Multiple correspondence analysis to study failures in a diverse population of a cable

Sachan, Swati; Zhou, Chengke; Wen, Rui; Su, Wubin ; Song, Chenjie

Published in:
IEEE Transactions on Power Delivery

DOI:
10.1109/TPWRD.2016.2615470

Publication date:
2017

Document Version
Author accepted manuscript

Multiple Correspondence Analysis to Study Failures in a Diverse Population of a Cable

Swati Sachan, Chengke Zhou *Senior Member, IEEE*, Rui Wen, Wubin Sun and Chenjie Song

***Abstract*--The study of failure behaviour of a diverse population of cables is challenging. Previous attempts have failed to capture the complexity of cable system failures due to an independent analysis of multiple failure causes or influential factors. In this paper, the Multiple Correspondence Analysis (MCA) is proposed for simultaneous analyses of multiple variables responsible for the cable failures and classification of cables into homogeneous groups in terms of past failure behaviour. The proposed classification method is less subjective as it gives equal consideration to all the cable features. The methodology has been applied to the main cable section and cable joint failure data of a diverse population of cables obtained from a Chinese utility company. The failure data have six categorical variables related to cable features and failure characteristics. The application of MCA provided an enriched view and understanding of failure behaviour by allowing visual exploration of the failure patterns and associations. Based on the past failure, the cable sections and joints were classified into three and four groups, respectively. The failure trend of each classified group is evaluated separately. Results show that failure history and trend of each classified group is different. Thus, they must be analyzed and treated differently in the forecasting or maintenance planning procedures.**

***Index Terms*--Study, performance, clustering, grouping, failure**

I. Introduction

The cable population within a utility always has a diverse list of features due to variation in designs, physical size and length, usage of different kind of accessories, and different installation and joining practices. The study of the failure behaviour of a diverse population of cables is very challenging. The most common interest among utilities is to have a broad understanding of cable failure behaviour so that their maintenance engineers can realistically forecast the failures and optimize maintenance strategies in terms of the cost and availability of the cable asset. The study of cable failures is a data driven approach where diagnostic tests, condition monitoring and failure data are utilized. Previous efforts have been made to apply diagnostic test and condition monitoring to enhance the reliability of cable assets. However, there has been limited success in its application to a wide population of cables, mainly due to constraints in the valid interpretation of the results and cost consideration [1]. Therefore, a detailed study of available failure data of a cable population is most useful in making an informed decision and judgment about the future failure possibilities.

Most utilities utilize failure mode and effect analysis (FMEA) for the detailed study the failure data. However, it does not consider the fact that the failure occurrence in a diverse population of cables is a multivariate process, in which the variables (failure cause or influencing factors) responsible for failures are often interrelated and correlated. The correlation between these variables is due to association, also called associative correlation [2]. The associative correlation in the multivariate failure process occurs because of the influence of some unobserved or undetectable failure causing phenomena. For example, it is observed that, degradation due to water tree increases over time in unjacketed XLPE cables [3]. However, degradation could additionally be influenced by other factors which sometimes are undetectable such as, manufacturing defect, poor installation, design issues, etc. The correlation between degradation due to water tree and service time is one of the associations. The variables responsible for failures therefore are dependent on each other due to associative correlation because of which they must be analyzed together. The Cox-Proportionality hazard model has been proposed by [4] for the simultaneous analysis of variables which has a significant effect on the failures of the power cables. In real failure situations the model mostly violates the assumption of non-collinearity between variables and constant effect of influential factors (variables) with time. Hence, a detailed study of the failure data of a population is important before application of statistical models.

A failure dataset can have information about cable design and constructional features (voltage rating, types of insulation, size, length, etc.), installation, past failure and maintenance history. All this information is presented by the set of variables. These variables can have multiple categories due to the presence of a diverse variety of cables in the population. For example, a population of cable can have different voltage levels, core cross-sections, and failure causes, etc. The increase in the number of cables and categorical variables within a data set increases the dimensionality of the dataset. The analysis of high dimensional data is very complex and time consuming [5]. The high dimensional data can be analyzed by utilizing data mining techniques which enable quick analysis of the data without any involvement of expensive human intervention.

The clustering is one of most famous data mining technique. It can group similar objects together. It has been

Swati Sachan and Chengke Zhou are with Glasgow Caledonian University, Glasgow, G4 0BA, United Kingdom (e-mail: swati.sachan@gcu.ac.uk; c.zhou@gcu.ac.uk).

Rui Wen, Wubin Sun and Chenjie Song are with the Suzhou Power Supply Company, State Grid Corporation of China, Suzhou, Jiangsu Province, China.

applied to group the load pattern of the diverse distributed network customers [6], assess the condition by classification of fault or health data [7] and failure mode identification [8]. Before grouping the correlated high dimensional data by clustering it can analysed and visualized by dimension reduction techniques, such as Principal Component Analysis (PCA) for numerical data and, Simple Correspondence Analysis (CA) and Multiple Correspondence Analysis (MCA) for categorical data. The MCA is an extension of CA which enables analysis of more than two categorical variables. It is originally utilized in the social science field [9][10] and biomedical engineering [11] as a practical exploratory tool which provides best and easiest way to gain useful insights of the information hidden in the high dimensional data. Additionally, it helps to understand the interrelations, patterns and association between categories of variables and enables identification of influential factors. The application of MCA is incipient in the field of electrical power engineering. Its counterpart PCA has been applied in recent research works such as fault classification in the power transmission line network [12] and earth fault protection [13].

This work presents the application of MCA to study the failure behaviour of a diverse population of a cable. It demonstrates the visual assessment of the underlying relational structure of the cable features and failure characteristics. In addition, the paper extends to quantify the past performance of each type of the cable in the population and classify the cables into groups.

II. Methodology

The Multiple Correspondence Analysis (MCA) allow the visualization of the structure of high dimensional data and the pattern of association between categorical variables in a two or three-dimensional plot. The visual interpretation of high dimensional data is not possible due to lack of human imagination beyond three dimensions.

A. Multiple Correspondence Analysis for Cable Data

Raw cable data is first organized in a $I \times J$ matrix where, I is number of cables with $i = 1\ to\ I$ and, J is number of variables with $j = 1\ to\ J$. Let the number of categories in a variable be p_j. The total number of categories of all the variables in a raw matrix is $P = \sum_{j=1}^{J} p_j$. The raw data matrix is then converted to an indicator matrix ($\boldsymbol{Z}$), in which data is represented in binary form (0-1). Following are the steps for MCA:

MCA 1:- Indicator Matrix

The indicator matrix ($\boldsymbol{Z}$) is a $I \times P$ binary matrix as shown in Table I. The element in the matrix is denoted by $z_{i,p}$ where, $i = 1\ to\ I$ and $p = 1\ to\ P$. It is either 1 or 0, depending on whether a cable i belongs or do not belongs to category p of a variable. The row sum is equal to number of variables J because each cable belongs to only one of the category of a variable. The column sum is equivalent to i_p, which shows the numbers of cables which belong to category p.

MCA 2:- Relative Frequency Matrix

TABLE I
INDICATOR MATRIX ($\boldsymbol{Z}$)

	Categorical Variables					
Cable	$\boldsymbol{Z_1}$		…	$\boldsymbol{Z_J}$		
1	Z_{11}	Z_{12}	…	$Z_{1p_j}\cdots$	Z_{1P}	J
.	.	.	…	.	.	.
$\boldsymbol{i}$	Z_{i1}	Z_{i2}	…	$Z_{1p_j}\cdots$	Z_{iP}	J
.	.	.	…	.	.	.
$\boldsymbol{I}$	Z_{I1}	Z_{I2}	…	$Z_{1p_j}\cdots$	Z_{IP}	J
	i_1	i_2	…	i_{p_j} …	i_P	IJ

The relative frequency matrix ($\boldsymbol{F}$) is obtained by dividing all the elements of indicator matrix by its grand total sum IJ as shown in Equation (1).

$$\boldsymbol{F} = \frac{\boldsymbol{Z}}{IJ} \tag{1}$$

The elements in frequency matrix are denoted by $f_{i,p}$. The row sum (row_i) and column sum (col_p) are called row mass and column mass, respectively. The expression for row and column mass is

$$row_i = \sum_{p=1}^{P} f_{i,p} = \frac{1}{J}, \quad 0 \le i \le I \tag{2}$$

$$col_p = \sum_{i=1}^{I} f_{i,p} = \frac{i_P}{IJ}, \quad 0 \le p \le P \tag{3}$$

MCA 3:- Chi-square Decomposition

The association between row (cable i) and column (variable categories p) is obtained by chi-square test of independence. It compares the observed value with the expected value. The relative frequency matrix is the chi-square table with $(I-1)(P-1)$ degree of freedoms. The observed value in the relative frequency table is $f_{i,p}$ and its expected value is the product of row and column masses ($row_i \times col_p$). The test statistics is

$$\sum_{i=1}^{I} \sum_{p=1}^{P} {S_{i,p}}^2 \tag{4}$$

where,

$${S_{i,p}}^2 = \frac{f_{i,p} - (row_i \times col_p)}{\sqrt{row_i \times col_p}} \tag{5}$$

here, $S_{i,p}$ is called standardized residual. From expression (5) we get matrix $\boldsymbol{S}$ of standardized residuals. If there is no association between cable and its categorical variables then standardized residual will be equal to zero. The non-zero value of standardized residual exhibit presence of some association.

MCA 4:- Singular Value Decomposition (SVD)

SVD restructures the high dimensional variable data to lower dimensional data space without any information loss by conversion of correlated variables to uncorrelated variables. SVD Let, K be the number of dimensions to which it can reduce the dimensionality of the data. The value of K is obtained by the following expression:

$$K = \min(I - 1, P - 1) \tag{6}$$

The matrix $\boldsymbol{S}$ is decomposed or factorized into three matrixes U, V and E by singular value decomposition, $i.e.$

$$\boldsymbol{S}_{I \times P} = \boldsymbol{U}_{I \times K}\, \boldsymbol{E}_{K \times K}\, \boldsymbol{V}^{T}_{K \times P} \tag{7}$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ are $I \times K$ and $K \times P$ matrices which are composed of Eigen vectors of rows (cables i) and columns (categorical variables p) and E is a $K \times K$ diagonal matrix in which diagonal is composed of Eigen values (λ) in descending order $\lambda_1 > \lambda_2 > \cdots > \lambda_K$. The computation of the SVD by characteristic polynomial is shown in [11]. The

variance of the whole cloud of the data is equivalent to 1.0 (or 100 % when converted into percentage). The sum of all Eigen values or variance of all the K dimensions is equivalent to 1:

$$\sum_{k=1}^{K} \lambda_k = 1 \quad (8)$$

The high Eigen values or variances of a dimension correspond to the thickest direction which has most variance, such dimensions must be retained because they represent of high magnitude of information. The small Eigen values correspond to the thinnest directions where the variance is least, such dimensions can be ignored because they have relatively little information.

MCA 5:- Principle Coordinates

The matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ are composed of Eigen vectors of unit length. It is necessary to resale the unit vector in $\boldsymbol{U}$ and $\boldsymbol{V}$ to get the principle coordinates which can be plotted to create a perceptual map. The principle coordinates for row (cable i) and column categories (variable categories p) are defined as:

$$\phi_{i,k} = \frac{u_{i,k}\,\lambda_k}{\sqrt{row_i}}, \quad 0 \le i \le I \text{ and } 0 \le i \le K \quad (9)$$

$$\theta_{p,k} = \frac{v_{p,k}\,\lambda_k}{\sqrt{col_p}}, \quad 0 \le p \le P \text{ and } 0 \le i \le K \quad (10)$$

where, $u_{i,k}$ and $v_{p,k}$ are elements of matrix $\boldsymbol{U}$ and $\boldsymbol{V}$ which are scaled by the Eigen values λ_k which is diagonal element of matrix $\boldsymbol{E}$. The expression (9) and (10) are divided by the row and column mass, respectively to transform the chi-square space to Euclidean space for the purpose of graphical representation of points. The first column of ϕ and θ is the first dimension of row (cable i) and column (variable categories p). Similarly, the second column of ϕ and θ is the second dimension of row and column and so on till last dimension K. Usually first two dimensions are utilized to study the structure of data by interpreting the association between variables. However, if the variance (λ) is low in first two dimensions then the third dimension is considered.

B. Classification of cables by performance score

By following the procedure in the previous section, most of the information from original high dimensional space can be condensed into first two dimensions which have the highest variability. The variability of the first dimension is higher than the second dimension, ($\lambda_1 > \lambda_2$) which produces elliptical shaped data as shown in Fig. 1. The mean of the data is the center $(0,0)$ of the ellipse. The cable design & constructional features and failure characteristics which are close to the mean have high failure tendency and high influence on failure, respectively. Therefore, the mean point is called ideal failure circumstance point. The distance between a category and mean point quantifies a category contribute towards the ideal failure circumstances. The distance from the center in ellipsoidal shaped data is found using Mahalanobis distance [2][14]. It weights the difference in variation along the axes of elongation of data by utilizing covariance and variance. The Euclidean distance is a special case of Mahalanobis distance when the variance of both the dimensions is same.

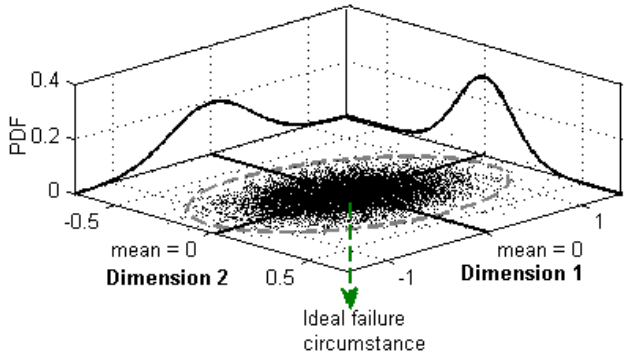

Fig. 1. Ideal failure circumstance point

Let, s_{12} be the covariance between dimension 1 and dimension 2 and, s_1 and s_2 are the variance of dimension 1 and dimension 2, respectively. The Mahalanobis distance of a category which has principle coordinate $(\theta_{p,1}, \theta_{p,2})$ from ideal failure circumstance point (mean, $\bar{\theta} = (0,0)$) is:

$$MD_P^2 = \frac{1}{1-r}\left[\left(\frac{\theta_{p,1}-\bar{\theta}}{s_1}\right)^2 + \left(\frac{\theta_{p,2}-\bar{\theta}}{s_2}\right)^2 - 2r\left(\frac{\theta_{p,1}-\bar{\theta}}{s_1}\right)\left(\frac{\theta_{p,2}-\bar{\theta}}{s_1}\right)\right] \quad (11)$$

where, $r = \frac{s_{12}}{s_1 s_2}$ is the correlation coefficient.

The Mahalanobis distance is utilized to quantify the performance of each type of cable in a diverse cable population. The diversity in cable population is due to a large variety of constructional features in the cables. Suppose, Q number of variables in the indicator matrix (explained in Part A) are constructional features and each variable has p_q number of categories where, $q = 1\ to\ Q$ as shown in Table II. The population can have total T different types of cable, where $T = p_1 \ldots \times p_q \times \ldots p_Q$.

TABLE II
CATEGORIES OF CONSTRUCTIONAL FEATURES

Variables ($q = 1$ to Q)									
$q = 1$			$q = 2$			…	$q = Q$		
Z_1	Z_2 …	Z_{p_1}	$Z_{(p_1+1)}$ …		$Z_{(p_1+p}$	…	$Z_{(\Sigma p_{Q-1}+1)}$ …		$Z_{(\Sigma p_Q)}$
p_1: number of categories			p_2: number of categories			…	p_Q: number of categories		

The classification of cables is important in the identification of the type of cables which had similar performance in the past. The Mahalanobis distance quantifies the performance of each cable type as a performance score. Usually, cables are classified based on one or two constructional features, for example, cables which have a similar cross section or voltage level or both [15]. The classification method proposed in this section classifies a population of cables into groups based on their past performance by providing equal consideration to all the constructional features.

Table III shows the cable type and performance score formulation. Each type of a cable belongs to one of the several possible categories of variables. The score of each cable type is obtained by the sum of Mahalanobis distance of all the categories. A low score reflects poor past performance whereas high score reflects good performance. The set of scores of each cable type is $SC = \{sc_1, sc_2, \ldots sc_t, \ldots, sc_T\},\ t = 1\ to\ T$.

TABLE III
TYPE OF CABLES

Type	Cable type	Performance Score (SC)
1	$Z_1, Z_{(p_1+1)},\ldots,Z_{(\sum p_{Q-1}+1)}$	sc_1:$MD_1 + MD_{(p_1+1)} + \cdots + MD_{(\sum p_{Q-1}+1)}$
2	$Z_1, Z_{(p_1+2)},\ldots ,Z_{(\sum p_{Q-1}+1)}$	sc_2:$MD_1 + MD_{(p_1+2)} + \cdots + MD_{(\sum p_{Q-1}+1)}$
	. . .	. . .
T	$Z_1, Z_{(p_1+p_2)},\ldots, Z_{(\sum p_Q)}$	sc_T: $MD_1 + MD_{(p_1+p_2)} + \cdots + MD_{(\sum p_Q)}$

The Agglomerative Hierarchical Clustering (HC) method is utilized to classify the types of cables into separate groups based on their past performance. Following are steps for Agglomerative Hierarchical Clustering (HC). Suppose, there are three scores sc_{t_1}, sc_{t_2} and $sc_{t_3} \in SC$ in three individual clusters A, B and C, respectively.

HC 1:- Compute the initial dissimilarity matrix of scores: The dissimilarity matrix for all pair of scores sc_t, $sc_{t'} \in SC$ is calculated by the following formula:

$$d_{tt'} = \sqrt{(sc_t - sc_{t'})^2} \quad (12)$$

The dissimilarity matrix obtained from the formula (12):

		A	B	C
		sc_{t_1}	sc_{t_2}	sc_{t_3}
A	sc_{t_1}	0	d_{t_1,t_2}	d_{t_3,t_1}
B	sc_{t_2}	d_{t_1,t_2}	0	$d_{t_3t_2}$
C	sc_{t_3}	d_{t_3,t_1}	d_{t_3,t_2}	0

HC 2:- Merge two clusters with minimum dissimilarity in scores: Suppose, the dissimilarity of scores in matrix have following order $d_{t_1t_2} < d_{t_1t_3} < d_{t_2t_3}$, then cluster A and B of score sc_{t_1} and sc_{t_2}, respectively are merged into a single cluster AB.

HC 3:- Update dissimilarity matrix after merging: The dissimilarity matrix must be updated to reflect the proximity of newly formed cluster and the remaining cluster. The Lance William formula of wards method must be utilized to calculate the dissimilarity between $sc_{t_{12}}$ and sc_{t_3}[16]:

$$d_{t_{12},t_3} = \frac{A+C}{A+B+C} d_{t_1t_3} + \frac{B+C}{A+B+C} d_{t_2t_3} - \frac{C}{A+B+C} d_{t_1t_2} \quad (13)$$

In equation (13) t_1, t_2 and t_3 are number of scores in cluster A, B and C, respectively. Go to HC2**.** Compute until one cluster remains.

HC 4:- The optimal number of clusters can be determined by R^2 index which measures the dissimilarity between clusters. If a cluster contains a group of homogenous cables then it could have a large difference from other clusters which contain different cables, when measured by R^2 [17]. In the presented work, the number of clusters is determined from the freely downloadable statistical software package "R".

III. APPLICATION OF PROPOSED METHODOLOGY TO A CABLE FAILURE DATASET

A. Failure data description

The detailed information about a population of pipe laid underground XLPE insulated cables has been collected from a China utility. The total distributed network of this population of cables is 13327.64 km which has a total of 1889 XLPE cable circuits. A total of 424 failures were observed in the period 2012 to 2014, out of which 194 were main cable sections and 230 were cable joint failures, shown in Table IV.

TABLE IV
YEARLY FAILURES OF XLPE CABLES

Year	Cable Section	Joint
2012	68	50
2013	77	95
2014	49	85
Total	194	230

The dataset has constructional features and failure characteristics of the cables. It has six variables and each variable has a set of categories, shown in Table V. All the variables in the data set are categorical variables. Two numerical variables, age and cable length have been converted into categorical variables because MCA is designed to analyze dataset which has categorical variables. Both age and cable length were binned in three categories A1, A2, A3 and L1, L2, L3, respectively. Three bins for both numerical variables were assumed most appropriate in this case because too few bins lead to too much bias whereas, too many bins lead to little bias with a loss of information and high variability.

The population consists of 82.68 % and 17.32 % of 10 kV and 20 kV cables, respectively. The number of failures in 10 kV cables is higher than 20 kV, shown in Fig 2. However, a greater proportion of 20 kV have failed compared to 10 kV. Here, the proportion is the number of failures in relation to the whole number of cables. Most cables have 400 mm^2, 500 mm^2 and 300 mm^2 core cross-sectionals accounting for 86.92%, 8.57% and 4.18%, respectively of the population. The remaining 0.32 % cables have 95 mm^2, 100 mm^2, 150 mm^2 and 240 mm^2 core

TABLE V
CATEGORICAL VARIABLE OF XLPE CABLES

Constructional Features			
Variables	Categories	Abbreviates	Composition of population %
Voltage level (kV)	10 kV	V10	82.68 %
	20 kV	V20	17.31 %
Core cross-sectional area (mm^2)	95	C95	0.053 %
	100	C100	0.053 %
	150	C150	0.053 %
	240	C240	0.159 %
	300	C300	4.18 %
	400	C400	86.92 %
	500	C500	8.57 %
Cable length (km)	0-10	L1	79.30 %
	10-20	L2	17.84 %
	20-35	L3	2.85 %

Failure Characteristics		
Variables	Categories	Abbreviates
Cause	Manufacturing	MFG
	Installation	INST
	Operational	OPER
	Environmental	ENVIR
	External damage	EXT_DMG
Mode	Open circuit	OC
	Conductor short circuit to ground	SH_GR
	Conductor to conductor short circuit	SH_CON
Age ($years$)	0-5	A1
	5-15	A2
	15-25	A3

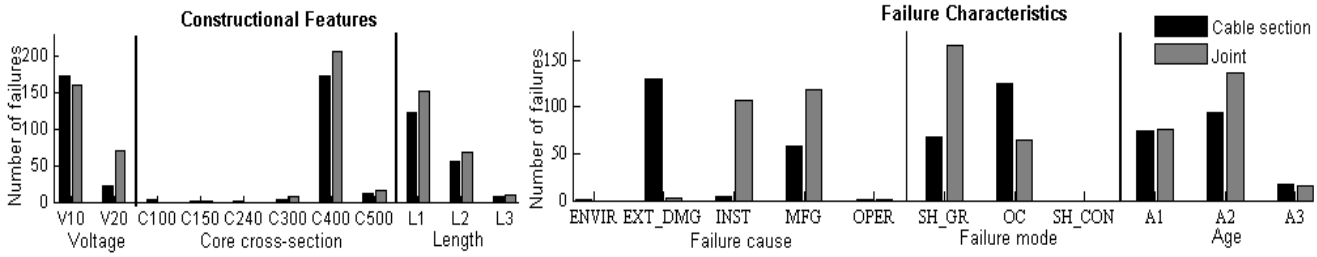

Fig. 2. Constructional features and Failure characteristics of failed XLPE cables

cross-section. The number of failures is highest in the cables with 400 mm^2 core cross section. The large cross-sectioned cables have a high number of failures; however, small cross-section cables have a high proportion of failures. The cable length is the third variable of cable constructional features. Longer cables have more joints and higher probability of degradation with the increase in length [18]. Here in this case Fig 2 shows that the number of failures decreases sharply with the increase in length, however, failure proportion of longer cable is much higher than compared shorter cables. The percentage (proportion) of XLPE cable population, which has failed in L1, L2 and L3 category is 18.22%, 36.49% and 35.29%.

A large number of cables have experienced failure at 0-5 years (A1) of age. Although, most cable had experienced failures are at 5-15 years (A2) of age. Cable sections have a high failure due to external damage and manufacturing defects whereas, joints failures were mostly due to installation and manufacturing defects.

B. Data preparation

The raw data matrix of both cable sections and joints consist of following six variables: voltage, core cross-section, length, failure cause, failure mode and age. The categories in each variable are listed in Table V. The raw data was cleaned before application of MCA. The variables which were common in all the cables such as, XLPE insulation (insulation type) and pipe buried installation (installation method) were excluded from the analysis. Also, some of the data points which were very rare or unique were treated as outliers. The outliers could bias the results thus; they must be removed before the application of MCA. The outliers removed from the data were one 240 mm^2, two 150 mm^2 and two 100 mm^2 core cross-section cables and one cable which had failure due to operational stress and environmental stress. After cleaning the raw data matrix was converted to indicator data matrix. An example of indicator matrix is shown in Table VI.

C. Results

Each row in indicator matrix corresponds to a cable and each column corresponds to a category of a variable. The cable section and joint have 189 × 16 and 228 × 16 (row × column), respectively, sized indicator matrix. The application of MCA has reduced the both dataset to a 10-

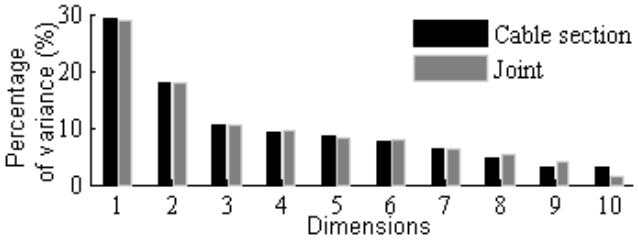

Fig. 3. Percentage of variance in explained by each dimension

dimensional data. The variance of first and second dimension of cable section is 29.23 % and 18.06% (hence a total of 47.29%), respectively, and, the variance of first and second dimension in the joint is 28.84 % and 17.92% (hence a total of 46.76%), respectively. All remaining eight dimensions in both cable section and joint do not have more than 11% of the variance, shown in Fig. 3. Therefore, first two dimensions are good enough to visually interpret the data from the perceptual map. It is important to note that only one or two dimensions can be plotted for visual interpretation more than two are hard to examine visually. The principle coordinates ϕ of rows (cables) and θ of columns (categories of variables) are obtained from Equation (9) and (10), respectively. The first two principle coordinates of ϕ and θ were utilized to plot two-dimensional perceptual map of cable section and joints, shown in Fig. 4.

Cables in perceptual map: - In map each dot corresponds to an individual cable. The cables which have similar profile in terms of constructional features and failure characteristics appear closer to each other and cables which have very different profile appear far away from each other. Some of the cables which have exactly same profile overlap each other in the map. Apart from this, cables which are in close proximity to the mean (0,0) have common profile, while cables which are far away from the mean have uncommon or rare profile. For example, in both cable section and joint map, the most common constructional features of the cables are 400 mm^2 core cross-sections (C400), 10 kV (V10) voltage level and length L1, the cables which appear in close proximity to the mean have either all or at least one of these constructional features. The rarest cables in the cable section map can be seen at bottom left and top right of 300 mm^2 (C300) and 500 mm^2 (C500) core cross section, respectively and the rarest cables in joint map are at left hand side of 300 mm^2 (C300) of core cross-section.

TABLE VI
INDICATOR MATRIX

Cable no.	V10	V20	C300	C400	C500	L1	L2	L3	EXT_DMG	INST	MFG	SH_GR	OC	A1	A2	A3
1	0	1	0	1	0	1	0	0	1	0	0	0	1	1	0	0
2	0	1	0	1	0	1	0	0	1	0	0	0	1	1	0	0
3	0	1	0	1	0	0	0	1	1	0	0	0	1	1	0	0
4	0	1	0	0	1	0	1	0	1	0	0	0	1	0	1	0

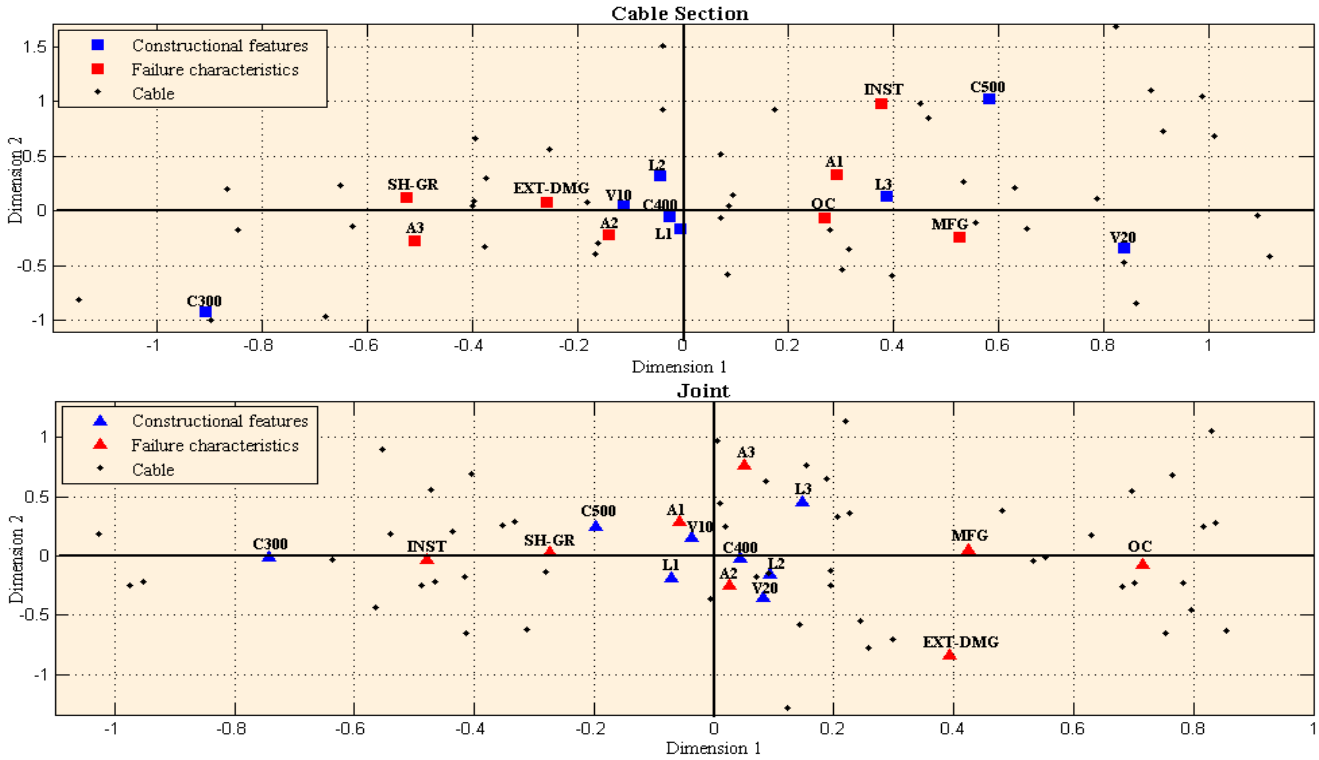

Fig. 4. 2-D perceptual map

Constructional features in perceptual map: -The distance of constructional features towards the mean (0,0) indicate the high failure tendency or frequency in the cables which have those features. In both cable section and joint map, length L1 is closest to the mean then L2 and L3 which indicates that failures were highest in cables with L1 length and lowest with L3 length. Similarly, voltage level 10 kV (V10) is closer to the mean than 20 kV (V20) and 400 mm^2 core cross section are very close to the mean than any other size of core cross-section.

Failure characteristics in perceptual map: - The distance of a failure characteristic towards the mean determines its contribution towards failure circumstances. A short distance from mean exhibits high contribution, as distance increases contribution decreases. In the cable section map, external damage (EXT_DMG) is in close proximity with the mean. In joint map installation (INST) and manufacturing defect (MFG) is close and almost equidistance from the mean which shows that both have almost same failure influence on joint failure. A pattern in failure age can be seen in both maps, the failure age A2 is close to the mean then A1 and finally A3 which indicates that number of failure were high in age A1, then age A2 and least at age A3.

The distance of a constructional feature and failure characteristics from the mean is shown in Table VII. This distance of a constructional feature from the mean is utilized to quantify the past performance of each type of the cable in the population.

TABLE VII
DISTANCE FROM THE MEAN (MAHALANOBIS DISTANCE)

Constructional features	Cable section	Joint	Failure characteristics	Cable section	Joint
V10	0.14	0.38	EXT_DMG	0.67	14.73
V20	8.01	1.86	MFG	3.00	2.16
C400	0.07	0.04	INST	14.88	2.59
C500	12.28	6.89	OC	0.69	6.69
C300	14.45	9.14	GR_SH	2.65	0.97
L1	0.52	0.84	A1	1.53	3.03
L2	2.62	4.24	A2	1.03	1.16
L3	14.02	11.16	A3	3.14	10.027

The constructional features voltage level, core cross section and cable length are not same in all the cables. This discrepancy is responsible for the large variety of cables in this population of the cables. There are two voltage, three core cross section and three cable length levels in the population. Thus, the cable population can have a total of 18 $(2 \times 3 \times 3)$ different types of cable. The Table VIII shows the performance score of 18 different types of cables. The cable type T12, T15, T17 and T18 does not exist in the population; therefore, they are removed after the classification of cables. The hierarchical clustering method was applied to the performance score of each cable type to classify them in separate groups. The cable section and

TABLE VIII
TYPES OF CABLES

Type	Constructional Features of each type of cable			Performance Score	
	Core cross section	Voltage	Length	Cable section	Joint
T1	C400	V10	L1	0.73	1.27
T2	C400	V10	L2	2.83	4.67
T3	C400	V10	L3	14.23	11.59
T4	C400	V20	L1	8.61	2.75
T5	C400	V20	L2	2.62	6.15
T6	C400	V20	L3	14.02	13.07
T7	C500	V10	L1	12.95	8.12
T8	C500	V10	L2	15.05	11.52
T9	C500	V10	L3	26.44	18.44
T10	C500	V20	L1	20.82	9.61
T11	C500	V20	L2	22.92	13
T12	C500	V20	L3	34.31	19.92
T13	C300	V10	L1	15.12	10.37
T14	C300	V10	L2	17.22	13.77
T15	C300	V10	L3	28.62	20.68
T16	C300	V20	L1	22.99	11.85
T17	C300	V20	L2	25.09	15.25
T18	C300	V20	L3	36.49	22.16

joints are classified in three and four groups, respectively, with the decreasing order of their past performance, G1<G2<G3<G4 in Fig 6 and 8.

The failures were observed only in the period 2012 to 2014. They were not recorded from the installation year 1994 to 2011. The unavailability of failure observations before the year 2012 indicates that, the data is left truncated. The observation period had two types of failure type A and B. Fig. 5 shows the type of failure data in the observation period and number cables which were installed in each year. The type A is the failure of the cables which were installed before the observation period and failed in the observation period and type B is the failure of the cables which were installed in the observation period and failed in the observation period. The failure of type C is not available. The type C is the failure of the cables which were installed in the unobserved period and failed in the unobserved period. Therefore, observation period has cables which failed at different ages.

The age-based failure trend of each group of the cable can be captured by the Power Law Non-Homogenous Poisson Process model (NHPP) [3][19]. The Power Law NHPP model has two parameters, shape (β) and scale(α). The shape parameter shows the failure trend. The age-based failure trend of the classified groups of cables in both cable section and joint has decreasing failure trend with $\beta < 1$. The group G2 in the joint has fairly constant failure trend $\beta \cong 1$. The decreasing failure trend with age as shown in Fig. 7 and 9 are consistent with the bath-tub curve reported by earlier researchers [20][21]. The classified groups of cables suffered from "infant mortality failures" in their early age and "random failures" after few years of service life. These groups have not yet manifested "wear-out" or "ageing failures", which is expected from the young population of cables. Therefore, it is interesting to observe that, in this cable population, the highest failures were not observed from the cables which were relatively older (old installations), as the population has not reached the age when aging-related failures would occur in volume. In future, most failure occurrences will be in the cables which are new. The failure causes of all groups are described in Table IX. All cable section groups had most failures due to external damage (EXT_DMG). Only group G2 of cable section had almost equal number of failures due manufacturing (MFG) and external damage (EXT_DMG). All joint groups had most failures due to installation (INST) and manufacturing defects (MFG). The cable type T1, T2, T4 and T5 are common in group G1 of both cable section and joint which clearly indicates that these cables had very high cable section and joint failures in the past three years (2012 to 2014).

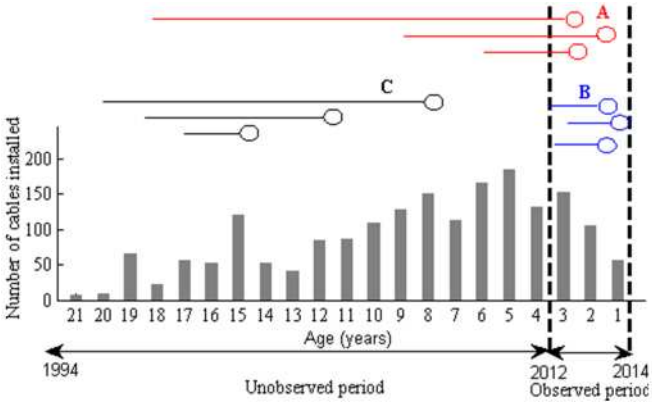

Fig. 5. Observed period and number of installations

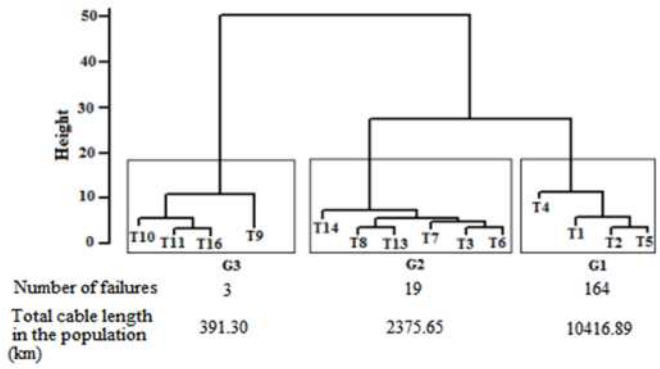

Fig. 6. Cable section groups

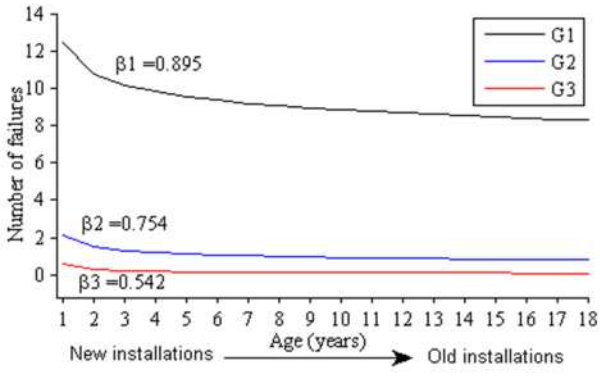

Fig. 7. Failure trend of cable section groups

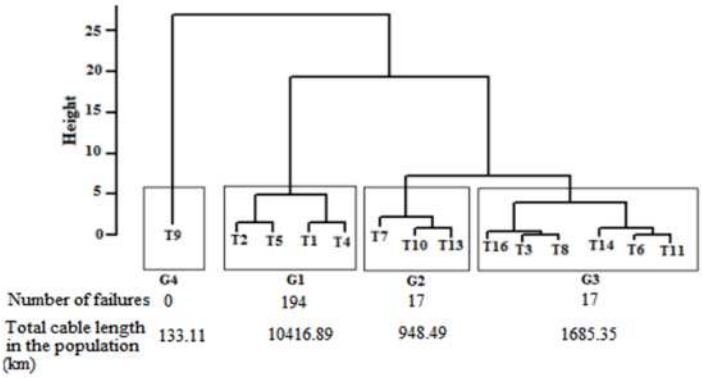

Fig. 8. Joint groups

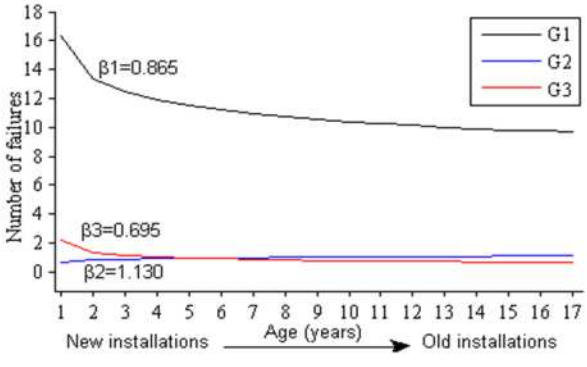

Fig. 9. Failure trend of joint groups

TABLE IX
FAILURE CAUSE OF EACH GROUP

	Cable Section (%)			Joint (%)			
Failure cause	G1	G2	G3	G1	G2	G3	G4
EXT_DMG	67.60	52.63	100.00	1.55	0.00	0.00	0.0
MFG	30.49	42.11	0.00	53.61	47.06	35.29	0.0
INST	1.83	5.26	0.00	44.85	52.94	64.71	0.0

IV. CONCLUSION

In this paper, it is shown how MCA helps to enrich the view and understanding of cable failure behaviour by allowing the user to visualize the preliminary pattern and associations which get obscured in high dimensional multivariate data.

The methodology presented in this paper can be used as a tool to study the failure behaviour of a diverse population of the cables. The MCA has been successfully implemented on the failure data of the Chinese utility company. The categorical failure data related to cable features and failure characteristics were analyzed together and performance of each type of the cable in the population was quantified. The quantified performance enabled the classification of cables. The proposed classification method is less subjective and can be used to pre-study cable failure data before executing statistical analysis for failure prediction and planning the maintenance strategy.

The MCA has certain limitations; it assumes that all the categorical variables which influence the cable failure are included in the analysis. It is an exploratory data analysis method, therefore; it is not suitable for statistical testing.

REFERENCES

[1] D. Wald, J. Perkel and N. Hampton, "Decision making & forecasting using the data available to utilities - Pitfalls, challenges and case studies of ways forward," *Jicable*, June 2015.

[2] R. L. Mason and J. C. Young, *Multivariate statistical process control with industrial applications*, American Statistical Society and Society of Industrial and Applied Mathematics, 2002, pp. 4-9.

[3] S. Sachan, C. Zhou, G. Bevan and B. Alkali, "Failure prediction of power cable using failure history and operational conditions," in *IEEE Properties and Applications of Dielectric Materials (ICPADM) Conf.*, pp.380-383, July 2015.

[4] Z. Tang, C. Zhou, W. Jiang, W. Zhou, X . Jing, J. Yu, B. Alkali and B. Sheng, "Analysis of significant factors on cable failure using the cox proportional hazard model," *IEEE Trans. Power Delivery*, vol. 29, pp. 951-957, March 2014.

[5] H. Samet, *Foundation of multidimensional and metric data structures*, 1st Edition, Morgan Kaufmann, 2006, pp. 488.

[6] T. Xu, H. Chiang, G. Liu and C. Tan ,"Hierarchical K-means method for clustering large-scale advanced metering infrastructure data," ," *IEEE Trans. Power Delivery*, vol. PP, pp. 1-1, September 2015.

[7] A. A. R. Kazemi, M. Vakilian, K. Niayesh and M. Lehtonen," Data mining of online diagnosed waveforms for probabilistic condition assessment of SF_6 circuit breaker," *IEEE Trans. Power Delivery*, vol. 30, pp.1354-1362, June 2015.

[8] S. G. Arunajadai, R. B. Stone and I. Y. Tumer,"Failure mode identification through clustering analysis," *Quality and Reliability Engineering International*, Vol. 20, Iss. 5, pp. 511-526, August 2004.

[9] M. R. D'Esposito, D. D. Stefano and G. Ragozini,"On the use of multiple correspondence analysis to visually explore affiliation networks," *Social Networks*, Vol. 38, pp. 28-40, July 2014.

[10] "Using multiple correspondence cluster analysis to map the competitive position of airlines," *Journal of Air Transport Management*, Vol. 17, Iss. 5, pp. 302-304, September 2011.

[11] J. C. G. D. Costa, R. M. V. R. Almeida, A. F. C. Infantosi and J. H. R. Suassuna,"A heuristic index for selecting similar categories in multiple correspondence analysis applied to living donor kidney transplantation," *Computer Methods and Programs in Biomedicine*, Vol. 90, pp. 217-229, June 2008.

[12] J. A. Jiang, C. L. Chuang, Y. C. Wang, C. H. Hung, J. Y. Wang, C. H. Lee and Y. T. Hsiao,"A hybrid framework for fault detection, classification, and location-Part 1: Concept, structure, and methodology," *IEEE Trans. Power Delivery*, vol. 26, pp.1988-1998, July 2011.

[13] Y. Wang, J. Zhou, Z. Li, Z. Dong and Y. Xu,"Discriminant- analysis-based single-phase earth fault protection using improved PCA in distribution systems," *IEEE Trans. Power Delivery*, vol. 30, pp.1974-1982, March 2015.

[14] R. G. Brereton, "The mahalanobis distance and its relationship to principle component scores," *Journal of Chemometrics*, Vol. 29, Iss. 3, PP. 143-145, March 2015.

[15] Y. Zhou and R. E. Brown, "A practical method for cable failure rate modelling," *IEEE PES Transmission and Distribution Conference and Exhibition*, pp. 794-798, May 2006.

[16] F. Murtagh and P. Legendre, "Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion?," *Journal of Classification*, Vol. 31, pp. 274-295, October, 2014.

[17] M. Halkidi, Y. Batistakis and M. Vazirgiannis , "Cluster validity methods: Part II," ACM SIGMOD Record, vol.31 pp. 19-27.

[18] T. Takeda, T. Okamoto and H. Suzuki, "A study of the actual failure situation of XLPE cable and the order of priority of degradation diagnosis," *Electrical Engineering in Japan*, Vol. 148, Iss. 4, pp. 50-58, September 2004.

[19] D. M. Louit, R. Pascual and A. K. S. Jardine,"A practical procedure for the selection of time-to-failure models based on the assessment of trends in maintenance data," *Reliability Engineering and System Safety*, Vol. 94, pp. 1618-1628, Octobar, 2009.

[20] J. Altamirano, T. Andrews, M. Begovic, Y. D. Valle, R. Harley, J. C. H. Mejia, T. P. Parker, "Diagnostic testing of underground cable systems," NEETRAC 04-211/04-212/09-166, December, 2010.

[21] M. Stotzel, M. Zdrallek, W. H. Wellssow, "Reliability calculation of MV distribution networks with regard to ageing in XLPE insulated cables," *IEEE Generation, Transmission and Distribution*, Vol. 148, Iss. 6, pp. 597-602, November, 2001.

Swati Sachan received her Ph.D. from Glasgow Caledonian University and is currently working as a research fellow at Center for Risk and Reliability Engineering at the University of Nottingham, UK. She holds Master's degree in Operational Research from The University of Edinburgh, UK. Her research interest is in the application of statistics and machine learning in utility asset management.

Chengke Zhou (FIET, SMIEEE and CEng) received his PhD degree from The University of Manchester, UK in 1994. He joined the School of Engineering and Built Environment at Glasgow Caledonian University (GCU) in 1994 and worked as Post-Doc Research Fellow, Lecturer and Senior Lecturer until August 2006 when he joined Heriot-Watt University as a Reader. In June 2007 he returned to GCU as a Professor. He has over 30 years research experience in power systems and partial discharge based HV plant condition monitoring and has acted as consultant to EDF Energy, Scottish Power plc and British Energy.

Rui Wen graduated with her BSc degree from the School of Electrical Engineering, Sichuan University in 2000. She is currently Head of the Department of Cable Operation, inspection and Maintenance, in China State Grid Suzhou Electrical Power Company.

Wubin Sun （born in1968） graduated with his B.Sc. degree in 2008 and is currently a group leader of the Department of Cable Operation, Inspection and Maintenance, Suzhou Power Supply Company, State Grid Corporation of China.

Chenjie Song （born in 1982） graduated with his B.Sc. degree in 2010 and is currently a maintenance engineer in the Department of Cable Operation, Inspection and Maintenance, Suzhou Power Supply Company, State Grid Corporation of China. His main research interest is the location of cable fault and estimation of cable life.