An OR Tool for Partitioning Datasets Into Uniform Segments

E.J. Emond
Central OR Team

DRDC CORA TN 2009–052
November 2009

Defence R&D Canada
Centre for Operational Research and Analysis

Central OR Team

National Defence Défense nationale

Canada

An OR Tool for Partitioning Datasets Into Uniform Segments

E.J. Emond
Central OR Team

DRDC – Centre for Operational Research and Analysis

Technical Note

DRDC CORA TN 2009–052

November 2009

Author

E.J. Emond

Approved by

Paul Massel
Central Operational Research Team Leader

Approved for release by

D.F. Redding
Chief Scientist

The information contained herein has been derived and determined through best practice and adherence to the highest levels of ethical, scientific and engineering investigative principles. The reported results, their interpretation, and any opinions expressed therein, remain those of the authors and do not represent, or otherwise reflect, any official opinion or position of DND or the Government of Canada.

Abstract

This report describes an operational research tool for dividing a set of scalar values into categories or clusters without having to specify the number of clusters in advance. The method has a theoretical underpinning based on a statistical test for outliers of a uniform distribution derived by D.A. Darling in 1952. The problem arose in an operational research study for the Canadian Department of National Defence in which it was necessary to segment a set of 82 options into categories based on a numerical score. The theoretical background is described and examples are given to illustrate the methodology.

This page intentionally left blank.

Table of contents

Abstract i

Table of contents iii

Tables iv

Figures iv

1 Introduction 1

1.1 Aim 2

2 Methodology 3

2.1 Testing Whether the Largest Observation is an Outlier 3

2.2 Testing the Null Hypothesis 3

2.3 Critical Values for Small Sample Sizes 4

2.3.1 Iterated Procedure to Find Ordered Categories 5

3 Simple Example 6

3.1 Determining Further Breakpoints Using an Iterated Procedure 7

4 A More Complicated Example 8

5 Discussion 12

6 Conclusion 13

References 14

List of abbreviations 15

Report Distribution 16

Tables

1 Critical Values for Small Values of n 5

2 Program Component Scores for Competitor 12 6

3 Capability Option Scores 8

Figures

1 Scores for Competitor 12 7

2 Histogram of Capability Option Scores 9

3 Score Clusters 11

1 Introduction

The Central Operational Research Team (CORT) has a mandate to create new tools and methodology in support of the Centre for Operational Research and Analysis (CORA). In a recent operational research study a requirement arose to categorize 82 options into an indeterminate number of segments based on a score derived from a variety of criteria. CORT was asked to recommend a methodology for this problem which would avoid arbitrary decisions by the analysts on the number of segments as well as the values of the break points. In short, the method must be repeatable given minimal stated assumptions.

Some analysts may recognize this problem as a one dimensional version of the classical clustering problem. However, the fact that there is only one dimension to the data means that the current problem is more amenable to statistical methods and in fact can be solved satisfactorily and repeatably with minimal assumptions.

Although it may seem like a backward step to go from measured values to ordered categories, there are cases where such problems come up. In a multicriteria decision problem for example, it may happen that there is a mix of variable types from pure measurements to pure rankings. In order to apply solution techniques that rely on rankings it may be necessary to create a set of ordered categories from measured data. The idea is to divide the range of the measured variable into segments within which the value of the variable, in the context of the decision problem, is considered to be constant, or at least not different enough to matter. Rather than arbitrarily divide the range of the measured variable into segments, we require a methodology that is repeatable and does not require any subjective input from the analyst.

One possible approach to this problem is to specify the number of segments and then find the optimal break locations by minimizing the overall variance of the dataset. This technique produces acceptable breakpoints but suffers from the obvious fault that the number of subgroups is not known, making it difficult to determine how many subgroups should be used. A related but more devastating problem is that this method will always produce the specified number of breakpoints in a dataset even when the dataset is distributed uniformly over its range and therefore should not be subdivided on statistical grounds.

The methodology described in this report avoids the problem of determining the number of subgroups by relying on an iterative statistical approach that only divides any dataset or part of a dataset if the hypothesis of uniformity has been rejected. The statistical part of the methodology is based on work by D.A. Darling in the 1950's on the problem of determining whether extreme values in a dataset are outliers. [1] Darling's results for the general case are somewhat complex to apply but for the uniform distribution simplify into a practical and useful test based on the standard Normal distribution.

1.1 Aim

The aim of this report is to document and illustrate a methodology for subdividing the range of a dataset into segments based on statistical grounds. The methodology determines the number of segments and only subdivides the range when the statistical test of hypothesis indicates that there is sufficient evidence to justify doing so.

2 Methodology

2.1 Testing Whether the Largest Observation is an Outlier

We start by considering the statistical problem of deciding whether the largest observation from a set of observations is an outlier. That is, has the largest observation come from a different distribution from that which generated the rest of the observations. Formally, we wish to test the hypothesis that all the observations are identically distributed against the alternative that the largest of them comes from a different distribution. The ultimate aim is to determine a set of categories within which the observations may be taken as uniformly distributed. Therefore we specify a uniform distribution for the null hypothesis.

Null Hypothesis: All the observations come from the same Uniform Distribution.

Alternate Hypothesis: The largest observation is an outlier.

Consider a random sample of n observations sorted from lowest to highest and written as $x_{(1)}, x_{(2)}, ..., x_{(n)}$ where the subscript in brackets indicates that the observations have been ordered. The technical term for the observations written this way is order statistics. We wish to test the null hypothesis that the observations are the order statistics of a sample of size n from a uniform distribution. The alternative hypothesis is that the largest observation is an outlier.

A technical issue is the determination of the endpoints of the Uniform distribution under the null hypothesis. For the outlier test which will be used in this report it is necessary that the lower endpoint be zero. In general this will not be the case but an easy remedy is to first transform the data by subtracting the smallest value from each data point. Note that this procedure is not performed if the lower endpoint is already zero. The transformation does not affect the analysis except to effectively reduce the sample size by one.

The outlier test which will be used in this report does not require specification of the upper endpoint of the Uniform distribution except to ensure that it is finite. We assume that there exists a finite upper bound for the data, denoted by B.

Given the above technical considerations, we can restate the test of hypothesis in a more specific manner. We denote the (possibly) transformed data values as $y_{(1)}, y_{(2)}, y_{(3)}, ..., y_{(n)}$, where $x_{(1)}$ has been subtracted from every observation in those cases where the lower endpoint is non-zero, and B is an unknown positive value.

H_0 : The observations $y_{(i)}$ are order statistics from a Uniform (0,B) distribution.

H_A : The largest observation $y_{(n)}$ does not come from this distribution.

2.2 Testing the Null Hypothesis

In [1] D.A. Darling showed that the distribution of z_n defined as the sum of n random variables from a uniform distribution on the interval $(0, B)$ divided by the largest observation

$y_{(n)}$ is the same as the distribution of the sum of a random sample of $n-1$ observations from a uniform distribution on the interval $(0,1)$ plus the constant 1. Kendall gives a more accessible derivation of this result in [2]. Note that the result holds whatever the value of B, so that we only require that the sample has a finite upper bound.

For even fairly small values of n, the sum of 1 plus $n-1$ uniform random variables on (0,1) can be approximated by the Normal distribution with mean value $(n+1)/2$ and variance $(n-1)/12$. The approximation is accurate to three figures for values of n greater than 10. (In the early days of computing, approximate Normally distributed random variables were generated by summing 12 uniform random variables.) The statistic z_n is thus ideal as a test statistic for testing the null hypothesis above.

Note that Darling's result holds for the case where the sample range is the interval $(0,B)$. As mentioned above, for the more general case where the lower bound is not necessarily zero, we must subtract the lowest sample value (first order statistic) from each observation. This has the effect of changing the mean value of the test statistic from $(n+1)/2$ to $n/2$ and the variance from $(n-1)/12$ to $(n-2)/12$.

Using Darling's statistic z_n defined above, we can test the null hypothesis as follows. If the largest observation $y_{(n)}$ is an outlier then the value of the test statistic z_n will be low relative to its distribution under H_0. We reject the null hypothesis at the 5 percent significance level when the following inequality holds.

$$\frac{y_{(1)}+y_{(2)}+\ldots+y_{(n)}}{y_{(n)}} \leq \frac{n}{2} - 1.645 * \sqrt{\frac{n-2}{12}} \tag{1}$$

Note that the value -1.645 is the lower 5th percentile for a standardized Normal distribution. If a 10 percent significance level is desired, the appropriate value in the statistical test would be -1.282.

For convenience the equation above may be rearranged to create a more recognizable test statistic as follows.

$$\frac{\frac{y_{(1)}+y_{(2)}+\ldots+y_{(n)}}{y_{(n)}} - \frac{n}{2}}{\sqrt{\frac{n-2}{12}}} \leq -1.645 \tag{2}$$

2.3 Critical Values for Small Sample Sizes

The Normal approximation discussed above can be used for all cases where N is at least 10. For smaller values of N, the following table of critical values has been created using Monte Carlo simulation with 100,000 iterations. Critical values are tabulated for significance levels of 1, 5 and 10 percent. Also given in the table for comparison purposes are the relevant values for the Normal approximation.

Table 1: *Critical Values for Small Values of n*

N	1 Percent (-2.326)	5 Percent (-1.645)	10 Percent (-1.282)
3	-1.699	-1.560	-1.391
4	-2.102	-1.682	-1.364
5	-2.216	-1.656	-1.309
6	-2.250	-1.646	-1.298
7	-2.263	-1.646	-1.296
8	-2.263	-1.650	-1.298
9	-2.271	-1.646	-1.292
10	-2.281	-1.648	-1.295

2.3.1 Iterated Procedure to Find Ordered Categories

If there is insufficient evidence to reject the initial test of hypothesis on the largest observation, the analysis is complete and we conclude that there is no statistical justification for further subdividing the observations. If subdivision into more ordered categories is required despite this finding, the analyst must find other means to determine the category boundaries. One possibility is to increase the Type I error level in the test of hypothesis described above. However, this should be done with caution because it increases the chance of dividing up datasets which are uniformly distributed.

If the null hypothesis is rejected and we find sufficient evidence to conclude that the largest observation is an outlier, then we continue the analysis in an iterative manner. The idea is to remove the largest observation and repeat the test of hypothesis on the reduced dataset. We continue in this way until we reach a point where the null hypothesis is accepted. At this point the remaining reduced set of observations is considered to be uniformly distributed and is therefore declared to be the first segment or category.

For example, suppose there are 100 observations and we apply the outlier test starting at the largest observation. Suppose further that this observation is declared an outlier based on the test of hypothesis. This observation is now temporarily set aside and we next apply the test to the reduced set of 99 observations to determine if the largest of these observations is also an outlier. We continue in this manner until we reach a point at which the largest observation of the reduced set of values is **not** declared an outlier. Suppose this point is reached at the 25th largest value. In that case the reduced set consisting of the first 75 ordered values is declared to be the initial segment. The other 25 values would then be analyzed to determine if further segmenting is warranted.

Before describing the procedure further, a simple example will help to illustrate the methodology.

3 Simple Example

The following example is taken from an analysis of judging performance at a recent world level sporting event. In this example, nine judges rated competitors on five criteria. The sum of the five criteria scores is called the Program Component score. Although some variability is expected between judges, it is hypothesized that the nine Program Component Scores should be approximately uniformly distributed for each competitor and in particular there should not be any extreme scores or outliers.

The nine Program Component scores for competitor 12 are given in the table below. The scores have been ordered from lowest to highest, ranging from 15.25 to 21. A graphical representation is given in Figure 1. Noting the relatively large gap between the highest value and the second highest, the question is whether or not this set of scores is compatible with an assumption of uniformity across the nine judges.

Table 2: *Program Component Scores for Competitor 12*

Judge Number	Program Component Scores
3	15.25
2	15.50
4	16.00
9	16.75
6	17.00
5	17.50
7	17.75
8	18.50
1	21.00

The question to be analyzed in this example is whether there is sufficient evidence to decide that the cumulative score of 21.0 for Judge 1 is an outlier, under the uniform distribution assumption.

The first step is to create the dataset $y_{(1)}, y_{(2)}, y_{(3)}, ..., y_{(9)}$ by subtracting the smallest total score, 15.25, from each of the 9 scores. We then test whether the largest remaining value can be considered an outlier using Darling's procedure. In this case the largest remaining value is 5.75, associated with Judge 1. We divide each of the 9 remaining values by 5.75 and sum them using equation (2) to find a test value of -1.793. Since this value is less than the critical value of -1.646 associated with a false alarm rate of 5 percent (from Table 1), we decide that the cumulative score associated with Judge 1 is an outlier. (The p-value is 0.0375.)

By applying the above procedure again to the remaining 8 scores, we verify that there is insufficient evidence to declare any other outliers. In other words, the remaining 8 scores are compatible with the assumption that they are a sample from a uniform distribution.

The conclusion for this example is that the cumulative scores for competitor 12 are not

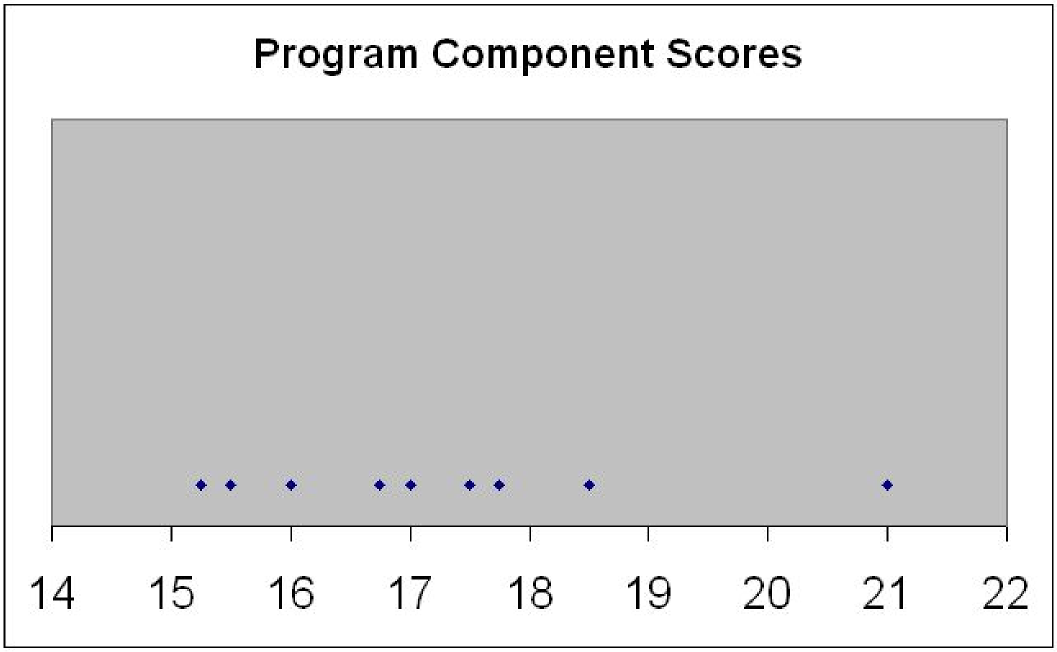

Figure 1: *Scores for Competitor 12*

consistent (at the 5 percent significance level) with the assumption of a uniform distribution of judges scores. In particular, the cumulative score for Judge 1 appears to be an outlier at a p-value of 0.0375.

Note that because of the Bonferroni Effect[1], this example by itself may not be conclusive evidence of bias or judging error on the part of Judge 1, but it may indicate the possibility of a problem. The result may indicate a requirement for further analysis of the scoring for other competitors to see if there is any additional evidence of problems with this judge.

3.1 Determining Further Breakpoints Using an Iterated Procedure

Suppose that the procedure based on identification of outliers has been applied and an initial segment has been identified. The set of values which are not in the first segment is considered next to determine if further segmentation is justified. In the simple example above there was only one value in the initial set of outliers so that no further segmentation was possible. In the case where there is more than one value outside of the initial segment we may apply the procedure again on this reduced set of values.

This iterated procedure will eventually end and will determine a number of segments. The following example will illustrate a case in which three segments are found.

[1] The Bonferroni Effect refers to the problem of maintaining a given overall false alarm rate when making multiple tests of hypotheses.

4 A More Complicated Example

The following example is from an actual analysis problem encountered during a study of capability assessment for the Canadian Department of National Defence. A set of 82 options was scored using various criteria. The 82 scores are listed in the table below.

Table 3: *Capability Option Scores*

0.000	0.000	0.000	0.000
0.000	0.000	0.000	0.023
0.047	0.047	0.059	0.064
0.080	0.080	0.093	0.098
0.100	0.113	0.120	0.135
0.140	0.149	0.200	0.203
0.210	0.261	0.269	0.285
0.286	0.322	0.330	0.340
0.360	0.363	0.364	0.379
0.397	0.415	0.433	0.437
0.515	0.533	0.545	0.550
0.757	0.800	0.830	1.057
1.283	1.455	1.461	1.461
1.497	1.627	1.851	1.872
1.882	1.933	1.950	2.082
2.082	2.082	2.190	2.207
2.288	2.323	2.337	2.345
2.410	2.461	2.611	2.730
2.772	2.920	3.126	3.598
3.664	3.664	3.772	3.897
3.897	3.970		

The question of interest in the analysis was to determine if there were any ’natural’ break points within this dataset and if so, how many. A histogram of the data is given in the Figure below.

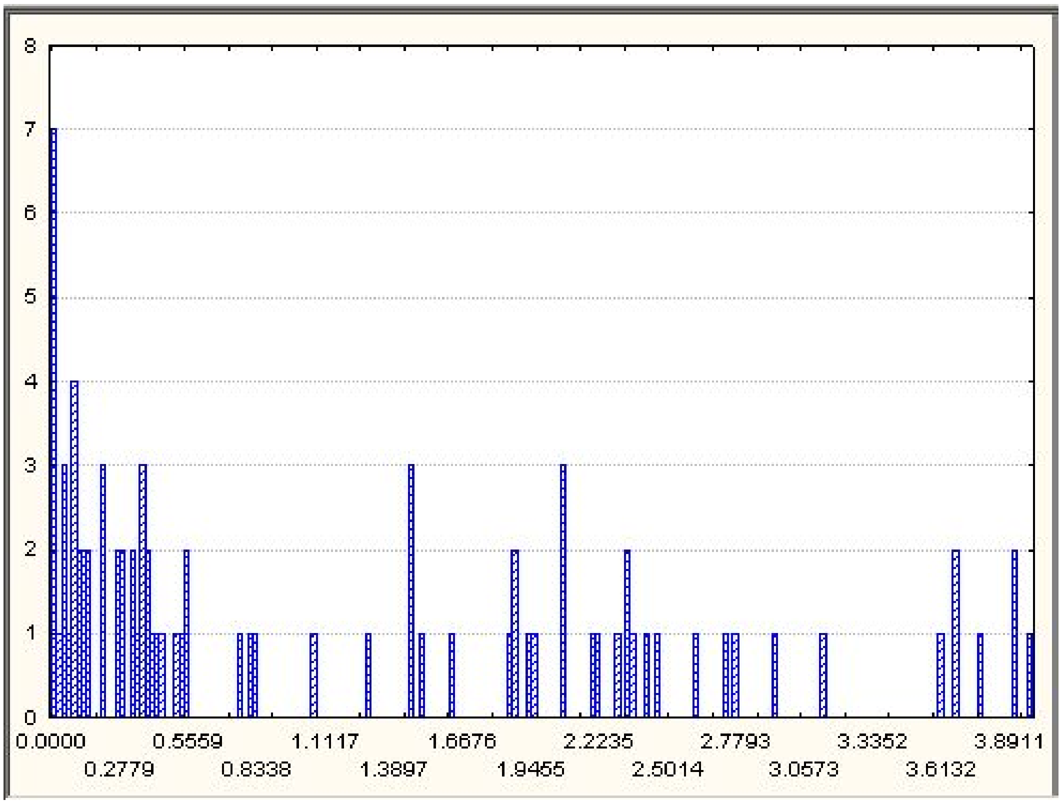

Figure 2: *Histogram of Capability Option Scores*

The method based on the work of Darling with the modifications discussed above was applied to this dataset to determine the number of subsets and possible break points. A p value (false alarm rate) of 0.05 was specified.

From Table 2 it can be seen that the first seven values in the dataset are zero. The initial pass over the other 75 values indicated as expected that all 75 were outliers with respect to the seven zeros. This result was expected since seven identical values at zero appears to be contrary to the assumption of uniformity. We therefore take the seven zero values as the first cluster and go on to consider whether the remaining 75 values are consistent with a uniform distribution.

The remaining 75 values ranged from 0.023 to 3.970. This set of values was analyzed separately. Every value from the largest to the 38th largest failed Darling's outlier test at the 5 percent significance level. The remaining 37 values however passed the test indicating that these values are consistent with the uniformity hypothesis. Therefore the set of 37 values ranging from 0.023 to 0.550 was identified as the next cluster.

After removing the first two clusters there were 38 values remaining, ranging from 0.757 to 3.970. Darling's procedure applied to the last value, 3.970, indicated that there was insufficient evidence to reject the uniformity hypothesis. Therefore the entire set of 38 values was taken to be the third and final cluster. (See Figure 3.)

In summary, the uniform outlier analysis indicated that the set of 82 capability scores could be divided into three clusters in a repeatable and non-arbitrary fashion. The three clusters had 7, 37 and 38 values respectively and each of the clusters represented a uniform distribution of values.

Again it should be noted that while this analysis gives an indication of a way to divide the scores into three clusters, each having a uniform distribution, the analyst is free to take this information and decide whether the particular problem requires further breakdown or not. The value of the method is that it gives a starting point without requiring arbitrary assumptions; in particular the number of clusters does not need to be specified in advance.

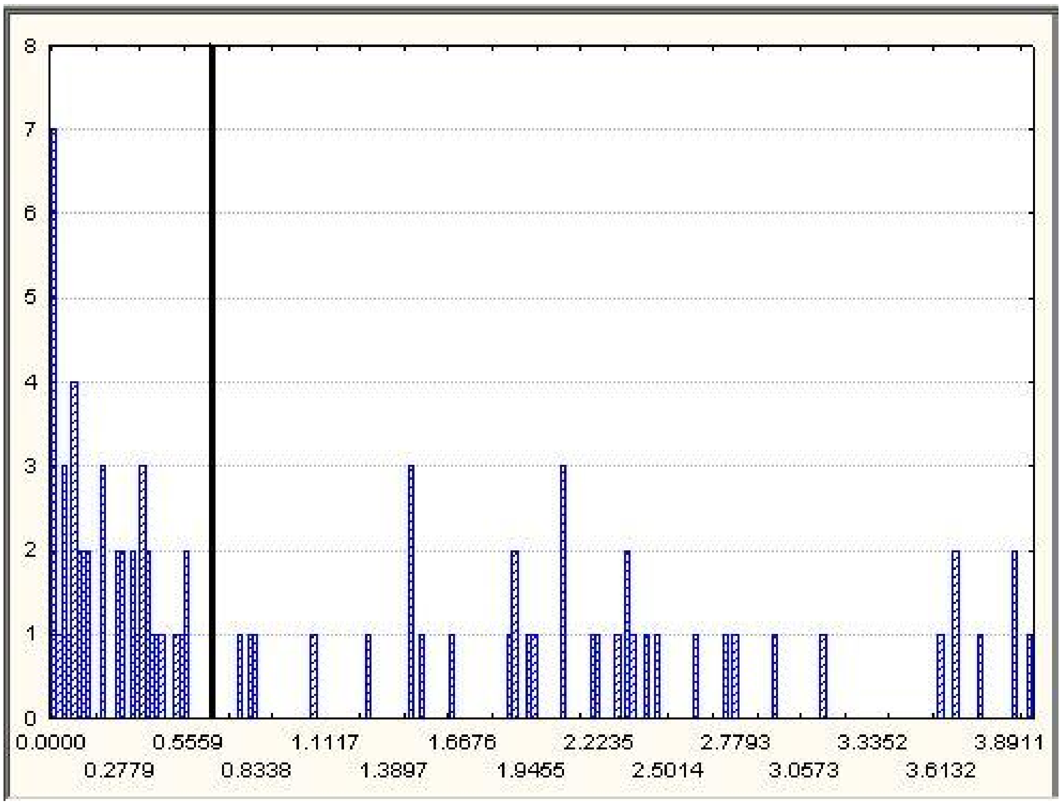

Figure 3: *Score Clusters*

5 Discussion

The methodology and examples outlined in this report provide the operational research analyst with a useful tool in situations where a dataset must be divided into ordered subsets but the number of subsets is not specified. The method based on Darling’s work subdivides a dataset into uniformly distributed subsets, but only if a null hypothesis of overall uniformity is rejected. This allows for a repeatable method to subdivide datasets with the property that the subdividing is only done if indicated by a statistical test of hypothesis.

Those who are interested in applying or learning more about the software are encouraged to contact the author.

6 Conclusion

The purpose of this report was to provide the theory and some examples outlining the process of determining the number of subsets into which a given dataset may be reasonably subdivided in a way which avoids an arbitrary predetermination of and is therefore repeatable. This tool has been useful in operational research studies within the Department of National Defence. It is simple to use and understand and as such may be a useful addition to the toolbox of quantitative analysts in CORA.

References

1. Darling, D. A. (1952). On A Test For Homogeneity and Extreme Values. *Ann. Math. Statist.*, **23**, 450–456.

2. Kendall, Maurice and Stuart, Alan (1979). The Advanced Theory of Statistics, Fourth ed. Vol. 2. New York: Macmillan.

List of abbreviations

CF	Canadian Forces
CORA	Centre for Operational Research and Analysis
DND	Department of National Defence
DRDC	Defence Research and Development Canada
OR	Operational Research

Report Distribution

Internal

DG CORA
CS and all Teams

CORA Library (2 copies)

Spares for Author (5 copies)

DRDKIM (2 copies)

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
   (NON-CONTROLLED GOODS)
   DMC A
   REVIEW: GCEC December 2013

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).

   An OR Tool for Partitioning Datasets Into Uniform Segments

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

   Emond, E.J.

5. DATE OF PUBLICATION (month and year of publication of document)

   November 2009

6a. NO. OF PAGES (total containing information. Include Annexes, Appendices, etc).

   22

6b. NO. OF REFS (total cited in document)

   2

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 Note

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 TN 2009–052

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 report describes an operational research tool for dividing a set of scalar values into categories or clusters without having to specify the number of clusters in advance. The method has a theoretical underpinning based on a statistical test for outliers of a uniform distribution derived by D.A. Darling in 1952. The problem arose in an operational research study for the Canadian Department of National Defence in which it was necessary to segment a set of 82 options into categories based on a numerical score. The theoretical background is described and examples are given to illustrate the methodology.

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).

clustering
outliers
uniform distribution
ordered categories

DEFENCE DÉFENSE

DRDC CORA

www.drdc-rddc.gc.ca