CONCURRENCY AND COMPUTATION: PRACTICE AND EXPERIENCE
Concurrency Computat.: Pract. Exper. 2015
Published online in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/cpe.3745

SPECIAL ISSUE PAPER

Using Spearman's correlation coefficients for exploratory data analysis on big dataset

Chengwei Xiao[1,*,†], Jiaqi Ye[1], Rui Máximo Esteves[2] and Chunming Rong[1]

[1]*Department of Electrical Engineering and Computer Science, University of Stavanger, Stavanger, Norway*
[2]*Department of Condition Monitoring, National Oilwell Varco, Stavanger, Norway*

SUMMARY

Correlation analysis is both popular and useful in a number of social networking research, particularly in the exploratory data analysis. In this paper, three well-known and often-used correlation coefficients, Pearson product–moment correlation coefficient, Spearman, and Kendall rank correlation coefficients, are compared from definition to application domain. Based on the characteristics of the pump's vibration dataset, the nonparametric and distribution-free Spearman rank correlation coefficient is introduced to analyze the relationship between the pump's working state and each of the 207′880 variables. The percentage of variables and exact variables' tables with high Spearman's correlation coefficients for states I and II, states I and III, states II and III, and three states in different files are obtained respectively, which has important valuation for the future research of the unsupervised machine learning system. Copyright © 2015 John Wiley & Sons, Ltd.

Received 3 April 2015; Revised 21 November 2015; Accepted 24 November 2015

KEY WORDS: exploratory data analysis; correlation analysis; Spearman correlation coefficient; *p*-value; vibration analysis; pump state

1. INTRODUCTION

Correlation analysis is one of the most common and most useful topics that has seen many interest in the social networking research, particularly in the domain of exploratory data analysis on a real-world big dataset. It is a data exploration technique for identifying and revealing the degree of association between one dependent variable and another one or more variables in a big or high-dimensional dataset. This information can be used for exploring and simplifying complex multivariate dataset and indicate possible factors that confound a relationship of interest.

The concept of correlation was first developed in 1846 by the French physicist Auguste Bravais; then, it was extended and proposed by Francis Galton on the book Natural Inheritance in 1889 [1]; after that, the famous mathematician and statistician Karl Pearson developed the theory of correlation by his 'product–moment' method and the relationship with the linear regression [2]. In 1904, Charles Spearman proposed a nonparametric rank correlation coefficient, which is a distribution-free version of the classical Pearson's product–moment correlation coefficient [2, 3]. An alternative to Spearman's rank correlation coefficient, which is called Kendall's tau rank correlation, was also defined by Maurice Kendall in 1948 [4]. Apart from the aforementioned three main types of correlation coefficients, there are also several other measures of dependence among random

*Correspondence to: Chengwei Xiao, Department of Electrical Engineering and Computer Science, University of Stavanger, Stavanger, Norway.
†E-mail: c.xiao@stud.uis.no

Copyright © 2015 John Wiley & Sons, Ltd.

variables, including distance correlation [5], Brownian correlation [6], polychoric correlation [7], randomized dependence coefficient [8], and so on.

In 1993, the Pearson product–moment correlation coefficient was used to perform the spatial data analysis before building regression models for the referenced data in the social and environmental sciences by R. Haining [9]. Jan Hauke and Tomas Kossowski compared the values and significance of the Pearson and Spearman correlation coefficients for the same dataset, which represent regional indices of socio-economic development in 2011 [2]. In order to analyze data, the statistical methods of Kolmogorov–Smirnov and Spearman rank correlation coefficient were used by S. H. Javaran, S. N. S. Sajadi, and M. Karamoozain in 2014 and got a conclusion that there was a significant and positive relationship between social responsibility of club with reputation and fans' dependency on team in the football premier league [10]. It is also being used to measure the strength of influence of the independent factors on project sustainability of non-government organizations-funded community projects in Kenya: a case of action aid-funded project in Makima location, Embu county [11]. All the aforementioned research shows the availability and reliability of Spearman rank correlation coefficient for the nonparametric or distribution-free cases.

The remainder of this paper is organized as follows. We first present in Section 2 the standard definition of the correlation coefficient and compare the Pearson, Spearman, and Kendall correlation coefficients from definition to application area. In Section 3, we give a detailed description for the vibration dataset from the pump and explain the importance of the vibrating analysis to detect the mechanical condition of pumps during the maintenance work for the purpose of preventing worse and more expensive failure. The data import and calculations of Spearman correlation coefficient between the pump's state and each of the 207′880 variables in the pumps' vibration dataset are processed by R programming language in Section 4. The calculation results for states I and II, states I and III, states II and III, and three states are discussed and analyzed in Section 5. Lastly, we provide a brief summary with conclusions.

2. COMPARISON OF THREE DIFFERENT CORRELATION COEFFICIENTS

2.1. *Standard definition of the correlation coefficient*

Suppose a consideration of two sets of N samples, X and Y, the basic correlation analysis is to find out whether the set of samples X is likely to predict another set of samples Y. And given that a predicted set $\hat{Y}$ is obtained by the use of the mathematical model $f(X)$, we define that error e is the difference between the actual value Y_i and the predicted value $\hat{Y}_i$, that is,

$$e_i = Y_i - \hat{Y}_i. \tag{1}$$

So that, the average square error of the prediction can be expressed by the following equation:

$$s_{yx}^2 = \frac{\sum e_i^2}{N} = \frac{\sum \left(Y_i - \hat{Y}_i\right)^2}{N}. \tag{2}$$

What we are trying to build is the mathematical relationship function or equation that is able to minimize the standard error s_{yx}. Given that $\overline{X}$, $\overline{Y}$ are the mean value of X_i and Y_i, respectively. And the standard deviation terms are s_x^2 and s_y^2, therefore, the standard definition of the correlation coefficient r can be expressed as follows:

$$r = \frac{\sum \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{N s_x s_y}. \tag{3}$$

As we can see from the aforementioned equation, the correlation coefficient r must be in the range from -1 to 1, which is able to provide help for determining the magnitude and direction of the

Copyright © 2015 John Wiley & Sons, Ltd.

DOI: 10.1002/cpe

relationship between two variables in pair. The sign of the correlation coefficient r shows that the direction is positive or negative; however, the numerical value of the correlation coefficient r represents the magnitude of the correlation coefficient. The bigger the absolute value of r is, the greater the relationship of the two variables is. The square value r^2, called coefficient of determination, is the measure of the proportion of the explained variance between 0 to 1, and $1-r^2$ is the proportion of variance unexplained. Another interesting property is that the correlation coefficient r will not change even if the set X and/or Y is shifted or rescaled. As a rule of thumb, the strength of the relationship between the variables can be categorized into four types, which can be shown in Table I.

Tests of significance tell analysts to reject or accept the low probability p-value to get at least as extreme results given the assumption that the null hypothesis H_0 is true [12–14]. p-value is always coupled to a significance level α for a given hypothesis test; typical values for α are 0.1, 0.05, and 0.01, which is always set ahead of time. Therefore, for example, given that a significance test under the significance level $\alpha=0.05$, if the p-value is to be 0.004, less than 0.05, it means that it is highly impossible (probability less than 0.05) to observe the outcome under the null hypothesis H_0, and we can get the conclusion that this outcome is significant at the 0.05 level.

As we all know, correlation coefficient is a simple statistical measure of the strength and direction between one dependent variable and one or more independent variables, and there are a number of different types of correlation coefficients under the different statistical hypothesis in certain application domain [15]. In the following subsections, we will make a comparison of three well-known and frequently used correlation coefficients: Pearson product–moment correlation coefficient r_p, Spearman rank correlation coefficient r_s, and Kendall rank correlation coefficient τ.

2.2. *Pearson product–moment correlation coefficient*

Pearson product–moment correlation coefficient r_p is a measure of strength and direction of only the linear relationship between two variables under the assumption that both the quantitative variables are normally distributed and measured on interval scales. It is defined as the sample covariance of the two variables divided by their sample standard deviations, which we can obtain based on the aforementioned standard correlation coefficient definition:

$$r_p = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\left(\sum (X_i - \overline{X})^2\right)\left(\sum (Y_i - \overline{Y})^2\right)}}. \tag{4}$$

As the definition involves the first moment of the mean-adjusted variables, thus, it is called product–moment. While the Pearson correlation coefficient r_p is a powerful tool, people often forget that r_p is only a measure of linear relationship. So that, if r_p is with a small value, it only represents that there is no linear relationship or a very weak linear relationship between the corresponding variables; it is possible that there are some other nonlinear relationships existing. Consequently, it is advisable to plot the scatter diagram before making a conclusion that no existence of a relationship. Under such circumstances, some reasonable transformation can be made to find out the linearity, for example, rank the variables.

Table I. Strength of relationship.

Value of r	Strength of relationship
−1.0 to −0.5 or 1.0 to 0.5	Strong
−0.5 to −0.3 or 0.3 to 0.5	Moderate
−0.3 to −0.1 or 0.1 to 0.3	Weak
−0.1 to 0.1	None or very weak

Copyright © 2015 John Wiley & Sons, Ltd.

DOI: 10.1002/cpe

2.3. Spearman rank correlation coefficient

Spearman rank correlation coefficient r_s is a nonparametric or distribution-free rank statistical measure of the strength and the direction of the arbitrary monotonic association between two ranked variables or one ranked variable and one measurement variable. In principle, Spearman's correlation coefficient is simply a special case of Pearson's coefficient under the situation that the samples are converted into ranks before doing the correlation coefficient calculations [2]. But it does not require to make any assumptions about the frequency distribution and the linear relationship between the two variables, nor measured on interval scale. The simple expression for r_s based on the difference between the two ranked variables is as follows:

$$r_s = 1 - \frac{6\sum d_i^2}{N\left(N^2 - 1\right)}, \tag{5}$$

where $d_i = X_i^{'} - Y_i^{'}$ is the difference between each pair of the ranked variables and N is the total number of the samples. It is a measure of a monotonic relationship that can be used when the characteristics of the in-pair variables (such as frequency distribution and/or linear distributed) make Pearson's r_s misleading or undesirable. Apart from the nonparametric privilege, the main advantage of this measure is that it is much more convenient to use because it does not require how the data rank, in ascending order or in descending order. As we can see from Figure 1, the two variables are monotonically related with Spearman correlation coefficient of 1, but Pearson correlation coefficient of only 0.88.

2.4. Kendall rank correlation coefficient

Kendall rank correlation coefficient τ is also a nonparametric measure for statistical dependence between two ordinal (or rank-transformed) quantities, which can be seen as an alternative to Spearman correlation coefficient. But it is more difficult to compute, because their underlying logic and computational formulae are quite different [2]. In most cases, Kendall's and Spearman's values are very close and would give rise to the same inferences, but it is relatively safer to adopt the lower value when discrepancies happen. Because Spearman sums the squared errors, it means it is more sensitive to error, whereas Kendall sums the absolute discrepancies, thus, it is insensitive to error. The Kendall correlation coefficient is defined as follows:

$$\tau = \frac{n_c - n_d}{\frac{1}{2}N(N-1)}, \tag{6}$$

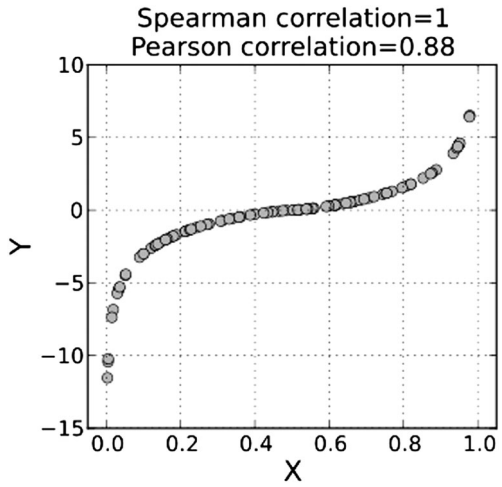

Figure 1. Example of Spearman correlation coefficient.

Copyright © 2015 John Wiley & Sons, Ltd.

DOI: 10.1002/cpe

where n_c is the number of concordant pairs, that is, both $x_i > x_j$ and $y_i > y_j$ or both $x_i < x_j$ and $y_i < y_j$. n_d is the number of discordant pairs, that is, both $x_i > x_j$ and $y_i < y_j$ or both $x_i < x_j$ and $y_i > y_j$. If $x_i = x_j$ or $y_i = y_j$, the pair is neither the concordant pair nor the discordant pair.

3. OVERVIEW OF THE DATASET

3.1. Kinematics and measurement technology

Vibrating analysis is a valuable tool to detect the mechanical condition of pumps in a predictive or preventive maintenance program in order to repair or replace the machinery before worse and more expensive failure happens. In general, it always involves three main components: amplitude, frequency, and direction to monitor the machine's condition. The comparison of the recorded vibration behavior with typical drive frequencies and that of potential damages is able to provide the information. The diagnostic information is often decided by the severity, which shows the bad degree of the problem, and the frequency represents the cause of the problem.

The measurement of the pump's vibration can be taken on every bearing location in three different planes (vertical, horizontal, and axial) and in three different amplitudes as follows [16]:

1. Displacement – Mils – peak–peak
   - determine the machine's movement,
   - turning speed vibration levels,
   - measure large sleeve-bearing machines normally, and
   - severity requires to know the frequency.

2 Velocity – in/s peak
   - broad range from 100–120′000.00 Cpm (cycles per minute),
   - commonly used for machine vibration analysis, and
   - velocity severity is independent of frequency.

3 Acceleration – Gs root mean square
   - determine high-frequency vibration problems because of damaged rolling element bearings or gears and
   - severity also requires to know the frequency.

When the vibration data are analyzed, the fast Fourier transform vibration spectrum is likely to be divided into a number of frequency ranges. There is always narrow band selective alarming in most commercial vibration analysis software, which can be used to determine the machine's condition.

In this study, the axial piston pump with nine pistons, which are arranged parallel to each other, is driven by a 380-KW motor with fundamental speed of 1492 rpm. The frequency of all the nine pistons and the high frequency of the swash plate are calculated based on the rotation speed. Each exciting frequency may indicate the failure. And one of the major exciting frequencies in this pump is the piston frequency, for example,

$$Fundamental\ \ frequency = 1492/60 = 24.86\text{Hz}$$
$$Piston\ \ frequency = 24.86{*}9 = 223.8\text{Hz}.$$

3.2. Resume of the tests

As we can see from Table II, all the dataset is from three pumps in three different states of lifetime, which can be defined as states I, II, and III, respectively. Each test data from any pump consists of four components: axial data, horizontal data, vertical data, and one more data from one motor sensor

Copyright © 2015 John Wiley & Sons, Ltd.

DOI: 10.1002/cpe

Table II. Resume of the tests.

Date	Pump no.	Test no.	Start time	End time	Flow (l/m)	Pressure (Bar)
May 13, 2014	1	1	17:50 h	18:38 h	220	200
May 14, 2014		2	18:46 h	20:20 h	200	273
		3	20:30 h	21:10 h	200	115
		4	08:39 h	11:19 h	100–180	202
May 14, 2014	2	1	12:48 h	13:35 h	190	190
		2	13:52 h	14:36 h	190	285
		3	14:38 h	15:23 h	120	200
		4	15:25 h	15:33 h	Vary	200
		5	15:33 h	16:42 h	220	200
		6	16:42 h	16:52 h	Vary	200
May 14, 2014	3	1	19:25 h	20:12 h	200	202
		2	20:18 h	21:04 h	235	292

Motor_DE_Magnet. In each component, there are six elements that can be further seen as six different categories of variables: overall_acc_2Hz–10kHz (two variables), over_vel_2Hz–1kHz (two variables), spec_acc_2Hz–50kHz (12′798 variables), spec_vel_2Hz–1kHz (6′398 variables), Surface Permanent Magnet (SPM) (two variables), and ts_vel_1kHz_10s (32′768 variables). In total, there are $4*(2+2+12{,}798+6398+2+32{,}768)=207'880$ independent variables and one dependent variable, which indicates the state of the pump.

The objective of this study is to find out a small number of variables, which has greater association with the state of the pump, from the huge number of all the 207′880 variables. If we would like to build an unsupervised learning system that aims to identify in which state the pump is in the future, this result is able to shorten the research range effectively and give vital reference to the analyst.

As the state of the pump is a ranked variable, and the distribution of each of the 207′880 variables is not known, thus, the nonparametric and distribution-free statistical measure, Spearman rank correlation coefficient, is used to conduct the exploratory data analysis.

4. DATA PROCESSING

Besides the huge number of variables, the number of the observations/samples is also great, for example, there are 26′491 observations in the file 'overall_acc_2Hz–10kHz' of the axial plane for pump 1, and there are 4′292′608 elements in the files 'ts_vel_1kHz_10s', although only 131 observations, but 32′768 variables. Therefore, the free software environment for statistical computing and graphics, R programming language, is introduced into the whole data processing of this project. The whole data processing involves two main steps: data visualization and Spearman correlation coefficient calculations.

4.1. Data visualization

Before calculating the Spearman's correlation coefficient and *p*-value between state of the pump and each of 207′880 variables, the ordinary scatter plot is used to describe the relationship between two variables. However, for the variables with even about 51′455 observations, there is a significant overlap among data points, which makes it difficult to discern where the concentration of data is greatest. Therefore, the *smoothScatter*( ) function in R is also applied to produce smoothed color density representations of the scatter plot, which is easy to read the concentration information of data.

At the left side of Figure 2 is scatter plot and scatter plot colored by smoothed densities about the pump's type and the variable '0-P' in the file 'axial_overll_acc_2Hz–10kHz' with 51′014 observations; however, the scatter plot and smoothed-colored scatter plot about ranks of pump's type and ranks of the variable '0-P' are on the right side. It looks like that there is no any special concentration of the observations, and later, we will know that the calculation results of the Spearman's correlation coefficient for three states are 0.4811, the one for states I and II is only 0.0860, the one for states I and III is higher with 0.6303, and the highest one is 0.7924 between states II and III.

Copyright © 2015 John Wiley & Sons, Ltd.

DOI: 10.1002/cpe

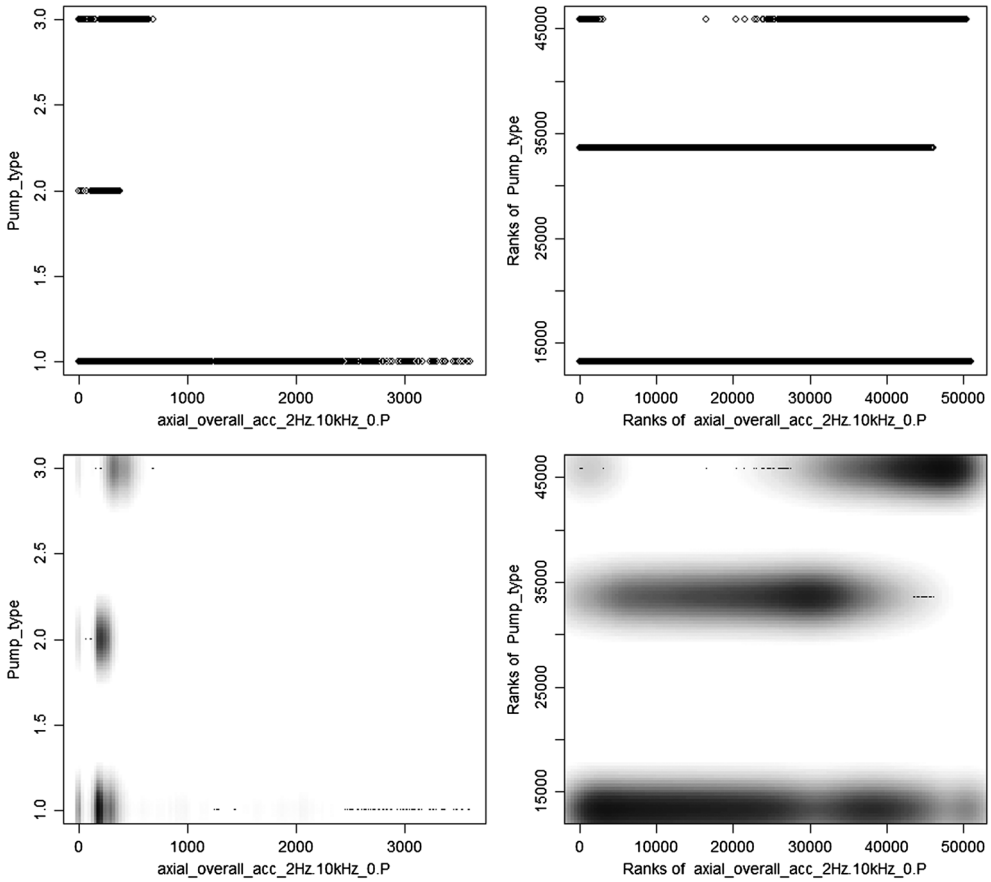

Figure 2. Scatter plot and scatter plot colored by smoothed densities with 51′014 observations.

At the left side of Figure 3, it is scatter plot colored by smoothed densities about the pump's type and the variable '25291.9375' in the file 'axial_spec_acc' with 224 observations, while the smoothed-colored scatter plot about ranks of pump's type and ranks of the variable '25291.9375' is on the right side. We can see that most of the observation mainly concentrates on the three points on the figure, which represents the three different states of the pump. And later, we will also know that the calculation result of the Spearman's correlation coefficient for three states is 0.8225 (p-value: 2.56E-56) and the ones for states I and II, for states I and III, and for states II and III are 0.7207 (p-value: 4.74E-32), 0.6689 (p-value: 4.64E-24), and 0.6286 (p-value: 0.6286E-10), respectively.

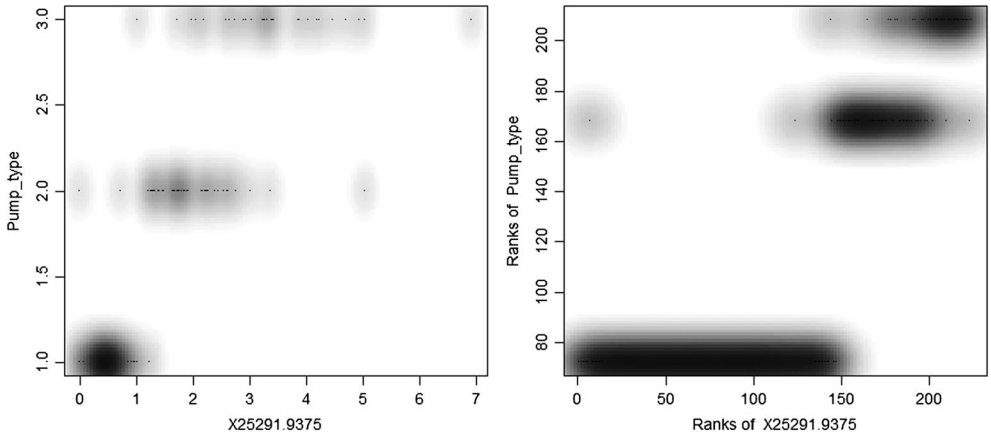

Figure 3. Scatter plot colored by smoothed densities with 224 observations.

Copyright © 2015 John Wiley & Sons, Ltd.

DOI: 10.1002/cpe

4.2. Spearman correlation coefficient calculations

After importing the whole dataset into R, the function *cor*( ) is used to calculate the Spearman correlation coefficient between the state of the pump (states I and II, states I and III, states II and III, and three states) and each of the variables, which can be seen in Figure 4. And the significance test for correlation between paired samples is also calculated at the same time. It is worth mentioning that as there are only three states of the pump in the last result, it means that there are identical values in the dataset, which are called ties. And the ranks are not unique anymore, hence, exact *p*-value cannot be calculated. In order to avoid the warning message, the logical parameter exact is set to false; the test of the significance adopts an Edgeworth series approximation method for large sample sizes.

5. RESULT ANALYSIS

The Spearman's correlation coefficient and *p*-value between state of the pump and each of 207′880 variables in three states, states I and II, states I and III, and states II and III have been calculated, respectively. And analysis of the correlation coefficients is as follows:

5.1. Spearman correlation coefficient for states I and II

As we can see from the histogram of correlation coefficient by Spearman for states I and III in Figure 5, the number of variables with high correlation coefficient, that is, whose absolute value is greater than 0.7 and *p*-value less than 0.05, is 504, while the number of variable with low correlation coefficient, whose absolute value is less than 0.3, is 178′569, accounting about 85.90% of the total variables. The histogram of high (abs >0.7 and *p*-value <0.05) and low (abs <0.3) correlation coefficient for states I and II can be seen in Figure 6; the five highest Spearman correlation coefficients are 0.7243, 0.7238, 0.7237, 0.7235, and 0.7233, while the five lowest ones are -0.6724, -0.6722, -0.6719, -0.6717, and -0.6713. All the variables come from the 'axial_spec_acc' file (28.97%), 'horizontal_spec_acc' file (68.06%), 'vertical_spec_acc' file (2.78%), and 'vertical_spec_vel' file (0.2%, only one variable), which can be seen in Figure 7.

5.2. Spearman correlation coefficient for states I and III

Histogram of Spearman correlation coefficient for states I and III in Figure 5 shows the distribution of the correlation coefficient by Spearman, in which the number of variables with high correlation coefficients, that is, whose absolute value is greater than 0.65 and *p*-value less than 0.05, is 1752, while the number of variable with low correlation coefficient, whose absolute value is less than 0.3, is 182′112, accounting about 87.60% of the total variables. The histogram of high (abs >0.65 and *p*-value <0.05) and low (abs <0.3) correlation coefficient for states I and III can be seen in Figure 8; the five highest Spearman correlation coefficients are 0.6679, 0.6685, 0.6722, 0.6735, and 0.6738, while the five lowest are -0.6396, -0.6392, -0.6373, -0.6369, and -0.6357. All the variables come from the 'axial_spec_acc' file (18.61%), 'horizontal_spec_acc' file (31.22%),

```
t <- data.frame(colnames(df_merge)[1],
                cor(df_merge_12[,1],df_merge_12[,2],method = "spearman"),
                cor.test(df_merge_12[,1],df_merge_12[,2],method = "spearman",exact=FALSE)$p.value,
                nrow(df_merge_12),
                cor(df_merge_13[,1],df_merge_13[,2],method = "spearman"),
                cor.test(df_merge_13[,1],df_merge_13[,2],method = "spearman",exact=FALSE)$p.value,
                nrow(df_merge_13),
                cor(df_merge_23[,1],df_merge_23[,2],method = "spearman"),
                cor.test(df_merge_23[,1],df_merge_23[,2],method = "spearman",exact=FALSE)$p.value,
                nrow(df_merge_23),
                cor(df_merge[,1],df_merge[,2],method = "spearman"),
                cor.test(df_merge[,1],df_merge[,2],method = "spearman",exact=FALSE)$p.value,
                nrow(df_merge),
                'axial_overall_acc')
```

Figure 4. Source code of the Spearman correlation coefficient's calculation.

Copyright © 2015 John Wiley & Sons, Ltd.

DOI: 10.1002/cpe

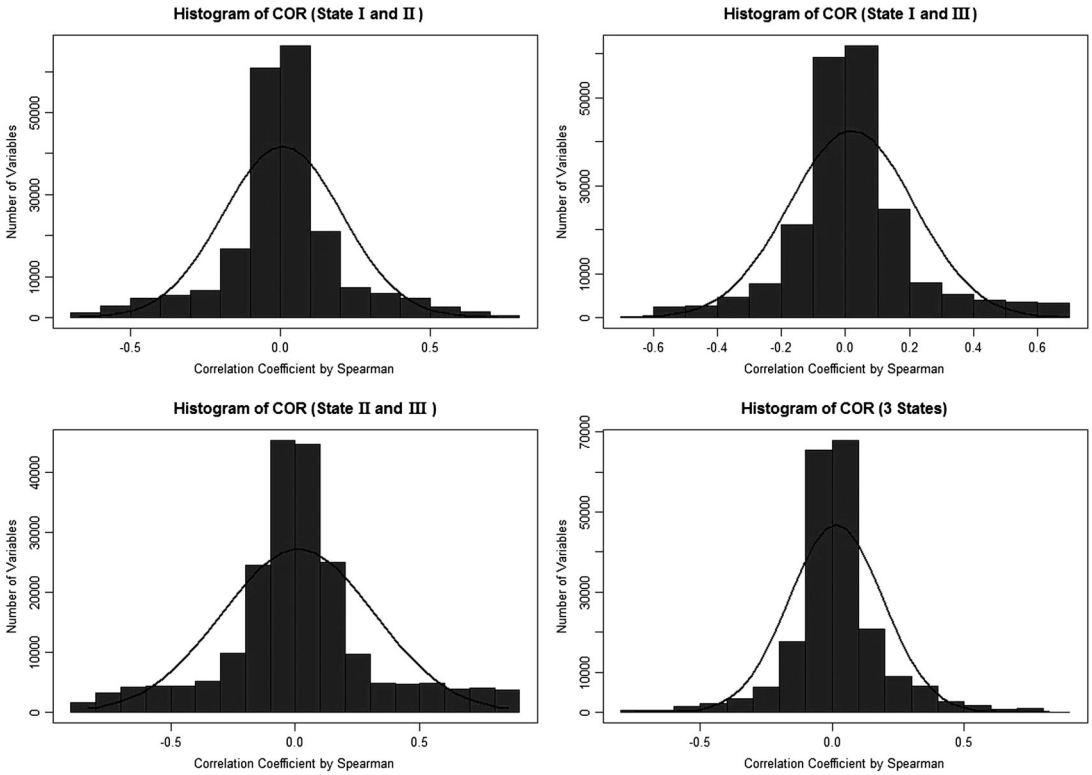

Figure 5. Histogram of Spearman correlation coefficients (207′880 variables).

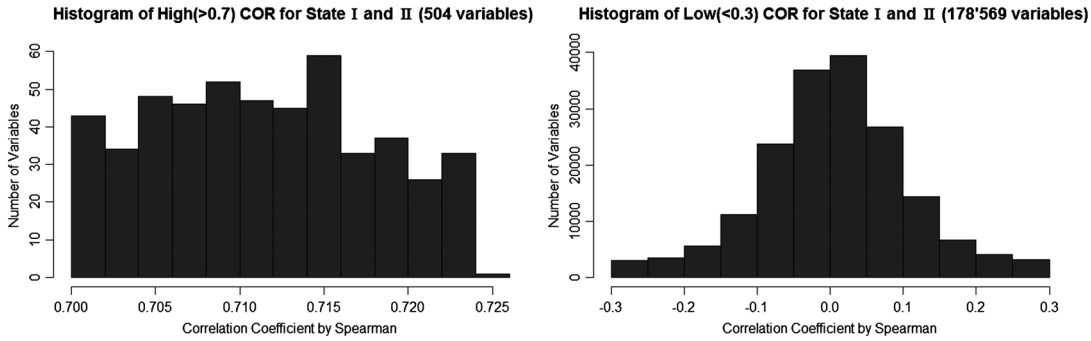

Figure 6. Histogram of high (abs 0.7 and *p*-value 0.05) and low (abs 0.3) Spearman correlation coefficients for states I and II.

'horizontal_spec_vel' file (46.92%), 'vertical_spec_acc' file (3.20%), and 'vertical_spec_vel' file (0.06%, only one variable), which can be seen in Figure 9.

5.3. Spearman correlation coefficient for states II and III

Histogram of Spearman correlation coefficients for states II and III in Figure 5 shows the distribution of the correlation coefficient by Spearman, in which the number of variables with high correlation coefficient, that is, whose absolute value is greater than 0.80 and *p*-value less than 0.05, is 5317, while the number of variable with low correlation coefficient, whose absolute value is less than 0.3, is 159′005, accounting about 76.49% of the total variables. The histogram of high (abs >0.8 and *p*-value <0.05) and low (abs <0.3) correlation coefficient for states II and III can be seen in Figure 10; the five highest Spearman correlation coefficients are 0.8559, 0.8559, 0.8549, 0.8540, and 0.8530, while the five lowest are -0.8326, -0.8316, -0.8307, -0.8278, and -0.8264. The

Copyright © 2015 John Wiley & Sons, Ltd.

DOI: 10.1002/cpe

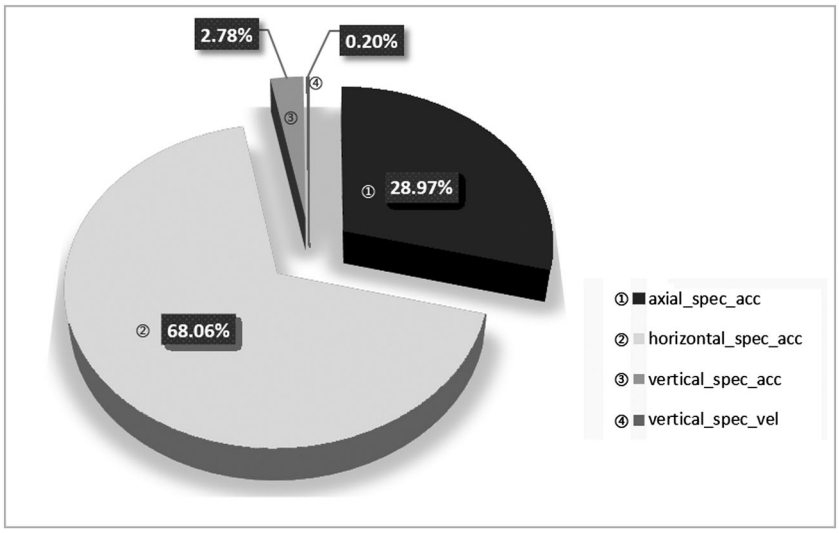

Figure 7. Percentage of variables with high Spearman correlation coefficients for states I and II in different files.

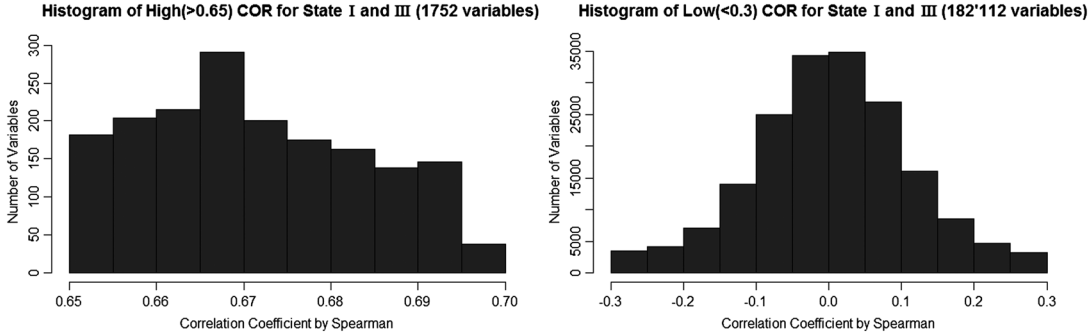

Figure 8. Histogram of high (abs 0.65 and *p*-value 0.05) and low (abs 0.3) Spearman correlation coefficients for states I and III.

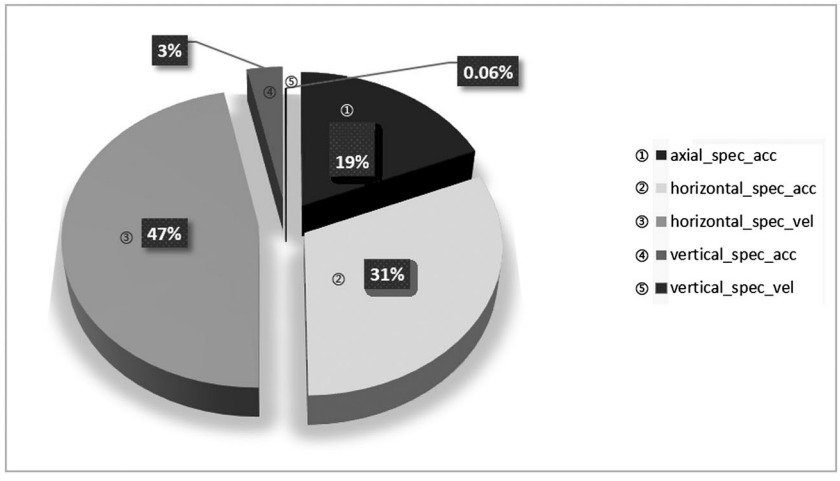

Figure 9. Percentage of variables with high Spearman correlation coefficients for states I and III in different files.

Copyright © 2015 John Wiley & Sons, Ltd.

DOI: 10.1002/cpe

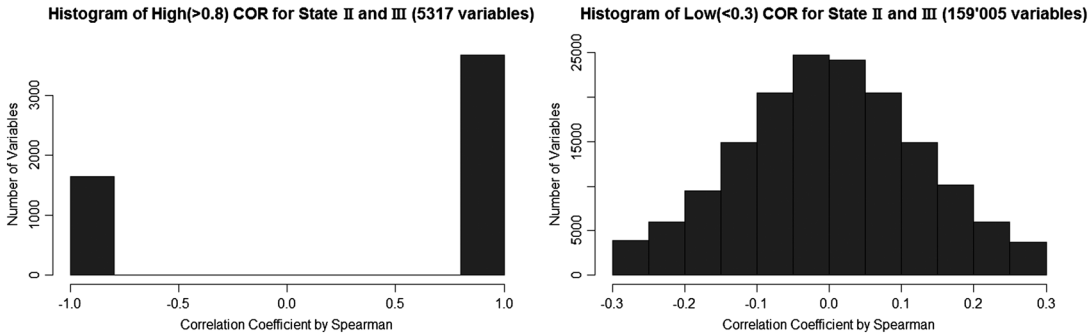

Figure 10. Histogram of high (abs 0.8 and p-value 0.05) and low (abs 0.3) Spearman correlation coefficients for states II and III.

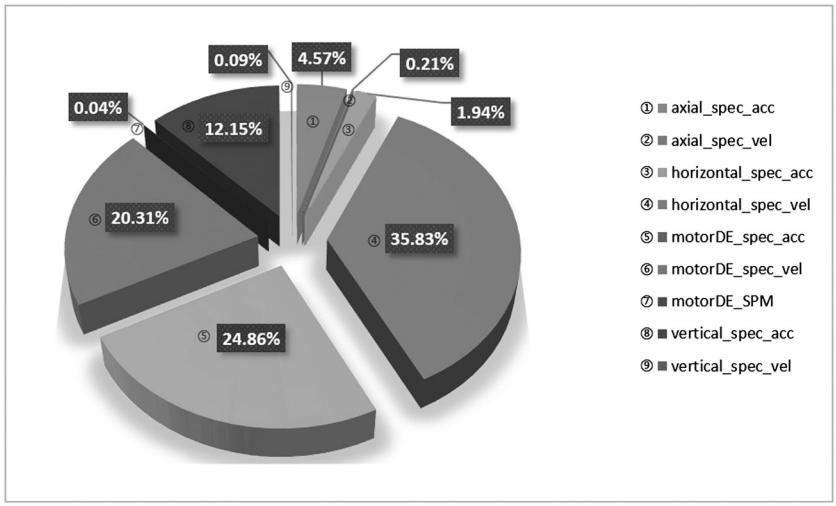

Figure 11. Percentage of variables with high Spearman correlation coefficients for states II and III in different files.

variables mainly come from the 'horizontal_spec_vel' file (35.83%), 'motorDE_spec_acc' file (24.86%), 'motorDE_spec_vel' file (20.31%), and 'vertical_spec_acc' file (12.15%), which can be seen in Figure 11.

5.4. Spearman correlation coefficient for three states

Histogram of correlation coefficients by Spearman in Figure 5 shows the distribution of the correlation coefficient, in which the number of variables with high correlation coefficient, that is, whose absolute value is greater than 0.80 and p-value less than 0.05, is 115, while the number of variable with low correlation coefficient, whose absolute value is less than 0.3, is 187′021, accounting about 89.97% of the total variables. The histogram of high ($\text{abs} > 0.8$ and $p\text{-value} < 0.05$) and low ($\text{abs} < 0.3$) correlation coefficient by Spearman for three states can be seen in Figure 12; the five highest Spearman correlation coefficients are 0.8225, 0.8208, 0.8191, 0.8180, and 0.8174, while the five lowest are −0.7484, −0.7480, −0.7474, −0.7470, and −0.7464. All the variables come from the 'axial_spec_acc' file (79.13%), 'horizontal_spec_acc' file (18.26%), 'vertical_spec_acc' file (1.74%), and 'vertical_spec_vel' file (0.87%, only one variable), which can be seen in Figure 13.

To sum up, through calculating the results of Spearman rank correlation coefficient for states I and II, states I and III, states II and III, and three states, we have 504 variables with high ($\text{abs} > 0.7$ and $p\text{-value} < 0.05$) correlation coefficient, 1752 variables with high ($\text{abs} > 0.65$ and $p\text{-value} < 0.05$) correlation coefficient, 5317 variables with high ($\text{abs} > 0.8$ and $p\text{-value} < 0.05$) correlation

Copyright © 2015 John Wiley & Sons, Ltd.

DOI: 10.1002/cpe

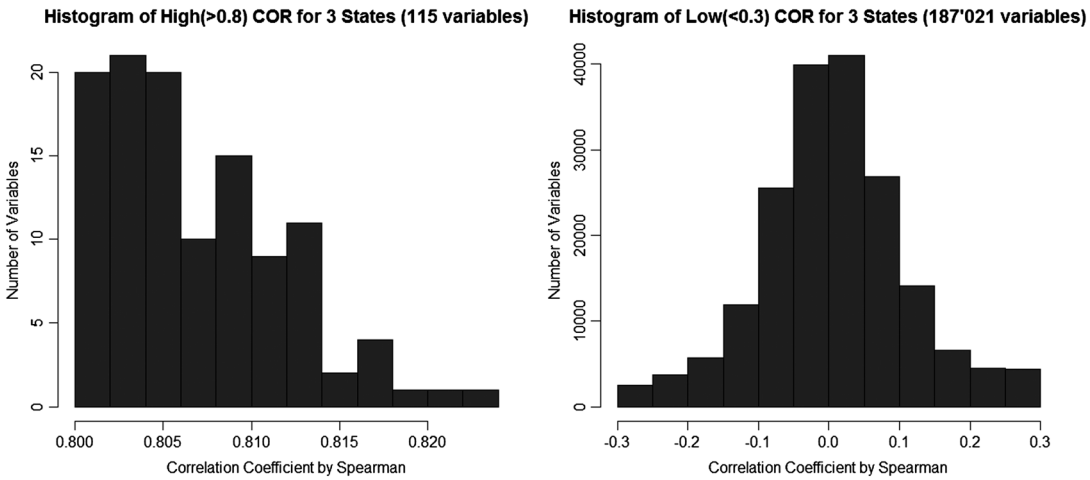

Figure 12. Histogram of high (abs 0.8 and *p*-value 0.05) and low (abs 0.3) correlation coefficients by Spearman for three states.

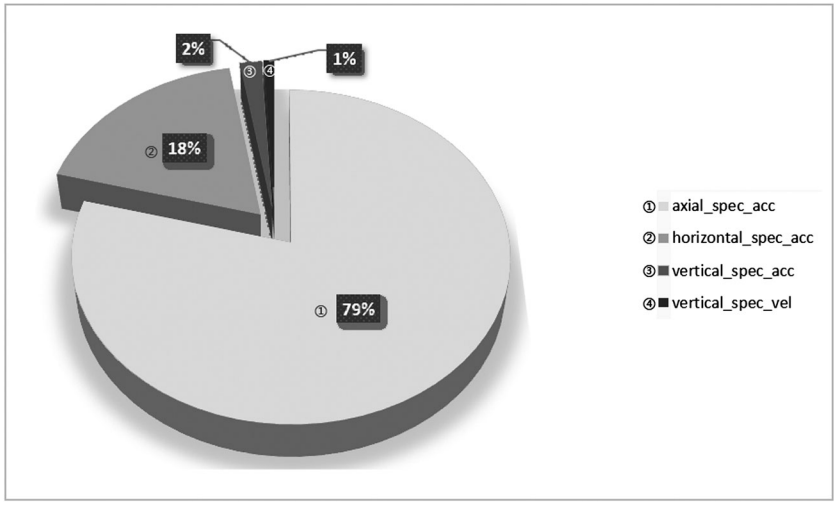

Figure 13. Percentage of variables with high Spearman correlation coefficients for three states in different files.

coefficient, and 115 variables with high (abs > 0.8 and p-value < 0.05) correlation coefficient, respectively.

In addition, through analyzing the high correlation coefficient results for four kinds of different combinations of states, take the Spearman correlation coefficient for states II and III as an example, we find that the variables with high (abs > 0.8 and p-value < 0.05) correlation coefficient mainly come from the 'horizontal_spec_vel' file (35.83%), 'motorDE_spec_acc' file (24.86%), 'motorDE_spec_vel' file (20.31%), and 'vertical_spec_acc' file (12.15%). These percentage values and exact variables' tables are of great importance for the future research of building predictive models.

6. CONCLUSIONS

Correlation analysis is both popular and useful in a number of social networking research, particularly in the exploratory data analysis. And correlation coefficient is a simple statistical measure of the strength and direction between one dependent variable and one or more independent variables. In this paper, three well-known and often-used correlation coefficients, Pearson product–moment correlation coefficient, Spearman, and Kendall rank correlation coefficients, are compared from the

Copyright © 2015 John Wiley & Sons, Ltd.

DOI: 10.1002/cpe

definition to advantages and disadvantages at first. Based on the characteristics of the pump's vibration dataset, the nonparametric and distribution-free Spearman rank correlation coefficient is introduced to do the correlation analysis on the association between the pump's state and each of the 207′880 variables. The percentage of variables and exact variables' tables with high Spearman's correlation coefficients for states I and II, states I and III, states II and III, and three states in different files are obtained respectively, which possess important reference valuation for transforming a larger number of variables into a much smaller set of variables while selecting and building the unsupervised machine learning system.

REFERENCES

1. Stigler SM. Francis Galton's account of the invention of correlation. *Statistical Science* 1989; **4**(2):73–79.
2. Hauke JKT. Comparison of values of Pearson's and Spearman's correlation coefficient on the same sets of data. *Quaestiones Geographicae* 2011; **30**(2):87–93.
3. Spearman C. The proof and measurement of association between two things. *The American Journal of Psychology* 1904; **15**(1):72–101.
4. Kendall MG. A new measure of rank correlation. *Biometrika* 1938; **30**(3):81–93.
5. Székely GJ, Rizzo ML, Bakirov NK. Measuring and testing dependence by correlation of distances. *The Annals of Statistics* 2007; **35**(6):2769–2794.
6. Székely GJ, Rizzo ML. Brownian distance covariance. *The annals of applied statistics* 2009; **3**(4):1236–1265.
7. Lee SY, Poon WY, Bentler PM. A two-stage estimation of structural equation models with continuous and polytomous variables. *British Journal of Mathematical and Statistical Psychology* 1995; **48**(2):339–358.
8. Lopez-Paz D, Hennig P, Schölkopf B. The randomized dependence coefficient. *Advances in Neural Information Processing Systems* 2013; 1–9.
9. Haining R. *Spatial Data Analysis in the Social and Environmental Sciences*. Cambridge University Press: Cambridge, 1993.
10. Javaran SH, Sajadi SAN, Karamoozain M. The relationship between the social responsibility of club with reputation and fans' dependency on the team in the football premier league. 2014.
11. Kariuki JM. Factors influencing sustainability of non government organizations funded community projects in Kenya: a case of action aid funded project in Makima location, Embu county. 2014.
12. Krzywinski M, Altman N. Points of significance: significance, *P* values and *t*-tests. *Nature Methods* 2013; **10**(11): 1041–1042.
13. Devore J. Probability and statistics for engineering and the sciences, Brooks/Cole Pub. Co., Pacific Grove: California, 2011.
14. Milton JS, Arnold JC. *Introduction to Probability and Statistics: Principles and Applications for Engineering and the Computing Sciences*. McGraw-Hill, Inc.: New York, 2002.
15. Onwuegbuzie AJ, Daniel LG. *Uses and Misuses of the Correlation Coefficient Annual Meeting of the Mid-South Educational Research Association* (Point Clear, AL, November 17–19, 1999).
16. Graney BP. Pump vibration analysis. *MISTRAS Products & Systems* 2011; 24–28.

Copyright © 2015 John Wiley & Sons, Ltd.

DOI: 10.1002/cpe