Semiquantitative spectrographic analysis and rank correlation in geochemistry

By F. J. Flanagan

Trace Elements Investigations Report 628

UNITED STATES DEPARTMENT OF THE INTERIOR
GEOLOGICAL SURVEY

Chemistry

This document consists of 17 pages.
Series A.

UNITED STATES DEPARTMENT OF THE INTERIOR

GEOLOGICAL SURVEY

SEMIQUANTITATIVE SPECTROGRAPHIC ANALYSIS AND RANK CORRELATION IN GEOCHEMISTRY*

By

F. J. Flanagan

October 1956

Trace Elements Investigations Report 628

This preliminary report is distributed without editorial and technical review for conformity with official standards and nomenclature. It is not for public inspection or quotation.

---

*This report concerns work done on behalf of the Division of Raw Materials of the U. S. Atomic Energy Commission.

USGS - TEI-628

CHEMISTRY

Distribution (Series A)	No. of copies
Argonne National Laboratory	1
Atomic Energy Commission, Washington	2
Division of Raw Materials, Albuquerque	1
Division of Raw Materials, Austin	1
Division of Raw Materials, Casper	1
Division of Raw Materials, Denver	1
Division of Raw Materials, Ishpeming	1
Division of Raw Materials, Phoenix	1
Division of Raw Materials, Rapid City	1
Division of Raw Materials, Salt Lake City	1
Division of Raw Materials, Spokane	1
Division of Raw Materials, Washington	3
Exploration Division, Grand Junction Operations Office	1
Grand Junction Operations Office	1
Technical Information Service Extension, Oak Ridge	6
Tennessee Valley Authority, Wilson Dam	1
U. S. Geological Survey:	
Fuels Branch, Washington	1
Geochemistry and Petrology Branch, Washington	20
Geophysics Branch, Washington	1
Mineral Deposits Branch, Washington	1
P. C. Bateman, Menlo Park	1
A. L. Brokaw, Grand Junction	2
N. M. Denson, Denver	1
V. L. Freeman, College	1
R. L. Griggs, Albuquerque	1
W. R. Keefer, Laramie	1
A. H. Koschmann, Denver	1
L. R. Page, Washington	1
Q. D. Singewald, Beltsville	1
F. N. Ward, Denver	1
A. E. Weissenborn, Spokane	1
TEPCO, Denver	2
TEPCO, RPS, Washington, (including master)	2
	63

CONTENTS

ILLUSTRATION

TABLES

SEMIQUANTITATIVE SPECTROGRAPHIC ANALYSIS AND RANK CORRELATION IN GEOCHEMISTRY

By F. J. Flanagan

ABSTRACT

The rank correlation coefficient, r_s, which involves less computation than the product moment correlation coefficient, r, can be used to indicate the degree of relationship between two elements. The method is applicable in situations where the assumptions underlying normal distribution correlation theory may not be satisfied. Semiquantitative spectrographic analyses can be used to calculate rank correlations between elements because the results are already partly ranked in an analytical report.

INTRODUCTION

One of the major problems in geochemistry is the origin of ore deposits. A logical approach to this problem is the relationship between elements in and around deposits; hence the search for associations between elements is important.

These associations between such elements are usually measured by the product moment correlation coefficient, r, and the significance or nonsignificance of the association is obtained by references to tables of correlation coefficients at the 1 percent and 5 percent levels of significance (Snedecor, 1946; Fisher and Yates, 1948) for different degrees of freedom. In practice, normality of the parent populations, independence of successive pairs of observations, and homogeneity of variances are generally, and very frequently unknowingly, assumed. However, because of its value to personnel in geological sciences, this use of the method should not be rejected in spite of both the tediousness of computations where a large number of pairs of observations are involved and the failure to validate the assumptions.

The lack of tables of significant correlation coefficients complicates the problem somewhat. In this case, however, t or F tables are more available than tables of significant correlation coefficients and we may still make our decision about the significance of r with only slight additional computation. As shown in most introductory texts, for example, Goulden (1952), the statistic t with n-2 degrees of freedom may be calculated by substitution in the formula

$$t_{n-2} = r\sqrt{(n-2)/1-r^2}$$

and the calculated t then compared with one in the table with n-2 degrees of freedom at a chosen confidence level. If one has only F tables available, one could use the square of the correlation coefficient, r^2, which is the proportion of the variation in one set of measurements which can be explained by their dependence on the other. By substitution in the formula

$$F = \frac{r^2(n-2)}{1-r^2}$$

one may calculate an F ratio which may be compared to values in the tables at chosen confidence levels and with degrees of freedom (1, n-2).

RANK CORRELATION

The calculation of correlation coefficients using the method of least squares is tedious when the data consist of two variables each with three significant figures, and when there are twenty or more pairs of such variables. Such a situation is common in geochemical work where many chemical analyses are reported to three significant figures.

Both computational labor and verification of the basic assumptions can be avoided by calculating Spearman's rank correlation coefficient. The method is nonparametric and the assumptions need not be made. Like the product moment correlation coefficient, the rank correlation coefficient, r_s, can assume values between $\pm$ 1, with larger absolute values indicating a higher degree of association. Computations are rapid and if the results are not sufficiently precise for the purpose at hand, for example, a general search for associations, they can be used as a priority basis for the more involved calculations of the usual correlation coefficient. The theoretical aspects of the method have been treated by Kendall (1948), and many texts, including one directed to the attention of chemists (Bennett and Franklin, 1954), give the algebraic derivation of the expression used to calculate the rank correlation coefficient.

The method of computation is briefly as follows: in a set of n samples on which elements x and y have both been determined, the determinations of element x can be arranged in increasing numerical value and ranked from 1, 2, 3, ... to n where rank 1 contains the least and rank n the greatest amount of element x. The y observations are similarly ranked. The difference between the ranks for x and y are taken for each sample, squared, and these squared rank differences $[(R.D.)^2]$ summed over the n samples. The Spearman rank correlation coefficient, r_s, may then be calculated from

$$r_s = 1 - \frac{6\Sigma(R.D.)^2}{n\ (n^2-1)} .$$

The null hypothesis that there is no association or that all permutations of one ranking relative to the other are equally likely is then set up and the significance of r_s tested by

$$t_{n-2} = r_s\sqrt{(n-2)/(1-r_s^2)}$$

as Student's t is a good approximation to the distribution of the function of r_s above for $n \geq 9$ (Kendall, 1948, p. 46-47). It might also be noted that for larger size samples ($n > 20$) the normal distribution is a good approximation to the distribution of $r_s\sqrt{(n-1)}$.

If, however, one substitutes for r_s in the above equation for t, the resulting expression can be rearranged to give

$$\Sigma(R.D.)^2 = \frac{n^3 - n}{6} \left[1 \pm \sqrt{\frac{t^2}{n-2+t^2}}\right].$$

Substitution of t with n-2 degrees of freedom into this last equation allows calculations of upper and lower limits for the sum of the squared rank differences and these have been tabled. The exact distribution of the sum of the squared rank differences has been studied by Kendall (1948), and Olds (1938, 1949) has tabled these sums for n from 5 to 30. Litchfield and Wilcoxon (1955) have recently extended these tables from $n = 31$ to $n = 40$ (table 1) and in addition have prepared a nomograph (fig. 1) from which, knowing n and $\Sigma(R.D.)^2$, the rank correlation coefficient, r_s, may be read directly.

SEMIQUANTITATIVE SPECTROGRAPHIC ANALYSES IN RANK CORRELATION

The spectrograph has been shown by Waring and Annell (1953) to be a useful tool for the semiquantitative determination of 68 elements. The results of the analysis were reported in brackets of powers of ten, that is, 0.01 to 0.1, 0.1 to 1.0, 1.0 to 10.0, and the authors estimated about 90 percent agreement with chemical results. As usually happens to analytical methods, either analysts further refine methods or the consumer of the analytical data requests refinements in the estimates. Hence the spectrographic laboratories of the U. S. Geological Survey now report semiquantitative determinations in both two and three brackets per power of ten. The

Table 1.--Critical values of $\Sigma(R.D.)^2$.

	Probability			Probability	
n*	0.05	0.01	n*	0.05	0.01
			21	873-2207	695-2385
			22	1022-2520	820-2722
			23	1187-2861	960-3088
			24	1370-3230	1115-3485
5	0-40		25	1570-3630	1287-3913
6	4-66	0-70	26	1789-4061	1475-4375
7	12-100	4-108	27	2028-4524	1681-4871
8	22-146	10-158	28	2287-5021	1906-5402
9	40-200	24-216	29	2569-5551	2149-5971
10	61-269	39-291	30	2873-6117	2414-6576
11	88-352	58-382	31	3199-6721	2700-7220
12	121-451	84-488	32	3550-7362	3008-7904
13	163-565	115-613	33	3926-8042	3338-8630
14	213-697	154-756	34	4328-8762	3693-9397
15	272-848	201-919	35	4757-9523	4073-10,207
16	342-1018	257-1103	36	5213-10,327	4476-11,064
17	423-1209	322-1310	37	5698-11,174	4908-11,964
18	515-1423	398-1540	38	6213-12,065	5366-12,912
19	621-1659	484-1796	39	6758-13,002	5853-13,907
20	740-1920	583-2077	40	7334-13,986	6367-14,953

*n = number of pairs ranked.

Use of table: If the observed total is equal to or less than the appropriate lower tabular value, or equal to or greater than the appropriate higher tabular value, the correlation is significant for that probability. High values correspond to negative correlations, and low values to positive correlations.

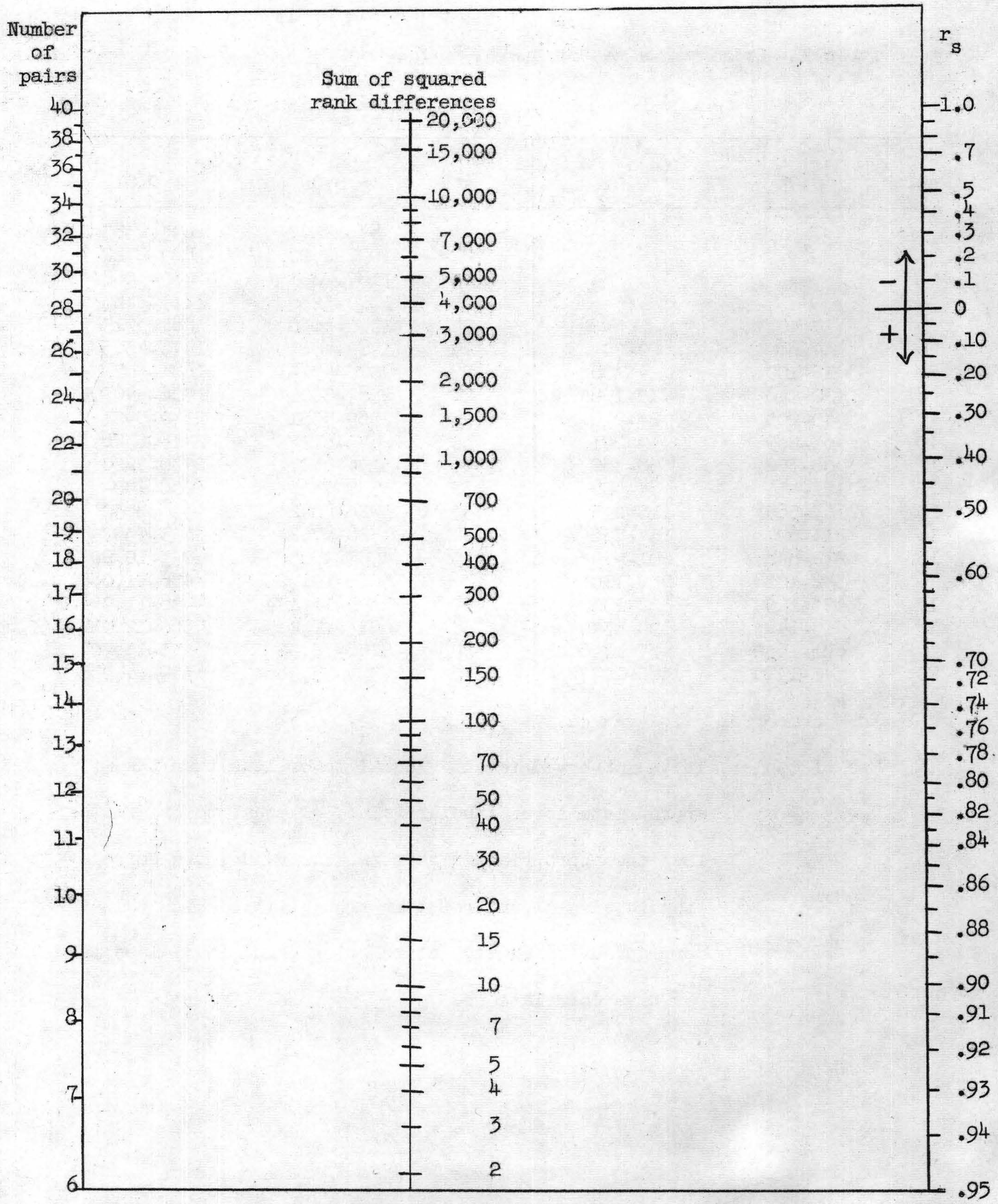

Figure 1.--Nomograph for r_s, the rank correlation coefficient (Litchfield and Wilcoxon, 1955).

first of these divides the samples into brackets of 0.1 to 0.5 and 0.5 to 1.0 for powers of ten, and 80 percent agreement is estimated. The three-bracket system is based on intervals of 1 to $\sqrt[3]{10}$, $\sqrt[3]{10}$ to $\sqrt[3]{10^2}$, and $\sqrt[3]{10^2}$ to $\sqrt[3]{10^3}$ for any power of ten and these may be represented numerically with no significance attached to the digits as 1 to 2.1, 2.1 to 4.6, and 4.6 to 10. Sixty percent agreement is estimated for this three-bracket reporting system.

The one-, two-, and three-bracket systems of reporting have the common property that they almost automatically rank the determinations. This is more obvious in the one-bracket system where, if there were one each of four determinations reported in the four brackets 0.001 - 0.01, 0.01 - 0.1, 0.1 - 1, and 1-10, it can be seen that the powers of ten of the lower bound of each bracket are -3, -2, -1, and zero, respectively. By adding a digit to the exponent of the lower bound of the lowest bracket so that the addition results in +1 for this bracket, correct ranks are automatically assigned. Hence, by adding four to the exponents above, the ranks of 1, 2, 3, and 4 are obtained respectively. This is an excellent procedure where there is one observation per bracket.

This simple procedure is not directly applicable to the two- or three-bracket systems nor to the one-bracket system where there are no entries in an intermediate bracket, or where there are replicate entries in one or more brackets. In these cases one must arrange the brackets in increasing numerical value and rank the observations in the usual fashion.

As these methods of reporting semiquantitative data partly rank the observations and the ranking itself is easily accomplished, the merits of using this kind of data for the calculation of rank correlation coefficients are obvious. One slight disadvantage of using semiquantitative spectrographic data is that the data are grouped and hence we have tied ranks.

As an example, a set of 10 observations of this kind might be distributed 1, 3, 4, and 2 in four brackets. As the measurement process does not distinguish among members of the group, ties are assigned the mean value of the ranks. The observation in the first bracket is ranked 1. In the second bracket the ranks would have been 2, 3, and 4 had the observations been slightly different. The mean of these three ranks is 3 and hence the observations in the second bracket are ranked 3, 3, and 3. The mean rank of the next bracket is 6.5, or $[\frac{5 + 6 + 7 + 8}{4}]$, and the four observations are each ranked 6.5. In a similar fashion the last two observations are each ranked 9.5. The rank correlation coefficient is affected by ties and Kendall (1948) gives corrections that can be applied. As the presence of tied ranks has been ignored in both table 1 and in the nomograph, it should be noted that the use of table 1 provides only an approximate test of significance and that values of r_s estimated from the nomograph are approximate.

The applicability of semiquantitative spectrographic analyses to rank correlation may be seen in table 2 which contains unpublished data for samples collected by A. F. Trites, Jr., U. S. Geological Survey. The data consist of chemical and semiquantitative analyses of copper, determined electrolytically, and manganese. The spectrographic data are reported in the three-bracket system.

Many geologic investigations require both chemical and semiquantitative spectrographic analyses for the same group of samples. If the investigator is interested in the degree of association between two elements which have been determined chemically, he can usually calculate the product moment

Table 2.--Ranks of chemical and spectrographic determinations of copper and manganese.1/

Copper					Manganese				
Electrolytic		Spectrographic			Chemical		Spectrographic		
Percent	Rank	Range in percent (3 bracket) 2/	Rank (3 bracket)	Rank (2 bracket)	Percent	Rank	Range in percent (3 bracket) 2/	Rank (3 bracket)	Rank (2 bracket)
0.03	3.5	.0X	3	3	0.014	2	$.0X^{-}$	3	6.5
0.028	2	.0X	3	3	0.029	5.5	.00X	1.5	1.5
0.24	12	.X	10.5	9.5	0.064	10	.0X	7	6.5
0.16	9.5	.X	10.5	9.5	0.021	3	.0X	7	6.5
1.81	16	X.	17	17	0.174	17	.X	17.5	15
0.15	8	.X	10.5	9.5	0.106	15	$.X^{-}$	14	15
0.74	14	$.X^{+}$	14.5	14.5	0.095	14	$.X^{-}$	14	15
0.10	6	$.X^{-}$	6.5	9.5	0.022	4	.0X	7	6.5
2.06	17	X.	17	17	0.041	7	.0X	7	6.5
1.0	15	$.X^{+}$	14.5	14.5	0.052	9	.0X	7	6.5
3.3	18	X.	17	17	0.206	18	.X	17.5	15
0.023	1	.0X	3	3	0.029	5.5	.0X	7	6.5
0.09	5	.0X	3	3	0.094	13	$.X^{-}$	14	15
0.03	3.5	.0X	3	3	0.007	1	.00X	1.5	1.5
0.22	11	.X	10.5	9.5	0.048	8	.0X	7	6.5
0.16	9.5	$.X^{-}$	6.5	9.5	0.077	11	$.0X^{+}$	11	11
0.41	13	.X	10.5	9.5	0.130	16	$.X^{-}$	14	15
0.14	7	.X	10.5	9.5	0.087	12	$.X^{-}$	14	15

1/ Unpublished data of A. F. Trites, Jr. Copper determinations by A. C. Horr and D. L. Skinner; manganese determinations by J. F. Wahlberg; and spectrographic determinations by R. C. Havens, all members of the U. S. Geological Survey.

2/ Ranges X^{-}, X and X^{+} for any power of ten correspond to intervals $1 - \sqrt[3]{10}$, $\sqrt[3]{10} - \sqrt[3]{10^2}$, and $\sqrt[3]{10^2} - \sqrt[3]{10^3}$ respectively.

correlation coefficient. If either or both the percentages of the pair of elements have been expressed as semiquantitative spectrographic data that are ranked data, the product moment coefficient cannot be calculated and the usual result is a rough visual comparison made between pairs. However, it is possible, where one or both sets of data are semiquantitative, to rank both sets of data and to estimate the degree of association by the rank correlation coefficient, r_s. An example of a calculation of the sum of squared rank differences, $\Sigma(R.D.)^2$, using the mixed type of data is shown in table 3.

As the method is applicable to continuous data (chemical analysis, for example) and to ranked (spectrographic) data, or to mixtures of both, it is of interest to calculate all possible correlation coefficients. Assuming that the basic assumptions associated with correlation have been validated the product moment correlation coefficient using the chemical data may be calculated. The four sets of data have been ranked as shown in table 2 and r calculated for the following ranked pairs, Cu_c - Mn_c, Cu_c - Mn_s, Cu_s - Mn_s, and Cu_s - Mn_c, where subscripts c and s refer to chemical and spectrographic data.

Under some assumptions the same data can be treated further. If in the original spectrographic data transformations of the type $0.0Z^{+} = 0.0X^{+}$ and $0.0Z = 0.0X^{-} + 0.0X$ are made, the three-bracket data may be changed to the two-bracket system of reporting and hence the original data used to illustrate ranks obtained from two-bracket data. In making use of this transformation the assumptions have been made that (1) there is equal accuracy of placing values in the correct brackets, (2) the lower bracket of the two-bracket system is identical to the lower two of the three-bracket system; and (3) the results of the transformation represent actual readings

Table 3.--Calculation of the sum of squared rank differences.

Copper		Manganese			
Percent	Rank	Range (3 bracket)	Rank	R.D.[1] (Cu-Mn)	$(R.D.)^2$
0.03	3.5	$.0X^{-}$	3	0.5	.25
0.028	2	.00X	1.5	0.5	.25
0.24	12	.0X	7	5	25.
0.16	9.5	.0X	7	2.5	6.25
1.81	16	.X	17.5	-1.5	2.25
0.15	8	$.X^{-}$	14	-6	36.
0.74	14	$.X^{-}$	14	0	0
0.10	6	.0X	7	-1	1
2.06	17	.0X	7	10	100
1.0	15	.0X	7	8	64
3.3	18	.X	17.5	0.5	.25
0.023	1	.0X	7	-6	36
0.09	5	$.X^{-}$	14	-9	81
0.03	3.5	.00X	1.5	2	4
0.22	11	.0X	7	4	16
0.16	9.5	$.0X^{+}$	11	-1.5	2.25
0.41	13	$.X^{-}$	14	-1	1
0.14	7	$.X^{-}$	14	-7	49

$\Sigma(R.D.)^2 = 424.50$

---

[1] Rank Differences

by a spectrographer. Having made the transformations, the spectrographic data for the two elements may be reranked as shown in table 2 and r_s calculated for them as well as for the pairs, Cu_c - Mn_s and Cu_s - Mn_c. The sums of squared rank differences and the rank correlation coefficients estimated from the nomograph are shown for each pair in table 4.

Table 4 shows that only one sum of squared rank differences (516) lies inside the region (515-1423) in which the correlation is not significant at the 0.05 level. However, this should not cause too much concern as the value is just inside the limit. One can suspect that the transformations made, plus the non-validated assumptions, have affected the rank correlation coefficients as may be noted by comparing the values for three-bracket results with those for two-bracket determinations.

Table 4.--$\Sigma(R.D.)^2$ and r_s.

Ranked pair	$\Sigma(R.D.)^2$	r_s
Cu_c - Mn_c	386.5	0.60
Cu_s - Mn_s (3 bracket)	392.5	0.59
Cu_c - Mn_s (3 bracket)	424.5	0.56
Cu_s - Mn_c (3 bracket)	391.0	0.60
Cu_s - Mn_s (2 bracket)	484.5	0.50
Cu_c - Mn_s (2 bracket)	516.0*	0.47
Cu_s - Mn_c (2 bracket)	405.0	0.58

Product moment correlation coefficient = 0.66

*Correlation is not significant.

Inspection of the coefficients in table 4 shows that all ranked pairs of types of data furnish estimates which agree fairly closely within themselves. With the exception of the two-bracket data obtained by transformation, these estimates are also in fairly good agreement with the product moment correlation coefficient. Hence, semiquantitative spectrographic data which are cruder measurements than chemical data can be used to obtain estimates of the degree of association by rank correlation.

Although the simplicity of calculation is destroyed where the adjustment for tied ranks is introduced (Kendall, 1948), the adjustment should be made where the data contain a large number of tied groups and where the significance of the association is of marginal value. The unadjusted and the adjusted coefficients are shown in table 5 (J. R. Rosenblatt, personal communication, 1956). Table 5 shows that the adjustment does not always have the effect of lowering the numerical value of r_s. For these data the adjustment has resulted in two coefficients being nonsignificant where only one of the unadjusted coefficients was not significant.

Table 5.--Rank correlation coefficients.

Ranked pair	r_s Unadjusted	r_s Adjusted
Cu_c - Mn_c	0.60	0.60
Cu_s - Mn_s (3 bracket)	0.59	0.56
Cu_c - Mn_s (3 bracket)	0.56	0.54
Cu_s - Mn_c (3 bracket)	0.60	0.58
Cu_s - Mn_s (2 bracket)	0.50	0.43
Cu_c - Mn_s (2 bracket)	0.47	0.43
Cu_s - Mn_c (2 bracket)	0.58	0.56

For rough estimates of the degree of association of pairs of elements in a group of samples, the Spearman rank correlation is attractive because of computational ease. Because of this simplicity its use is to be recommended with the caution that the user should refer to Kendall's adjustment where a large number of ties occur and where the significance of the sum of the squared rank differences or of the coefficient r_s tends to be marginal.

The author is indebted to Joan R. Rosenblatt of the Statistical Engineering Laboratory of the National Bureau of Standards for the adjusted rank correlation coefficients and for helpful discussion of the subject. This study is part of a program conducted by the U. S. Geological Survey on behalf of the Division of Raw Materials of the U. S. Atomic Energy Commission.

REFERENCES

Bennett, C. A., and Franklin, N. L., 1954, Statistical analysis in chemistry and the chemical industry, New York, John Wiley and Sons, 724 p.

Fisher, R. A., and Yates, Frank, 1948, Statistical tables for biological, agricultural and medical research, New York, Hafner Publishing Co., 112 p.

Goulden, C. H., 1952, Methods of statistical analysis, New York, John Wiley and Sons, 467 p.

Kendall, M. G., 1948, Rank correlation methods, London, Charles Griffin and Co., Ltd., 160 p.

Litchfield, J. T., Jr., and Wilcoxon, Frank, 1955, The rank correlation method: Anal. Chemistry, v. 27, p. 299-300.

Olds, E. G., 1938, Distribution of sums of squares of rank differences for small numbers of individuals: Annals Math. Statistics, v. 9, p. 133-148.

__________1949, The 5 percent significance levels for sums of squares of rank differences and a correction: Annals Math. Statistics, v. 20, p. 117-118.

Snedecor, G. W., 1946, Statistical methods applied to experiments in agriculture and biology, Ames, Iowa, Iowa State College Press, 485 p.

Waring, C. L., and Annell, C. S., 1953, Semiquantitative spectrographic method for the analysis of minerals, rocks, and ores: Anal. Chemistry, v. 25, p. 1174-1179.