36	0.00	0.05	0.00	0.00	0.13	0.00	0.06
$ML02$	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.46	0.00	0.14	0.12	0.00	0.11	0.00	0.00	0.09	0.00	0.18
$ML03$	0.20	0.42	0.00	0.29	0.00	0.00	0.00	0.00	0.04	0.00	0.86	0.51	0.04	0.06	0.00	0.00	0.05	0.00	0.03
$ML04$	0.03	0.11	0.00	0.03	0.00	0.00	0.01	0.00	0.38	0.00	0.23	0.29	0.01	0.23	0.00	0.00	0.29	0.01	0.41
$ML05$	0.00	0.00	0.00	0.00	0.00	0.10	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.09	0.06	0.01	0.02
$ML06$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00
$ML07$	0.00	0.37	0.00	0.21	0.19	0.00	0.00	0.08	0.00	0.00	0.20	0.14	0.02	0.00	0.23	0.00	0.00	0.00	0.00
$ML08$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.04	0.00
$ML09$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$ML10$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.12	0.00	0.00	0.18	0.00	0.35
$ML11$	0.13	0.33	0.03	0.32	0.01	0.04	0.64	0.02	0.91	0.01	0.36	0.49	0.47	0.53	0.05	0.02	0.58	0.19	0.94
$ML12$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.27	0.00	0.00	0.12	0.00	0.13
$ML13$	0.00	0.00	0.00	0.00	0.61	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$ML14$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00
$ML15$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$ML16$	0.03	0.03	0.01	0.01	0.00	0.18	0.33	0.00	0.09	0.68	0.04	0.07	0.03	0.25	0.00	0.17	0.68	0.62	0.19
$ML17$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00
$ML18$	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.22	0.01	0.00	0.00
$ML19$	0.04	0.87	0.01	0.84	0.03	0.00	0.00	0.72	0.01	0.00	0.50	0.28	0.16	0.00	0.88	0.00	0.02	0.00	0.00
$ML20$	0.00	0.00	0.00	0.00	0.47	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
$ML21$	–	0.07	0.17	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.05	0.00	0.01	0.00	0.00	0.04	0.00	0.00
$ML22$	0.07	–	0.00	0.99	0.00	0.00	0.00	0.19	0.08	0.00	0.86	0.79	0.52	0.01	0.37	0.00	0.03	0.01	0.02
$ML23$	0.17	0.00	–	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$ML24$	0.01	0.99	0.00	–	0.00	0.00	0.00	0.05	0.00	0.00	0.63	0.62	0.11	0.00	0.08	0.00	0.02	0.00	0.00
$ML25$	0.00	0.00	0.00	0.00	–	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$ML26$	0.00	0.00	0.00	0.00	0.00	–	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00
$ML27$	0.00	0.00	0.00	0.00	0.00	0.00	–	0.00	0.00	0.00	0.00	0.00	0.00	0.15	0.00	0.00	0.44	0.00	0.02
$ML28$	0.00	0.19	0.00	0.05	0.00	0.00	0.00	–	0.00	0.00	0.02	0.00	0.00	0.00	0.07	0.00	0.00	0.00	0.00
$ML29$	0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.00	–	0.00	0.17	0.17	0.00	0.05	0.00	0.00	0.09	0.00	0.17
$ML30$	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	–	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.06	0.00
$ML31$	0.11	0.86	0.00	0.63	0.00	0.00	0.00	0.02	0.17	0.00	–	0.96	0.63	0.05	0.10	0.00	0.04	0.00	0.05
$ML32$	0.05	0.79	0.00	0.62	0.00	0.00	0.00	0.00	0.17	0.00	0.96	–	0.87	0.04	0.01	0.00	0.07	0.01	0.03
$ML33$	0.00	0.52	0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.63	0.87	–	0.00	0.00	0.00	0.03	0.00	0.00
$ML34$	0.01	0.01	0.00	0.00	0.00	0.00	0.15	0.00	0.05	0.00	0.05	0.04	0.00	–	0.00	0.00	0.08	0.01	0.49
$ML35$	0.00	0.37	0.00	0.08	0.00	0.00	0.00	0.07	0.00	0.00	0.10	0.01	0.00	0.00	–	0.00	0.00	0.00	0.00
$ML36$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	–	0.01	0.00	0.00
$ML37$	0.04	0.03	0.00	0.02	0.00	0.03	0.44	0.00	0.09	0.06	0.04	0.07	0.03	0.08	0.00	0.01	–	0.82	0.20
$ML38$	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.01	0.00	0.01	0.00	0.00	0.82	–	0.00
$ML39$	0.00	0.02	0.00	0.00	0.00	0.00	0.02	0.00	0.17	0.00	0.05	0.03	0.00	0.49	0.00	0.00	0.20	0.00	–

Table B.2: Part b of the Similarity Matrix of the breakdown duration data for the 39 machines involved in DuntonL01 engine assembly line, estimated using the method described in Chapter 4.

Group	Machines
G01	ML01, ML02, ML04, ML29
G02	ML03, ML21
G03	ML05, ML06
G04	ML07, ML19, ML35
G05	ML08, ML15
G06	ML10, ML11, ML34, ML39
G07	ML13, ML20, ML25
G08	ML16, ML37, ML38
G09	ML18, ML36
G10	ML22, ML24, ML31, ML32, ML33
G11-G19	(Single machine groups) ML09, ML12, ML14, ML17, ML23, ML26, ML27, ML28, ML30

Table B.3: Grouping results of the 39 machines based on the Similarity Matrix given in Tables B.1 and B.2, using the Arrows Classification method with threshold $p_0 = 0.10$.

References

[1] E. K. Al-Hussaini and A. H. Abdel-Hamid. Accelerated life tests under finite mixture models. *Journal of Statistical Computation and Simulation*, 76(8):673–690, 2006.

[2] M. S. Aldenderfer and R. K. Blashfield. Cluster analysis and archaeological classification. *American Antiquity*, 43:502–505, 1978.

[3] M. R. Anderberg. *Cluster Analysis for Applications*. Academic Press, New York, 1973.

[4] T. W. Anderson. *The Statistical Analysis of Time Series*. John Wiley, New York, 1994.

[5] T. Y. Anderson. On the distribution of the two-sample Cramér-von Mises criterion. *The Annals of Mathematical Statistics*, 33:1148–1159, 1962.

[6] H. E. Ascher and H. Fiengold. *Repairable Systems Reliability: Modelling, Inference, Misconceptions and Their Causes*. Marcel Dekker, New York, 1984.

[7] J. Banks, J. S. Carson, and B. L. Nelson. *Discrete-Event System Simulation*. Prentice Hall, New Jersey, 2nd edition, 1996.

[8] J. Banks, J. S. Carson, B. L. Nelson, and D. Nicol. *Discrete-Event System Simulation*. Prentice Hall, New Jersey, 3nd edition, 2001.

[9] D. N. Baron and P. M. Fraser. Medical applications of taxonomic methods. *British medical bulletin*, 24(3):236–240, September 1968.

[10] R. R. Barton, S. E. Chick, R. C. H. Cheng, S. G. Henderson, A. M. Law, B. W. Schmeiser, L. M. Leemis, L. W. Schruben, and J. R. Wilson. Panel discussion on current issues in input modeling: panel on current issues in simulation input modeling. In *Proceedings of the 2002 Winter Simulation Conference*, pages 353–369. Winter Simulation Conference, 2002.

[11] C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7):719–725, 2000.

[12] B. Biller and B. L. Nelson. Answers to the top ten input modeling questions. In E. Yücesan, C. H. Chen, J. L. Snowdon, and J. M. Charnes, editors, *Proceedings of the 2002 Winter Simulation Conference*, pages 35–40, 2002.

[13] B. Biller and B. L. Nelson. Modeling and generating multivariate time-series input processes using a vector autoregressive technique. *ACM Trans. Model. Comput. Simul.*, 13(3):211–237, 2003.

[14] B. Biller and B. L. Nelson. Fitting time-series input processes for simulation. *Oper. Res.*, 53(3):549–559, 2005.

[15] W. Binroth and R. K. Haboush. Stochastic system modelling applied to an industrial production system. In *International Congress and Exposition*, Detroit, MI, USA, March 1978.

[16] K. M. Blache and A. B. Shrivastava. Reliability and maintainability of machinery and equipment for effective maintenance. In *International Congress and Exposition*, pages 19–23, Detroit, MI, USA, March 1993.

[17] H. P. Bloch and F. K. Geitner. *Practical Machinery Management for Process Plants Volume 2 - Machinery Analysis and Troubleshooting*. Gulf Publishing, 2nd edition, 1994.

[18] G. E. P. Box and G. M. Jenkins. *Time Series Analysis: Forecasting and Control*. Holden-Day series in time series analysis and digital processing. Holden-Day, San Francisco, 1976.

[19] G. E. P. Box and G. C. Tiao. *Bayesian Inference in Statistical Analysis*. Wiley Classics Library. John Wiley and Sons, New York, May 1992.

[20] A. J. Boyce. Mapping diversity: A comparative study of some numerical methods. In A. J. Cole, editor, *Numerical Taxonomy*, pages 1–30. Academic Press, New York, 1969.

[21] T. C. Bradford and K. F. Martin. Modelling the breakdown behavior of transfer line machines for use in computer simulation. *International Journal of Modelling and Simulation*, 13, 1993.

[22] J. R. Bray and J. T. Curtis. An ordination of the upland forest communities of s. wisconsin. *Ecological Monographs*, 27:325–349, 1957.

[23] J. A. Buzacott and L. E. Hanifin. Models of automatic transfer. lines with inventory banksa review and comparison. *AIIE. Transactions*, 10:197–207, 1978.

[24] A. J. Cain and G. A. Harrison. An analysis of the taxonomist's judgment of affinity. *Proc. Zool. Soc.*, 131:85–98, 1958.

[25] A. S. Carrie. *Simulation of Manufacturing Systems*. John Wiley and Sons, New York, NY, USA, 1988.

[26] J. D. Carroll, L. A. Clark, and W. S. DeSarbo. The representation of three-way proximity data by simple and multiple tree structure models. *Journal of Classification*, 1:25–74, 1984.

[27] A. D. S. Carter. *Mechanical Reliability*. Halsted Pr, 1986.

[28] C. R. Cash, B. L. Nelson, D. G. Dippold, J. M. Long, and W. P. Pollard. Evaluation of tests for initial-condition bias. In *Proceedings of the 1992 Winter simulation conference*, pages 577–585, New York, NY, USA, 1992. ACM.

[29] R. B. Cattell. A note on correlation clusters and cluster search methods. *Psychometrika*, 9:169–184, 1944.

[30] R. B. Cattell. r_p and other coefficients of pattern similarity. *Psychometrika*, 14:279–298, 1949.

[31] R. B. Cattell and M. A. Coulter. Principles of behavioural taxonomy and the mathematical basis of the taxonome computer program. *British Journal of Mathematical and Statistical Psychology*, 19:237–269, 1966.

[32] F. T. S. Chan. Using simulation to predict system performance: a case study of an electro-phoretic deposition plant. *Integrated Manufacturing Systems*, 6, 1995.

[33] R. C. H. Cheng. Bayesian model selection when the number of components is unknown. In D. J. Medeiros, E. F. Watson, J. S. Carson, and M. S. Manivannan, editors, *Proceedings of the 1998 Winter Simulation Conference*, pages 653–659. IEEE, 1998.

[34] R. C. H. Cheng. Analysis of simulation output by resampling. *International Journal of Simulation Systems, Science & Technology*, 1:51–58, 2001.

[35] R. C. H. Cheng and C. S. M. Currie. Prior and candidate models in the Bayesian analysis of finite mixtures. In S. Chick, P. J. Sánchez, D. Ferrin, and D. J. Morrice, editors, *Proceedings of the 2003 Winter Simulation Conference*, pages 392–398. IEEE, 2003.

[36] H. Chipman and R. Tibshirani. Hybrid hierarchical clustering with applications to microarray data. *Biostatistics*, 7(2):286–301, April 2006.

[37] L. W. Condra. *Reliability Improvement with Design of Experiments*. Marcel Dekker, New York, 2nd edition, 2001.

[38] R. M. Cormack. A review of classification. *Journal of the Royal Statistical Society*, 134:321–367, 1971.

[39] P. J. Crosby and D. R. Murton. Simulation of machine tool breakdowns. Project report for the degree of bachelor of engineering, University of Bath, June 1990.

[40] C. S. M. Currie. *Bayesian Sampling Methods in Epidemic and Finite Mixture Models*. PhD thesis, University of Southampton, November 2004.

[41] C. S. M. Currie and L. Lu. *Intelligent Patient Management*, volume 189/2009, chapter Optimal Scheduling Using Length-of-Stay Data for Diverse Routine Procedures, pages 193–205. Springer Berlin, Heidelberg, 2009.

[42] R. B. D'Agostino and M. A. Stephens. *Goodness-of-fit techniques*. Statistics: textbooks and monographs; 68. Marcel Dekker, New York, 1986.

[43] D. J. Davis. An analysis of some failure data. *Journal of the American Statistical Association*, 47(258):113–150, 1952.

[44] A. C. Davison and D. V. Hinkley. *Bootstrap Methods and Their Application*. Cambridge University Press, Cambridge, 1997.

[45] G. De Soete. A least squares algorithm for fitting an ultrametric tree to a dissimilarity matrix. *Pattern Recognition Letters*, 2:133–137, 1984.

[46] D. Defays. An efficient algorithm for a complete link method. *The Computer Journal*, 20(4):364–366, 1977.

[47] B. Deler and B. L. Nelson. Input modeling and its impact: modeling and generating multivariate time series with arbitrary marginals and autocorrelation structures. In *Proceedings of the 2001 Winter Simulation Conference*, pages 275–283, Washington, DC, USA, 2001. IEEE Computer Society.

[48] L. R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26:297–302, 1945.

[49] E. Diday and J. V. Moreau. Learning hierarchical clustering from examples - application to the adaptive construction of dissimilarity indices. *Pattern Recognition Letters*, 2:365–378, 1984.

[50] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. FL: CRC Press, Boca Raton, 1994.

[51] B. Epstein and M. Sobel. Life testing. *Journal of the American Statistical Association*, 48(263):486–502, 1953.

[52] M. Evans, N. Hastings, and B. Peacock. *Statistical Distributions*. John Wiley, New York, 3nd edition, 2000.

[53] B. S. Everitt. *Cluster Analysis*. Halstead Press, London, 1974.

[54] C. E. Feltner and S. A. Weiner. Models, myths and mysteries in manufacturing. *Industrial Engineering*, 17(7):66–76, July 1985.

[55] A. H. Fielding. *Cluster and Classification Techniques for the Biosciences*. Cambridge University Press, Cambridge, 2007.

[56] L. Fisher and J. W. Van Ness. Admissible clustering procedures. *Biometrika*, 58(1):91–104, 1971.

[57] G. S. Fishman. Estimating sample size in computing simulation experiments. *Management Science*, 18(1):21–38, 1971.

[58] G. S. Fishman. *Concepts and Methods in Discrete Event Digital Simulation*. Wiley, New York, 2004.

[59] J. J. Fortier and H. Solomon. Clustering procedures. In *Proceedings of Symp. Multiv. Analysis*, pages 493–506, 1966.

[60] A. V. Gafarian, C. J. Ancker, and T. Morisaku. Evaluation of commonly used rules for detecting steady-state in computer simulation. *Naval Research Logistics Quarterly*, 25:511–529, 1978.

[61] S. Gallivan and M. Utley. Modelling admissions booking of elective inpatients into a treatment center. *IMA Journal of Management Mathematics*, 16:305–315, 2005.

[62] J. A. Gengerelli. A method for detecting subgroups in a population and specifying their membership. *Journal of Psychology*, 55:457–468, 1963.

[63] E. I. George and R. E. McCulloch. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

[64] L. George. *The Bathtub Curve Doesn't Always hold water*. American Society for Quality.

[65] S. Ghosh and S. G. Henderson. Chessboard distributions. In *Proceedings of the 2001 Winter Simulation Conference*, pages 385–393, Washington, DC, USA, 2001. IEEE Computer Society.

[66] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London, 1996.

[67] P. W. Glynn. Initial transient problem for steady-state output analysis. In *Proceedings of the 2005 Winter Simulation Conference*, pages 739–740, 2005.

[68] A. Goel and R. J. Graves. Electronic system reliability: Collating prediction models. *IEEE Transactions on Device and Materials Reliability*, 6(2):258–265, June 2006.

[69] D. Goldsman, L. W. Schruben, and J. J. Swain. Tests for transient means in simulation time series. *Naval Research Logistics*, 41:171–187, 1994.

[70] A. D. Gordon. A review of hierarchical classification. *Journal of the Royal Statistical Society*, 150:119–137, 1987.

[71] G. Gordon. *System Simulation*. Prentice Hall, New Jersey, 1969.

[72] J. C. Gower. Multivariate analysis and multidimensional geometry. *The Statistician*, 17(1):13–28, 1967.

[73] J. C. Gower. A survey of numerical methods useful in taxonomy. *Acarologia*, 11:357–376, 1969.

[74] J. Grabmeier and A. Rudolph. Techniques of cluster algorithms in data mining. *Data Mining and Knowledge Discovery*, 6(4):303–360, 2002.

[75] P. J. Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

[76] B. Grün and F. Leisch. Fitting finite mixtures of generalized linear regressions in R. *Computational Statistics & Data Analysis*, 51(11):5247–5252, July 2007.

[77] L. E. Hanifin. *Increased Transfer Line Productivity Utilizing Systems Simulation*. PhD thesis, University of Detroit, 1975.

[78] L. E. Hanifin and S. G. Liberty. Improved efficiency of transmission case machining: A gpss-v simulation of a transfer line. In *Automotive Engineering Congress and Exposition*, pages 23–27, Detroit, MI, USA, February 1976.

[79] J. A. Hartigan. Representation of similarity matrices by trees. *Journal of the American Statistical Association*, 62(320):1140–1158, 1967.

[80] M. O. Hill, R. G. H. Bunce, and M. W. Shaw. Indicator species analysis, a divisive polythetic method of classification, and its application to a survey of native pinewoods in scotland. *The Journal of Ecology*, 63(2):597–613, 1975.

[81] K. Hoad, S. Robinson, and R. Davies. Automating estimation of warm-up length. *Unpublished*, 2008.

[82] F. R. Hodson, P. H. A. Sneath, and J. E. Doran. Some experiments in the numerical analysis of archaeological data. *Biometrika*, 53(3/4):311–324, 1966.

[83] W. J. Hopp and M. L. Spearman. *Factory Physics: Foundations of Manufacturing Management*. Illinois, Chicago, 2nd edition, 2001.

[84] O. C. Ibe and A. S. Wein. Availability of systems with partially observable failures. *IEEE Transactions on Reliability*, 41(1), 1992.

[85] I. Ikonen. Simulation of transfer line machine breakdowns. Master's thesis, University of Cranfield, September 1994.

[86] P. T. Jackway and B. S. DeSilva. A methodology for initialization bias reduction in computer simulation output. *Asia-Pacific Journal of Operational Research*, 9:87–100, 1992.

[87] M. Jambu. *Classification automatique pour l'analyse des données. I - Méthodes et algorithmes*. Dunod, Paris, 1978.

[88] N. Jardine and R. Sibson. *Mathematical Taxonomy*. Wiley, London, 1971.

[89] A. Jayaraman and A. K. Gunal. Applications of discrete event simulation in the design of automotive powertrain manufacturing systems. In *Proceedings of the 1997 Winter simulation Conference*, pages 758–764, Washington, DC, USA, 1997. IEEE Computer Society.

[90] N. L. Johnson and S. Kotz. *Continuous Univariate Distributions*, volume 2 of *Distributions in Statistics*. John Wiley, New York, 1970.

[91] R. L. Johnson and D. D. Wall. Cluster analysis of semantic differential data. *Educational and Psychological Measurement*, 29:769–780, 1969.

[92] E. Kay. The effectiveness of preventive maintenance. *International Journal of Production Research*, 14, 1976.

[93] M. G. Kendall. Discrimination and classification. In *Proceedings of Symp. Multiv. Analysis*, pages 165–185, 1966.

[94] M. E. Kuhl, E. K. Lada, N. M. Steiger, M. A. Wagner, and J. R. Wilson. Introduction to modeling and generating probabilistic input processes for simulation. In *Proceedings of the 2006 Winter simulation Conference*, pages 19–35. Winter Simulation Conference, 2006.

[95] M. E. Kuhl, E. K. Lada, N. M. Steiger, M. A. Wagner, and J. R. Wilson. Introduction to modeling and generating probabilistic input processes for

simulation. In *Proceedings of the 2007 Winter simulation Conference*, pages 63–76, Piscataway, NJ, USA, 2007. IEEE Press.

[96] E. K. Lada, M. A. Wagner, N. M. Steiger, and J. R. Wilson. Introduction to modeling and generating probabilistic input processes for simulation. In *Proceedings of the 2005 Winter simulation Conference*, pages 41–55. Winter Simulation Conference, 2005.

[97] J. Ladbrook. Breakdowns modelling - an inquest. Master in philosophy thesis, University of Birmingham, September 1998.

[98] J. Ladbrook and A. Januszczak. Ford's power train operations: changing the simulation environment. In *Proceedings of the 2001 Winter Simulation Conference*, pages 863–869, Washington, DC, USA, 2001. IEEE Computer Society.

[99] P. G. Lamoureux. Electronic reliability. *Quality*, page 45, September 1991.

[100] G. N. Lance and W. T. Williams. A generalized sorting strategy for computer classifications. *Nature*, 212:218, 1966.

[101] G. N. Lance and W. T. Williams. A general theory of classificatory sorting strategies i. hierarchical systems. *Computer Journal*, 9:373–380, 1967.

[102] Lanner Group, The Oaks, Clews Road, Redditch, UK. *WITNESS 2008 User Manual*, 2008.

[103] A. M. Law. *Simulation Modeling and Analysis*. McGraw-Hill series in industrial engineering and management science. McGraw-Hill, New York, 4th edition, 2007.

[104] A. M. Law and M. G. McComas. Pitfalls to avoid in the simulation of manufacturing systems. *Industrial Engineering*, 21(5):28–31, 1989.

[105] A. M. Law and M. G. McComas. Expertfit distribution-fitting software: how the expertfit distribution-fitting software can make your simulation models more valid. In *Proceedings of the 2003 Winter Simulation Conference*, pages 169–174. Winter Simulation Conference, 2003.

[106] A. M. Law, M. G. McComas, and S. G. Vincent. The crucial role of input modeling in successful simulation studies. *Industrial Engineering*, pages 55–57, July 1994.

[107] L. Leemis. Input modeling. In *Proceedings of the 1998 Winter Simulation Conference*, pages 15–22, Los Alamitos, CA, USA, 1998. IEEE Computer Society Press.

[108] L. Leemis. Input modeling techniques for discrete-event simulations. In *Proceedings of the 2001 Winter Simulation Conference*, pages 63–73, 2001.

[109] L. Leemis. Input modeling. In *Proceedings of the 2003 Winter Simulation Conference*, pages 14–24. Winter Simulation Conference, 2003.

[110] F. Leisch. Flexmix: A general framework for finite mixture models and latent class regression in r. *Journal of Statistical Software*, 11(8):1–18, 10 2004.

[111] X. Ma and A. K. Kochhar. Prediction of the system efficiency of balanced automatic transfer lines. In *Proceedings of the Institution of Mechanical Engineers, Part B: Management and Engineering Manufacture*, pages 51–59, Jun 1988.

[112] X. Ma and A. K. Kochhar. A comparison study of two tests for detecting initialization bias in simulation output. *Simulation*, 61(2):94–101, 1993.

[113] P. S. Mahajan and R. G. Ingalls. Evaluation of methods used to detect warm-up period in steady state simulation. In *Proceedings of the 2004 Winter Simulation Conference*, pages 663–671. Winter Simulation Conference, 2004.

[114] H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18:50–60, 1947.

[115] G. J. McLachlan and D. Peel. *Finite Mixture Models*. Wiley series in probability and statistics. Applied probability and statistics section. Wiley, New York, 2000.

[116] L. L. McQuitty. Hierarchical linkage analysis for the isolation of types. *Educational and Psychological Measurement*, 20:55–67, 1966.

[117] L. L. McQuitty. Multivariate analysis and multidimensional geometry. *Educational and Psychological Measurement*, 26:825–831, 1966.

[118] L. L. McQuitty. Expansion of similarity analysis by reciprocal pairs for discrete and continuous data. *Educational and Psychological Measurement*, 27:253–255, 1967.

[119] J. G. Miller and J. Anderson. Computer simulation of high speed battery manufacturing systems. In *Society of Manufacturing Engineers Conference in Food Processing*, pages 20–30, Dearborn Michigan, October 1992.

[120] G. W. Milligan. A review of monte carlo tests of cluster analysis. *Multivariate Behavioral Research*, 16:379–407, July 1981.

[121] R. M. Needham. A method of using computers in information classification. In C. Popplewell, editor, *Information Processing 62: Proceedings of IFIP Congress 1962*, pages 284–287. North-Holland, Amsterdam, 1963.

[122] J. A. Nelder and R. Mead. A simplex method for function minimization. *Computer Journal*, 7:308–313, 1965.

[123] B. L. Nelson and M. Yamnitsky. Input modeling tools for complex problems. In *Proceedings of the 1998 Winter Simulation Conference*, pages 105–112, Los Alamitos, CA, USA, 1998. IEEE Computer Society Press.

[124] F. Nixon. *Managing to Achieve Quality and Reliability*. McGraw-Hill European series in management and marketing. McGraw-Hill, New York, 2nd edition, 1971.

[125] I. Noy-Meir. Divisive polythetic classification of vegetation data by optimized division on ordination components. *The Journal of Ecology*, 61(3):753–760, 1973.

[126] K. Pawlikowski. Steady-state simulation of queueing processes: a survey of problems and solutions. *ACM Computing Surveys*, 22(2):123–170, 1990.

[127] Porter and Finke. Reliability prediction models for microcircuits. In *Proceedings of the ninth Reliability and Maintainability Conference*, pages 567–569, Detroit, MI, USA, July 1970.

[128] F. Proschan. Theoretical explanation of observed decreasing failure rate. *Technometrics*, 5:375–383, 1963.

[129] F. Proschan. Theoretical explanation of observed decreasing failure rate. *Technometrics*, 42(1):7–11, 2000.

[130] S. Richardson and P. J. Green. On bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society. Series B (Methodological)*, 59(4):731–792, 1997.

[131] S. Richardson, L. Leblond, I. Jaussent, and P. J. Green. Mixture models in measurement error problems, with reference to epidemiological studies. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 165(3):549–566, 2002.

[132] S. Robinson. A statistical process control approach for estimating the warm-up period. In E. Yücesan, C. H. Chen, J. L. Snowdon, and J. M. Charnes, editors, *Proceedings of the 2002 Winter Simulation Conference*, pages 439–446, 2002.

[133] S. Robinson. *Simulation: the Practice of Model Development and Use*. John Wiley, Chichester, 2004.

[134] S. Robinson. A statistical process control approach to selecting a warm-up period for a discrete-event simulation. *European Journal of Operational Research*, 176:332–346, 2007.

[135] S. Robinson, T. Alifantis, R. Hurrion, J. Ladbrook, J. Edwards, and T. Waller. Modelling and improving human decision making with simulation. In *Proceedings of the 2001 Winter Simulation Conference*, pages 913–920, Washington, DC, USA, 2001. IEEE Computer Society.

[136] K. Roeder and L. Wasserman. Practical bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*, 92(439):894–902, 1997.

[137] D. J. Rogers, H. S. Fleming, and G. Estabrook. Use of computers in studies of taxonomy and evolution. In T. Dobzhansky, M. K. Hecht, and W. C. Steere, editors, *Evolutionary Biology*, volume 1, pages 169–196. Appleton Century Crofts, New York, 1967.

[138] F. J. Rohlf. Adaptive hierarchical clustering schemes. *Systematic Zoology*, 19(1):58–82, 1970.

[139] F. J. Rohlf. Single-link clustering algorithms. In P. R. Krishnaiah and L. N. Kanal, editors, *Handbook of Statistics Volume 2. Classification, Pattern Recognition, and Reduction of Dimensionality*, pages 267–284. North Holland Publishing Company, Amsterdam, 1982.

[140] M. Rohrer and B. Strong. Automotive applications of discrete event simulation. *Automotive Manufacturing and Production*, July 1997.

[141] J. Rosser. *Unpublished Ford Internal Communication*. Ford Dunton Technical Center, 1985.

[142] Carson J. S. Convincing users of model's validity is challenging aspect of modeler's job. *Industrial Engineering*, 18:74–85, 1986.

[143] L. W. Schruben. Detecting initialization bias in simulation output. *Operations Research*, 30(3):569–590, 1982.

[144] L. W. Schruben, H. Singh, and L. Tierney. Optimal tests for initialization bias in simulation output. *Operations Research*, 31(6):1167–1178, 1983.

[145] M. J. Shepherd and A. J. Willmott. Cluster analysis on the atlas computer. *The Computer Journal*, 11(1):57–62, 1968.

[146] R. Sibson. Order invariant methods for data analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(3):311–349, 1972.

[147] P. H. A. Sneath. The application of computers to taxonomy. *Journal of General Microbiology*, 17:201–226, 1957.

[148] R. R. Sokal and C. D. Michener. A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin*, 28:1409–1438, 1958.

[149] R. R. Sokal and P. H. A. Sneath. *Principles of Numerical Taxonomy*. Freeman, London, 1963.

[150] P. Spavin. *Some Notes on Modelling Breakdowns Using WITNESS Version 5*. ISTEL Visual Interactive Systems Limited, 1985.

[151] M. A. Stephens. EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, 69:730–737, 1974.

[152] G. J. Székely and M. L. Rizzo. Hierarchical clustering via joint between-within distances Extending ward's minimum variance method. *Journal of Classification*, 22:151–183, 2005.

[153] L. C. Thomas, D. B. Edelman, and J. N. Crook. *Credit Scoring and Its Applications*. SIAM Monographs on Mathematical Modeling and Computation. Society for Industrial and Applied Mathematics, Philadelphia, 2002.

[154] J. W. Van Ness. Admissible clustering procedures. *Biometrika*, 60(2):422–424, 1973.

[155] G. Vassilacopoulos. Testing for initialization bias in simulation output. *Simulation*, 52(4):151–153, 1989.

[156] A. O. F. Venton and T. R. Ross. Component based prediction for mechanical reliability. In *Mechanical Reliability in the Process Industries*. Mechanical Engineering Publications Ltd., Edmunds, Suffolk England, 1984.

[157] S. Vincent. Input data analysis. In Banks Jerry, editor, *Handbook of Simulation*, pages 55–91. John Wiley and Sons, New York, 1998.

[158] S. G. Vincent and A. M. Law. Unifit ii: total support for simulation input modeling. In *Proceedings of the 1991 Winter Simulation Conference*, pages 136–142, Washington, DC, USA, 1991. IEEE Computer Society.

[159] S. G. Vincent and A. M. Law. Unifit ii: total support for simulation input modeling. In *Proceedings of the 1994 Winter Simulation Conference*, pages 409–414, San Diego, CA, USA, 1994. Society for Computer Simulation International.

[160] M. A. F. Wagner and J. R. Wilson. Using univariate Bézier distributions to model simulation input processes. In G. W. Evans, M. Mollaghasemi, E. C. Russell, and W. E. Biles, editors, *Proceedings of the 1993 Winter Simulation Conference*, pages 365–373. IEEE, 1993.

[161] M. A. F. Wagner and J. R. Wilson. Recent developments in input modeling with Bézier distributions. In J. M. Charnes, D. J. Morrice, D. T. Brunner, and J. J. Swain, editors, *Proceedings of the 1996 Winter Simulation Conference*, pages 1448–1456. IEEE, 1996.

[162] A. P. Waller and J. Ladbrook. Virtual worlds: experiencing virtual factories of the future. In *Proceedings of the 2002 Winter Simulation Conference*, pages 513–517. Winter Simulation Conference, 2002.

[163] G. F. Watson. Mil reliability: a new approach. *IEEE Spectrum*, 29:46–49, August 1992.

[164] P. D. Welch. The statistical analysis of simulation results. In S. S. Lavenberg, editor, *The Computer Performance Modeling Handbook*, Proceedings of the Society of Photo-Optical Instrumentation Engineers (Spie). Academic Press, New York, 1983.

[165] K. P. White Jr. An effective truncation heuristic for bias reduction in simulation output. *Simulation*, 69(6):323–334, 1997.

[166] K. P. White Jr., Michael J. C., and Stephen C. S. A comparison of five steady-state truncation heuristics for simulation. In *Proceedings of the 2000*

Winter simulation Conference, pages 755–760, San Diego, CA, USA, 2000. Society for Computer Simulation International.

[167] W. T. Williams and J. M. Lambert. Multivariate methods in plant ecology i. association analysis in plant communities. *Journal of Ecology*, 47:83–101, 1959.

[168] J. R. Wilson and A. A. B. Pritsker. A survey of research on the simulation startup problem. *Simulation*, 31(2):55–58, 1978a.

[169] J. R. Wilson and A. A. B. Pritsker. Evaluation of startup policies in simulation experiments. *Simulation*, 31(3):79–89, 1978b.

[170] D. Wishart. An algorithm for hierarchical classifications. *Biometrics*, 25(1):165–170, 1969.

[171] L. Yan and J. R. English. Economic cost modeling of environmental-stress-screening and burn-in. *IEEE Transactions on Reliability*, 46(2), 1997.

[172] E. Yucesan. Randomization tests for initialization bias in simulation output. *Naval Research Logistics*, 40:643–663, 1993.