GILA RIVER SEDIMENT ANALYSIS

P.Y. Julien[1], *P.W. Mielke*[2], *A. Paris*[3]

Introduction

The Multi-Response Permutation Procedure (MRPP) is used to identify 3 groups of particle size distributions that can be generically called *silt*, *sand* and *gravel* (Section 1). In Section 2, a Fortran executable code that applies this method has been written to determine which group a new generic sample belongs to. A range index has also been defined to determine how close the generic sample is to the range of the measured data. The explanation of how the code works is presented in Section 3, while application examples can be found in Section 4.

1 Three sediment groups

The particle size distributions of 41 samples were provided. We selected 29 distributions and sorted them in 3 distinct groups characterized by different ranges of the typical grain sizes: d_{10}, d_{16}, d_{50}, d_{84} and d_{90} (Figure 1):

- Group A (*silt*): 7 particle size distributions
- Group B (*sand*): 8 particle size distributions
- Group C (*gravel*): 14 particle size distributions

The values of d_{50} and d_{84} for these 29 particle size distributions separated into the 3 groups, are reported in Table 1. Since we did not have the d_{50} for 5 particle size distributions belonging to Group A, we extrapolated them linearly from the log-scale graph, as shown in Figure 1.

[1] Prof., Dept. of Civ. and Env. Engrg., Colorado State Univ., Fort Collins, CO 80523
[2] Emer. Prof., Dept. of Stat., Colorado State Univ., Fort Collins, CO 80523
[3] Vis. Scientist, Dept. of Civ. and Env. Engrg., Colorado State Univ., Fort Collins, CO 80523

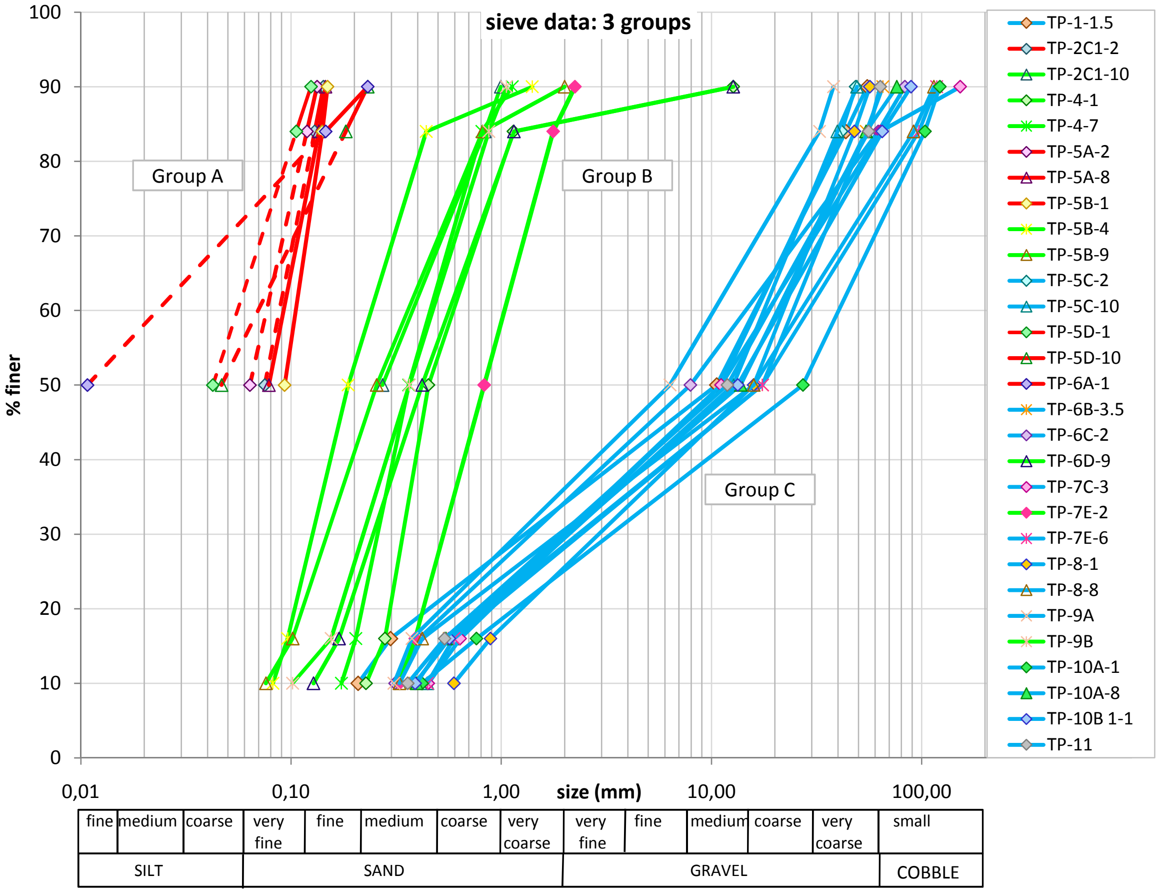


Figure 1: *Particle size distributions: Group A, Group B and Group C.*

Group A	(mm)	TP-2C1-2	TP-5A-2	TP-5A-8	TP-5B-1	TP-5D-1	TP-5D-10	TP-6A-1
	d50	0.075 (*)	0.064 (*)	0.08	0.09	0.042 (*)	0.047 (*)	0.011 (*)
	d84	0.13	0.12	0.14	0.14	0.11	0.18	0.15

Group B	(mm)	TP-2C1-10	TP-4-1	TP-4-7	TP-5B-4	TP-5B-9	TP-6D-9	TP-7E-2	TP-9B
	d50	0.27	0.45	0.36	0.19	0.26	0.42	0.83	0.37
	d84	0.83	1.14	0.81	0.44	0.81	1.15	1.77	0.87

Group C	(mm)	TP-1-1.5	TP-5C-2	TP-5C-10	TP-6B-3.5	TP-6C-2	TP-7C-3	TP-7E-6	TP-8-1
	d50	10.58	11.11	12.70	17.46	7.94	11.11	17.46	15.87
	d84	43.54	41.56	39.69	54.43	62.09	63.50	96.98	47.87
	(mm)	TP-8-8	TP-9A	TP-10A-1	TP-10A-8	TP-10B 1-1	TP-11		
	d50	15.87	6.35	27.21	13.76	13.33	11.91		
	d84	91.44	32.66	103.63	54.43	64.91	55.88		

Table 1: *d_{50} and d_{84} of the 3 grain size distribution groups .(*) indicates an extrapolated value.*

2 Methods and Procedures

2.1 Group Identification with MRPP procedure

We used the MRPP to calculate the P-value that is the probability of the data to be as extreme or more extreme than the observed data (see the details of the method from Mielke & Berry 2007 in the Appendix). We computed the P-value using the logarithmic values of d_{50} and d_{84} first separately and then together. We obtained that the couple d_{50}-d_{84} leads to the smallest P-value (P-value$< 10^{-6}$), therefore we decided that these two parameters should be used together to identify which group the testing sample belongs to. It should be noticed that the value of the probability is really small indicating that the three groups are well-sorted and that the MRPP is well-suited to this analysis.

The MRPP procedure is then applied to determine which group a new sample characterized by a d_{50} and a d_{84} belongs to. These data are added to each sediment size group in turn and the P-value is computed each time. The sample then belongs to the group that gets the smallest P-value.

2.2 Range Index

For each group, H_{d_x} and L_{d_x} are respectively the highest and the lowest sediment size values for that d_x (Table 2) and $\bar{R}_{d_x}$ is the logarithmic average value of the range computed as:

$$\bar{R}_{d_{50}} = \frac{\log H_{d_{50}} + \log L_{d_{50}}}{2} \tag{1}$$

$$\bar{R}_{d_{84}} = \frac{\log H_{d_{84}} + \log L_{d_{84}}}{2} \tag{2}$$

	d₅₀		d₈₄	
Group	**L_{d50}**	**H_{d50}**	**L_{d50}**	**H_{d50}**
A	0.011	0.09	0.11	0.18
B	0.19	0.84	0.44	1.77
C	6.35	27.21	32.66	103.6

Table 2: *H_{d_x} and L_{d_x} values for each group for both d_{50} and d_{84} .*

For a new sample characterized by its d_{50} and d_{84}, a range index has been defined to measure how far the sample is from the measured range of the group determined by the MRPP procedure. In particular we defined 3 different range indexes for d_{50}, d_{84} and d_{50}-d_{84}:

- the d_{50} range index:

$$I_{d_{50}} = \frac{\log H_{d_{50}} - \bar{R}_{d_{50}}}{\left|\log d_{50} - \bar{R}_{d_{50}}\right|} * 100 \tag{3}$$

- the d_{84} range index:

$$I_{d_{84}} = \frac{\log H_{d_{84}} - \bar{R}_{d_{84}}}{\left|\log d_{84} - \bar{R}_{d_{84}}\right|} * 100 \tag{4}$$

- the composite range index:

$$I = \frac{I_{d_{50}} + I_{d_{84}}}{2} \tag{5}$$

Finally the composite index, I, is the average percentage value between the two range indexes, $I_{d_{50}}$ and $I_{d_{84}}$, and ranges between 0 and 100 %. We also established that if d_x falls within the group range values, between H_{d_x} and L_{d_x}, then I assumes the maximum value, i.e. 100 %.

For example, the characteristic grain sizes for the sample TP-2C-7.5 are:

$$d_{50} = 0.93 \text{ mm and } d_{84} = 79.02 \text{ mm}$$

Since the MRPP suggests that this sample belongs to Group C, we focus on this group. According to Table 2:

$$L_{d_{50}} = 6.35 \text{ mm and } H_{d_{50}} = 27.21 \text{ mm}$$

$$L_{d_{84}} = 32.66 \text{ mm and } H_{d_{84}} = 103.63 \text{ mm}$$

so we can compute:

$$\bar{R}_{d_{50}} = \tfrac{\log 27.21 + \log 6.35}{2} = 1.119$$

$$\bar{R}_{d_{84}} = \tfrac{\log 103.63 + \log 32.66}{2} = 1.765$$

Then the range indexes are:

$$I_{d_{50}} = \tfrac{\log 27.21 - 1.119}{|\log 0.93 - 1.119|} * 100 = 27.47\%$$

$$I_{d_{84}} = 100\%$$

$$I = \tfrac{27.47 + 100}{2} = 63.73\%$$

Since $d_{50} = 0.93$ mm is quite far from the range [6.35-27.21] mm, the d_{50} range index results to be small: $I_{d_{50}} = 27.47\%$. On the other hand, $d_{84} = 79.02$ mm is included inside the relative range [32.66-103.63] mm so the d_{84} range index assumes the highest value: $I_{d_{84}} = 100\%$. Finally, the procedure associates this sample with Group C, but the level of confidence is low: $I = 63.73\%$.

3 The Fortran Code

The executable Fortran program, *Code.exe*, has been created. The code can be presented schematically in 3 parts:

1. **user input**:

    - d_{50} and d_{84} of the new sample in mm: these values must be entered by the user on the display interface

2. **code**:

    - asks for the input data d_{50} and d_{84}
    - uses the file *OBS.D* based on the observed data for the 3 groups (Table 1)
    - adds the new sample data to each group A, B and C of observed data and computes the related 3 P-values using MRPP
    - identifies the group with the smallest P-value
    - computes the range indexes, $I_{d_{50}}$ and $I_{d_{84}}$ and the composite index, I
    - loops back for new input

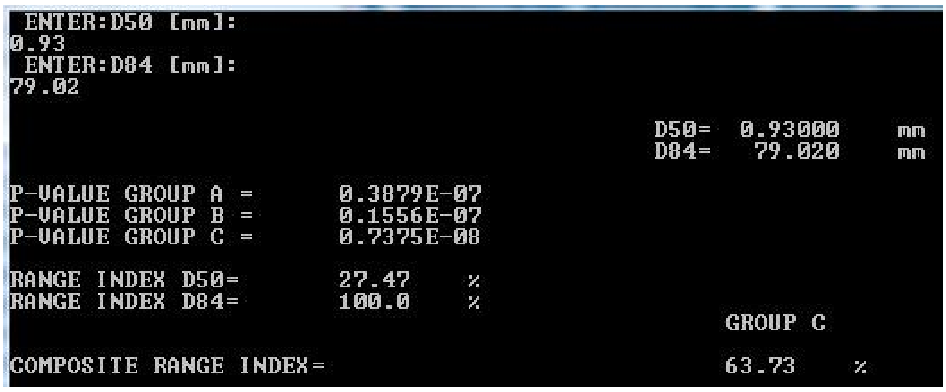
3. **outputs**:

    - *OBSA.D, OBSB.D* and *OBSC.D*: these files are produced and then used by the code as input data for the MRPP calculations
    - the following results are reported on the program interface:
        - the d_{50} and d_{84} of the new sample
        - the P-values for each group: A, B and C
        - the computed range indexes values (%): $I_{d_{50}}$ and $I_{d_{84}}$
        - the group which the sample belongs to
        - the composite index value(%): I
    - code generates the output file *OUTPUT.D*

4 Application

4.1 Application example

Figure 2 presents a code interface example. The first 4 lines request *input* from the user. The code asks for the d_{50} and d_{84} of the new sample in mm. Then the codes writes down these grain sizes on the right side and computes the P-values for Group A, B and C applying the MRPP procedure. The smallest P-value is related to Group C meaning that the sample belongs to this group. The values of the range indexes are reported. Finally on the right side the group to which the sample belongs is indicated and associated to its composite range index.



```
 ENTER:D50 [mm]:
0.93
 ENTER:D84 [mm]:
79.02

                                                        D50=  0.93000     mm
                                                        D84=    79.020    mm

P-VALUE GROUP A =       0.3879E-07
P-VALUE GROUP B =       0.1556E-07
P-VALUE GROUP C =       0.7375E-08

RANGE INDEX D50=        27.47    %
RANGE INDEX D84=        100.0    %
                                                            GROUP C

COMPOSITE RANGE INDEX=                                      63.73    %
```

Figure 2: *Code interface example: results related to sample TP-2C-7.5.*

4.2 Application to the additional dataset

We used the most recent 18 sieve data (e-mails from STANTEC: 05-07-2009 and 05-11-2009) to determine the group to which they belong (A, B or C) and to compute the percentage that describes how far they are from the closest groups (I [%]). Table 3 reports the samples identification code joined by the colour of the group we expect it to belong to and the values of d_{50} and d_{84} for each sample. The MRPP results in terms of group A, B, C and their related colours. The values of the composite index, I, and the values of the range indexes, $I_{d_{50}}$ and $I_{d_{84}}$ are then listed. The samples TP-7G-1 and TP-7G-10 did not have all the information required so they could not be classified.

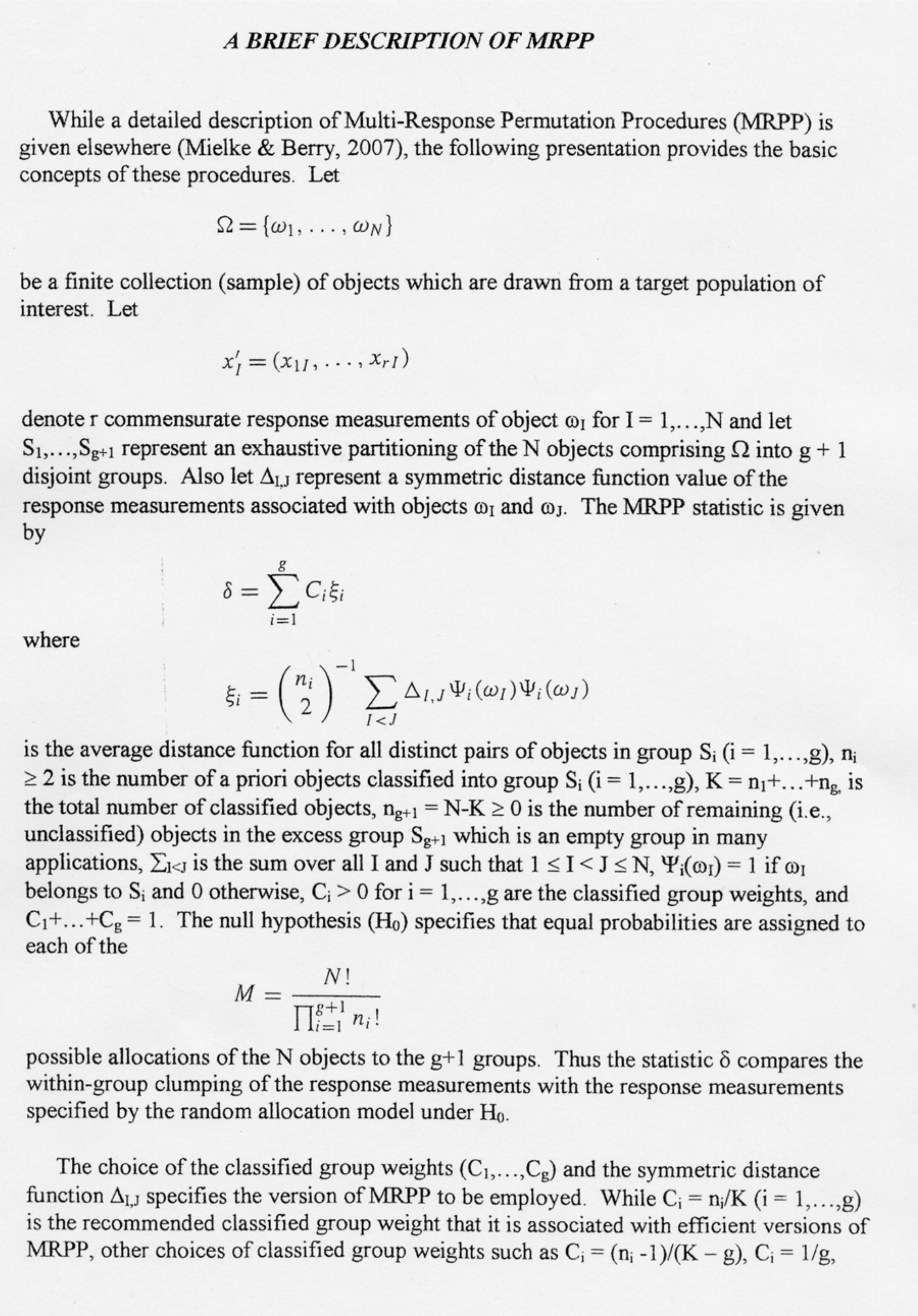
The MRPP predicts the correct group for all 16 samples. For instance the samples TP-2A-10, TP-2B-8.5, TP-2C-1, TP-2D-3, TP-2E-1, TP-7A-1, TP-7B-1, TP-7F-1 and TP-7F-10 clearly belong to their respective groups as the composite range index values are equal to 100 %. The samples that did not clearly belong to any of the groups, have lower composite index values.

	TP-2A-1	TP-2A-10	TP-2B-1	TP-2B-3	TP-2B-8.5	TP-2C-1	TP-2C-7.5	TP-2D-3	TP-2E-1	TP-7A-1	TP-7A-11	TP-7B-1	TP-7B-10	TP-7F-1	TP-7F-10	TP-7G-1	TP-7G-4	TP-7G-10
d50 (mm)	0.11	0.019 (*)	0.41	0.27	0.28	0.32	0.93	17.10	20.32	0.72	1.08	0.69	2.00	0.36	0.36		0.118 (*)	14.60
d84 (mm)	0.17	0.13	9.52	0.42	0.58	0.46	79.02	36.83	84.66	1.45	38.10	1.73	40.64	0.59	0.74		0.24	

MRPP	A	A	B	B	B	B	C	C	C	B	C	B	C	B	B		A	
I	**91.98**	**100**	**64.6**	**96.87**	**100**	**100**	**63.73**	**100**	**100**	**100**	**64.56**	**100**	**69.32**	**100**	**100**		**62.81**	
I_{d50}	83.97	100	100	100	100	100	27.47	100	100	100	29.11	100	38.64	100	100		79.51	
I_{d84}	100	100	29.2	93.73	100	100	100	100	100	100	100	100	100	100	100		46.12	

Table 3: *Application of the model to the recent samples. (*) indicates an extrapolated value.*

Appendix



A BRIEF DESCRIPTION OF MRPP

While a detailed description of Multi-Response Permutation Procedures (MRPP) is given elsewhere (Mielke & Berry, 2007), the following presentation provides the basic concepts of these procedures. Let

$$\Omega = \{\omega_1, \ldots, \omega_N\}$$

be a finite collection (sample) of objects which are drawn from a target population of interest. Let

$$x'_I = (x_{1I}, \ldots, x_{rI})$$

denote r commensurate response measurements of object ω_I for $I = 1,\ldots,N$ and let $S_1,\ldots,S_{g+1}$ represent an exhaustive partitioning of the N objects comprising Ω into $g + 1$ disjoint groups. Also let $\Delta_{I,J}$ represent a symmetric distance function value of the response measurements associated with objects ω_I and ω_J. The MRPP statistic is given by

$$\delta = \sum_{i=1}^{g} C_i \xi_i$$

where

$$\xi_i = \binom{n_i}{2}^{-1} \sum_{I<J} \Delta_{I,J} \Psi_i(\omega_I) \Psi_i(\omega_J)$$

is the average distance function for all distinct pairs of objects in group S_i ($i = 1,\ldots,g$), $n_i \geq 2$ is the number of a priori objects classified into group S_i ($i = 1,\ldots,g$), $K = n_1+\ldots+n_{g,}$ is the total number of classified objects, $n_{g+1} = N\text{-}K \geq 0$ is the number of remaining (i.e., unclassified) objects in the excess group S_{g+1} which is an empty group in many applications, $\Sigma_{I<J}$ is the sum over all I and J such that $1 \leq I < J \leq N$, $\Psi_i(\omega_I) = 1$ if ω_I belongs to S_i and 0 otherwise, $C_i > 0$ for $i = 1,\ldots,g$ are the classified group weights, and $C_1+\ldots+C_g = 1$. The null hypothesis (H_0) specifies that equal probabilities are assigned to each of the

$$M = \frac{N!}{\prod_{i=1}^{g+1} n_i!}$$

possible allocations of the N objects to the g+1 groups. Thus the statistic δ compares the within-group clumping of the response measurements with the response measurements specified by the random allocation model under H_0.

The choice of the classified group weights ($C_1,\ldots,C_g$) and the symmetric distance function $\Delta_{I,J}$ specifies the version of MRPP to be employed. While $C_i = n_i/K$ ($i = 1,\ldots,g$) is the recommended classified group weight that it is associated with efficient versions of MRPP, other choices of classified group weights such as $C_i = (n_i - 1)/(K - g)$, $C_i = 1/g$, and $C_i = n_i(n_i-1)/[n_1(n_1-1)+\ldots+n_g(n_g-1)]$ have also been considered. In many applications of MRPP the symmetric distance function is given by

$$\Delta_{I,J} = \begin{cases} d_{I,J} & \text{if } d_{I,J} \le B \\ B & \text{otherwise} \end{cases}$$

where

$$d_{I,J} = \left[\sum_{h=1}^{r} (x_{hI} - x_{hJ})^2 \right]^{\frac{v}{2}} ,$$

$B > 0$ is a specified truncation constant, and $v > 0$ is a specified power constant (note that $\Delta_{I,J}$ is ordinary Euclidean distance when B is ∞ and $v = 1$). The choice of B is purely subjective since its use includes the detection of multiple clumping of objects of a single group. Whereas the symmetric distance functions discussed here are confined to simple variations of Euclidean distance, many alternative choices of symmetric distances functions to include applications such as cyclic and autoregressive data are possible. Incidentally, the choice of $C_i = (n_i-1)/(K-g)$, $N = K$, and $v = 2$ includes one-way ANOVA as a special case. Since $v = 2$, the analysis space of MRPP is not a metric space in this case since the triangle inequality property of a metric is not satisfied. Because the data space is most likely a Euclidean space (i.e., $v = 1$), the analysis space of one-way ANOVA with $v = 2$ usually yields counter-intuitive interpretation problems for an investigator when one or more extreme values occur because the data and analysis spaces are not equivalent. This last comment unfortunately affects most of the presently used statistical techniques because they are based on $v = 2$. The reason for this problem with non-equivalent data and analysis spaces is primarily due to the fact that the majority of statistical techniques based on $v = 2$ were developed to simplify computations between 1900 and 1960 at a time that present day electronic computers were not available. As a consequence the statistical techniques used here are based on $v = 1$ where the data and analysis spaces are congruent to one another in moat cases.

The statistical inference of permutation tests are based on P-values. The exact P-value and two approximate P-values termed resampling and moment P-values are commonly used. The exact P-value is the probability of having a randomly selected statistic as or more extreme than the observed statistic under H_0 that all M possible allocations occur with equal chance. A resampling P-value approximation is the probability of having a without replacement randomly selected statistic among L allocations from the M possible allocations under H_0 (1,000,000 is a common choice for L). A moment P-value approximation is based on fitting the distribution of the statistic to a distribution such as the Pearson type III distribution which is characterized by its first three exact moments. In this approximation the first three exact moments are obtained from the empirical distribution of all M values of the MRPP statistic under H_0. While a special case of the Pearson type III distribution is the normal distribution (which is based on the first two exact moments), the empirical distribution in question is usually very skewed and consequently the Pearson type III is used to compensate for this property. While the exact P-value is the desired result, most samples yield a value of M that makes its evaluation unfeasible. The resampling P-value approximation is usually very good when the exact P-value is not too small. However, if M is very large and the exact P-value is very small (i.e., examples such as a P-value = 10^{-15}), then the moment P-value approximation at least yields a crude approximation of the exact distribution which is obtainable even very extreme conditions. A simple classification for assigning a single object to one of g groups is first to obtain the P-values for each of the g ordered groups (i.e., $p_1,\ldots,p_g$) and second to select the group associated with the smallest value among the g P-values ($p_1,\ldots,p_g$). If the difference between each pair of two or more of the smallest P-values is very small, then the evidence for choosing one group over another is admittedly very weak. As previously mentioned, further details corresponding to MRPP along with many earlier references are given in Mielke and Berry (2007).

REFERENCE

Mielke, P.W., and K.J. Berry (2007). *Permutation Methods: A Distance Function Approach.* (2nd Ed.). New York: Springer Verlag.