BRANDO
BReakpoint Analysis with Nonparametric Data Option

E.J. Emond
Central OR Team

A.E. Turnbull
Central OR Team

DRDC CORA TM 2006–40
November 2006

Defence R&D Canada
Centre for Operational Research and Analysis

Central OR Team

National Defence Défense nationale

Canada

BRANDO
BReakpoint Analysis with Nonparametric Data Option

E.J. Emond
Central OR Team

A.E. Turnbull
Central OR Team

DRDC – Centre for Operational Research and Analysis

Technical Memorandum

DRDC CORA TM 2006–40

November 2006

Author

A.E. Turnbull

Approved by

Paul Massel
Central Operational Research Team Leader

Approved for release by

Jocelyn Tremblay
Chief Scientist - DRDC CORA

Abstract

This paper reports on the mathematical details and software solution to the Operational Research problem of comparing the central values of multiple datasets using nonparametric statistics. The work is an extension of the iterative post-hoc analysis method described by Emond and Massel (2003). The original methodology required an assumption of Normality for all of the datasets, an assumption that is often not met. An additional analysis method has now been developed in which the data are ranked allowing for a nonparametric solution. Using a proxy for the likelihood function on the ranked datasets, this method finds the most likely separation points between them after a nonparametric analysis of variance has indicated that differences exist. In addition to the theoretical methodology and explanation, examples are given to demonstrate the practicality of the process. Of particular interest is an example which illustrates the application of the methodology to the analysis of ordered categorical data often found in surveys.

Résumé

Ce rapport révèle les détails mathématiques et la solution logicielle au problème en recherche opérationnelle que représente la comparaison des valeurs centrales entre elles des multiples ensembles de données à l'aide de statistiques non paramétriques. Ces travaux sont un prolongement d'une méthode d'analyse itérative ultérieurement décrite par Emond et Massel (2003). La méthodologie originale prévoyait l'établissement d'une hypothèse de normalité pour tous les ensembles de données, une hypothèse qui n'est pas souvent vérifiée. Une autre méthode d'analyse a été élaborée dans laquelle les données sont classées en tenant compte d'une solution non paramétrique. En utilisant un calcul par approximation pour la fonction de vraisemblance des ensembles de données classés, cette méthode établit les points de séparation les plus probables entre ces ensembles de données lorsque cette différence a été démontrée par une analyse non paramétrique de la variance. Ce rapport présente, en plus de la méthodologie et l'explication théoriques, des exemples pour établir la valeur concrète du processus. Un exemple d'un intérêt particulier illustre l'application de la méthodologie à l'analyse de données nominales ordonnées que l'on retrouve souvent dans des enquêtes.

This page intentionally left blank.

Executive summary

Background

Operational Research analysts often require a simple method to analyze multiple datasets. To assist in this problem the Central OR Team has developed a methodology for general use and has also translated it into the form of a software tool called BRANDO. The methodology was initially developed in response to a requirement to determine whether differences between sets of results observed in operational research studies were statistically significant. Existing statistical procedures were inconclusive because they sometimes gave intransitive results. In addition there were circumstances where existing procedures were not able to find differences despite initial tests indicating differences existed.

Previous work on this subject has been detailed for both a Normal distribution of data (Emond and Massel [1]) as well as a Binomial distribution (Hunter and Emond [2]).

Principal results

A solution was found to the problem of sorting a nonparametric distribution of data. This solution allows for non-normal dataset distributions by ranking the data rather than analysing them in their raw form. The initial step is to rank the data within the K datasets. The modified data is then tested against the null hypothesis that the K datasets are all from the same distribution versus the alternate hypothesis that one or more of the datasets is from a different distribution than the remainder of the datasets.

If the null hypothesis is accepted we conclude that there is insufficient evidence to declare that any of the datasets differs significantly from the others, and if the null hypothesis is rejected we conclude that at least one of the datasets differs significantly from the others. When this is the case a post-hoc procedure is employed, a term which refers specifically to the problem of determining how to best split the group of K datasets into subgroups in which the datasets are not significantly different. The subgroups are then looked at individually with the null hypothesis test and, if necessary, the post-hoc procedure is applied. This process is repeated until the null hypothesis has been accepted for all subgroups.

Significance of results

The nonparametric model for analysing the non-normal data described in this paper has very desirable properties: unlike the Normal model there is no requirement for an assumption of normality within the datasets; it ensures that there are no intransitive results; and the methodology is easy to use and understand. The software tool adds to this ease of use by removing any need for the analyst to perform the many small calculations.

E.J. Emond, A.E. Turnbull; 2006; BRANDO
BReakpoint Analysis with Nonparametric Data Option; DRDC CORA TM 2006–40;
DRDC – Centre for Operational Research and Analysis.

Sommaire

Contexte

Les analystes en recherche opérationnelle requièrent souvent une méthode d'analyse simple pour les multiples ensembles de données. Pour les aider à régler ce problème, l'équipe centrale en RO a élaboré une méthodologie d'usage général et a transformé celle-ci en logiciel-outil, appelé BRANDO. Cette méthodologie a d'abord été élaborée en réponse au besoin de déterminer la valeur statistique des différences entre des ensembles de résultats d'études de recherche opérationnelle. Les méthodes statistiques actuelles étaient peu concluantes en raison de leurs résultats intransitifs. En outre, par moment les méthodes existantes ne permettaient pas de reconnaître les différences malgré le fait que les premières analyses les avaient soulevées.

Les travaux antérieurs sur le sujet ont été exposés en détail dans la Distribution normale de données (Emond et Massel [1]) et dans la Distribution binomiale (Hunter et Emond [2]).

Principaux résultats

Le problème de classement d'une distribution non-paramétrique de données a été résolu. Cette solution permet les distributions non normales des ensembles de données en classant les données plutôt qu'en les analysant sous leur forme brute. La première étape consiste à classer les données dans les ensembles de données K. Les données modifiées sont ensuite confrontées à une hypothèse nulle voulant que les ensembles de données K proviennent tous de la même distribution et à l'hypothèse alternative voulant qu'un ensemble ou plus de données provienne d'une distribution différente des autres ensembles de données.

Si l'hypothèse nulle est vérifiée, nous concluons que les preuves sont insuffisantes pour affirmer qu'un ensemble de données quelconque diffère particulièrement des autres, et si l'hypothèse nulle est rejetée, nous concluons qu'au moins un des ensembles de données diffère particulièrement des autres. Si tel est le cas, une méthode ultérieure est employée et cherche particulièrement à déterminer la meilleure façon de séparer le groupe d'ensembles de données K en sous-groupes dans lesquels les ensembles de données ne seraient pas particulièrement différents. Les sous-groupes sont ensuite évalués individuellement à l'aide du test de l'hypothèse nulle et, si nécessaire, la méthode ultérieure est utilisée. Cette procédure est répétée jusqu'à ce que l'hypothèse nulle soit vérifiée pour tous les sous-groupes.

L'importance des résultats

Le modèle non paramétrique d'analyse de données non normales présenté dans ce rapport possède des particularités tentantes : contrairement au modèle normal, l'hypothèse de normalité dans les ensembles de données n'est pas une exigence ; le modèle permet de garantir aucun résultat intransitif ; et finalement, il est facile à utiliser et à comprendre. Le logiciel-outil ajoute à cette convivialité en évitant à l'analyste de nombreux calculs.

E.J. Emond, A.E. Turnbull; 2006; BRANDO
BReakpoint Analysis with Nonparametric Data Option; DRDC CORA TM 2006–40;
RDDC – Centre d’analyse et de recherche opérationnelle.

This page intentionally left blank.

Table of contents

Tables

Figures

This page intentionally left blank.

1 Introduction

This paper reports on recent improvements to an operational research (OR) tool developed by the Central OR Team of the Centre for Operational Research and Analysis. The tool was initially developed in response to a requirement to determine whether differences between sets of results observed in operational research studies were statistically significant. Existing statistical procedures were problematical for OR purposes because they sometimes gave intransitive results. In addition there were circumstances where existing procedures were not able to find differences despite initial tests indicating differences existed.

The first report on this work, "A Post-ANOVA Methodology for Finding Subgroups with Equal Means Using Maximum Likelihood" [1], was published in June, 2003. It documented an iterative method combining analysis of variance with a maximum likelihood function to find separation points between Normally distributed datasets. Additional work performed on this tool incorporated a binomial data model. This was documented the following year in "Separation Point Analysis Method" (SPAM) [2].

The SPAM tool was successfully used in several studies but the assumption of normality was found to be unsustainable in many cases. Therefore a nonparametric version of the tool was developed and is documented in this report. The GUI-based software tool has been renamed BRANDO which stands for BReakpoint Analysis with Nonparametric Data Option. The methodology described in this report is able to handle a variety of non-normal data as well as ordered categorical data. This added capability enables the use of BRANDO in the analysis of survey data. An example from a recent survey is provided in this report.

The general problem at which the BRANDO software is aimed was described in reference [2] from which the following excerpt is taken.

"*An important quantitative problem in operational research is the comparison of differences between group means in the presence of sampling error or other random effects. Such an analysis is composed of two separate parts. First it must be determined that the observed differences between the group averages are unlikely to be due to chance. This is done by applying formal statistical hypothesis testing. The null hypothesis is that the groups have a common mean and that the observed differences are due to chance variation. The alternate hypothesis is the logical negation of this, namely that there are at least two subgroups with different means. If the null hypothesis is accepted, the analyst concludes that there is insufficient evidence to reject the possibility that the observed differences between the group means are due to chance. Other than calculating the overall average, the analysis is complete.*

"*In the case that the null hypothesis is rejected, the analyst is now faced with determining which subgroups have different means. ...*

"*Most of the commonly used pairwise-comparison procedures, including the Tukey HSD (honest significant difference) test, are perplexing for the practitioner because they produce intransitive results in many cases. As a simple example of intransitivity consider a case* *with three groups labelled A, B and C with population means* μ_A, μ_B, *and* μ_C *respectively. We assume for convenience in the discussion that the sample means are ordered and that the null hypothesis of common means has been rejected. Because the test that indicates that the null hypothesis should be rejected is independent of the post hoc analysis, a common result of the post hoc analysis of this case when using pairwise comparisons is that the equalities* $\mu_A = \mu_B$ *and* $\mu_B = \mu_C$ *may be indicated while simultaneously the equality* $\mu_A = \mu_C$ *is rejected. Although this phenomenon is understandable for a theoretical / statistical point of view, it leaves the analyst in the uncomfortable position of reporting an intransitive result.*

"*Other statistical methods that avoid pairwise comparisons may also be difficult for the operational research analyst because of another problem. ... it may happen that despite rejecting the null hypothesis of equal means, the post hoc procedure fails to indicate any differences. This again leaves the analyst in an uncomfortable position of stating that there are differences but being unable to determine what they are.*

"*The methodology used in this report has been developed with the goal of providing the operational research analyst with a simple method for post hoc analysis that is statistically sound and above all practical. This means that it must be easy to understand and apply and must give results that avoid the problems of intransitivity and insensitivity noted above.* "

1.1 Aim

The aim of this report is to document the theoretical basis of the nonparametric components of BRANDO and to give guidelines for the appropriate use of the Normal, Binomial, and nonparametric data models. Detailed examples will be given to illustrate the appropriate use of the nonparametric model.

2 Methodology

This section describes the methodology for testing the hypothesis of equality for several datasets and for splitting a group of datasets into their most likely subgroups. An example of the methodology with numbers is found in section 4 (Examples).

2.1 The Nonparametric Data Model

The data to be analyzed consists of K datasets, where K is a positive integer greater than 1. We will index the K datasets with the letter k so that k is an integer index ranging from 1 to K.

Dataset k consists of N_k values where N_k is a positive integer. We will index the N_k values in each dataset with the letter i so that i is an integer index ranging from 1 to N_k. The individual data values in dataset k are denoted as x_{ki}.

Let N_{tot} be the total number of observations in the K datasets:

$$N_{tot} = \sum_{k=1}^{K} N_k \tag{1}$$

In the nonparametric model we make no distributional assumptions about any of the individual datasets. The basic idea is to use the rank order of the values instead of the values themselves. This allows us to perform a mathematically sound analysis while avoiding assumptions which may not be warranted in specific cases. In particular we do not need to assume Normality.

2.2 Kruskal-Wallis Nonparametric ANOVA by Ranks

We begin by testing the null hypothesis that the K datasets are identically distributed versus the alternate hypothesis that at least one of the datasets is different. We are particularly interested in differences in the central values.[1]

H_O: All datasets identically distributed
H_A: At least one dataset is different

The test of hypothesis is based on the average ranks of the values in the K datasets. Under the null hypothesis, all the observations come from the same distribution so that the average rank of the values in each dataset should be about the same. The extent to which the average ranks differ from their expected value is an indicator of whether or not the null hypothesis is true.

[1] Differences in distribution around the central value may also be of interest but would require a different statistical analysis.

We first consider the N_{tot} observations from the K datasets as one group. We rank all the observations from lowest to highest and associate each observed value x_{ki} with its rank. Tied observations are given the average rank of the tied group. We then compute the average rank associated with the N_k values in each of the K datasets.

More specifically, let r_{ki} denote the rank value associated with observation x_{ki} when all N_{tot} observations are considered as one group. The smallest observation is given a rank value of 1 and the largest observation is given a rank value of N_{tot} . Let $r_{k.}$ be the average rank of the N_k values in dataset k.

$$r_{k.} = \frac{1}{N_k}\sum_{i=1}^{N_k} r_{ki} \tag{2}$$

Under the null hypothesis we expect that all K average rank values will be approximately equal to $r_{..}$.

$$r_{..} = \frac{N_{tot}+1}{2} \tag{3}$$

The Kruskal-Wallis test statistic, H, is defined as follows:

$$H = \frac{12}{N_{tot}(N_{tot}+1)}\sum_{k=1}^{K} N_k(r_{k.} - r_{..})^2 \tag{4}$$

A Monte-Carlo simulation will be used to evaluate the distribution of the test statistic under the null hypothesis. Therefore we can simplify our calculations by dropping the multiplicative constant in Equation 4. We redefine the Kruskal-Wallis test statistic as:

$$H' = \sum_{k=1}^{K} N_k(r_{k.} - r_{..})^2 \tag{5}$$

The null hypothesis is rejected for large values of the test statistic in accordance with the discussion above.

Let H* be the value of the test statistic obtained from the average ranks of the input data. In order to evaluate the distribution of the test statistic under the null hypothesis we use a Monte-Carlo simulation. For each iteration we select K groups of size N_1 , N_2 , ... , N_K , at random from the N_{tot} rank values r_{11} , r_{12} , ... and calculate the corresponding value of H′ . For a given level of Type I error[2], α, we reject the null hypothesis if the proportion, p, of Monte-Carlo test statistic values greater than H* is less than α.

[2] See discussion of Type I error below.

2.3 Post-Hoc Analysis for the Nonparametric Model

If the null hypothesis is accepted we conclude that there is insufficient evidence to declare that any of the datasets differs significantly from the others at the chosen value of Type I error.

If the null hypothesis is rejected, we conclude that at least one of the datasets differs significantly from the others. We must therefore employ a post-hoc procedure, a term which refers specifically to the problem of determining how to split the group of K datasets into subgroups in which the datasets are not significantly different.

There are several post-hoc procedures in the statistical literature. A popular choice is Tukey's Honest Significant Difference method. The problem with all of the usual post-hoc procedures from an Operational Research perspective is that the more conservative procedures may not always give an answer while others may give an intransitive answer.

The following post-hoc procedure has been developed to get around these problems. The procedure is statistically sound in that it maintains the desired probability of Type I error or false alarm rate[3]. Intransitivity is avoided by first ordering the datasets with respect to the parameter of interest and only considering breakpoints which preserve this order. We guarantee an answer in every case by using a maximum likelihood procedure to find the most likely breakpoint within the ordered datasets whenever the initial analysis of variance test indicates that at least one breakpoint exists.

2.4 BRANDO: BReakpoint Analysis with Nonparametric Data Option

Logically, the minimum constraint that rejection of the null hypothesis places on the post-hoc procedure is that there must be at least two subgroups within the group of K datasets. Let the K datasets be ordered and relabelled by increasing value of the average rank calculated for the Kruskal-Wallis test which resulted in the null hypothesis being rejected. We will consider the K-1 possible ways to break the K datasets into exactly two subgroups while maintaining the order of the datasets. The first breakpoint is between datasets 1 and 2, forming two subgroups consisting of dataset 1 by itself, and datasets 2 through K. The second breakpoint is between datasets 2 and 3, forming two subgroups consisting of datasets 1 and 2 together and datasets 3 through K. The $(K-1)^{th}$ breakpoint is between datasets K-1 and K, forming two subgroups consisting of datasets 1 through K-1 and dataset K by itself.

In contrast to the Normal and Binomial data models, we cannot use the maximum likelihood method to determine the most likely breakpoint in the nonparametric case because we do not have an explicit likelihood equation. Therefore we will use the Kruskal-Wallis test statistic to determine the most likely breakpoint.

The Kruskal-Wallis test statistic given above measures the extent to which a group of

[3] See discussion of Type I error below.

datasets have the same average rank. The test statistic is zero only when all the datasets have exactly the same average rank. The test statistic increases monotonically as the weighted sum of the squared differences between the dataset average ranks and the overall average rank increases. For any possible breakpoint, we have two resulting datasets. For each of these datasets independently, we calculate the observed p-value of the Kruskal-Wallis test. The product of these two observed p-values is a measure of the relative likelihood of the given breakpoint as compared to the other possible breakpoints. We choose the breakpoint for which this product is maximum.

Note that when there is only one dataset in a group, the observed p-value will be 1 by definition because all datasets (one) will have average rank equal to the overall average rank so that the Kruskal-Wallis test statistic value will be zero and hence the p-value will be 1.

Having divided the K groups into two subgroups, the procedure continues by testing the null hypothesis of identically distributed datasets for each of the subgroups (where appropriate) and using the same methodology to subdivide as outlined above. This procedure continues until the null hypothesis is accepted in all cases.

2.5 Monte-Carlo Simulations

The breakpoint analysis is carried out using a Monte-Carlo simulation to estimate the p-values for each dataset. A problem arises in dealing with large numbers of datasets in that the p-values are sometimes so small that the Monte-Carlo simulation never comes up with any test statistic values which exceed the observed value. Increasing the number of iterations for the Monte-Carlo simulation is not feasible since the required numbers could take virtually forever to run.

In order to get around this problem we can make use of the fact (proven in Annex A) that the Kruskal-Wallis test statistic follows approximately a Gamma distribution. Given the results of a large number of Monte-Carlo iterations, we can use these to fit a Gamma distribution and thereby estimate the area to the right of the observed test statistic. The area under a Gamma distribution to the left of a given value is a standard mathematical function, the Incomplete Gamma. We therefore use the parameters of our fitted Gamma distribution from the Monte-Carlo results to evaluate the Incomplete Gamma function and subtract this value from 1 to get the required p-value. Mathematical details of this procedure are given in Annex A.

2.6 Probability of Type I Error

In order to understand why the BRANDO methodology is statistically sound, consider the probability of Type I error: rejecting the null hypothesis when it should have been accepted. Type I error is also known as the false alarm rate. When comparing multiple datasets using a pairwise approach, a fundamental problem is that as the number of pairwise comparisons increases, the probability of false alarms also increases. A Bonferroni procedure can be used to decrease the sensitivity of the pairwise tests to maintain a given level of Type I error overall, but this may lead to undesirable increases in Type II error - accepting the null hypothesis when it should have been rejected. Therefore, the best statistical practice is to initially perform an omnibus test of the null hypothesis (that all the datasets are from the same parent distribution or have the same central value). If the null hypothesis is rejected, one of the many so-called post hoc procedures is then applied to find which of the datasets should be grouped together.

This procedure is basic to the BRANDO methodology. An initial omnibus test is applied to test the null hypothesis that all the datasets are the same or at least have the same central value. Only if this hypothesis is rejected does BRANDO continue on to find the most likely way to arrange the datasets into subgroups. Therefore BRANDO maintains the false alarm rate or Type I error rate at the desired level and only identifies breakpoints when the initial test so indicates.

In addition, BRANDO initially finds only the single most likely breakpoint. This is because the rejection of the original null hypothesis (that all datasets are the same) logically implies only that there is at least one breakpoint. Having broken the group of datasets into two separate subgroups, BRANDO iteratively re-applies the omnibus test of hypothesis to both of the subgroups.

In order to see how this iterative application of the omnibus test of the null hypothesis preserves the desired false alarm rate, consider a case where K datasets are independently drawn from the same distribution and suppose we set a false alarm rate (probability of Type I error) at 5 percent or 0.05. In the one case in 20 that BRANDO rejects the null hypothesis, an initial breakpoint will be found creating two subsets. However, when the two subsets are tested, each will independently have a false alarm rate of at most 0.05 so that BRANDO will only find multiple breakpoints with probability at most 0.004875. In practice, multiple breakpoints occur less frequently based on a Monte-Carlo evaluation using 10 datasets from a standard Normal distribution. In 1000 iterations with a 95% level of significance, 51 false alarms were found, none of which gave more than a single break point. Thus the iterative testing procedure ensures that the false alarm rate in BRANDO is at the desired level and in particular the probability of finding more than one breakpoint when none actually exists is vanishingly small.

3 Model Selection

This section explains when each model should be used. There is a description of the assumptions and requirements that are made for all the model types and some examples of each are given. A more extensive set of examples, that include numerical values, of the Normal and nonparametric models can be found in section 4.

3.1 When to Use the Normal Model

The Normal model assumes N_k elements within K subsets where K is greater than or equal to 2. It is also required that the values within each subset are approximately normally distributed. The Normal model works well as long as the datasets are unimodal and there are no outliers. BRANDO applies the D'Agostino K^2 test for Normality whenever the Normal data model is chosen.

The Normal model may be used if all K datasets under consideration pass the omnibus D'Agostino K^2-test for Normality. This test is the preferred omnibus test for Normality when there is no prior information about the alternative distribution. The term omnibus refers to the testing of the shape of the distribution for non-normality in terms of both skewness and kurtosis. To accommodate all sample sizes, the measures of skewness and kurtosis are transformed using functions sensitive to the number of observations. See references [3, 4] for details.

It is difficult to test small datasets for Normality because moderate numbers of observations are required to determine if shape properties such as skewness and kurtosis are significantly different from those expected with a Normal distribution. Therefore it is recommended that small datasets be analyzed in BRANDO using the nonparametric model which has no size restrictions. More specifically, if any of the K datasets has 7 or fewer observations then the nonparametric model should be used. Note that the omnibus D'Agostino K^2-test for Normality breaks down with 7 or fewer observations meaning that BRANDO will not test any dataset for Normality unless it has a minimum of 8 observations.

A Normal data distribution often results from replicated measurements from independent trials. This includes the example of an open-fire range at a war game where several (N_k) replications are performed for each game, and each of K equipment options is tested in the game.

3.2 When to Use the Binomial Model

The Binomial model is appropriate for data that are proportions (e.g. 23 out of 50) where there are at least two datasets under consideration. The observations must be independent. The purpose of the analysis is to determine whether or not the proportions can be considered equal for all datasets.

An example of the use of the Binomial model is the evaluation of various sensor equipment types within a detailed computerized war game. The resulting data are the proportion of total targets that each sensor type detected. See reference [2] for more details and examples of the use of the Binomial model.

3.3 When to Use the Nonparametric Model

Since no distribution assumptions are made the nonparametric model is universally applicable. The disadvantage of the nonparametric approach is that the analysis will be less sensitive to small differences relative to the Normal model. The nonparametric model assumes the same data parameters as the Normal model with the exception that the assumption of normality is not required.

In addition since the nonparametric model uses only rank data it can be used to analyse ordered categorical data from surveys. An example of this is given in the Examples section.

3.4 Decision Flow Diagram for the Models

To better summarize the choices provided by BRANDO a decision flow diagram of the more pertinent points of the model types is shown in Figure 1 below.

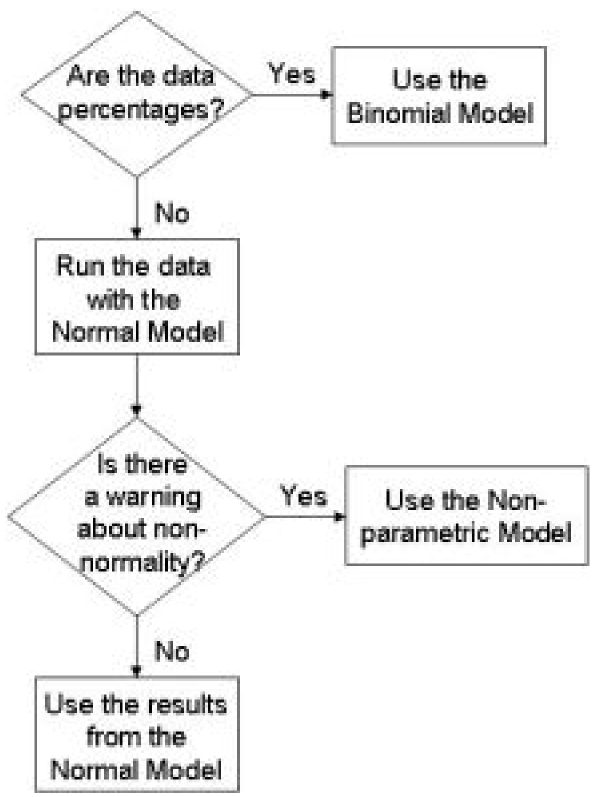

Figure 1: *A decision flow diagram for all the model types.*

4 Examples

This section outlines several numeric examples that demonstrate the use of BRANDO.

4.1 Example of Nonparametric Analysis

The data in this example were collected from an experiment on the effectiveness of equipment that was being analysed for acquisition by the Canadian Department of National Defence. A total of 5 equipment types was tested and the observations consisted of the number of targets successfully defeated by each. The problem was to determine whether the effectiveness of the equipment types differed significantly.

In Table 1 the columns are listed as A, B, C, D, and E representing the five equipment types under consideration. For each equipment type there are 20 values which represent the number of targets successfully defeated in 20 independent iterations of a computerized war game.

Table 1: *Data for Example 1.*

A	B	C	D	E
22	24	24	24	24
21	22	24	24	24
25	23	25	21	25
21	23	24	22	24
20	23	25	18	23
22	22	22	21	21
19	24	24	25	24
23	24	25	25	25
17	22	24	21	25
23	20	24	20	25
23	23	24	22	25
21	22	23	24	24
19	23	22	16	24
22	23	24	23	24
22	24	23	22	24
23	19	23	23	24
18	18	24	23	24
24	24	24	23	24
19	24	23	24	24
20	23	25	19	25

Two of the datasets were found to be non-normal when tested with the D'Agostino K^2 test. Specifically datasets B and E were found to be non-normal, as is more readily seen from the histograms in Figure 2.

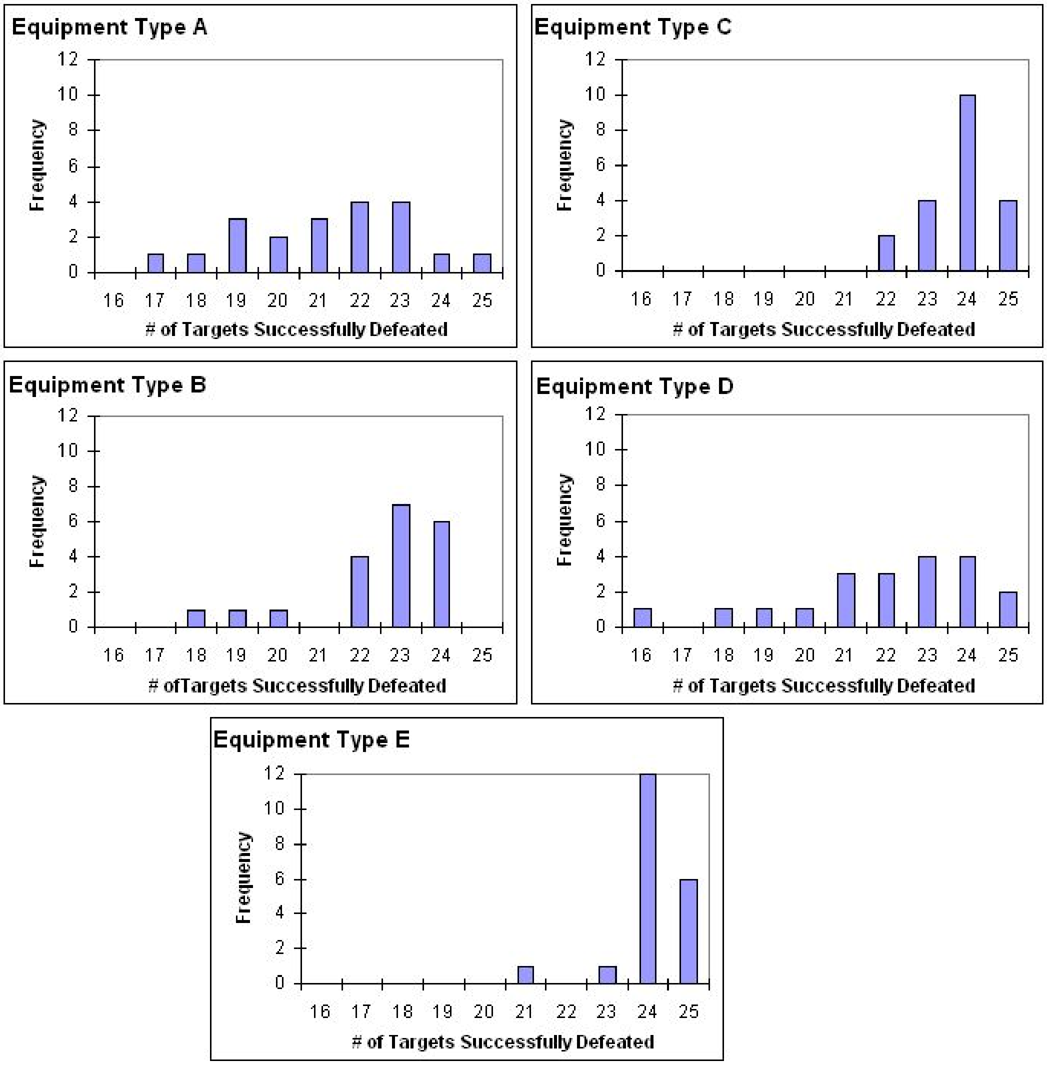

Figure 2: *Histogram of the all the Equipment Types.*

The question to be answered is whether or not there is sufficient evidence to indicate that there are differences between the effectiveness of the equipment types. In statistical terms it is necessary to either accept or reject the null hypothesis that all the datasets are statistically equivalent and any differences in their effectiveness are attributable to chance variation. If the null hypothesis is rejected then the datasets must be divided into subgroups. The division between the groups is determined by finding the most likely breakpoint between the sets when ordered by their average rank value.

The analysis is more easily performed by the graphical user interface (GUI) software but is reviewed in this report so that the reader has a better understanding of the process. The GUI's modifiable parameters were set to 5,000 iterations for the Monte Carlo test of the null hypothesis (H_o) and a 95% significance level.

As defined in the section on Methodology, this nonparametric data example has the following parameters:

$$N_{tot} = \sum_{k=1}^{K} N_k \tag{6}$$

As N_k is equal to 20 for each dataset and there are K = 5 datasets then N_{tot} is equal to 100.

The first step is to rank the data from Table 1 where all equal values are given the same rank, a number that is the mean rank of the tied group. As an example a dataset of:

< 19 23 22 16 24 18 18 24 23 24 >

would generate the following values for each datapoint's rank:

< 4.0 6.5 5.0 1.0 9.0 2.5 2.5 9.0 6.5 9.0 >

All 100 values were ranked among themselves. They were then sorted back into the original equipment type datasets. The resulting ranked data are shown in Table 2. The five columns are labelled with the equipment types as in the previous table. The table also includes the average rank of the 20 items in each equipment type dataset at the bottom of each column.

Under the null hypothesis we expect that all 5 average rank values will be approximately equal to $r_{..}$ which is defined below.

$$r_{..} = \frac{N_{tot}+1}{2} = \frac{100+1}{2} = 50.5 \tag{7}$$

The value of the test statistic, using equation 5, is given as 28,581.0.

A total of 5,000 similar calculations are run with the values within each column being randomly distributed. This Monte Carlo simulation produces a large number of results against which our test statistic is compared.

Table 2: *The Ranked Data, and Dataset Average Ranks.*

A	B	C	D	E
28.0	71.0	71.0	71.0	71.0
18.0	28.0	71.0	71.0	71.0
94.0	44.5	94.0	18.0	94.0
18.0	44.5	71.0	28.0	71.0
12.5	44.5	94.0	4.0	44.5
28.0	28.0	28.0	18.0	18.0
8.0	71.0	71.0	94.0	71.0
44.5	71.0	94.0	94.0	94.0
2.0	28.0	71.0	18.0	94.0
44.5	12.5	71.0	12.5	94.0
44.5	44.5	71.0	28.0	94.0
18.0	28.0	44.5	71.0	71.0
8.0	44.5	28.0	1.0	71.0
28.0	44.5	71.0	44.5	71.0
28.0	71.0	44.5	28.0	71.0
44.5	8.0	44.5	44.5	71.0
4.0	4.0	71.0	44.5	71.0
71.0	71.0	71.0	44.5	71.0
8.0	71.0	44.5	71.0	71.0
12.5	44.5	94.0	8.0	94.0
28.2	43.7	66.0	40.675	73.925

In this case the test statistic was much greater than any of the values derived from the Monte Carlo iterations and therefore the null hypothesis is rejected. Mathematically a rejection is determined to occur when the p-value, or percentage of Monte Carlo iterations greater than the test statistic, is less than the significance level of 0.05. As the p-value is calculated to be 0 the null hypothesis is rejected and it is concluded that there is sufficient evidence to determine that there are differences between the performance of the five equipment types.

After the null hypothesis has been rejected it is necessary to determine the point in the data at which a break is most likely to occur. For each possible breakpoint the two resulting subgroups of datasets are analysed using the BRANDO methodology and then the most likely breakpoint is selected. After sorting the datasets in order of their average rank the four possible breakpoint groupings and their likelihoods are listed in Table 3.

As the third value is the largest, a breakpoint between B and C is therefore most likely. Therefore the two subgroups are now comprised of A-B-D and C-E.

The new null hypothesis is then tested on the both of the subgroups. A p-value of 0.072 is found for the first grouping of A, B, and D, and a p-value of 0.206 is calculated for datasets C and E. As the p-value for both subgroups is found to be greater than 0.05 the null hypothesis for both is accepted and a final graphed solution is shown in Figure 3.

Table 3: *Likelihood values for the 4 possible breakpoints.*

Breakpoint	Subgroup1	Subgroup2	Likelihood value
1	A	D, B, C, E	0.0000
2	A, D	B, C, E	0.0002
3	A, D, B	C, E	0.0211
4	A, D, B, C	E	0.0001

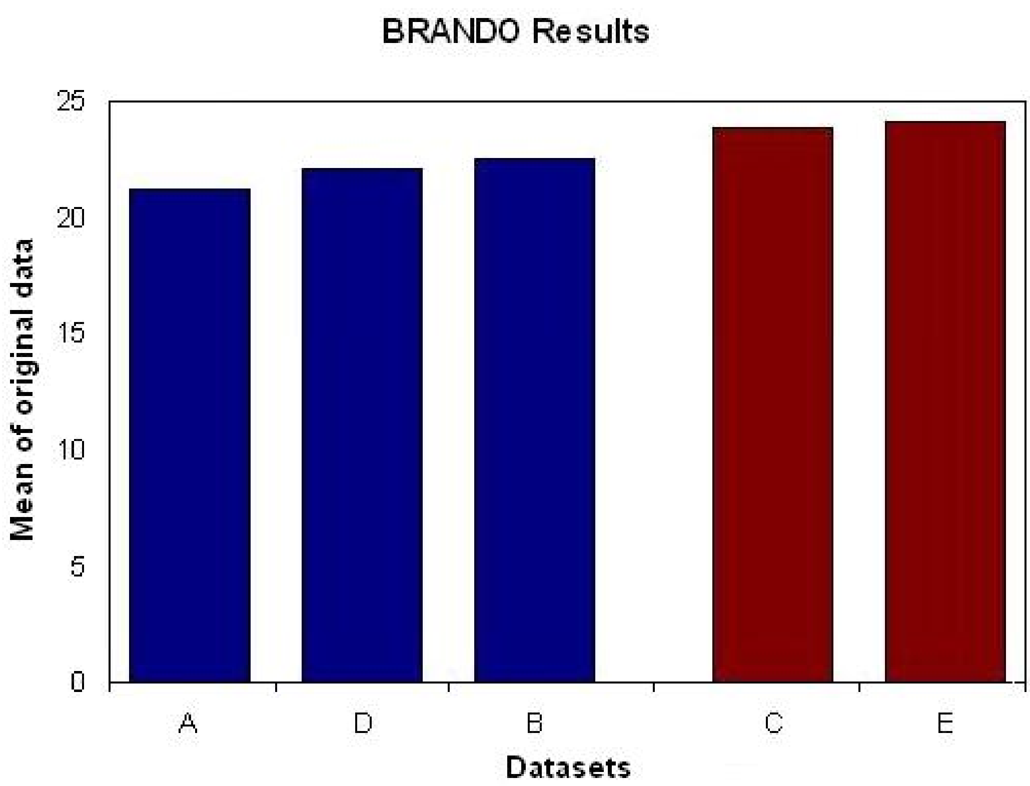

Figure 3: *Final graph of BRANDO results.*

In summary the analysis indicates that the effectiveness ratings of the five equipment types are not the same. There is insufficient evidence to indicate any difference between the C and E equipment types. There was also insufficient evidence to indicate any difference between the A, B, and D equipment types. However the two subgroups have a significantly different response distribution from each other at the 5 percent significance level [4].

4.2 More Complex Nonparametric Analysis

This example provides a similar analysis to the previous example but it is hoped that the more substantial and complex data included here will provide insight to the reader on the potential for the powerful analysis that is available with the nonparametric model.

A total of 13 vehicles were tested for their ability to complete a task. Each task was performed 20 times. The problem was to determine whether the performance of the vehicle types differed significantly.

Of the 13 vehicles that were tested for their performance, two were found to have test data that were non-normal. Figure 4 gives the histograms of these datasets and shows the outlier values.

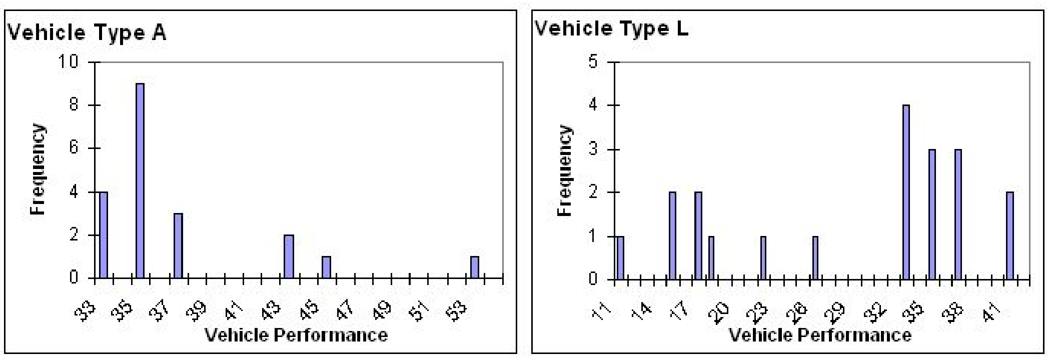

Figure 4: *Histogram of the Non-normal Data for Vehicle Performance.*

The results of each test are reported in Table 4. The alphabetical columns represent the thirteen vehicle types under consideration.

Two of the datasets, A and L, were found to be non-normal when tested with the D'Agostino K^2 test. Therefore, use of the nonparametric version of BRANDO is indicated.

In statistical terms, we wish to test the null hypothesis that the performance of the thirteen vehicles is about the same and that the observed differences in the response pattern are attributable to chance variation. We test this hypothesis by ranking all the test results and

[4] We reject the hypothesis that the two subgroups have equal response distributions with a false alarm or Type I error rate of 5 percent. ($\alpha = 0.05$)

Table 4: *A Table of the Raw Data for 13 Vehicle Options.*

A	B	C	D	E	F	G	H	I	J	K	L	M
37	75	97	66	78	65	78	53	49	35	43	15	15
33	59	55	33	17	19	17	45	54	54	39	18	52
53	63	61	47	45	47	45	53	43	55	43	41	51
35	95	65	77	59	77	59	72	59	57	56	26	56
45	47	59	45	15	17	15	15	43	67	43	35	43
43	65	61	33	27	33	27	47	23	14	19	37	14
35	47	43	11	11	11	11	11	43	11	14	11	11
33	33	33	35	33	35	33	33	33	41	33	33	41
43	53	33	47	51	39	51	53	33	33	33	15	18
35	61	45	56	61	43	61	57	33	33	45	17	35
35	61	57	31	59	47	59	43	55	30	43	37	11
33	93	62	54	73	33	73	59	41	41	54	41	11
35	59	53	59	59	63	59	59	51	33	51	33	33
35	35	59	41	35	35	35	35	13	12	13	35	12
35	57	79	57	56	69	56	58	59	57	58	22	57
33	93	83	55	59	33	69	59	58	46	31	37	31
37	53	49	41	37	33	37	37	61	33	35	33	29
37	47	43	15	47	20	47	47	63	18	37	35	37
35	62	71	67	55	61	55	57	56	42	43	17	35
35	75	103	71	67	71	67	67	63	45	61	33	55

calculating the average rank of each. Under the null hypothesis each of the average rank values should be about the same.

The value of the Kruskal-Wallis test statistic was calculated to be 372,593.38. Based on 5,000 Monte-Carlo simulation iterations the probability under the null hypothesis of observing a value as large as 372,590 with the given data is less than 0.0001. We therefore reject the null hypothesis and conclude that there is sufficient evidence to declare differences between all of the vehicle types in terms of the tests performed.

The next step is to find the most likely breakpoint between the vehicle types when ordered by their average rank values. With 13 datasets there are twelve possible breakpoints to be considered. Using the BRANDO post-hoc methodology Table 5 gives the 'likelihood' values for the possible breakpoints.

Since Breakpoint 6 has the highest likelihood value, it is selected as the first breakpoint.

The iterative process continues by considering the two subgroups created by the first breakpoint. The first subgroup consists of datasets L, M, J, A, F, and K and the second consists of datasets I, D, H, E, G, C, and B. We rank the data within each subgroup once again and test the null hypothesis for both subgroups.

Table 5: *Likelihood values for the 12 possible breakpoints.*

Breakpoint	Subgroup1	Subgroup2	Likelihood value
1	L	M, A, J, K, F, D, I, E, G, H, C, B	0.0000
2	L, M	A, J, K, F, D, I, E, G, H, C, B	0.0000
3	L, M, A	J, K, F, D, I, E, G, H, C, B	0.0000
4	L, M, A, J	K, F, D, I, E, G, H, C, B	0.0002
5	L, M, A, J, K	F, D, I, E, G, H, C, B	0.0010
6	L, M, A, J, K, F	D, I, E, G, H, C, B	0.0021
7	L, M, A, J, K, F, D	I, E, G, H, C, B	0.0002
8	L, M, A, J, K, F, D, I	E, G, H, C, B	0.0000
9	L, M, A, J, K, F, D, I, E	G, H, C, B	0.0000
10	L, M, A, J, K, F, D, I, E, G	H, C, B	0.0000
11	L, M, A, J, K, F, D, I, E, G, H	C, B	0.0000
12	L, M, A, J, K, F, D, I, E, G, H, C	B	0.0000

The value of the Kruskal-Wallis test statistic for the data in the first subgroup is 12,712 and the probability under the null hypothesis of observing a value as large as this was found to be 0.054 based on 5,000 Monte-Carlo iterations. As a result of the probability being greater than 0.05 the null hypothesis is accepted and it is concluded that there is insufficient evidence to declare that the vehicle types are significantly different for L, M, J, A, F, and K.

The second subgroup was found to have a Kruskal-Wallis test statistic of 21,600 and the probability under the null hypothesis of observing a value as large as this was found to be 0.0394 based on 5,000 Monte-Carlo iterations. As the probability is less than 0.05 the null hypothesis is rejected and it is concluded that there is sufficient evidence to declare that the performance of the vehicle types in this subgroup is significantly different. A further analysis of the most likely breakpoint is required.

With seven vehicle types remaining there are six possible breakpoints to be considered. Table 6 gives the 'likelihood' values for the six possible breakpoints using the BRANDO post-hoc methodology.

Table 6: *Likelihood values for the 6 possible breakpoints.*

Breakpoint	Subgroup1	Subgroup2	Likelihood value
1	I	D, H, E, G, C, B	0.0458
2	I, D	H, E, G, C, B	0.0461
3	I, D, H	E, G, C, B	0.0599
4	I, D, H, E	G, C, B	0.1082
5	I, D, H, E, G	C, B	0.7494
6	I, D, H, E, G, C	B	0.2520

Since Breakpoint 5 has the highest likelihood value it is selected as the second breakpoint.

The iterative procedure continues by considering the two subgroups created by the second breakpoint. The first grouping of I, D, H, E, and G and the second grouping of C and B are ranked once again and the null hypothesis is tested for both distributions.

The value of the Kruskal-Wallis test statistic for the data in the I, D, H, E, and G subgroup is 315.725 and the probability under the null hypothesis of observing a value as large as this was found to be 0.9866 based on 5,000 Monte-Carlo iterations. The value of the Kruskal-Wallis test statistic for the data in the C and B subgroup is 13.225 and the probability under the null hypothesis of observing a value as large as this was found to be 0.7722 based on the same number of iterations.

As a result of the probability being greater than 0.05 for both of the subgroups, the null hypothesis is accepted for both and it is concluded that there is insufficient evidence to declare that the performance of the vehicle types is significantly different for (I, D, H, E, and G) or (C and B). This completes the analysis as there are no further break points to be found. A final graphed solution is shown in Figure 5.

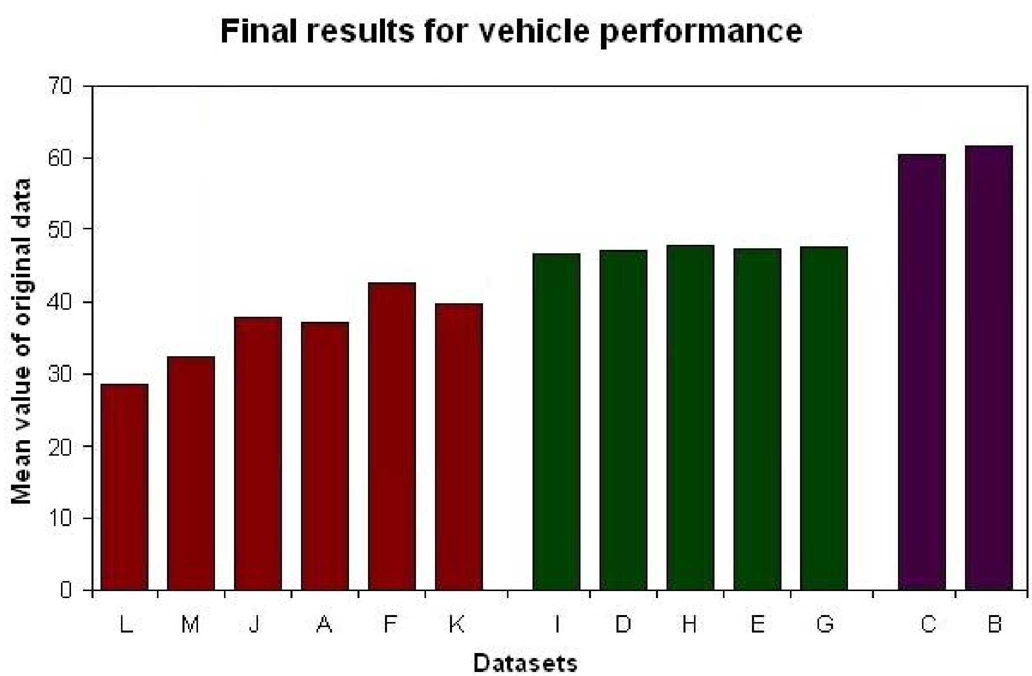

Figure 5: *Final graph of BRANDO nonparametric model results for Vehicle Performance.*

In summary the analysis indicates that the performance ratings are not the same for all thirteen of the vehicle types. The L, M, J, A, F, and K types were found to be equivalent. The I, D, H, E, and G types were equivalent, and the C and B types were also equivalent to each other. However the three subgroups are different from each other at the 5 percent significance level.

4.2.1 Comparison Between the Nonparametric and Normal Models

A comparison between the solution to the nonparametric and Normal models is made to underline the importance of using the correct model.

The values in the datasets for vehicle types A and L are clearly not normal and the assumption of normality with the data has not been met. The results from the Normal model may therefore be erroneous. The Normal model results are reviewed and a comparison is made below to the nonparametric results.

We applied the Normal model [1] to the datasets from Table 4. The results are shown in Figure 6.

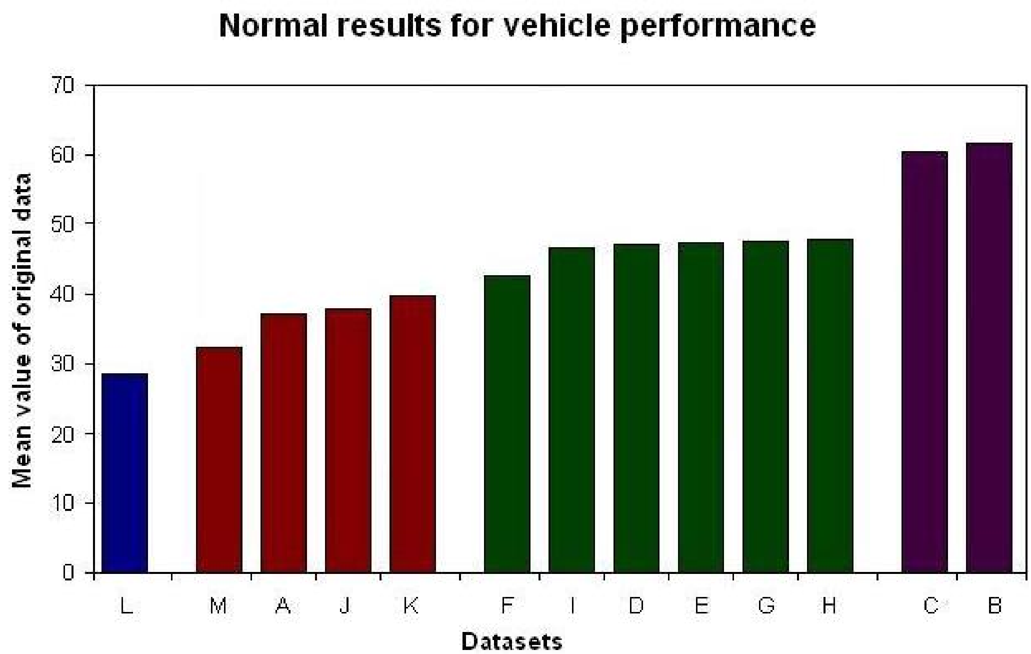

Figure 6: *Final graph of BRANDO Normal model results for Vehicle Performance.*

When these results are compared to the nonparametric model results in Figure 5 there is a distinctive difference in the dataset divisions. Most noticeable is the separation of dataset L from the rest of the datasets. This is a division that highlights the importance of using the nonparametric model with non-normal data. The order and division points of the datasets in the middle are also different but not so distinctive as with dataset L.

In total there are three breakpoints found with the Normal model, in comparison to the nonparametric model's two breakpoints. Dataset L is divided from the rest of the group with the Normal model analysis as a result of its non-normal distribution, an erroneous result that is avoided with the ranking of data. An overview of the dataset divisions from the two models is given below.

Nonparametric Model Datasets:		L M J A F K	I D H E G	C B
Normal Model Datasets:	L	M A J K	F I D E G H	C B

It should be noted that this differentiation could be critical to a decision. If the analyst was testing the fuel consumption component of vehicle performance, where a smaller value is more desirable, there is a definite difference between the results of the two models. The Normal model reports that vehicle L is distinctly different from the other results and further to that there are only four equivalent vehicles in the next group. When the assumption of normality is removed the nonparametric model reports that six vehicles, including L, M, J, A, F, and K, are all equivalent.

It is often the circumstance that the Normal model has more breakpoints when there is a difference between the model results. When the assumption of normality is removed there is less certainty in the ability to divide up the datasets.

4.3 Analysis of Ordered Categorical Data

This example is taken from the analysis of a post-exercise user survey in the Canadian Department of National Defence. A total of 33 participants from four separate organizations in the Canadian Forces involved in Command Post Exercise Virtual Warrior IV, October 2005, responded to a survey question using a 5 point scale ranging from strong agreement to strong disagreement. The problem was to determine whether the distributions of the responses differed significantly across the four organizations.

This type of data is known as ordered categorical data. Although it is often common practice and even appealing to assign scores to the ordered categories, mathematically it is incorrect and illogical to use these scores as if they were numbers. However, since the Kruskal-Wallis test relies only on the relative ranking of the scores rather than the actual values, it is possible to analyze ordered categorical data using the Kruskal-Wallis test. In the event that significant differences are found, we can use BRANDO to determine the subgroups because BRANDO also uses only relative rank data in its nonparametric mode.

Table 7 gives the coded responses observed from the 33 completed surveys. A score of 1 indicates strong agreement, 2 indicates moderate agreement, 3 indicates neutrality, 4 indicates moderate disagreement and 5 indicates strong disagreement.

The question to be answered is whether or not there is sufficient evidence from the survey results to indicate that there are differences between the distribution of the responses that were obtained from the four organizations.

In statistical terms we wish to test the null hypothesis that the response patterns across the four organizations are about the same and that the observed differences in the response patterns are attributable to chance variation. We can test this hypothesis by ranking all 33

Table 7: *Survey Responses*

Organization Code	Response Code
A	2 (Moderate Agreement)
A	2 (Moderate Agreement)
B	2 (Moderate Agreement)
B	3 (Neutral)
B	1 (Strong Agreement)
B	2 (Moderate Agreement)
B	2 (Moderate Agreement)
B	1 (Strong Agreement)
B	1 (Strong Agreement)
B	1 (Strong Agreement)
B	2 (Moderate Agreement)
B	2 (Moderate Agreement)
B	1 (Strong Agreement)
B	2 (Moderate Agreement)
B	2 (Moderate Agreement)
B	2 (Moderate Agreement)
B	2 (Moderate Agreement)
B	1 (Strong Agreement)
B	2 (Moderate Agreement)
C	2 (Moderate Agreement)
C	3 (Neutral)
C	2 (Moderate Agreement)
C	3 (Neutral)
C	2 (Moderate Agreement)
C	2 (Moderate Agreement)
C	1 (Strong Agreement)
C	2 (Moderate Agreement)
C	2 (Moderate Agreement)
C	2 (Moderate Agreement)
C	1 (Strong Agreement)
D	1 (Strong Agreement)
D	1 (Strong Agreement)
D	1 (Strong Agreement)

responses and calculating the average rank of the responses for each organization. Under the null hypothesis each of the four average rank values should be about the same.

As a technical point, note that usually when applying the Kruskal-Wallis test we require that the minimum group size be 5 or more. The reason for this is that the distribution of the Kruskal-Wallis test statistic is usually approximated by a Chi-square distribution and the approximation requires a minimum group size of 5 or more. However, BRANDO uses a Monte-Carlo simulation to determine the distribution of the Kruskal-Wallis test statistic and is valid no matter how small the subgroups. In addition, the requirement to correct for tied values when using the Chi-square approximation is also not necessary when using a Monte-Carlo approach. The advent of fast and inexpensive computing power has made this possible.

The first step in the analysis is to rank the survey responses. We will arbitrarily assign the lowest rank to the category 'Strong Agreement'. Tied groups are given the average rank of the group. For example, there were 11 Strong Agreement responses so that the average rank assigned to them is 6.0, the average of the first 11 rank values. Tables 8 and 9 show the results of this process.

Using equation 5 and the values in Table 9, we can calculate the value of the modified Kruskal-Wallis test statistic. The value is 519.936. Based on 10,000 Monte-Carlo simulation iterations, the probability under the null hypothesis of observing a value as large as 519.936 with the given data is 0.0486. We therefore reject the null hypothesis (at the 0.05 significance level) and conclude that there is sufficient evidence to declare differences between the four organizations in terms of their responses to this survey question.

The next step is to find the most likely breakpoint between the four organizations when ordered by their average rank values as in Table 9. With four organizations there are three possible breakpoints to be considered. Using the BRANDO post-hoc methodology, Table 10 gives the likelihood values for the three possible breakpoints.

Since Breakpoint 1 has the highest likelihood value, it is selected as the first breakpoint.

The iterative procedure continues by considering the two sets of organizations created by the first breakpoint. The first set consists of organization D by itself so there is no further analysis required. The second set consists of organizations B, C, and A. We next test the new null hypothesis that the distribution of survey responses for these three organizations are about the same.

Table 11 gives the results of ranking the 30 responses from organizations B,C and A using the same procedure as discussed earlier. Table 12 gives the resulting average rank values for each of the three organizations.

The value of the Kruskal-Wallis test statistic for the data in Table 12 is 103.008 and the probability under the null hypothesis of observing a value as large as 103.008 was found to

Table 8: *Rank Values*

Organization	Response Code	Rank Value
A	2	21.0
A	2	21.0
B	2	21.0
B	3	32.0
B	1	6.0
B	2	21.0
B	2	21.0
B	1	6.0
B	1	6.0
B	1	6.0
B	2	21.0
B	2	21.0
B	1	6.0
B	2	21.0
B	2	21.0
B	2	21.0
B	2	21.0
B	1	6.0
B	2	21.0
C	2	21.0
C	3	32.0
C	2	21.0
C	3	32.0
C	2	21.0
C	2	21.0
C	1	6.0
C	2	21.0
C	2	21.0
C	2	21.0
C	1	6.0
D	1	6.0
D	1	6.0
D	1	6.0

Table 9: *Average Rank Values*

Organization	Number of Responses	Average Rank
D	3	6.0
B	17	16.353
C	11	20.273
A	2	21.0
Total	33	17.0

Table 10: *Likelihood values for the 3 possible breakpoints.*

Breakpoint	Subgroup1	Subgroup2	Likelihood value
1	D	B, C, A	0.4253
2	D, B	C, A	0.0393
3	D, B, C	A	0.0319

be 0.423 based on 10,000[5] Monte-Carlo iterations. Therefore we accept the null hypothesis and conclude that there is insufficient evidence to declare that the distribution of survey responses are significantly different for organizations A, B and C. This competes the analysis since there are no further break points to be found.

In summary the analysis indicates that the response patterns are not the same for all four organizations. More specifically, Organization D has a significantly different response distribution from the other three at the 5 percent significance level. Furthermore, there is insufficient evidence to indicate any difference in the survey response distributions between organizations A, B and C.

[5]For further guidance on the number of Monte-Carlo runs required see Annex B

Table 11: *Rank Values*

Organization	Response Code	Rank Value
A	2	18.0
A	2	18.0
B	2	18.0
B	3	29.0
B	1	4.5
B	2	18.0
B	2	18.0
B	1	4.5
B	1	4.5
B	1	4.5
B	2	18.0
B	2	18.0
B	1	4.5
B	2	18.0
B	2	18.0
B	2	18.0
B	2	18.0
B	1	4.5
B	2	18.0
C	2	18.0
C	3	29.0
C	2	18.0
C	3	29.0
C	2	18.0
C	2	18.0
C	1	4.5
C	2	18.0
C	2	18.0
C	2	18.0
C	1	4.5

Table 12: *Average Rank Values*

Organization	Number of Responses	Average Rank
B	17	13.882
C	11	17.545
A	2	18.0
Total	30	15.5

5 Discussion

The methodology and examples reviewed in this report provide the analyst with a straightforward yet robust solution to the problem of statistical analysis of datasets.

The methodology is notable as it avoids intransitivities and allows for nonparametric data. Applying the method in a GUI-based software provides the analyst with a rapid solution that is free of calculation errors. The methodology can be applied to any nonparametric sets of data that need to be statistically analysed. Common uses include the need to find differentiation between equipment that is to be purchased and the optimization of equipment ratios in particular scenarios. More unconventional uses include the analysis of ordinal (survey) data.

Those who are interested in learning more about the software are recommended to contact the Central OR Team (CORT) for a copy in addition to consulting Annex B for an overview of how to use the GUI-based software.

6 Conclusion

The purpose of this report was to describe the process of determining the most likely separation point(s) amongst a group of datasets that have been shown to have non-normal distributions. The work is in addition to earlier analysis done with normal and binomial data. All of these analysis tools have been incorporated into a graphical user interface known as the BReakpoint Analysis with Nonparametric Data Option (BRANDO) toolkit. The nonparametric is a powerful addition to the toolkit as it is able to handle data that is non-normal including ordinal data. The BRANDO methodology should be a very useful tool for the OR scientist, or indeed any researcher, in their effort to determine convincingly which sets of results from a given study differ significantly from the other results.

References

1. Emond, E.J. and Massel, P.L. (2003). A Post-ANOVA Methodology for Finding Subgroups with Equal Means Using Maximum Likelihood.

2. Hunter, D.G. and Emond, E.J. (2004). Seperation Point Analysis Method (SPAM).

3. Jurgen A. Doornik and Henrik Hansen (24 November, 1994). An Omnibus Test for Univariate and Multivariate Normality (Online). www.doornik.com/research/normal2.pdf (1 December 2005).

4. Geoffrey Poitras, Simon Fraser University (31 January, 2005). More on the Correct Use of Omnibus Tests for Normality (Online). www.sfu.ca/~poitras/el$.pdf (1 December 2005).

Annex A
Gamma Distribution of the Kruskal-Wallis Test Statistic

In order to show that the Kruskal-Wallis test statistic has approximately the Gamma distribution for large datasets we proceed as follows.

We are given K datasets with N_1 , N_2 , ... , N_K observations respectively and a total of N_{tot} observations in all. We rank all N_{tot} observations from lowest to highest and associate each value x_{ki} with its rank value r_{ki} . Tied observations are given the average rank of the tied group.

Before proceeding, we first clarify what is meant by large datasets. With no ties in any of the observations, each of the K datasets should have a minimum of 5 values in order for the Gamma approximation to be valid. When there are large numbers of tied observations the Gamma approximation may be less accurate. In such a situation it would be prudent to require at least 5 distinct values in each dataset in order to ensure that the Gamma approximation is valid.

Let us assume that the above conditions on the size of the datasets and the number of ties have been met. We consider a randomly selected subset of size N_k from the N_{tot} rank values. We denote the mean value of this randomly selected subset by $r_{k.}$. Since $r_{k.}$ is the average of k independently chosen rank values the Central Limit Theorem states that as k approaches infinity the distribution of $r_{k.}$ may be approximated by a Normal distribution. The mean and variance (derived from the mean and variance of the discrete uniform distribution of the rank values) of the limiting Normal distribution are given below.

Mean Value:

$$E(r_{k.}) = r_{..} = \frac{N_{tot}+1}{2} \tag{A.1}$$

Variance:

$$Var(r_{k.}) = \frac{N_{tot}^2 - 1}{12N_k} \tag{A.2}$$

Using a random selection process we subdivide the N_{tot} rank values into K subsets of size N_1, N_2, ... N_K respectively. For each of the K subsets the mean rank and its variance are determined from the above equations. The mean rank values are standardized by subtracting their mean and dividing by their standard deviation.

Standardized Mean Rank:

$$r_{k.}^{St} = \frac{r_{k.} - \frac{N_{tot}+1}{2}}{\sqrt{\frac{N_{tot}^2-1}{12N_k}}} \tag{A.3}$$

Because of the random selection process the mean rank values from the K subsets are independent except that their sum is a constant. Since each of the standardized dataset mean rank values has approximately a standard Normal distribution, the sum of their squares has approximately a Chi-squared distribution with K-1 degrees of freedom. If we denote this sum of squares by Y and simplify, we get the following.

$$Y = \sum_{k=1}^{K} (r_{k.}^{St})^2 = \sum_{k=1}^{K} \left[\frac{r_{k.} - \frac{N_{tot}+1}{2}}{\sqrt{\frac{N_{tot}^2-1}{12N_k}}} \right]^2 \tag{A.4}$$

$$Y = \sum_{k=1}^{K} \left[\frac{(r_{k.} - r_{..})^2}{\frac{N_{tot}^2-1}{12N_k}} \right] \tag{A.5}$$

$$Y = \frac{12}{(N_{tot}^2 - 1)} \sum_{k=1}^{K} N_k (r_{k.} - r_{..})^2 \tag{A.6}$$

This latter expression, except for the multiplicative constant, is the same as the Kruskal-Wallis test statistic defined in section 2 of the main report. Therefore we have shown that in cases where N_{tot} is large and where there is not an unreasonable number of ties, the Kruskal-Wallis test statistic has approximately a Chi-squared distribution which is a particular type of Gamma distribution.

A.1 Estimating the Parameters of the Gamma Distribution

The general form of the Gamma distribution is given below. It has two parameters, α and β which we will estimate using the method of moments. (Note that a slightly better estimate of α and β may be obtained using numerical maximum likelihood methods but for the present purposes this extra accuracy is not needed.)

Gamma Distribution: $f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}}$ *for* $x \geq 0$

The mean and variance of the Gamma distribution are as follows.

Mean of the Gamma Distribution: $E(x) = \beta\ \alpha$

Variance of the Gamma Distribution: $Var(x) = \beta^2\ \alpha$

Given a sample of M randomly generated Kruskal-Wallis test statistic values: v_1, v_2, ... , v_M, from a Monte-Carlo simulation, we can fit a Gamma distribution using the method of moments by equating the sample mean and variance to the theoretical mean and variance as follows. This procedure is accurate for large values of M which are typically on the order of a thousand or more in Monte-Carlo simulations.

The sample mean and variance are calculated first.

Sample Mean: $\bar{v} = \frac{1}{M}\sum_{i=1}^{M} v_i$

Sample Variance: $s^2(v) = \frac{1}{M-1}\sum_{i=1}^{M} (v_i - \bar{v})^2$

Given the sample mean and variance we estimate α and β as follows.

$$\beta_{est} = \frac{s^2(v)}{\bar{v}} \qquad \text{and} \qquad \alpha_{est} = \frac{\bar{v}}{\beta_{est}} \tag{A.7}$$

A.2 Example

To illustrate the fitting of a gamma distribution to the Kruskal-Wallis test statistic values from a Monte-Carlo simulation, we will consider the following simple example. We have three datasets, each consisting of 10 observations from independent trials. The initial data values are given in Table A.1 along with their means.

Table A.1: *Data Values and Means*

Dataset 1	Dataset 2	Dataset 3
3.0	11.0	8.0
3.0	10.0	10.0
2.0	11.0	14.0
4.0	6.0	3.0
7.0	11.0	12.0
8.0	5.0	14.0
0.0	13.0	11.0
0.0	1.0	13.0
2.0	10.0	2.0
2.0	11.0	12.0
Mean Value: 3.1	Mean Value: 8.9	Mean Value: 9.9

Since each dataset has at least 5 distinct values, we have a sufficiently large example to illustrate the fitting of a Gamma distribution. The first step is to replace all the observations by their ranks within the total of 30 observations, with ties handled by assigning the mean rank of the tied group. Table A.2 gives the resulting rank values and dataset average ranks.

The overall average rank, $r_{..}$, is 15.5. The weighted sum of squares of the dataset average ranks minus 15.5 is H* = 928.55.

We next perform a Monte-Carlo simulation. For each iteration we randomly subdivide the 30 rank values from Table A.2 into three sets of 10 values each. We calculate the average

Table A.2: *Rank Values and Average Ranks*

Dataset 1	Dataset 2	Dataset 3
9	22	15.5
9	18	18
5.5	22	29.5
11	13	9
14	22	25.5
15.5	12	29.5
1.5	27.5	22
1.5	3	27.5
5.5	18	5.5
5.5	22	25.5
Average Rank: 7.8	Average Rank: 17.95	Average Rank: 20.75

rank for each of the three sets and find the resulting value of the Kruskal-Wallis test statistic, H* . For 1000 iterations the mean value and variance of the Kruskal-Wallis test statistic were 142.38 and $(135.40)^2$ respectively. The estimates for α and β calculated using the equations given above are 128.77 and 1.106 respectively.

Figure A.1 displays the histogram for 1000 Monte-Carlo iterations and the fitted Gamma distribution.

Note that only one Monte-Carlo value out of 1000 (i.e. 0.001) was as large as the observed test statistic value of 928.55. For a Gamma distribution with $\alpha = 128.77$ and $\beta = 1.106$ the probability of observing a value as large as 928.55 is found to be 0.00097 using the Incomplete Gamma function. This illustrates the accuracy of the fitted distribution.

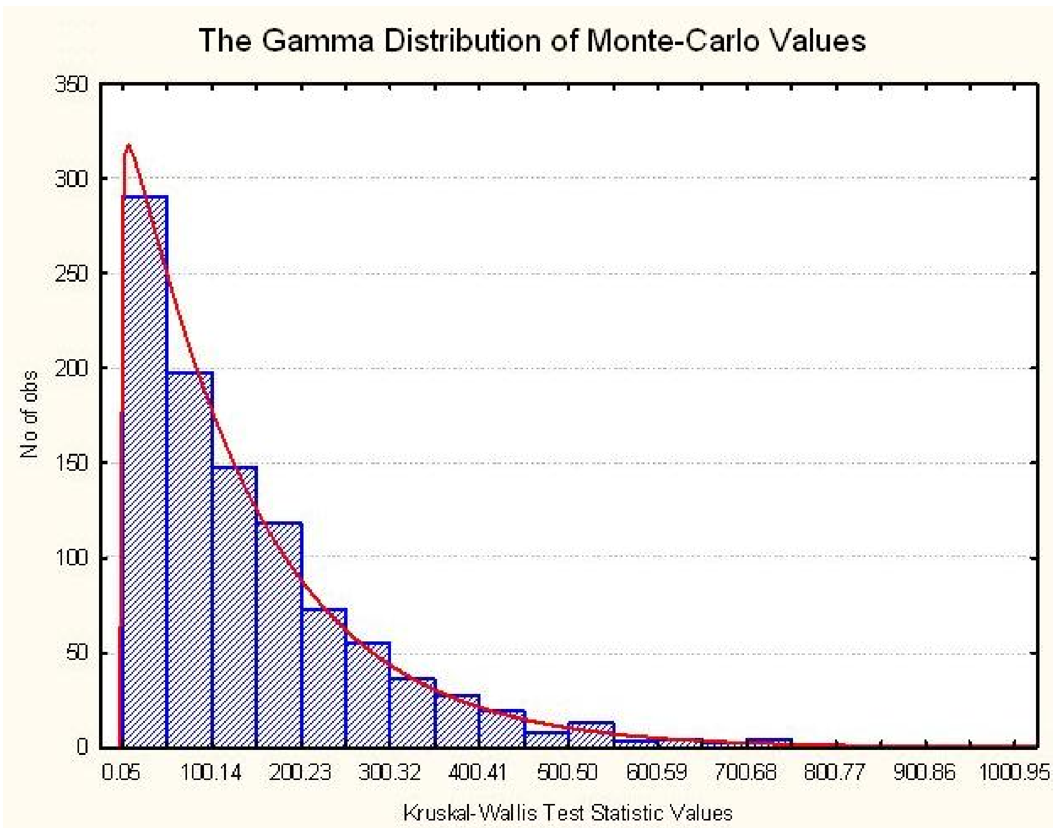

Figure A.1: *Gamma Distribution of Monte-Carlo Values.*

Annex B
Use of BRANDO software

For a detailed overview of how to use the BRANDO software the user is directed to the Separation Point Analysis Method (SPAM) document written by CORT [2]. Variations and additions have since been made to the software and are outlined below.

The software's graphical user interface (GUI) is started through MATLAB. The figure and executable file should both be in a directory easily accessible to the software. When the MATLAB tool has been started the user must type BRANDO at the prompt so that a GUI window opens and is available to the user.

The most significant improvement to this new version of software is the addition of the nonparametric model option. When starting the nonparametric model it is necessary to include the required number of Monte Carlo runs. A point of similarity between the Normal and nonparametric models is the necessity of having the data sorted in such a way that they are acceptable to the software.

If there is any question of the data's normalcy there is a quick solution. The Normal model automatically runs the D'Agostino K^2-test for normality and will return an error if the condition has not been met for all datasets. The data must then be run with the nonparametric model. If the data is tested with the Normal model and no problems are found then the assumption of normality is valid.

The reliability of the p-value at values near the significance level is tested by determining the likely (within 95 %) Monte-Carlo range of the p-value. The range is determined from the equations below for a number of Monte-Carlo runs, M.

p_1 = *p-value solved with Monte-Carlo iterations*

p_2 = *1 - significance level*

$$P_{avg} = (p_1 + p_2)/2 \tag{B.1}$$

$$P_{pm} = 1.96 * \sqrt{\frac{(P_{avg}) * (1 - P_{avg})}{M}} \tag{B.2}$$

If the absolute difference between the calculated p-value and that found from (1 - significance level) is less than P_{pm} then the results are unreliable and more Monte-Carlo iterations are required. All these calculations are done in the software and, if required, a warning is produced for the user suggesting more iterations be tried.

List of abbreviations

BRANDO	BReakpoint Analysis with Nonparametric Data Option
CF	Canadian Forces
CORA	Centre for Operational Research and Analysis
DND	Department of National Defence
DRDC	Defence Research and Development Canada
GUI	Graphical User Interface
OR	Operational Research
SPAM	Separation Point Analysis Method

Report Distribution

Internal

DG CORA
CORA(CS)
DOR(MLA)
DOR(Joint)

D Strat HR
PORT
DMGOR
CORT
JSORT
SPORT
CEFCOM ORT
CANCOM ORT
CANSOFCOM ORT
TDP ORT
MORT
DASOR
LFORT

CORA Library (2 copies)
DRDKIM (2 copies)

Authors (2 copies)

External

DRDC Valcartier CS
DRDC Valcartier - C. Fortier
DRDC Ottawa CS
DRDC Toronto CS
DRDC Atlantic CS
DRDC Suffield CS

NORAD
DRDC (V) ORT
MARLANT - P. Saunders
MARPAC - P. Sutherland
1 CAD HQ - C. Hunter
LFDTS - F. Cameron

DOCUMENT CONTROL DATA

(Security classification of title, body of abstract and indexing annotation must be entered when document is classified)

1. ORIGINATOR (the name and address of the organization preparing the document. Organizations for whom the document was prepared, e.g. Centre sponsoring a contractor's report, or tasking agency, are entered in section 8.) DRDC – Centre for Operational Research and Analysis NDHQ, 101 Col By Drive, Ottawa ON K1A 0K2	2. SECURITY CLASSIFICATION (overall security classification of the document including special warning terms if applicable). UNCLASSIFIED

3. TITLE (the complete document title as indicated on the title page. Its classification should be indicated by the appropriate abbreviation (S,C,R or U) in parentheses after the title).

BRANDO
BReakpoint Analysis with Nonparametric Data Option

4. AUTHORS
(Last name, first name, middle initial. If military, show rank, e.g. Doe, Maj. John E.)

Emond, E.J. ; Turnbull, A.E.

5. DATE OF PUBLICATION (month and year of publication of document) November 2006	6a. NO. OF PAGES (total containing information. Include Annexes, Appendices, etc). 48	6b. NO. OF REFS (total cited in document) 4

7. DESCRIPTIVE NOTES (the category of the document, e.g. technical report, technical note or memorandum. If appropriate, enter the type of report, e.g. interim, progress, summary, annual or final. Give the inclusive dates when a specific reporting period is covered).

Technical Memorandum

8. SPONSORING ACTIVITY (the name of the department project office or laboratory sponsoring the research and development. Include address).

DRDC – Centre for Operational Research and Analysis
NDHQ, 101 Col By Drive, Ottawa ON K1A 0K2

9a. PROJECT OR GRANT NO. (if appropriate, the applicable research and development project or grant number under which the document was written. Specify whether project or grant). N/A	9b. CONTRACT NO. (if appropriate, the applicable number under which the document was written).
10a. ORIGINATOR'S DOCUMENT NUMBER (the official document number by which the document is identified by the originating activity. This number must be unique.) DRDC CORA TM 2006–40	10b. OTHER DOCUMENT NOs. (Any other numbers which may be assigned this document either by the originator or by the sponsor.)

11. DOCUMENT AVAILABILITY (any limitations on further dissemination of the document, other than those imposed by security classification)

(X) Unlimited distribution
( ) Defence departments and defence contractors; further distribution only as approved
( ) Defence departments and Canadian defence contractors; further distribution only as approved
( ) Government departments and agencies; further distribution only as approved
( ) Defence departments; further distribution only as approved
( ) Other (please specify):

12. DOCUMENT ANNOUNCEMENT (any limitation to the bibliographic announcement of this document. This will normally correspond to the Document Availability (11). However, where further distribution beyond the audience specified in (11) is possible, a wider announcement audience may be selected).

13. ABSTRACT (a brief and factual summary of the document. It may also appear elsewhere in the body of the document itself. It is highly desirable that the abstract of classified documents be unclassified. Each paragraph of the abstract shall begin with an indication of the security classification of the information in the paragraph (unless the document itself is unclassified) represented as (S), (C), (R), or (U). It is not necessary to include here abstracts in both official languages unless the text is bilingual).

This paper reports on the mathematical details and software solution to the Operational Research problem of comparing the central values of multiple datasets using nonparametric statistics. The work is an extension of the iterative post-hoc analysis method described by Emond and Massel (2003). The original methodology required an assumption of Normality for all of the datasets, an assumption that is often not met. An additional analysis method has now been developed in which the data are ranked allowing for a nonparametric solution. Using a proxy for the likelihood function on the ranked datasets, this method finds the most likely separation points between them after a nonparametric analysis of variance has indicated that differences exist. In addition to the theoretical methodology and explanation, examples are given to demonstrate the practicality of the process. Of particular interest is an example which illustrates the application of the methodology to the analysis of ordered categorical data often found in surveys.

14. KEYWORDS, DESCRIPTORS or IDENTIFIERS (technically meaningful terms or short phrases that characterize a document and could be helpful in cataloguing the document. They should be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location may also be included. If possible keywords should be selected from a published thesaurus. e.g. Thesaurus of Engineering and Scientific Terms (TEST) and that thesaurus-identified. If it not possible to select indexing terms which are Unclassified, the classification of each should be indicated as with the title).

nonparametric data
BRANDO
SPAM
breakpoint analysis
post-hoc
ANOVA
separation point analysis
statistically significantly different

DEFENCE DÉFENSE

DRDC CORA

www.drdc-rddc.gc.ca