Modern Environmental Science and Engineering (ISSN 2333-2581)
January 2020, Volume 6, No. 1, pp. 128-138
Doi: 10.15341/mese(2333-2581)/01.06.2020/012
Academic Star Publishing Company, 2020
www.academicstar.us

The Use of Nonparametric Statistical Inference for Studying the Effects of Construction Waste

Aqli Mursadin

Civil Engineering Graduate Program, Engineering Faculty, Lambung Mangkurat University, Banjarmasin 70123, Indonesia

Abstract: Minimizing construction waste can help achieve the environmental, economic, and social benefits of sustainable construction. Types of waste may include those known as non-value adding activities. Studies on the effects of construction waste on project performance are important to enable mitigation actions. Most of such studies, however, are based on perception surveys. This has led to problems in deriving valid information using parametric methods during the statistical analysis of the response. These problems are mainly related to the assumptions concerning the underlying distribution and the categorical nature of the data. This paper explores a class of nonparametric methods for analyzing survey data concerning the effects of construction waste on project performance. It includes a number of nonparametric tests and post-hoc procedures for repeated measures. Data concerning seven types of construction waste on the generation of material waste from past study are used for this purpose. The results show that consistent outcomes and inferences can be made using different nonparametric methods. A recommendation on which nonparametric methods to use is given.

Key words: construction, nonparametric statistics, waste

1. Introduction

Three dimensions of sustainable construction are environmental, economic, and social dimensions [1, 2]. Minimizing construction waste can potentially help achieve sustainability–that is through the improvement of project performance. By construction waste it means construction activities that consume resources and produce no values. They include the following seven activities: overproduction, waiting, transport, extra processing, inventory, motion, and defects [3].

Studies that reveal construction waste contributions to material waste generation have been reported [4-6]. Other studies have also been carried out to establish how types of construction waste are related to client decisions in construction projects in the Indonesian Provinces of South and East Kalimantan [7] and their effects on project lateness [8]. These studies rely on surveys of perception. Also, the requirement is that the respondents have to have a certain level of construction experience and knowledge.

This creates two problems concerning the validity of information derived from the statistical analysis of the data. Firstly, the limited population of respondents having the above-mentioned qualification usually does not allow a sufficiently large size of data for estimation purposes. Since statistical modelling is mostly based on normality or asymptotic normality for parameter estimation, a small sample size easily leads to large errors of estimation. Secondly, the categorical nature of responses in perception surveys means that those responses are qualitative [9] and may not necessarily represent any underlying continuous variables. Unlike their parametric counterparts, nonparametric analysis techniques are usually insensitive to such limitations.

As much as it is important to understand how construction waste affects project performance and construction sustainability as a whole, it is also critical

Corresponding author: Aqli Mursadin, Doctor, Senior Lecturer. E-mail: a.mursadin@ulm.ac.id.

to ensure that mitigation actions for improving the performance are taken based on valid statistical inference. For this particular reason, the purpose of this paper is to give a recommendation on which nonparametric methods that can be used for comparing the effects of various types of construction waste.

2. Literature Review

2.1 Construction Waste and Project Performance

Construction waste can be defined as construction activities that consume resources and produce no values, which in particular include overproduction, waiting, transport, extra processing, inventory, motion, and defects [3]. These are also known as non-value adding activities. A study in the Indonesian Provinces of South and Central Kalimantan [8] reveals that the majority of these types of waste contribute to late project completion. There is also an overwhelmingly apparent connection between these types of waste and project performance if the latter is measured in terms of the amount of material waste generated during construction [4-6]. The obvious conclusion is that construction waste contributes significantly to material waste generation.

A definition of construction material is given by [10]. It is any form of material on the site apart from earth material that cannot be used for the purpose of the project and has to be removed from the site or used for other purposes within the site. Produced in various types of construction work from demolition to renovation and new building projects [11], construction material waste contributes an estimated 35 percent of the entire waste produced around the world [12]. It is, however, quite surprising that among relatively many research studies on what generates construction material waste [4-6, 12-17] very few have attempted to look at the significant roles of non-value adding activities.

2.2 Nonparametric Methods for Repeated Measures

Data from studies of material waste or the effects of construction waste on project performance as mentioned above are usually used to compare effects from different waste sources and to discover several sources that contribute the most. Another use is for prediction purposes. In this case it requires that a model be built and tested. If it is suspected that there are some underlying and unobservable factors behind the sources and that the relationships between these factors are to be investigated, then a structural equation model (SEM) may be taken into consideration [18]. Whichever intended purpose the data are gathered for, the categorical nature of responses and the usually small sample size make the data relatively incompatible with parametric analysis methods.

In Ref. [4], for instance, perception of the respondents is obtained by asking them to give a score to each of predefined material waste sources using an ordinal scale. Here, a score is given as a measure of the contribution of the corresponding source to the generation of material waste. The score mean for each of the sources is computed as if the corresponding responses are continuous values. The resulting means are then used to rank the sources. This is not an appropriate use of sample mean. This statistic is meaningless if the sample is taken by observing a discrete variable. Also, unlike its median, the sample mean is too sensitive to outliers, especially if the size is small. A similar treatment of data is found in Ref. [5]. A regression model is also developed in Ref. [12]. Any parametric tests on the model may not be valid if, for instance, the assumption of normality is required for the residuals since the sample size may not be large enough for that purpose. This paper, however, focuses on repeated measures of effects only.

When it comes to comparing effects of several different treatments (such as different material waste sources), a one-way analysis of variance (ANOVA) with a repeated measures design or a similar model followed by some post-hoc tests is perhaps the simplest way to perform. ANOVA, however, assumes that the measurement errors are independently and normally

distributed. Consequently, it means that the response values correspond to an underlying continuous variable. It also assumes that a common variance is shared between those different treatments (this is known as homoscedasticity). If any of these assumptions is violated, then the analysis results are not valid. A nonparametric alternative to this type of ANOVA, where the same respondents rate different sources of waste, is the Friedman test [19]. This is actually a nonparametric test for a randomized block design, and a repeated measures design is basically a randomized block one.

The table format for this design is shown in Table 1. In that table, the score by the i-th respondent or block given as a response to the j-th treatment is replaced by its rank, r_{ij}, where $i = 1, \ldots, n$, $j = 1, \ldots, k$, and n and k are the number of respondents or blocks and the number of treatments, respectively. The ranks are obtained within each block by arranging the scores of all treatments in that block in an increasing order. The position of a score in the resulting order is its rank. Hence, for the i-th respondent or block, $r_{ij} \in [1, k]$. A treatment, for instance, may be used to represent a source of material waste or a type of construction waste. It is assumed throughout this paper that the responses are given as nonnegative categorical values in an ordinal scale and a higher score of response corresponds to a stronger effect.

Table 1 Table Format for repeated measures.

Respondent/ block	Treatments					
	1	**2**	**…**	***j***	**…**	***k***
1	r_{11}	r_{12}	…	r_{1j}	…	r_{1k}
2	r_{21}	r_{22}	…	r_{2j}	…	r_{2k}
.	.	.	…	.	…	.
.	.	.	…	.	…	.
.	.	.	…	.	…	.
i	r_{i1}	r_{i2}	…	r_{ij}	…	r_{ik}
.	.	.	…	.	…	.
.	.	.	…	.	…	.
.	.	.	…	.	…	.
n	r_{n1}	r_{n2}	…	r_{nj}	…	r_{nk}

The corresponding test statistic is

$$Q_{\mathrm{F}} = \frac{12(k-1)\left(\sum_{j=1}^{k} R_j^2 - k(k+1)^2 n^2/4\right)}{kn(k^2-1) - \sum_{\forall j}\sum_{\forall r} t_{j(r)}(t_{j(r)}^2 - 1)} \tag{1}$$

where R_j is the sum of ranks in the j-th treatment and $t_{j(r)}$ is the number of scores having the same rank r due to that treatment. A significant difference between the treatments corresponds to a significantly large Q_{F} value. A significant test can be based on an exact distribution of Q_{F} as in Ref. [19] for $k = 3$ and $n \le 8$, and $k = 4$ and $n \le 4$. For other values of k and n, the test can use a chi-square distribution with k–1 degrees of freedom as an approximate distribution of Q_{F}.

A modified version of this statistic as suggested by Iman and Davenport [20] is given as

$$Q_{\mathrm{FID}} = \frac{(n-1)\,Q_{\mathrm{F}}}{(k-1)\,n - Q_{\mathrm{F}}} \tag{2}$$

This statistic follows an F distribution with (k–1) and (k–1)(n–1) degrees of freedom.

Another approach, the rank transformation suggests the use of ANOVA parametrically on the ranks of the scores [21]. The ranks are determined globally across all blocks and treatments, they are also known as RT-1 type ranks. This ANOVA with rank transform (or ANOVA on ranks) is basically a parametric test applied on RT-1 type ranks. The corresponding test statistic is

$$F = \frac{(n-1)}{kn} \frac{k\sum_{j=1}^{k}\left(\sum_{i=1}^{n} r_{ij}\right)^2 - \left(\sum_{i=1}^{n}\sum_{j=1}^{k} r_{ij}\right)^2}{\sum_{i=1}^{n}\sum_{j=1}^{k}(r_{ij} - \bar{r}_{i\cdot} - \bar{r}_{\cdot j} + \bar{r}_{\cdot\cdot})^2} \tag{3}$$

which follows an F distribution with (k–1) and (k–1)(n–1) degrees of freedom, where respectively, $\bar{r}_{i\cdot}$, $\bar{r}_{\cdot j}$, and $\bar{r}_{\cdot\cdot}$ are the rank average in the i-th block, the j-th treatment, and across all blocks and treatments.

The above-mentioned tests do not consider the spread of scores within different treatments, while

responses belonging to some groups may be more homogeneous than others. Taking this spread of values into consideration can be done by assigning weights to the treatments as in the Quade test [23]. The corresponding test statistic is given by

$$Q_{\mathrm{W}} = \frac{(n-1)A}{\left(n\,(n+1)\,(2n+1)\,(k-1)\,k\,(k+1)/72\right) - A} \tag{4}$$

which follows an F distribution with (k–1) and (k–1)(n–1) degrees of freedom, where

$$A = \frac{1}{n}\sum_{j=1}^{k}\left(\sum_{i=1}^{n} s_{ij}\right)^2 \tag{5}$$

n, k, and r_{ij} are as before, and

$$s_{ij} = w_i\left(r_{ij} - \frac{k+1}{2}\right) \tag{6}$$

which is known as the *weighted adjusted-rank*. The value of w_i is the weight assigned to the i-th block. Notice how r_{ij} is adjusted before weighted. It is actually the i-th block's adjusted rank relative to other blocks. These ranks of blocks are determined based on the range of the original scores within each of them, that is, the difference between the maximum and the minimum of the scores. Arranging these range values in an increasing order reveals w_i's for i = 1, …, n.

An important criterion in selecting a test is its power. The power of a test is the probability that the test *correctly* classifies a result as *not* significant. García et al. [24] conducted a series of experimental studies to measure the power of several tests for repeated measures. According to the results, Quade test tends to be more powerful than Friedman test.

If the responses (or the residuals) are normally and independently distributed, then ANOVA is the most powerful. Other tests for repeated measures are usually compared to a corresponding ANOVA model in terms of power. Given the same data, the power of a nonparametric test for repeated measures relative to a corresponding ANOVA model is given as its *asymptotic relative efficiency* (*ARE*). For Friedman test, this *ARE* is a function of the number of treatments (k) as follows [25]

$$ARE = 0.955\frac{k}{k+1} \tag{7}$$

For k = 7, for instance, Friedman test is 83.56 percent as powerful as ANOVA with randomized block design.

2.3 Post Hoc Tests

If a significant difference is detected between the treatments, a post hoc test is usually carried out to compare every two treatments with each other such that all treatments can be put in an order of increasing effect. Paired comparison such as this is what should have been performed in Ref. [4] instead of the ranking of the score means. Further, a treatment with the lowest rank sum (or the lowest weighted adjusted-rank sum in the case of the Quade test) can be used as a baseline to which every other treatment will be compared. This makes sure that the maximum number of comparisons is k(k–1)/2.

A famous nonparametric paired test is the Wilcoxon signed-rank test. It looks at the differences between two dependent samples or paired observations drawn from a continuous population [19]. However, it has an issue when some of the differences are zero, for which the corresponding pairs are usually ignored in the computation and the value of n has to be reduced accordingly. This is not a good practice when the number of blocks is already small. Another issue is the presence of tied differences, namely, differences having the same value.

A simpler Dunn test which is based on mean-ranks is an alternative for this purpose. However, it has been shown that a comparison result between two treatments using this test is also affected by other treatments [26].

A special attention is given to a less well known comparison test, the Conover-Iman test [27]. It takes rank sums (or weighted adjusted-rank sums) and every single rank (or weighted adjusted-rank) into consideration. According to this test, a significantly large value of

$$T = \frac{1}{U}\frac{\left|R - R_{\mathrm{B}}\right|}{\sqrt{\left(\frac{2n(k-1)}{kn-k-n+1}\right)\left(1-\frac{V}{n(k-1)}\right)}} \tag{8}$$

corresponds to a particular treatment having higher total scores than those of the baseline if $R > R_{\mathrm{B}}$. Here, R_{B} and R are, respectively, the baseline sum of ranks and the sum of ranks of the particular treatment,

$$U = \sqrt{\frac{1}{k-1}\left(\sum_{i=1}^{n}\sum_{j=1}^{k} r_{ij}^2 - \frac{kn(k+1)^2}{4}\right)} \tag{9}$$

$$V = \frac{1}{U^2}\sum_{j=1}^{k}\left(R_j - \frac{n(k+1)}{2}\right)^2 \tag{10}$$

and r_{ij}, R_j n, and k are as before. If the difference is not significant, T in (8) follows a Student's t distribution with $kn–k–n+1$ degrees of freedom. This is a two-tail significance test. For Quade test, s_{ij} and S_j are used in place of r_{ij} and R_j in (9) and (10), respectively, where S_j is the sum of weighted adjusted-ranks in the j-th treatment and s_{ij} is as given in (6). Also, for Quade test, the difference $R–R_{\mathrm{B}}$ is replaced with $S–S_{\mathrm{B}}$, where S_{B} is the baseline sum of weighted adjusted-ranks.

Every paired comparison between a treatment and the baseline produces a p-value. This value is a measure of significance of the test. The more significant the difference between the two rank sums is, the smaller the p-value becomes. The increasing order in which the resulting p-values can be arranged is the decreasing order of the effects. However, for the sake of accuracy, before any conclusion can be drawn from the test, every single one of these p-values needs to be checked against an adjusted significant level. A fixed significant level α (e.g., 0.05) is chosen for all the k–1 tests as the probability of falsely concluding that the difference is significant.

The Bonferroni adjustment is a simple one-step procedure for this purpose [28]. According to this procedure, if

$$p\text{-value} < \frac{\alpha}{k-1} \tag{11}$$

then the test result is significant. Several other procedures may also be used. Suppose that p-value$_{(h)}$ is the h-th smallest among the k–1 p-values, where $h \le k$–1. Using the adjustment procedure by Holm [29], if

$$p\text{-value}_{(h)} < \frac{\alpha}{k-h} \tag{12}$$

then the corresponding test result is significant. This procedure is repeated for all available h's. Holland and Copenhaver [30] proposed another adjustment, namely,

$$p\text{-value}_{(h)} < 1-(1-\alpha)^{k-h} \tag{13}$$

Meanwhile, Finner [31] also suggested the use of

$$p\text{-value}_{(h)} < 1-(1-\alpha)^{(k-1)/h} \tag{14}$$

A two-step procedure by Li [32] starts by checking if

$$p\text{-value}_{(k-1)} < \alpha \tag{15}$$

If that is true, then all the test results are significant. Otherwise, it proceeds to the second step to see, for all $h \le k$–1, whether

$$p\text{-value}_{(h)} < \alpha\left(p\text{-value}_{(k-1)}\right)\big/(1-\alpha) \tag{16}$$

If that is the case, then the corresponding test result is significant.

According to Ref. [24], Bonferroni adjustment is likely to be the worst in terms of power. Procedures by both Holm and Holland tend to behave similarly to each other, while those by Finner and Li usually perform the best. However, when the blocks are very similar to each other, the latter two are likely to incorrectly classify the corresponding results as not significant. In such a case, the adjustment by Li even performs worse than that of Bonferroni.

Once the treatments can be grouped into being significantly or not significantly different from the baseline, the above procedures are executed once again, this time with a new baseline and without the old one. The new baseline is chosen the same way as before and the number of treatments becomes k–1. This is repeated until all the $k(k–1)/2$ comparisons are made or no more

differences between treatments can be found, whichever happens first.

2.4 Measures of Effect Size

The effect of each of the waste types on the response in a repeated measures design cannot be measured directly (unlike, for instance, the case of regression models). This is the effect of variation *within* a particular type. However, the effect of variation *between* the types in this design can be calculated. This can be used to indicate the proportion of variation within the response that can be explained by variation between the types. This class of *effect size* is different from correlation coefficients. Although together they indicate the proportion of variance explained, correlation coefficients would — in this case — measure the effect of variation within a particular type *on* variation within the response.

An obvious effect size for Friedman test is Kendall's *W*, which is the ratio of the variability among types and the maximum possible variability [19]. In fact, this is a normalized form of Friedman's Q_F from Eq. (1) and given as follows:

$$W = \frac{Q_F}{n(k-1)} \tag{17}$$

with values ranging from 0 to 1. For ANOVA on ranks, two measures of effect size for ANOVA can be used, namely [33],

$$\eta_P^2 = \frac{k\sum_{j=1}^{k}\left(\sum_{i=1}^{n} r_{ij}\right)^2 - \left(\sum_{i=1}^{n}\sum_{j=1}^{k} r_{ij}\right)^2}{k\sum_{j=1}^{k}\left(\sum_{i=1}^{n} r_{ij}\right)^2 - \left(\sum_{i=1}^{n}\sum_{j=1}^{k} r_{ij}\right)^2 + kn\,SSE} \tag{18}$$

and

$$\omega^2 = \frac{k\sum_{j=1}^{k}\left(\sum_{i=1}^{n} r_{ij}\right)^2 - \left(\sum_{i=1}^{n}\sum_{j=1}^{k} r_{ij}\right)^2 - kn(n-1)s_E^2}{kn\sum_{i=1}^{n}\sum_{j=1}^{k} r_{ij}^2 - \left(\sum_{i=1}^{n}\sum_{j=1}^{k} r_{ij}\right)^2 + kn\,s_E^2} \tag{19}$$

where

$$SSE = \sum_{i=1}^{n}\sum_{j=1}^{k}(r_{ij} - \bar{r}_{i.} - \bar{r}_{.j} + \bar{r}_{..})^2 \tag{20}$$

and

$$s_E^2 = \frac{SSE}{(k-1)(n-1)} \tag{21}$$

The bigger these measures are, the bigger the effect of variation between the types on the response becomes.

There are no clear measures of effect size for Quade test. However, if ANOVA is performed on the weighted adjusted-ranks, then Eqs. (18) and (19) can be used as the measures.

3. Methods

Data from a study of construction material waste generation as reported in Ref. [6] will be used for demonstration. In that study, 26 respondents were randomly chosen in the Banjar Regency (South Kalimantan) from a limited population of construction professionals having a minimum of 5 years of experience in the field. They were asked to judge the relative amount of material waste generated in relation with seven types of construction waste. Basically, this was a study on the effects of construction waste. The responses were given in a 5-point ordinal scale (1 = extremely small, 2 = small, 3 = moderate, 4 = large, 5 = extremely large).

Repeated measures design is a clever choice of model for such a noticeably small number of respondents. With $n = 26$ and $k = 7$, it virtually inflates the size to 182.

The three ANOVA-like tests will be performed on the data, namely, Friedman test, ANOVA on ranks, and Quade test. A session of ANOVA will also be run with no verification of the usual assumptions (normality, independence, and homoscedasticity). The test statistic for ANOVA is given by Eq. (3) with the ranks being replaced with the corresponding original responses. If

the result from any of the above tests for repeated measures is significant, a series of post hoc tests and *p*-value adjustment procedures will follow. Conover-Iman test and five *p*-value adjustment procedures will be employed for this purpose.

Comparisons between different tests or procedures at every phase of the analysis will also be conducted. The purpose of the comparisons is to examine the consistency and validity of the tests or procedures relative to each other.

4. Results

Table 2 shows the data using a format given in Table 1. Here, every row and every column corresponds to a respondent and a waste type, respectively.

4.1 Repeated Measures

Table 3 shows the test results. Each of the *p*-values indicates that the result is significant for the corresponding test. Together, the tests consistently conclude the significance of the effects. The values of effect size also show a degree of agreement. Interestingly, both ANOVA on ranks and Quade test produce relatively bigger values of effect size.

The assignment of ranks (or weighted adjusted-ranks) in the cells of Table 2 varies from one test to another. This variation results in different orders of effects supposedly brought by the waste types to the relative amount of material waste.

Fig. 1 shows the sums of ranks or weighted adjusted-ranks or scores of the tests for repeated measures. If the sums of ranks (or weighted adjusted-ranks) in different types of waste are arranged in an increasing order for every test, that order will show the relative strength of a waste type in generating material waste. The order for ANOVA is based on the original responses. Table 4 shows the resulting orders.

All the tests show that some types of waste affect the generation of material waste differently from one another. Both Friedman test and ANOVA on ranks result in exactly the same order of waste types.

Table 2 Data from Ref. [6] in repeated measures design.

Respondent	Responses on waste types*						
	A	B	C	D	E	F	G
1	1	2	1	1	2	2	1
2	1	1	1	1	1	1	1
3	4	1	2	3	3	2	4
4	2	1	1	2	3	1	3
5	2	2	2	2	2	2	2
6	2	1	1	3	1	1	2
7	2	3	2	2	2	2	2
8	3	3	4	4	2	3	4
9	2	3	1	3	3	2	3
10	2	2	2	3	2	3	2
11	3	2	3	1	1	2	2
12	3	2	1	2	3	2	2
13	4	2	3	3	2	1	4
14	3	2	2	3	2	2	3
15	2	2	1	2	1	1	1
16	1	1	1	2	1	2	1
17	1	2	1	1	1	2	1
18	4	1	1	4	3	1	4
19	3	2	2	3	3	4	2
20	3	2	2	3	2	2	2
21	3	2	2	3	3	4	2
22	4	2	2	3	3	2	4
23	3	1	2	2	1	2	2
24	2	1	3	3	4	4	4
25	2	2	2	2	2	3	2
26	2	1	1	1	2	2	1

* A = overproduction, B = waiting, C = transport, D = extra processing, E = inventory, F = motion, G = defects

Table 3 Test results.

Test	Test statistics	*p*-value	Effect size*
Friedman	3.818	0.0014	0.127
ANOVA on ranks	4.715	0.0002	0.159 0.075
Quade	31.608	0.0000	0.151 0.074
ANOVA	4.150	0.0007	0.142 0.063

* For ANOVA on ranks, the Quade test, and ANOVA, the first measure of effect size is ηp^2, the second is ω^2.

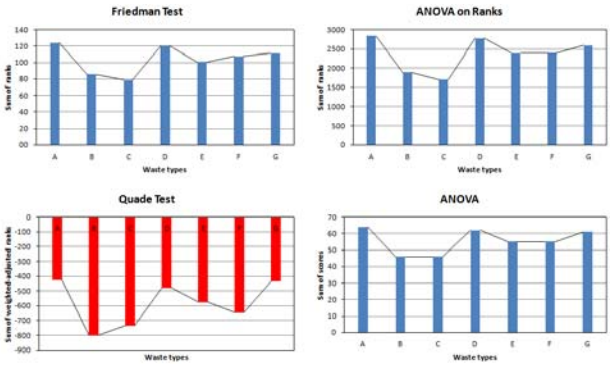

Fig. 1 Sums of ranks or weighted adjusted-ranks or scores (A = overproduction, B = waiting, C = transport, D = extra processing, E = inventory, F = motion, G = defects).

Table 4 Resulting orders of waste types.

Test	Waste types* in an increasing order of effect strength
Friedman	C, B, E, F, G, D, A
ANOVA on ranks	C, B, E, F, G, D, A
Quade	B, C, F, E, D, G, A
ANOVA	C, B, E, F, G, D, A

* A = overproduction, B = waiting, C = transport, D = extra processing, E = inventory, F = motion, G = defects

Overproduction is found as having the strongest effect and transport as having the weakest. Waiting does not seem to strongly affect material waste generation compare to inventory, which along with motion and defects exhibit moderate effects. Extra processing shows a strong effect right below that of overproduction. The same order is also produced by ANOVA.

On the other hand, Quade test produces a slightly different order. This is due to the fact that Quade test incorporates information concerning the spread of scores. Overproduction is still the strongest in terms of effect, while other waste types swap positions with each other. Inventory and motion remain moderate. Extra processing becomes less important and swaps positions with defects. Also, waiting becomes the weakest in terms of effect followed by transport.

4.2 Paired Comparisons and p-Value Adjustment

Now, it is time to delve deeper into the differences between the types of waste to establish how significant they actually are. A post hoc test by Conover-Iman [27] on every order followed by five *p*-value adjustment procedures gives results as shown in Table 5.

The number of paired comparisons indicates how many times Conover-Iman test is performed. The maximum number is 18 resulting in four groups of

Table 5 Post hoc results.

Test	Group* of waste types (adjustment proc.)	# of paired comp.
Friedman	C, B, E, F, G B, E, H, G, D, A (Bonferroni)	11
	C, B, E, F B, E, F, G, D, A (Holm, Li)	11
	C, B B, E, F, G E, F, G, D, A (Holland)	15
	C, B, E B, E, F, G, D E, F, G, D, A (Finner)	15
ANOVA on ranks	C, B, E, F, G, D, A (Bonferroni, Holm)	6
	C, B, E B, E, F E, F, G, D, A (Holland)	15
	C, B, E, F B, E, F, G E, F, G, D, A (Finner)	15
	C, B, E, F, G B, E, F, G, D E, F, G, D, A (Li)	15
Quade	B, C C, F, E F, E, D E, D, G, A (Bonferroni)	18
	B, C C, F F, E, D E, D, G, A (Holm)	18
	B, C C, F F, E E, D, G, A (Holland, Finner, Li)	18

* A = overproduction, B = waiting, C = transport, D = extra processing, E = inventory, F = motion, G = defects

waste types. The second largest number is 15 and it is related to three groups as the result. Two groups are produced after 11 comparisons, and one after only 6.

Each of the adjustment procedures puts different types of waste into groups. Within any given group, the waste types are not significantly different from each other in terms of effect. For instance, Bonferroni procedure on the result of Conover-Iman test following Friedman test identifies two distinct groups of waste types. Transport, waiting, inventory, motion, and defects are not significantly different from each other in terms of effect, and neither are waiting, inventory, motion, defects, extra processing, and overproduction. Some of the procedures produce exactly the same groups for the same test. Both Holm and Li procedures, for instance, result in transport, waiting, inventory, and motion in one group, *and* waiting, inventory, motion, defects, extra processing, and overproduction in the other.

The results demonstrate different abilities by the procedures in separating waste types based on their effects. The more groups a procedure can produce, the more sensitive the procedure is to the difference between types. However, it also depends on the results from the corresponding test. In general, Holland and Finner procedures are likely to be more sensitive than the others as they use up the maximum number of comparisons for any given test. On the other hand, Bonferroni and Holm procedures are the least sensitive.

4.3 Recommendation of Methods

Results from repeated measures show that the tests are consistent with each other. There is no abrupt change between the orders of waste types that they produce. In fact, the difference is only due to information of the spread of scores being incorporated into the computation by Quade test.

Hence, the nonparametric tests recommended for studying of effects of construction waste with repeated measures are

- Friedman test and ANOVA on ranks if the spread of data is not so severe, and
- Quade test if the spread of data is severe.

It should be followed by Conover-Iman test as the post hoc test.

Further, the grouping created by *p*-value adjustment is useful for scrutinizing the differences among construction waste types more quantitatively. It enables judgement on how different a certain set of waste types are from the others.

In general, these procedures perform relatively consistent with each other. In particular, however, Both Holland and Finner procedures tend to result in relatively more groups than the others. Li procedure performs rather moderately in terms of the number of resulting groups.

Hence, the *p*-value adjustment procedures recommended for the above nonparametric tests and their corresponding post hoc results are

- Holland and Finner procedures if finer grouping among waste types is an objective, and
- Bonferroni, Holm, and Li procedures if finer grouping among waste types is not an objective.

5. Conclusion

The use of three nonparametric tests for repeated measures has been explored. The results have been demonstrated on the effects of construction waste. Conover-Iman test has also been used to obtain the corresponding post hoc results. A series of *p*-value adjustment procedures have been shown to result in inferences concerning the grouping of different waste types. In general, these procedures perform relatively consistent with each other. Some procedures, however, tend to produce finer grouping than the others. Finally, a recommendation on which tests and procedures to be used has been given.

References

[1] B. Stubbs, *Plain English Guide to Sustainable Construction*, London, UK: Constructing Excellence, 2007, pp. 3-6.

[2] H. Danso, Identification of key indicators for sustainable construction materials, *Res. & Dev. Material Sci.* 3 (2018) (4) Article ID 6916258.

[3] J. E. Diekman, M. Krewedl, J. Balonick, T. Stewart and S. Won, *Application of Lean Manufacturing Principles to Construction*, Austin, TX: University of Austin, 2004, pp. 60-61.

[4] Aiyetan and J. Smallwood, Materials management and waste minimization on construction sites in Lagos State, Nigeria, in: *Proc. The 4th Int'l Conf. on Eng., Project and Prod. Management*, Bangkok, Thailand, 23-25 October 2013.

[5] O. Fadiya, P. Georgakis and E. Chinyio, Quantitative analysis of the sources of construction waste, *J. Construction Eng.* (2014) 1-9.

[6] Mursadin and Isra, Modeling the relationship between material waste generation and NVAAs in construction work, in: *Proc. 3rd International Conference on Emerging Trends in Academic Research*, Banjarmasin, Indonesia, 26–27 September 2016.

[7] Mursadin, Modelling contractor response to client-induced construction waste, in: *The 4th International Conference on Asset and Facility Management*, Padang, Indonesia, 4-5 October 2017.

[8] Mursadin, Consensus of construction professionals on the contribution of construction waste to project lateness: A case study in the Indonesian Provinces of South Kalimantan and Central Kalimantan, in: *The International Conference on Engineering, Technologies, and Applied Sciences*, Bandar Lampung, Indonesia, 18-20 October 2018.

[9] Agresti, *An Introduction to Categorical Data Analysis* (2nd ed.), Hoboken, NJ: John Wiley & Sons, Inc, 2007, pp. 1-3.

[10] L. L. Ekanayake and G. Ofori, Construction material waste source evaluation, in: *Proc. Strategies for a Sustainable Built Environment*, Pretoria, South Africa, August 2000.

[11] P. Nowak, M. Steiner, and U. Wiegel, Waste management challenges for the construction industry, *Construction Information Quarterly* 11 (2009) (1) 8.

[12] J. Solis-Guzman, M. Marrero, M. V. Montes-Delgado, and A. Ramirez-de-Arellano, A Spanish model for quantification and management of construction waste, *Waste Management* 29 (2009) (9) 2542-2548.

[13] K. Cochran, T. Townsend, D. Reinhart and H. Heck, Estimation of regional building-related C&D debris generation and composition: Case study for Florida, US, *Waste Management* 27 (2007) (7) 921-931.

[14] T. Y. Hsiao, Y. T. Huang, Y. H. Yu, and I. K. Wernick, Modeling materials flow of waste concrete from construction and demolition wastes in Taiwan, *Resources Policy* 28 (2002) (1-2) 39-47.

[15] Martinez-Lage, F. M. Abella, C. V. Herrero and J. L. P. Ordonez, Estimation of the annual production and composition of C&D debris in Galicia, *Waste Management* 30 (2010) (4) 636-645.

[16] A. D. S. Wimalasena, J. Y., Ruwanpura and J. P. A. Hettiaratchi, On-site construction waste management, in: *The Intrnational Conf. on Sustainable Built Environment (ICSBE-2010)*, Kandy, Sri Lanka, 13-14 December 2010.

[17] W. Zhao, R. B. Leeftink and V. S. Rotter, Evaluation of the economic feasibility for the recycling of construction and demolition waste in China — The case study of Chongqing, *Resources, Conservation and Recycling* 54 (2010) (6) 377-389.

[18] R. B. Kline, *Principles and Practice of Structural Equation Modeling* (3rd ed.), New York, NY: Guilford Publications, 2010, pp. 230-264.

[19] J. D. Gibbons and S. Chakraborti, *Nonparametric Statistical Inference* (4th ed.), New York, NY: Marcel Dekker, Inc., 2003, pp. 453-462.

[20] R. L. Iman and J. M. Davenport, Approximations of the critical region of the Friedman statistic, *Communications in Statistics* 9 (1980) (6) 571-595.

[21] W. J. Conover and R. L. Iman, Rank transformations as a bridge parametric and nonparametric statistics, *The American Statistician* 35 (1981) (3) 124-129.

[22] E. Brunner and M. L. Puri, Nonparametric methods in factorial designs, *Statistical Papers* 42 (2001) (1) 1-52.

[23] Quade, Using weighted rankings in the analysis of complete blocks with additive block effects, *J. American Statistical Assoc.* 74 (1979) (367) 680-683.

[24] S. García, A. Fernández, J. Luengo and F. Herrera, Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power, *Information Sciences* 180 (2010) 2044-2064.

[25] W. Hager, Some common features and some differences between the parametric ANOVA for repeated measures and the Friedman ANOVA for ranked data, *Psychology Science* 49 (2007) (3) 209-222.

[26] Benavoli, G. Corani and F. Mangili, Should we really use post-hoc tests based on mean-ranks?, *J. Machine Learning Res.* 17 (2016) 1-10.

[27] W. J. Conover and R. L. Iman, On multiple-comparisons procedures, Report LA-7677-MS, Los Alamos, NM: Los Alamos Scientific Laboratory, 1979.

[28] J. Dunn, Multiple comparisons among means, *J. Am. Statistical Assoc.* 56 (1961) (296) 52-64.

[29] S. Holm, A simple sequentially rejective multiple test procedure, *Scandinavian J. Stat.* 6 (1979) (2) 65-70.

[30] S. Holland and M. D. Copenhaver, An improved sequentially rejective Bonferroni test procedure, *Biometrics* 43 (1987) (2) 417-423.

[31] H. Finner, On a monotonicity problem in step-down multiple test procedures, *J. Am. Statistical Assoc*. 88 (1993) (423) 920-923.

[32] J. Li, A two-step rejection procedure for testing multiple hypotheses, *J. Statistical Planning and Inference* 138 (2008) (6) 1521-1527.

[33] C. Howell, Statistical Methods for Psychology (7th ed.), Belmont, CA: Cengage Wadsworth, 2010, pp. 343-348.