Open Journal of Statistics, 2012, 2, 408-414
http://dx.doi.org/10.4236/ojs.2012.24049 Published Online October 2012 (http://www.SciRP.org/journal/ojs)

Regularities in Sequences of Observations

Mahkame Megan Khoshyaran
Economics Traffic Clinic (ETC), Paris, France
Email: megan.khoshyaran@wanadoo.fr

Received July 18, 2012; revised August 20, 2012; accepted September 2, 2012

ABSTRACT

The objective of this paper is to propose an adjustment to the three methods of calculating the probability that regularities in a sample data represent a systemic influence in the population data. The method proposed is called data profiling. It consists of calculating vertical and horizontal correlation coefficients in a sample data. The two correlation coefficients indicate the internal dynamic or inter dependency among observation points, and thus add new information. This information is incorporated in the already established methods and the consequence of this integration is that one can conclude with certainty that the probability calculated is indeed a valid indication of systemic influence in the population data.

Keywords: Systematic Influence; Theory of Large Samples; Analysis of the Variance Principle; Multiple Regression; Data Profiling; Vertical Correlation Coefficient; Horizontal Correlation Coefficient

1. Introduction

Suppose that in a sequence of observations one observes a striking regularity; for example suppose that the values arrange themselves in an increasing or decreasing order of magnitude, or a maximum or a minimum is indicated. Many questions arise. Is the observed regularity a general phenomenon, or is it true only of the sequence of the data set sampled. Is the observed regularity due to the particular sequence sampled or is it due to sampling from a random sequence. In other words, in recurrent sampling, is it reasonable to believe that approximately the same general results will occur. Is it the manner of sampling that creates artificial regularities. The occurrence of regularity in a data set that results from random sampling is highly improbable; thus regularity in a sample data is a justification for regarding regularity as a true representative of the population data. The assumption is that unless the probability of random occurrence is small, there is no objective proof that there exists an actual regularity in the population data.

To explore regularities in random sample data sets many researchers have made significant contributions, [1-6]. For example, assuming that the sequence of individual numerical values is available, they have applied various tests based on characteristics of a random sequence. For example, they concluded that the number of maxima in a sequence of unrelated numbers is one-third of the number of data points. The deviation of any sequence of data in any characteristic from what is assumed for a random sample of sequences implies that there is a systematic influence, the extent of which depends on the magnitude of deviation and the number of data points in a sample. In general, random sampling of data is not a sufficient criteria for proving systematic influence. It is shown that unless there are a large number of data points, the proof of the existence of systematic influence remains unresolved.

Up to now, the attempts to determine the probability of getting a short sequence of terms having a strict appearance of regularity have proven to be rather misleading. Given the uncertainties researchers have modified the analysis of regularities in random samples. In the new approach a sequence of averages of groups of individual observations is obtained in a systematic way. For example, random samples are drawn any number of times. The averages of each sequence of data sample are calculated. These averages form a composite sequence that can be used in testing the systematic influence in samples. The statistical significance of such a sequence of averages can be determined by comparing the variance of the individual observations in a random sample computed directly with that calculated from the variances of the averages, [7-14]. This analysis of variance principle can be applied to a general case where the values of the independent variable are related to each of a number of correlated independent variables. This is a problem of multiple curvilinear correlation, where a sequence of averages of the dependent variable is computed with respect to each independent variable and correlated to the constant values of the other independent variables, [15-17].

A method of testing the statistical significances of each sequence of averages as well as the composite significance of all of the sequences is derived.

Although the use of a sequence of averages is a logical approach this method is highly uncertain and in some cases inapplicable. Analysis of variance principle, and the multiple curvilinear correlation have a solid logic, they provide approximate indications of any systematic influence. The main shortcoming of these models is that they do not detect the source of variability in a data set. The focus of the three models is on the variability within and the correlation among averages in a sample. To address this shortcoming of the three approaches, a modification to these models is proposed. The modification consists of detecting the correlation among individual observations both within and across groups in a data sample, or in another word, data profiling. This aim is achieved by calculating vertical (ρ_v) and horizontal (ρ_h) correlation coefficients, and incorporating them in the calculations. The precise definitions of these variables are given and the manner in which they are integrated into the three models are demonstrated in the following sections.

2. The Approach Based on the Theory of Large Samples

Commonly, the values of sequences in a data sample are averages of measurements or numbers grouped in some systematic fashion. There are many readily available methods that calculate the probability of such systematic grouping of data. These methods are based on calculating the variability between the averages and within the groups. These methods are extended to special cases where the regularities of a sequence are periodic, [7-14]. The method based on the theory of large numbers developed by [9] consists of computing the standard deviation of the groups means multiplied by the square root of the number of observations $\left(\sqrt{n}\sigma_n\right)$, and the standard deviation of the entire series in a data sample (σ_o). Let m = number of columns or groups, n = number of entries per column, $\bar{y}_s = a$ group mean, and $\bar{y}$ = the grand mean, then the group standard deviation is given by the following:

$$\sqrt{n}\sigma_n = \sqrt{n}\left[\sqrt{\frac{\sum\left(\bar{y}_s-\bar{y}\right)^2}{m}}\right] \tag{1}$$

The sample standard deviation is calculated using the following equation:

$$\sigma_o = \sqrt{\frac{\sum\left(\bar{y}_s-\bar{y}\right)^2}{mn}} \tag{2}$$

It is assumed that the difference in value between (σ_o) and (σ_n) should be less than the sampling error. If this principle holds then, Cox employed the criterion of significance. If the error of standard deviation of (σ_o) is small, then the standard error of the ratio ($\frac{\sqrt{n}\sigma_n}{\sigma_o}$) should be proportional to (σ_n). Cox assumes that if there are systematic influences, then the expression $\left[\frac{\sqrt{n}\sigma_n}{\sigma_o}-1\right]$ should on the average equal zero given the theory of large samples, and the standard error $\left[\frac{\sqrt{n}\sigma_n}{\sigma_o}\sqrt{2m}\right]$.

Practice has shown that this method that is based on the theory of large samples is often inapplicable. One way to circumvent this problem is to introduce vertical and horizontal correlation coefficients. Correlation coefficients show the variations between observations, and across groups. A minor change of notation is introduced. The $(\bar{y}_s)$ that represented the group mean is modified to $\left(\bar{y}_j, j=1,\cdots,m\right)$ to reflect the mean by group of the observations or vertical means. Horizontal means $\left(\bar{y}_i, i=1,\cdots,n\right)$ reflect observation means. Each observation is represented by $\left(y_{ij}, i=1,\cdots,n; j=1,\cdots,m\right)$. The vertical (ρ_v) and horizontal (ρ_h) correlation coefficients are calculated using the following formulas:

$$\rho_v = \frac{\sum_{i=0}^{n}\left(\bar{y}_{i.}-\bar{y}_j\right)\left(\bar{y}_{(i+1)}-\bar{y}_j\right)}{\sqrt{\sum_{i=0}^{n}\left(\bar{y}_{i.}-\bar{y}_j\right)^2}} \text{ for } j=1,\cdots,m \tag{3}$$

and

$$\rho_h = \frac{\sum_{j=0}^{m}\left(\bar{y}_j-\bar{y}_i\right)\left(\bar{y}_{(j+1)}-\bar{y}_i\right)}{\sqrt{\sum_{j=0}^{m}\left(\bar{y}_{.j}-\bar{y}_i\right)^2}} \text{ for } i=1,\cdots,n \tag{4}$$

If the ratio $\left(\frac{\max\left(\rho_v\right)}{\max\left(\rho_h\right)}\right)$ is equal to one, then the indication is that each observation is related to the other, both within each column and across columns; in other words, there is evidence of systematic influence or systematic regularity. On the other hand, if the ratio $\left(\frac{\max\left(\rho_v\right)}{\max\left(\rho_h\right)}\right)$ is either less than one or greater than one, then the evidence points to the contrary, which translates into the lack of any systematic influence. Thus, in general if the ratio $\left(\frac{\sqrt{n}\sigma_n}{\sigma_o}\right)$ is equal to $\left(\frac{\max\left(\rho_v\right)}{\max\left(\rho_h\right)}\right)$, equal to one, then there is absolute certainty that the polation data exhibits systematic influence. The reverse case where the ratio $\left(\frac{\sqrt{n}\sigma_n}{\sigma_o}\right)$ is equal to $\left(\frac{\max(\rho_v)}{\max(\rho_h)}\right)$, is either less than or greater than one, then there is no systematic influence in the population data.

The approach based on the theory of large samples looks at the sample data from the macroscopic level, meaning sample averages and sample standard deviations. Data profiling explores the data set from the microscopic level, meaning the vertical and the horizontal correlation coefficients. Data profiling method adds new information which allows for an efficient and accurate detection of systemic influence. To state this formally, let $L^0 = (\Omega, A, P, R)$ be the space of almost surely random sets, where (Ω) is the set of all random sets, and (*A*) is a subset with σ-algebra. It can be stated that the sample data $\left(y_{ij}, i = 1,\cdots,n; j = 1,\cdots,m\right)$, and $\left(y_{ij} \in L^0\right)$ exhibits systemic influence if and only if the probability that

$$\left(\frac{\sqrt{n}\sigma_n}{\sigma_o} \xrightarrow[n\to\infty]{p=1} \left(\frac{\max(\rho_V)}{\max(\rho_h)}\right)\right);$$

$$\left(\frac{\sqrt{n}\sigma_n}{\sigma_o} \in L^0 \text{ and } \frac{\max(\rho_V)}{\max(\rho_h)} \in L^0\right)$$

exists and is equal to 1, or

$$\left(P\left(\left|\frac{\max(\rho_V)}{\max(\rho_h)} - \frac{\sqrt{n}\sigma_n}{\sigma_o}\right| \ge t\right) \to 0 \quad \forall t > 0\right),$$

where (*t*) is some constant. Data profiling assigns to the space (L^0) a metric (*d*) which is associated with the probability of convergence. Let

$$d\left(\frac{\sqrt{n}\sigma_n}{\sigma_o}, \frac{\max(\rho_V)}{\max(\rho_h)}\right) = E\left(\frac{\left|\frac{\max(\rho_V)}{\max(\rho_h)} - \frac{\sqrt{n}\sigma_n}{\sigma_o}\right|}{1 + \left|\frac{\max(\rho_V)}{\max(\rho_h)} - \frac{\sqrt{n}\sigma_n}{\sigma_o}\right|}\right);$$

then it is easy to notice that (*d*) represents a distance in $\left(L^0\right)$, and is invariant under any transformation (no matter which subset of random sets is used). If $\left(\frac{\sqrt{n}\sigma_n}{\sigma_o}\right)$ and $\left(\frac{\max(\rho_v)}{\max(\rho_h)}\right)$ are true representations of data at the two levels (macroscopic and microscopic respectively), then one would expect

$$d\left(\frac{\sqrt{n}\sigma_n}{\sigma_o}, \frac{\max(\rho_v)}{\max(\rho_h)}\right) \to 0 \text{ as } n \to \infty$$

if and only if

$$\left(P\left(\left|\frac{\max(\rho_v)}{\max(\rho_h)} - \frac{\sqrt{n}\sigma_n}{\sigma_o}\right| \ge t\right) \to 0 \quad \forall t > 0\right).$$

In fact the convergence of (*d*) to zero causes the convergence of the probability. This is due to the Bienaymé-Tchebychev or Markov inequality and the fact that as

$$(n \to \infty) t \mapsto \frac{t}{(1+t)}.$$

Inversely if

$$\left(P\left(\left|\frac{\max(\rho_v)}{\max(\rho_h)} - \frac{\sqrt{n}\sigma_n}{\sigma_o}\right| \ge t\right) \to 0\right)$$

holds then let for any (*ε* > 0)

$$\left(\frac{t}{(1+t)} \le \frac{\varepsilon}{2} \;;\; t > 0\right)$$

and

$$\left(P\left(\left|\frac{\max(\rho_v)}{\max(\rho_h)} - \frac{\sqrt{n}\sigma_n}{\sigma_o}\right| \ge t\right) \le \frac{\varepsilon}{2} \text{ for } n \ge n_0, n_0 = 1\right),$$

then for $n \ge n_0$

$$\begin{aligned} & d\left(\frac{\sqrt{n}\sigma_n}{\sigma_o}, \frac{\max(\rho_v)}{\max(\rho_h)}\right) \\ & = \int_{\left(\left|\frac{\max(\rho_v)}{\max(\rho_h)} - \frac{\sqrt{n}\sigma_n}{\sigma_o}\right| \ge t\right)} \frac{\left|\frac{\max(\rho_v)}{\max(\rho_h)} - \frac{\sqrt{n}\sigma_n}{\sigma_o}\right|}{1 + \left|\frac{\max(\rho_v)}{\max(\rho_h)} - \frac{\sqrt{n}\sigma_n}{\sigma_o}\right|} \mathrm{d}P \\ & + \int_{\left(\left|\frac{\max(\rho_v)}{\max(\rho_h)} - \frac{\sqrt{n}\sigma_n}{\sigma_o}\right| < t\right)} \frac{\left|\frac{\max(\rho_v)}{\max(\rho_h)} - \frac{\sqrt{n}\sigma_n}{\sigma_o}\right|}{1 + \left|\frac{\max(\rho_v)}{\max(\rho_h)} - \frac{\sqrt{n}\sigma_n}{\sigma_o}\right|} \mathrm{d}P \\ & \le P\left(\left|\frac{\max(\rho_v)}{\max(\rho_h)} - \frac{\sqrt{n}\sigma_n}{\sigma_o}\right| \ge t\right) + \frac{t}{(1+t)} \le \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon \end{aligned}$$

The conclusion is that $\left(\frac{\max(\rho_v)}{\max(\rho_h)}\right)$ assures almost surely the detection of systemic influence in a data set.

3. The Approach Based on the Method of Analysis of Variance

This method finds the probability that any variation in between averages is purely random, [18]. An outline of the procedure follows:

n_s = number of entries in column (s)

$N=\sum n_s$ is the total number of entries
a = a reasonable estimate of $\bar{y}$
$h=\left(\bar{y}-a\right)$
The mean variance between columns is calculated:

$$V_s=\frac{\sum_{j=0}^{m} n_s\left(\bar{y}_s-a\right)^2-Nh^2}{\left(m-1\right)=n_1} \tag{5}$$

The residual variance is calculated using the formula:

$$V_r=\frac{\sum\left(y-a\right)-\sum_{j=0}^{m} n_s\left(\bar{y}_s-a\right)^2}{\left(N-m\right)=n_2} \tag{6}$$

Let $Z=\log_e\left(\sqrt{\frac{V_s}{V_r}}\right)$, then the probability of no systematic influence is found from tables, [18-20] given (*Z*), and the degrees of freedom $\left(n_1\right)$, and $\left(n_2\right)$. The method of analysis of variance looks into the variability between column means and the variability of individual observations from the corresponding mean within each column. This method has a shortcoming in that it does not look at the corresponding correlations between individual observations in each column and across groups. Data profiling allows for a better analysis and detection of internal or systematic variability. To account for data profiling, the formula for (*Z*) should be modified in the following way:

$$Z=\log_e\left(\sqrt{\left(\frac{V_s}{V_r}\right)}\right)+\log_e\left(\sqrt{\left(\frac{\max\rho_v}{\max\rho_h}\right)}\right).$$

The addition of a log of the fraction of vertical and horizontal correlation coefficients has one major effect; it either augments the value of (*Z*), in which case lowers the probability of systematic influence or lowers the value of (*Z*), in which case raises the probability of systematic influence.

4. The Approach Based on Multiple Regression

Up to this point, we have been dealing with one independent variable only. [17] generalizes the method of analysis of variance to many independent variables which may be mutually correlated. In other words, the group averages are given as a multiple regression of (*K*) independent variables. He thus modifies the mean variance between columns $\left(V_s\right)$, and the residual variance $\left(V_r\right)$, using the multiple regression method. The outline of the procedure is as follows:

M = Total number of columns (groups) to be averaged with respect to all the independent variables

K = Number of independent variables

Ky' = Value of an observation corrected with respect to all except the *k*th independent variable

$\begin{pmatrix}\bar{y}_s^1\\ \cdot\\ \bar{y}_s^K\end{pmatrix}$ corrected group averages of the dependent variable w.r.t. to K independent variable

$\begin{pmatrix}\bar{y}_1\\ \cdot\\ \bar{y}_K\end{pmatrix}\begin{matrix}\text{weighted average of }\bar{y}_s^1\\ \cdot\\ \text{weighted average of }\bar{y}_s^K\end{matrix}$

The overall variance between columns is calculated:

$$V_s=\frac{\sum_{j=0}^{m} n_s\left(\bar{y}_s^1-\bar{y}_1\right)^2+\sum_{j=0}^{m} n_s\left(\bar{y}_s^2-\bar{y}_2\right)^2+\cdots+\sum_{j=0}^{m} n_s\left(\bar{y}_s^K-\bar{y}_K\right)^2}{\left(M-K\right)=n_1} \tag{7}$$

The residual variance is calculated using the formula:

$$V_r=\frac{\sum\sum\left(ky'-\bar{y}_s^k\right)}{N+\left(K-1\right)-M=n_2} \tag{8}$$

The probability that data being random is obtained as before from $Z=\log_e\left(\sqrt{\frac{V_s}{V_r}}\right)$, and the degrees of freedom $\left(n_1\right)$, and $\left(n_2\right)$; and the probability of systematic influence is thus $\left(1-Z=\log_e\left(\sqrt{\frac{V_s}{V_r}}\right)\right)$. The shortcoming of the generalized method of the analysis of variance is that although it tries to look more closely at individual data sets, it does not look at the strength of the relationship between each individual data points. Data profiling in this case allows for adjusting for this shortcoming. The vertical and horizontal correlation coefficients, $\left(\rho_v\right)$, $\left(\rho_h\right)$ are modified to adjust to the (*K*) independent variables. Let $\left(\rho_v^1,\cdots,\rho_v^K\right)$ be the vertical correlation coefficients calculated for the K independent variables, and $\left(\rho_h^1,\cdots,\rho_h^K\right)$ be the horizontal correlation coefficients calculated for the K independent variables. New correlation coefficients are introduced: $\left(\rho_v^1\left(r\right),\cdots,\rho_v^K\left(r\right)\right)$ represent residual vertical correlation coefficients of (*Ky'*) adjusted observations, and $\left(\rho_h^1\left(r\right),\cdots,\rho_h^K\left(r\right)\right)$ represent residual vertical correlation coefficients of (*Ky'*) adjusted observations. For each independent variable (*k*), the vertical and horizontal correlation coefficients, $\left(\rho_v\right)$, $\left(\rho_h\right)$ are calculated as before:

$$\rho_v^k=\frac{\sum_{i=0}^{n}\left(y_i^k-\bar{y}_j^k\right)\left(y_{(i+1)}^k-\bar{y}_j^k\right)}{\sqrt{\sum_{i=0}^{n}\left(y_i^k-\bar{y}_j^k\right)^2}}\text{for } j=1,\cdots,m; \tag{9}$$

and $k=1,\cdots,K$

and

$$\rho_h^k = \frac{\sum_{j=0}^{m}\left(y_j^k - \bar{y}_i^k\right)\left(y_{(j+1)}^k - \bar{y}_i^k\right)}{\sqrt{\sum_{j=0}^{m}\left(y_j^k - \bar{y}_i^k\right)^2}} \text{ for } i = 1,\cdots,n \text{ and } k = 1,\cdots,K \tag{10}$$

The overall variance between columns is then modified as follows:

$$V_s' = \frac{\sum_{j=0}^{m} n_s\left(\bar{y}_s^1 - \bar{y}_1\right)^2 \times \left[\frac{\max\left(\rho_v^1\right)}{\max\left(\rho_h^1\right)}\right] + \sum_{j=0}^{m} n_s\left(\bar{y}_s^2 - \bar{y}_2\right)^2 \times \left[\frac{\max\left(\rho_v^2\right)}{\max\left(\rho_h^2\right)}\right]}{\left(M-K\right) = n_1} \tag{11}$$
$$\frac{+\cdots+\sum_{j=0}^{m} n_s\left(\bar{y}_s^K - \bar{y}_K\right)^2 \times \left[\frac{\max\left(\rho_v^K\right)}{\max\left(\rho_h^K\right)}\right]}{\left(M-K\right) = n_1}$$

In order to modify the residual variance, residual vertical and horizontal correlation coefficients are calculated using (ky'), the value of an observation which is corrected to constant values of all the rest except the kth independent variable given in McEwen's generalized method of the analysis of variance, [17]. The residual vertical correlation coefficients are calculated:

$$\rho_v^k\left(r\right) = \frac{\sum_{i=0}^{n}\left(ky_i' - \bar{y}_j^k\right)\left(ky_{(i+1)}' - \bar{y}_j^k\right)}{\sqrt{\sum_{i=0}^{n}\left(ky_i' - \bar{y}_j^k\right)^2}} \text{for } j = 1,\cdots,m; \text{and } k = 1,\cdots,K \tag{12}$$

and the residual horizontal correlation coefficients are given by:

$$\rho_h^k\left(r\right) = \frac{\sum_{j=0}^{m}\left(ky_j' - \bar{y}_i^k\right)\left(ky_{(j+1)}' - \bar{y}_i^k\right)}{\sqrt{\sum_{j=0}^{m}\left(ky_j' - \bar{y}_i^k\right)^2}} \text{ for } i = 1,\cdots,n \text{ and } k = 1,\cdots,K \tag{13}$$

The residual variance is modified as:

$$V_r' = \frac{\sum\sum\left(ky' - \bar{y}_s^k\right)\times\left[\frac{\max\left(\rho_v^k\left(r\right)\right)}{\max\left(\rho_h^k\left(r\right)\right)}\right]}{N + \left(K-1\right) - M = n_2} \tag{14}$$

The probability that data exhibits systematic influence is obtained using $Z = \log_e\left(\sqrt{\frac{V_s'}{V_r'}}\right)$ and the degrees of freedom (n_1), and (n_2) as is already explained.

5. An Example: Sunspot Numbers

In this section the validity of the improvement in the form of data profiling is tested. For this purpose the data set used in [17] is revisited and the probability of the existence of systematic influences in the data is calculated once given the proposed analysis of variance method, which is already demonstrated in [17], and once with a modified version. Consider the data corresponding to sunspot numbers arranged with respect to a trial cycle of length 11 years, *i.e.* from 1749 to 1826. The sunspot numbers exceeding 99 are excluded. The data is shown in a matrix form as (**Table 1**):

The averages $\bar{y}_s = \bar{y}_j$ are given:

$$\bar{y}_s = (52.5,\ 43.2,\ 26.0,\ 21.5,\ 13.5,\ 6.5,\ 7.5,\ 12.7,\ 24.5,\ 30.5,\ 43.2)$$

The number of columns is (m = 11). The number of observations in each column is (n_s = 4). The number of observations of the dependent variable is (N = 44). The overall average is $\bar{y}$ = 25.59. The degrees of freedom

Table 1. Sunspot numbers arranged with respect to a trail cycle of 11 years, 1749-1826.

	1749-1759	1794-1804	1805-1815	1816-1826
1	81	41	42	46
2	83	21	28	41
3	48	16	10	30
4	48	6	8	24
5	31	4	3	16
6	12	7	0	7
7	10	15	1	4
8	10	34	5	2
9	32	45	12	9
10	48	43	14	17
11	54	48	35	36

$\left(n_1\right)$, and $\left(n_2\right)$ are respectively $\left(n_1 = 11-1=10\right)$, and $\left(n_2 = 44-11=33\right)$. The averages $\left(\bar{y}_s\right)$ decrease up to the 6th column, and then increase from then on. To calculate the probability that the sample data is indicative of the population data, and thus there are cyclic effects, the (Z) statistic is calculated. The statistic (Z) is calculated using the mean variance between the columns $\left(V_s\right)$, and the residual variance (V_r). $V_s = 956.32$, and $V_r = 265.0$.

The statistic $Z = \log_e\left(\sqrt{\frac{V'_s}{V'_r}}\right) = \sqrt{\frac{956.32}{265.0}} = 0.6408$.

The value of (Z) corresponding to the 20, 5, 1, and 0.1 percent points are respectively 0.19, 0.38, 0.54, and 0.71. Since (Z = 0.64) is greater than 0.54, then the probability of random effects is 0.01, which makes the probability of systematic influence to be 0.99. Though the results seem to point in favor of systematic influence or the existence of cycles, the evidence is not conclusive. To find out if the sample obtained implies cyclic appearance of sun spots, the data profiling method is tested. The vertical and horizontal correlation coefficients are calculated given Equations (3) and (4). The vertical averages $\left(\bar{y}_j, j = 1,2,3,4\right)$ are calculated as: ($\bar{y}_j = 41.5,\ 25.4,\ 14.3,\ 21.0$). The two statistics (ρ_v), and (ρ_h) are calculated.

$$\left(\rho_v = 13.82,\ 5.40,\ 4.06,\ 1.41,\ 1.12,\ 1.85,\ 1.13,\ -3.85,\ 6.00,\ 12.22,\ 15.99\right)$$

$$(\ \rho_h = -3280.54,\ 54.19,\ -1214.77,\ -1322.41)$$

The max of (ρ_v), and (ρ_h) are calculated as well.

$$\max\left(\rho_v\right) = 15.996053$$

$$\max\left(\rho_h\right) = 54.188689$$

The ratio $\left(\frac{\max\left(\rho_v\right)}{\max\left(\rho_h\right)}\right)$ is calculated as:

$\left(\frac{\max\left(\rho_v\right)}{\max\left(\rho_h\right)}\right) = 0.2951917$.

The value $\log_e\left(\sqrt{\left(\frac{\max\left(\rho_v\right)}{\max\left(\rho_h\right)}\right)}\right)$ is equal to 0.2586587.

The modified value of the statistic (Z) is then obtained by adding the two values of

$$Z = \log_e\left(\sqrt{\left(\frac{V_s}{V_r}\right)}\right) + \log_e\left(\sqrt{\left(\frac{\max\left(\rho_v\right)}{\max\left(\rho_h\right)}\right)}\right)$$

which then would give (0.6408 + 0.2587) = 0.8995. Since the value (0.8995) is higher than (0.71), it indicates that the probability that the population data is random is less than 0.001 which is less than 0.1 indicating with certainty that the number of sunspots is cyclic. The existence of systemic influence is indisputable. Applying the approach based on the method of large samples, the statistic $\frac{\sqrt{n}\sigma_n}{\sigma_o} = 8$ is obtained. There is a large discrepancy between this statistic and the adjustment proposed in Section 2, $\left(1-\left(\frac{\max\left(\rho_v\right)}{\max\left(\rho_h\right)}\right) = 0.71\right)$. The statistic $\frac{\sqrt{n}\sigma_n}{\sigma_o}$ is thus inapplicable. The statistic $Z = \log_e\left(\sqrt{\left(\frac{V_s}{V_r}\right)}\right) = 0.6669$ calculated using the approach based on multiple regression is a slight improvement over the statistic obtained using the method of analysis of variance (Z = 0.6408).

Using data profiling method, the statistic Z is corrected to (Z = 1.0). As in the case of the analysis of variance method, it can be stated with absolute certainty that there is indeed a systemic influence in the sample data.

6. Conclusion

The objective is to derive conclusions about the randomness of observations in a population given that the sample data set exhibits strict regularities. Three methods are analyzed and their shortcomings are indicated. An improvement to the three methods is suggested and formulated. The improvement comes in the form of data profiling which in essence is the integration of vertical and horizontal correlation coefficients in the equations. Through a simple example, it is shown that data profiling is indeed a compliment of the original formulation.

REFERENCES

[1] L. Besson, "On the Comparison of Methodological Data with Results of Chance," *Journal of Monthly Weather Review*, Vol. 48, 1920, pp. 89-94.

[2] H. W. Clough, "A Statistical Comparison of Meteorological Data with Data of Random Occurrence," *Journal of Monthly Weather Review*, Vol. 49, No. 3, 1921, pp. 124-132. doi:10.1175/1520-0493(1921)49<124:ASCOMD>2.0.CO;2

[3] W. L. Crum, "A Measure of Dispersion for Ordered Series," *Journal of American Statistical Association Quarterly Publication*, Vol. 17, 1921, pp. 969-975.

[4] E. W. Wooland, "On the Mean Variability in Random Series," *Journal of Monthly Weather Review*, Vol. 53, No. 3, 1925, pp. 107-111. doi:10.1175/1520-0493(1925)53<107:OTMVIR>2.0.CO;2

[5] H. Working, "A Random Difference Series for Use in the Analysis of Time Series," *Journal of American Statistical Association Quarterly Publication*, Vol. 24, 1934, pp. 11-24. doi:10.1080/01621459.1934.10502683

[6] W. O. Kermack and A. G. McKendrick, "A Measure of Dispersion for Ordered Series," *Journal of the Proceedings of the Royal Society Edinburgh*, Vol. 57, 1937, pp. 228-240.

[7] D. Alter, "A Group or Correlation Periodogram with Application to the Rainfall of the British Iles," *Journal of Monthly Weather Review*, Vol. 55, No. 210, 1927, pp. 263-266. doi:10.1175/1520-0493(1927)55<263:AGOCPW>2.0.CO;2

[8] C. Chree, "Periodicities Solar and Meteorological," *Journal of the Royal Meteorological Society*, Vol. 85, 1924, pp. 87-97.

[9] J. B. Cox, "Periodic Fluctuations of Rainfall in Hawaii" *Proceedings of the American Society of Civil Engineers*, Vol. 87, 1924, pp. 461-491.

[10] E. L. Dodd, "The Probability Law for the Intensity of a Trail Period with Data Subject to the Gaussian Law," *Bulletin of the American Mathematical Association Society*, Vol. 33, 1927, pp. 681-684. doi:10.1090/S0002-9904-1927-04451-2

[11] S. Kuznets, "Random Events and Cyclical Oscillations," *Journal of the American Statistical Association*, Vol. 24, 1929, pp. 258-275. doi:10.1080/01621459.1929.10503048

[12] R. W. Powell, "Successive Integration as a Method of Finding Long Period Cycles," *Annals of the Mathematical Statistics*, Vol. 1, No. 2, 1930, pp. 123-136. doi:10.1214/aoms/1177733127

[13] K. Stumpff, "Grunlagen und Methoden der Periodenforschung," Springer, Berlin, 1925.

[14] G. T. Walker, "On Periodicity—Criteria for Reality," *Memorandum of the Royal meteorological Society*, Vol. 3, No. 25, 1930, pp. 97-101.

[15] C. F. McEwen and E. L. Michel, "The Functional Relation of One Variable to Each of a Number of Correlated Variables Determined by a Method of Successive Approximations to Group Averages," *Proceedings of the American Academy of Arts and Sciences*, Vol. 55, No. 8, 1919, pp. 89-133.

[16] C. F. McEwen, "The Minimum Temperature, a Function of the Dew Point and Humidity, at 5 p.m. of the Preceding Day; Method of Determining This Function by Successive Approximations to Group Averages," *Monthly Weather Review Supplement*, No. 16, 1920, pp. 64-69.

[17] C. F. McEwen, "The Reality of Regularities Indicated in Sequences of Observations," *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, San Francisco, 13-18 August 1945, pp. 229-238.

[18] R. A. Fisher, "Statistical Methods for Research Workers," 4th Edition, Biological Monographs and Manuals, London, 1932.

[19] G. U. Yule and M. G. Kendall, "An Introduction to the Theory of Statistics" 11th Edition, Charles Griffin and Company Ltd., London, 1937.

[20] R. A. Fisher and F. Yates, "Statistical Tables for Biological, Agricultural, and Medical Research," Oliver and Boyd, London, 1938.