Grouping of California State Wildland Fire Accident Data Based on Discriminant Classification and Cluster Analysis

Dayakar L. Naik[1], Pallavi Sharma[2] and Shantanu Awasthi[3]

Abstract: In the current study, the grouping of wildland fire accident data corresponding to state of California is performed based on discriminant analysis and cluster analysis. Descriptive statistics of the selected variables of fire accident data is demonstrated through Q-Q plots and normality of data is verified. Stepwise reduction technique is adopted to reduce the number of variables. Principal component analysis (PCA) is performed to reduce the dimensionality of the data. MANOVA is performed on the dimensionally reduced data to verify the assumptions of covariance equality and classification rule is determined. Discriminant rule based classification is performed on the data to classify them into one of the 9 causes of the fire. Cluster analysis is performed to determine the number of groups. The accuracy of discriminant based classification is evaluated to be 36% when three variables were chosen. The Ward's method resulted in 3 number of clusters and complete linkage resulted in 6 number of clusters.

Keywords: Wildland Fire; Principal Component Analysis; MANOVA; Discriminant Analysis, Cluster Analysis.

[1]Graduate Research Assistant, Dept. of Civil & Env.Engg., North Dakota State University, ND58105.
[2]Graduate Student, Dept. of Computer Science, North Dakota State University, ND58105.
[3]Graduate Research Assistant, Dept. of Mathematics, North Dakota State University, ND58105.

1. Introduction

Fire is a dynamic force of nature that can be devastating if not controlled. Among the various kinds of fire accidents that occur, wildland fires are nearly as impossible to prevent and difficult to control [1].It plays a key role in shaping ecosystems by serving as an agent of renewal and change. On an average 67000 fire incidents occur every year in USA resulting in 800,000acres of land burnt[2].These wildland fires can be deadly, destroying homes, wildlife habitat and timber, and polluting the air with emissions harmful to human health. The U.S. spends over $5 billion dollars to fight fires each year[3].Wildfires are so extreme that they can burn at a temperature more than 2,000 degrees Fahrenheit. Wildfires can grow big that they change the weather patterns of a given area. They also spread fast on the ground, moving twice as fast an average human can run. One way to minimize the occurrence of such fire accidents is to investigate the cause of fire as it greatly aids in taking appropriate control and safety measures. Therefore, there is a need to understand the various attributes that influences the initiation and spread of fire.

Wildland fires are severe and devastating in few states of US. Among these states, state of California is ranked one in terms of risk for households followed by Texas and Colorado, while it is ranked number two in terms of number of fire accidents and number of acres burnt [2].In year 2016, 7000 accidents were reported in the state of California with 560,000 acres of land burned and they have resulted in 8900 million dollars financial loss [2].Accident that occurred in Monterey county is one among the noted fire accidents in year 2016 that resulted in 132,000 acres burnt. While the fire causes for majority of accidents are reported, there are many more incidents where the fire cause is not documented. It might be possible that the fire cause is one among the already known fire causes or might be some new cause. This can be discovered using multivariate statistical techniques.

In the current study, the undetermined fire cause of the fire accidents that occurred in California state are investigated using the multivariate statistical technique. Specifically, the discriminant rule based classification and cluster analysis is performed to assign the known fire causes or new fire cause. Although this can be performed on the full dataset available for all the states, the current study is only limited to the state of California due to severity of wild land fires in this state. However, the proposed methodology can be extended to the full dataset with the sufficient amount of computational resources provided. The rest of the manuscript is organized as follows: the severity of wildland fire in California is described in Section 1, the methodology adopted to assign the fire causes or newly formed clusters is described in Section 2, results obtained from the analysis are provided in Section 3, followed by conclusions in Section 4.

2. Methodology

The aim of the current study is to group the wildland fire accident data that occurred in the state of California, USA. based on the discriminant rule based classification and the cluster analysis. In the case of discriminant rule based classification, the number of groups is already known priori, whereas in the case of cluster analysis the number of groups is not known priori. Therefore, more precisely, the first objective can be framed as to '*identify*' the group of the given wildland fire accident data from the known groups and the second objective is to '*partition*' the data into unknown 'n' number of groups. These objectives are accomplished by following a methodology that is described in detail in this section. The methodology adopted in this study involves six steps: 1) data acquisition and description, 2) step-wise variable reduction 3) principal component analysis (PCA), 4) MANOVA, 5) discriminant rule based classification and 6) hierarchical clustering (agglomerative).

2.1. Data Description

Prior to the application of any multivariate statistical method, there are three steps that needs to be accomplished: 1) data acquisition from the reliable source, 2) data cleaning using appropriate tools and 3) evaluation of descriptive statistics of the cleaned data. While the first two steps are described in this section, the third step is discussed in detail in the Section 3.1.

2.1.1. Data Acquisition

Data acquisition is the process of gathering data from either a primary source or a secondary source. Primary source is the one that collects the data on site through field works, interviews and surveys etc., while the secondary source includes books, mass media products and web information etc. In the current study, the data is acquired in the form of compact disc (CD) from a primary source, the United States Fire Administration (USFA), a division of the Federal Emergency Management Act. (FEMA) that is managed by the Department of Homeland Security (DHS) [4].The data obtained consists of reported wild land fire accidents that occurred throughout the United States and includes details such as location and time of the accident, various causes of fire accident, measured physical variables during accident (air temperature, relative humidity, flame length etc.), financial losses, injuries and fatalities due to accidents etc. More details about the data is presented in the Section 2.1.2. Although the data is available for all the states, the current study is restricted only to the analysis of the accidents that occurred in the state of California, due to lack of computational resources.

2.1.2. Data Cleaning and Integration

Data cleaning is the process of correcting the inconsistencies in the data with the aim of generating a more organized and structured data[5]. Very often, the data acquired from the primary source (called as raw data) consists of missing information, typographic errors and inconsistent format etc. It is almost impossible to perform any kind of statistical analysis on such unorganized inconsistent data. Therefore, the data needs to be corrected for any such existing errors and inconsistencies using tools such EXCEL®, MATLAB®[6]and SQL® etc. Sometimes, variables corresponding to the same incident may be present in different files and in such cases, there is a need to integrate both the files based on 'schema matching' and this process is known

as *data integration*. The rest of this section describes the data cleaning and integration performed in this study.

In the current study, the wildland fire accidents that occurred in the past 3 years (2013-15) corresponding to the state of California is considered. This dataset primarily consists of 71 features like Incident Date, Fire Cause, Human Factors, Heat Source, Weather Type, Danger Rate etc. Some of these features are qualitative in nature, that are out of scope of this study and hence disregarded. While rest of the variables are quantitative in nature, only specific attributes are chosen that seemed to be descriptive i.e. the variables that can best describe the incident. Among the 71 features, 12 features are chosen (State, Fire Department ID, Incident Date, Incident Number, Fire Cause, Wind Speed, Air Temperature, Relative Humidity, Acres Burned, Elevation, Flame length, spread rate) that seemed to be of relative importance in the current study. In order to account for the effects that resulted from wildland fire accidents, another dataset corresponding to 'Basic Fire Incident' is merged with the prior dataset. This dataset includes information pertaining to property loss, Content loss and death count of fire fighters and civilians. The losses from Basic Incident dataset corresponding to incidents present in Wild land data set were fetched using inner join in Tableau. The two datasets were joined based on Incident date, Incident number, Fire Department Id and Incident type related to Natural Vegetation fire (Incident type 140, 141, 142 and 143). The resulting dataset includes 27466 data points with 14 features grouped by fire Cause. There are 10 fire causes specified in the dataset (1-'Natural source', 2-'Equipment', 3-'Smoking', 4-'Open/ Outdoor Fire', 5- 'Debris/vegetation burn', 6- 'Structure(exposure)', 7- 'Incendiary', 8- 'Misuse of fire', 0- 'Other cause' and U– 'Undetermined') and these serve as grouping variable for our further analysis.

The dataset retrieved in the previous step includes missing data and inconsistent data format. Among the available variables, air temperature, wind speed and relative humidity are set as bench mark and the rows with missing data were deleted. However, the missing values in the remaining columns are replaced by the average values of the corresponding columns (like 1200 ft. for elevation, 2 ft./hr. for spread rate, 4 ft. for flame length etc.). The final dataset resulted in 11051 data points for 13variables grouped by fire cause. The data points for fire Cause 0 to 8 were separated from fire Cause 'U' for their use as training dataset. The data points for fire Cause 'U' (Undetermined) were used as testing dataset. The final dataset for the further analysis thus consists of 7491 data points in training set and 3560 data points in testing set. Hereafter, the results to be reported are based on training dataset only, unless otherwise mentioned. Followed by the process of data cleaning and data integration, the data is ready for statistical analysis.

2.2. Stepwise Reduction of Variables

Stepwise reduction (also known as stepwise MANOVA)is a commonly adopted procedure that is implemented to reduce the number of dependent/redundant variables during linear or quadratic discriminant analysis of the data [7].This procedure includes addition or deletion of variables in each step based on the likelihood ratio test statistic Λ known as Wilk's lambda test. When the Wilk's lambda value is near zero it indicates high discrimination and those close to one indicate poor discrimination[8] . In the current study, as the dependence between the 13 variables of the wildland fire accident data is not known priori, it is important to perform the stepwise reduction

of variables to determine the redundant variables. This task is accomplished using the program written in SAS®[9]environment by implementing the command '*PROC=STEPWISE*'. Initially, this analysis begins with all the 13 variables that are taken into consideration and after the complete process only redundant variables are obtained as the output. The complete procedure of stepwise reduction is explained elsewhere. The corresponding SAS® program that is developed is presented in the Appendix B.

Assumptions: The 13 variables follow Multivariate Normal Distribution for each fire cause. The samples taken are random and independent. The 9 populations from which observation vectors are drawn have a common covariance matrixΣ.

2.3. Principal Component Analysis

Principal component analysis (PCA) is the well-established dimensional reduction technique implemented to represent the data as a linear combination of variables such that the maximum variance is obtained. Although the variable reduction is performed in the previous step, the PCA is conducted for the sake of completeness and which might also aid in further reduction of variables, if possible, and identification of any presence of clusters of data on the biplot. Biplot is the scatter plot of the one principal component versus the other principal component. This step is accomplished using the program written in RStudio® and SAS® environment using the command '*PROC=PRINC*'. Prior to the execution of the code, the data is standardized to avoid the higher variance of some variables observed in the covariance matrix. The original data is then transformed using the computed principal components and the biplot is generated. The SAS® and R code corresponding to the PCA is presented in Appendix B.

2.4. Box's M Test

Box's M test is used to identify the homogeneity of the covariance matrix ($\Sigma_1 = \Sigma_2 = \Sigma_3 = \cdots = \Sigma_k$) across the '$k$' groups. The equivalence of the covariance matrix is an important assumption while performing MANOVA and discriminant classification. Specifically, in discriminant rule based classification, this assumption plays a vital role in deciding whether the discrimination classification is linear or quadratic. Therefore, it is required to check if at least one of the covariance matrices among the nine fire causes significantly differ from the other. In the current study, Box's M test is conducted on the data set with the nine variables (obtained after stepwise reduction) using a program written in SAS® environment. All the nine fire causes are assumed to be independent and follow multivariate normal distribution (MVN) while performing this test. The SAS® program computes the test statistic 'u' (shown in Eq.1) based on the given data and it is compared to the approximate chi-square distribution (χ^2) to make a decision on acceptance or rejection of the null hypothesis: 'Covariance matrices of each fire cause does not significantly differ from each other$-\mathrm{H}_o$: $\Sigma_0 = \Sigma_2 = \Sigma_3 = \cdots = \Sigma_8$'.

$$\mathrm{u} = -2(1 - \mathrm{c}_1)\ln(\mathrm{M}) \sim \chi^2\left[\frac{1}{2}(\mathrm{k} - 1)\mathrm{p}(\mathrm{p} + 1)\right] \quad (1)$$

$$c_1 = \left(\sum_{i=1}^{k}\frac{1}{v_1} - \frac{1}{\sum_i v_i}\right)\left(\frac{2p^2+3p+1}{6(p+1)(k-1)}\right) \quad (2)$$

$$-2lnM = v(\ k\ \ln|S_{pl}| - \sum_{i=1}^{k} \ln|S_i|\ ) \quad (3)$$

Assumptions: The 9 variables follow Multivariate Normal Distribution for each fire cause. The samples taken are random and independent.

2.5. Discriminant Classification

Discriminant rule based classification is one of the well-established data classification methods that is used to assign the new set of observations (whose group is unknown) into one of the known groups. This method primarily relies on the evaluation of discrimination function that maximizes the standardized distance between the means (transformed) of the sample groups. In other words, the discriminant function which is a linear combination of the 'p' variables, consists of discriminant or transformation coefficients 'a' that transforms the mean of each group such that the distance between them is maximized. From this idea, a distance function (shown in Eq. 4) is derived using maximum likelihood ratio method, that compares the distance between the new observation ($\boldsymbol{y}$) and the mean of each group ($\bar{\boldsymbol{y}}_{\boldsymbol{i}}$). Later the new observation is assigned to the group with smallest distance.

$$D_i^2(y) = (\boldsymbol{y} - \bar{\boldsymbol{y}}_{\boldsymbol{i}})' \boldsymbol{S}_{pl}^{-1} (\boldsymbol{y} - \bar{\boldsymbol{y}}_{\boldsymbol{i}}) \ \ldots . i = 0,2,..8 \quad (4)$$

$$D_i^2(y) = (\boldsymbol{y} - \bar{\boldsymbol{y}}_{\boldsymbol{i}})' \boldsymbol{S}_{i}^{-1} (\boldsymbol{y} - \bar{\boldsymbol{y}}_{\boldsymbol{i}}) \ldots . i = 0,2,..,8 \quad (5)$$

In discriminant function based classification, there are two types of distance functions: (1) linear discriminant rule (Eq. 4.) and (5) quadratic discriminant rule (Eq. 5). The first rule is based on the assumption that the covariance matrices of all groups are equal ($\Sigma_0 = \Sigma_2 = \Sigma_3 = \cdots = \Sigma_8$), and the second rule is based on the assumption that the covariance matrices are not equal ($\Sigma_0 \neq \Sigma_2 \neq \Sigma_3 \neq \cdots \neq \Sigma_8$). Also, it is assumed that the observations should exhibit normal distribution[10]. In the current study, the equivalence of covariance matrix is first verified using Box's M test as mentioned in previous section and the appropriate classification rule is determined. This rule is implemented in the further step to assign the group to new observation.

Assumptions: The 9 variables follow Multivariate Normal Distribution for each fire cause. The samples taken are random and independent

2.6.Cluster Analysis

Cluster analysis is a natural grouping technique that is implemented on the dataset to group the observations that are similar to each other[7].Such similar dataset groups are referred to as clusters, and each cluster is dissimilar to the other cluster. In the context of this study, similar and dissimilar are referred to distance between the observations i.e. the distance between the similar data points is smaller when compared to the distance between the dissimilar data points. While

distance based cluster analysis is very popular, there are other methods such as density based methods, grid based methods and mixed distribution methods. In the current study, only distance based method is implemented to perform the cluster analysis on the dataset. Among the existing distance based methods such as hierarchical clustering (single linkage, complete linkage, average linkage, centroid linkage, ward's method) and nonhierarchical method (K-means), Ward's method and complete linkage method are chosen to accomplish the cluster analysis of the dataset. Although other methods could have also been chosen, one method that is sensitive to outliers and one method insensitive to outliers is preferred in this study. The other methods are not implemented in the current study due to time constraint.

In the current study, RStudio® in built function 'NbCluster' is used to perform cluster analysis on the data set whose fire causes are unknown. This function facilitates with the option of distance function ('*Euclidean*') and hierarchical clustering method during the analysis. The dataset, with 9 variables, whose fire causes are unknown, is given as an input to the cluster function and cluster analysis is performed using a) ward's method and (b) complete linkage method. The distance option is chosen as 'Euclidean distance'. The clusters generated as an output from the cluster function is presented in the form of a dendrogram.

3. Results

3.1. Descriptive Statistics

Descriptive statistics consists of a set of techniques that are used to summarize and characterize the measurements of given data. Generally, descriptive statistics includes (a) the basic statistical measures such as mean, median, range, standard deviation and covariance of the data, (b) identifying the type of distribution and (c) recognizing the patterns or correlations among variables. In the current study, a summary of basic statistical measure of California wildland fire accident data is obtained using RStudio®[11]open source software and is shown in **Table 1**. From **Table 1**, (a) the mean values of 'air temperature', 'elevation', 'flame length', 'relative humidity', 'spread rate', 'wind speed' are 74ºF, 1315 ft., 2.22 ft., 38, 4.314ft./hr. and 4.423 miles/hr. respectively, and (b) the range of values for the same variables are 0-119ºF, 0-32400ft., 0-80 ft., 0-100, 0-800ft./hr. and 0-147 miles/hr. respectively. Also for these variables the mean and median values are comparable which reveals the distribution is non-skewed. Unlike the above 6 variables, three other variables 'acres burned', 'content loss' and 'property loss' exhibit skewness and rest of the variables ('fire fighter deaths', 'fire fighter injuries', 'other deaths' and 'other injuries') are ignored as they are zero.

The covariance matrix and correlation matrix are important in describing variation of the data and strength of association among the variables and hence are provided in **Table 2**. and **Table 3**. respectively. Among the 13 variables, a moderate correlation is observed between 'acres burned' and 'spread rate' (positive), 'air temperature' and 'relative humidity' (negative), 'content loss' and 'property loss' (positive), 'flame length' and 'spread rate' (positive), 'wind speed' and 'flame length' (positive) and weak correlation is observed among remaining variables for all the 7244 data points. Later, the box plots for variables ('air temperature', 'elevation', 'flame length', 'relative humidity', 'spread rate' and 'wind speed') with respect to all the 9 fire causes are

generated in MATLAB® and depicted in **Figure 1**Box plots are used to describe the distribution of data based on the five-number summary: minimum, first quartile, median, third quartile, and maximum. From Figure 1, variables 'elevation', 'flame length' and 'relative humidity' are observed to have more number of outliers and almost all the variables are skewed. For the sake of brevity, only the summary of 'air temperature' from **Figure 1**(a) is interpreted and the same applies to the rest of variables. From **Figure 1**(a), (i) the range of 'air temperature' values for causes '0' to '8' are 178.92ºF, 174.17ºF, 174.17ºF, 182ºF, 166.25ºF, 171ºF, 91.84ºF, 174.17ºF and 171ºF (ii) the first quartile for causes '0' to '8' are: 99.4ºF, 96.24ºF, 112ºF, 110ºF, 83.5ºF, 88.32ºF, 88.32ºF, 91.5ºF and 96.24ºF (iii) the median for causes '0' to '8' lies at: 117ºF, 112ºF, 126ºF, 127ºF, 96ºF, 102ºF, 104ºF, 105ºF and 113ºF (iv) the third quartile for causes '0' to '8' are: 134ºF, 126ºF, 137ºF, 139ºF, 115ºF, 116.8ºF, 123ºF, 123ºF and 131ºF respectively.

The rest of the variables ('acres burned', 'content loss' and 'property loss') which are the consequences of the wildland fire accidents are analyzed using pie charts to identify the major cause of the fire accident and shown in **Figure2**. From **Figure2** it is identified that cause '0' is the dominant cause of the 'acres burned' followed by cause '2'; cause '0' is the dominant cause of the 'content loss' followed by cause '5' and cause '3';cause '0' is the dominant cause of the 'property loss' followed by cause '3' and cause '7'. It is important to identify the distribution of the data as most of the multivariate statistical methods are framed based on the assumption that the data exhibits multivariate normal distribution (MVN). Therefore, the normality of the fire accident data associated with each variable and for each fire cause is verified. To verify the normality of the variables in the data, Q-Q plots and bi-scatter plots are used very often. In the current study, these plots are generated for the wildland fire accident data using MATLAB® and are shown in **Figure 3** (only for cause '0'). The Q-Q plots for remaining fire causes are shown in Appendix. From **Figure 3**, it is observed that the variables 'air temperature', 'elevation', 'relative humidity' and 'wind speed' follow normal distribution after box-cox transformation (for cause '0'), while rest of the variables fail to exhibit normality. The complete set of Q-Q plots as shown in Appendix **Figure A.1**-**Figure A.8** reveal that the above-mentioned variables exhibit normality for all the wildland fire causes.

3.2. Stepwise Reduction of Variables

Stepwise reduction of variables is a dimensionality reduction procedure based on Wilk's Λ statistic. In the current study, this procedure is accomplished through SAS® software, as explained in Section 2.2, and the results are presented in **Table 4**. From **Table 4**, it is observed that only 9 out of 13 variables are retained. These variables, ranked in the ascending order, are 'relative humidity', 'Elevation', 'Air temperature', 'property loss', 'spread rate', 'acres burnt', 'content loss', 'wind speed' and 'flame length'. The variables that are removed are either highly correlated and redundant or are not at all relevant [12]. From the results of correlation matrix (see **Table 3**), it is clear that none of the variables are highly correlated, and variables 'fire fighter deaths', 'fire fighter injuries', 'other injuries' and 'other deaths' are mostly zeros. Therefore, the results from stepwise reduction are justifiable that the later variables are removed as they are irrelevant and the variables that are fairly correlated are retained. For further analysis only these 9 variables will be used and remaining 4 variables are ignored.

3.3. Principal Component Analysis

Although the stepwise reduction of variables method reduced the number of variables from 13 to 9, PCA is still performed to check if there is further possibility of dimensional reduction and identify the possible clusters on biplot. The eigen values and corresponding principal components (PC 1 to PC 9) are generated using the RStudio® programming software and shown in **Table 5**. and **Table 6**. Number of principal components to be retained is determined through scree plot as shown in **Figure 4.** Based on the eigen value proportion provided in **Table 5**, six principal components are retained with a cut-off of 80% i.e. six principal components suffice to explain the 80% variance in the data. The first six principal components in **Table 6**. are interpreted to identify the correlation between the variables and principal components. Examining PC 1 reveals that none of the 9 variables is highly correlated with PC 1. Variables 'acres burned', 'flame length' and 'property loss' are moderately correlated with PC 1 and the PC1 increases with an increase in the magnitude of these variables. PC 2 is fairly correlated with the 'air temperature', 'content loss' and 'property loss'. With an increase in 'air temperature' PC 2 increases and with an increase in 'content loss' and 'property loss' PC 2 decreases. PC 3 is highly correlated with the 'elevation' and 'relative humidity' and increases with an increase in the magnitude of 'elevation' and 'relative humidity'. Biplot of the first two principal components is shown in **Figure 5.** From **Figure 5,** any specific clusters or groups are not observed. However, variables 'property loss' and 'content loss'; 'air temperature', 'wind speed' and 'spread rate' are observed to be closely correlated.

3.4. Box's M Test

For the standardized data, Box's M test is performed to verify the assumption of equality of covariance ($\Sigma_0 = \Sigma_2 = \Sigma_3 = \cdots = \Sigma_8$). The resulting table with test statistic for the complete training set (variables = 9 and groups = 9) is shown in **Table 7**. From **Table 7**, the value of test statistic is greater than the χ^2approximation and hence the null hypothesis stated in Section 3.4 is rejected. Based on the M test, there is enough evidence to conclude that the covariance matrices of all fire causes differ significantly from each other.

3.5. Discriminant Classification

As mentioned in Section 2.5, discriminant classification rule is selected based on the assumption of equivalence of covariance matrix ($\Sigma_0 = \Sigma_2 = \Sigma_3 = \cdots = \Sigma_8$). This assumption is verified through Box's M test, in the previous step, and the data involved in the current study exhibited unequal covariance matrices when 9 variables were selected. Therefore, this study resorts to the implementation of quadratic discriminant rule provided in Eq. 5. The prior probabilities in this study are chosen as 0.33 for "0", 0.08 for "1", 0.13 for "2", 0.04 for "3", 0.06 for "4", 0.19 for "5", 0.01 for "6", 0.12 for "7" and 0.04 for "8" fire cause and the resulting equation is modified and provided in Eq. 6.

$$Q_i(\boldsymbol{y}) = \ln(\boldsymbol{p_i}) - \frac{1}{2}\ln(|\boldsymbol{S_i}|) - \frac{1}{2}(\boldsymbol{y} - \bar{\boldsymbol{y}}_{\boldsymbol{i}})'\boldsymbol{S}_i^{-1}(\boldsymbol{y} - \bar{\boldsymbol{y}}_{\boldsymbol{i}}) \ldots . i = 0, 2, .., 8 \qquad (6)$$

Prior to the allocation of appropriate fire cause to the new observations (testing set), the quadratic classification rule is implemented on the already existing data (training set) whose fire causes are known, to verify the accuracy of the classification method. This task is accomplished through SAS® program which generates cross-validation error rate (CVER) and apparent error rate (AER). In this study, only the cross-validation error rate is reported as it is widely accepted validation method in the statistical community. The cross-validation error rate for the current problem resulted in 25% accuracy when 9 variables are chosen. It is observed that most of the observations are allocated to fire cause '0'.

Although there are 9 variables after dimensionality reduction, only few variables might be enough for the classification purpose in order to avoid confusion matrix that results in poor accuracy. Therefore, an attempt is made in this study to identify the set of variables (among the existing 9 variables) that improves the accuracy of classification. This is accomplished by eliminating each variable one after the other until the accuracy is improved. The set of variables considered and the accuracy corresponding to each set is shown in **Table 10**. It is important to note that the Box's M test needs to be performed for each new set of variables that are considered for the classification problem and is shown in **Table 9**. From **Table 9**, it is observed that the accuracy is improved when set of variables ("relative humidity", "elevation" and "air temperature") are considered and the cross-validation error rate for the current problem is shown in **Table 10**. Also, it is interesting to note that, irrespective of variable selection, the classification rule remained quadratic. Based on the results obtained in **Table 9**, only "relative humidity", "elevation" and "air temperature" variables are chosen for allocating the groups to the new observations. As mentioned before, there are 3560 observations whose fire cause is unidentified and RStudio® program is used to assign the fire causes to the new observations. The summary of the fire causes assigned to the new observations is presented in **Table 11**. Most of the new observations are allocated to fire cause '4'.

3.6. Cluster Analysis

In the current study, cluster analysis is performed on the dataset whose fire cause is unknown, assuming that the fire causes are not one among the already existing fire causes. As described in Section 2.6, the cluster built in function 'NbCluster' from RStudio® is used to cluster the data. The clusters generated using both 'Ward's method' and 'complete linkage' method is illustrated in **Figure 6** and **Figure 7** respectively. The clusters generated from both the methods are observed to be different. Complete linkage results in one sided clustering. From **Figure 7** it is clearly observed that 6 number of clusters are clearly distinguished, whereas from **Figure 6** it is observed that 3clusters are present.

Prior to computation of optimal number of clusters for new observations, 'NbCluster' function is implemented on the already available dataset. It resulted in 10 optimal number of clusters while known fire causes are 9. For the new observations, the optimal number of clusters are determined using 12 different indices available in RStudio® software package and are tabulated in **Table 12**. Ward's method resulted in 3 clusters as an optimal number and complete linkage method resulted in 6 clusters. It can be concluded that there is almost no dissimilarity among the

observations. Irrespective of the cause of fire, the effect and the result are same. Few groups of clusters that are formed can be attributed to outliers.

4. Conclusions

1. The variables ‘air temperature’, ‘elevation’, ‘flame length’, ‘relative humidity’, ‘spread rate’ and ‘wind speed’ are observed to follow non-skewed distribution based on the mean and median comparison.
2. Stepwise reduction of variables resulted in 9 number of variables.
3. Principal component analysis after stepwise reduction did not show any significant improvement.
4. Box’s M test revealed that the covariance matrices are unequal.
5. Discriminant based classification using quadratic rule resulted in 36% accuracy and for the new observations most of them are classified into fire cause ‘4’.
6. Cluster analysis using Ward’s method resulted in 3 number of clusters and complete linkage resulted in 6 number of clusters.

Acknowledgements

We would like to acknowledge Dr. Megan Orr for the valuable suggestions and discussions provided through this course and project.

References

[1] Wildland Fire | US Forest Service. 2017.
[2] Facts + Statistics: Wildfires | III. Â© 2015 Munich Re. NatCatSERVICE; National Interagency Fire Center. As of June 2015.; Name, Location.
[3] Forest Service Wildland Fire Suppression Costs Exceed $2 Billion | USDA. 2017.
[4] U.S. fire statistics. 2017.
[5] Ganti V, Sarma AD. Data cleaning: A practical perspective. Synthesis Lectures on Data Management. 2013;5:1-85.
[6] MathWorks I. MATLAB: the language of technical computing. Desktop tools and development environment, version 7: MathWorks; 2005.
[7] Rencher AC. Methods of multivariate analysis: John Wiley & Sons; 2003.
[8] Krishnaswamy K, Sivakumar AI, Mathirajan M. Management research methodology: integration of principles, methods and techniques: Pearson Education India; 2009.
[9] Institute S. SAS/STAT user's guide: version 6: Sas Inst; 1990.
[10] Izenman AJ. Modern multivariate statistical techniques. Regression, classification and manifold learning. 2008.
[11] Team R. RStudio: integrated development for R. RStudio, Inc, Boston, MA URL http://www rstudio com. 2015.
[12] Beebe LH, Gossler SM. Evaluating Research Articles from Start to Finish. Issues in Mental Health Nursing. 2011;32:792-.

Table 1. Descriptive Statistics for Overall Data

	Acres Burned	Air Temperature	Content Loss	Elevation	Fire Fighter Deaths	Fire Fighter Injuries	Flame Length
Minimum	0	0	0	0	0	0	0
1st Quartile	0	65	0	300	0	0	1
Median	0	75	0.00	1100	0	0	2
Mean	81.06	74	3664	1315	0	0.0013	2.22
3rd Quartile	1	86	0	1700	0	0	2
Maximum	70868	119	8000000	32400	0	1	80

	Other Deaths	Other Injuries	Property Loss	Relative Humidity	Spread Rate	Wind Speed
Minimum	0	0	0	0	0	0
1st Quartile	0	0	0	23	0	1
Median	0	0	0.00	34	2	3
Mean	0.002	0.0011	6217	38.01	4.314	4.423
3rd Quartile	0	0	0	50	4	5
Maximum	2	1	1500000	100	800	147

Table 2. Covariance matrix for Overall Data

	Acres Burned	Air Temperature	Content Loss	Elevation	Fire Fighter Deaths	Fire Fighter Injuries	Flame Length
Acres Burned	1384678	825.0659	7304867	-1161.65	0	-0.0906	1087.65
Air Temperature	825.066	305.8444	-18139.9	-166.761	0	-0.00374	10.7386
Content Loss	7304867	-18139.9	1.56E+10	728521.2	0	-4.62538	6892.14
Elevation	-1161.65	-166.761	728521.2	1926403	0	-0.7787	-11.342
Fire Fighter Deaths	0	0	0	0	0	0	0
Fire Fighter Injuries	-0.0906	-0.00374	-4.62538	-0.7787	0	0.001333	0.00238
Flame Length	1087.65	10.73859	6892.144	-11.3419	0	0.002377	14.9418
Other Deaths	1.52886	-0.00081	15.97915	0.428683	0	-2.70E-06	0.00610
Other Injuries	-0.08617	0.001598	22.81505	0.89074	0	-1.40E-06	-0.0008
Property Loss	8729544	74728.36	2.39E+09	64289.16	0	1.586427	64601.3
Relative Humidity	-1683.57	-117.106	-77966.4	318.8906	0	-0.01069	-10.438
Spread Rate	8336.01	30.73465	5261.62	-508.92	0	0.021076	21.2253
Wind Speed	500.007	11.37391	13304.46	-188.341	0	-0.0003	4.09126

	Other Deaths	Other Injuries	Property Loss	Relative Humidity	Spread Rate	Wind Speed
Acres Burned	1.52886	-0.08617	8729544	-1683.57	8336.011	500.0072
Air Temperature	-0.00081	0.00159	74728.36	-117.106	30.73465	11.37391
Content Loss	15.97915	22.81505	2.39E+09	-77966.4	5261.62	13304.46
Elevation	0.42868	0.89074	64289.16	318.8906	-508.92	-188.341
Fire Fighter Deaths	0	0	0	0	0	0
Fire Fighter Injuries	-0.00000	-0.00000	1.58642	-0.01069	0.02107	-0.0003
Flame Length	0.00610	-0.00077	64601.28	-10.438	21.22533	4.09126
Other Deaths	0.00226	0.00026	217.8636	0.00479	0.03181	-3.38E-03
Other Injuries	0.00026	0.00106	20.19571	0.00733	-0.00194	-1.80E-04
Property Loss	217.8636	20.19571	3.61E+09	-115475	53860.64	38464.77
Relative Humidity	0.00479	0.00733	-115475	451.0659	-30.1286	-17.3625
Spread Rate	0.03181	-0.00194	53860.64	-30.1286	325.6823	11.60299
Wind Speed	-0.00338	-0.00018	38464.77	-17.3625	11.60299	26.19276

Table 3. Correlation matrix for Overall Data

	Acres Burned	Air Temperature	Content Loss	Elevation	Fire Fighter Deaths	Fire Fighter Injuries	Flame Length
Acres Burned	1	0.04009	0.04962	-0.00071	N/A	-0.00210	0.23911
Air Temperature	0.04009	1	-0.00829	-0.00687	N/A	-0.00586	0.15885
Content Loss	0.04962	-0.00829	1.00E+00	0.00419	N/A	-0.00101	0.01425
Elevation	-0.00071	-0.00687	0.00419	1	N/A	-0.01536	-0.00211
Fire Fighter Deaths	N/A	N/A	N/A	N/A	1	N/A	N/A
Fire Fighter Injuries	-0.00210	-0.00586	-0.00101	-0.01536	N/A	1	0.01683
Flame Length	0.23911	0.15885	0.01425	-0.00211	N/A	0.01683	1
Other Deaths	0.02729	-0.00097	0.00268	0.00648	N/A	-1.54E-03	0.03316
Other Injuries	-0.00224	0.00279	0.00558	0.01964	N/A	-1.20E-03	-0.00608
Property Loss	0.12344	0.07110	3.18E-01	0.00077	N/A	0.00072	0.27810
Relative Humidity	-0.06736	-0.31528	-0.02934	0.01081	N/A	-0.01378	-0.12714
Spread Rate	0.39254	0.09738	0.00233	-0.02031	N/A	0.03198	0.30426
Wind Speed	0.08302	0.12707	0.02078	-0.02651	N/A	-0.00158	0.20680

	Other Deaths	Other Injuries	Property Loss	Relative Humidity	Spread Rate	Wind Speed
Acres Burned	0.02729	-0.00224	0.12344	-0.06736	0.39254	0.08302
Air Temperature	-0.00097	0.00279	0.07110	-0.31528	0.09738	0.12707
Content Loss	0.00268	0.00558	0.31816	-2.93E-02	0.00233	0.02078
Elevation	0.00648	0.01964	0.00077	0.01081	-0.02031	-0.02651
Fire Fighter Deaths	N/A	N/A	N/A	N/A	N/A	N/A
Fire Fighter Injuries	-0.00153	-0.00119	0.00072	-0.01378	0.03198	-0.00158
Flame Length	0.03316	-0.00608	0.27810	-0.12714	0.30426	0.20680
Other Deaths	1	0.17036	0.07616	0.00474	0.03703	-0.01388
Other Injuries	0.17036	1	0.01028	0.01057	-0.00328	-0.00110
Property Loss	0.07616	0.01028	1	-9.05E-02	0.04966	0.12506
Relative Humidity	0.00474	0.01057	-0.09047	1	-0.07860	-0.15973
Spread Rate	0.03703	-0.00328	0.04966	-0.07860	1	0.12562
Wind Speed	-0.01388	-0.00110	0.12506	-0.15973	0.12562	1

Table 4. Stepwise Selection Summary for Overall Data

Stepwise Selection Summary										
Step	**Number In**	**Entered**	**Removed**	**Partial R-Square**	**F Value**	**Pr > F**	**Wilks' Lambda**	**Pr < Lambda**	**Average Squared Canonical Correlation**	**Pr > ASCC**
1	1	RELHUMID		0.0717	69.8	<.0001	0.9283	<.0001	0.0089	<.0001
2	2	ELEV		0.0593	57	<.0001	0.8733	<.0001	0.0163	<.0001
3	3	AIRTEMP		0.0578	55.42	<.0001	0.8229	<.0001	0.0233	<.0001
4	4	PROPLOSS		0.033	30.86	<.0001	0.7957	<.0001	0.0272	<.0001
5	5	SPREADR		0.0195	18.02	<.0001	0.7801	<.0001	0.0294	<.0001
6	6	ACREB		0.0149	13.67	<.0001	0.7685	<.0001	0.0312	<.0001
7	7	CONTLOSS		0.0093	8.52	<.0001	0.7613	<.0001	0.0323	<.0001
8	8	WINDS		0.0093	8.44	<.0001	0.7543	<.0001	0.0335	<.0001
9	9	FLAMEL		0.0072	6.55	<.0001	0.7489	<.0001	0.0343	<.0001

Table 5. Eigen values of the Correlation Matrix generated by Principal Component Analysis

Eigen values of the Correlation Matrix				
	Eigen value	**Difference**	**Proportion**	**Cumulative**
1	2.5738	1.2349	0.286	0.286
2	1.3389	0.2359	0.1488	0.4348
3	1.1030	0.0745	0.1226	0.5573
4	1.0284	0.2078	0.1143	0.6716
5	0.8206	0.1353	0.0912	0.7628
6	0.6852	0.1385	7.61E-02	0.8389
7	0.5467	0.0583	0.0608	0.8997
8	0.4884	0.0739	0.0543	0.9539
9	0.4145		0.0461	1

Table 6. Eigen vectors of the Correlation matrix generated by Principal Component Analysis

Eigenvectors									
	Prin1	**Prin2**	**Prin3**	**Prin4**	**Prin5**	**Prin6**	**Prin7**	**Prin8**	**Prin9**
Acres Burned	0.4268	-0.1411	-0.0416	-0.2337	0.1304	-0.6472	0.0546	0.4957	0.2382
Air Temperature	0.3114	0.4443	-0.0918	0.4390	-0.1012	0.0137	0.7021	0.0108	0.0289
Content Loss	0.3418	-0.4995	0.1081	0.3314	-0.0581	0.2912	-0.0659	-0.1688	0.6253
Elevation	0.1099	0.3831	0.6201	0.4319	0.0909	-2.16E-01	-0.4622	0.0257	-0.0229
Flame Length	0.4282	0.0624	0.1943	-0.4185	-0.0477	-1.98E-01	0.0701	-0.7372	-0.0999
Property Loss	0.4070	-0.4513	7.18E-02	0.2169	-0.0312	0.1308	-0.0005	0.1599	-0.7311
Relative Humidity	-0.2474	-0.2055	0.6389	-0.1809	0.4337	0.1156	0.4944	0.0952	0.0208
Spread Rate	0.3101	0.2934	0.2355	-0.4461	-0.3510	0.5298	-0.0834	0.3827	0.0757
Wind Speed	0.2924	0.2244	-0.2947	-0.0453	0.8040	0.3145	-0.1719	-0.0159	0.0003

Table 7. Chi Square Approximation for Test of Equivalence of Covariance Matrices (M-test)

CHI-SQUARE APPROXIMATION			
c1	**u**	**X^2_{crit}**	**X^2p-val**
0.0119	9806.5761	405.2435	0

Table 8. Error rate for Quadratic Classification Function

Quadratic Classification		
Variables included	**Resubstitution Accuracy (%)**	**Cross Validation Accuracy (%)**
9	72.19	72.82
8	72.78	73.25
7	73.28	73.67
6	71.00	71.43
5	63.64	64.01
4	63.70	63.96
3	64.92	65.02
2	66.25	66.35

Table 9. Chi Square Approximation for Test of Equivalence of Covariance Matrices (M-test)

Number of Variables	CHI-SQUARE APPROXIMATION		
	c1	u	X^2_{crit}
9	0.0119	9806.5761	405.2435
8	0.0106	9515.1357	328.5804
7	0.0093	9308.4814	259.9144
6	0.008	6689.8343	199.2442
5	0.0067	4148.8089	146.5674
4	0.0054	3797.0247	101.8795
3	0.0041	1994.3612	65.1708
2	0.0027	945.6169	36.415

Table 10. Cross Validation Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into FIREC										
From FIREC	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Total**
0	1700	64	14	0	70	515	5	38	14	2420
	70.25	2.64	0.58	0	2.89	21.28	0.21	1.57	0.58	100
1	206	148	2	0	13	198	0	5	16	588
	35.03	25.17	0.34	0	2.21	33.67	0	0.85	2.72	100
2	737	5	9	0	20	99	0	8	41	919
	80.2	0.54	0.98	0	2.18	10.77	0	0.87	4.46	100
3	203	0	6	0	14	39	0	0	12	274
	74.09	0	2.19	0	5.11	14.23	0	0	4.38	100
4	161	18	3	0	69	151	0	13	26	441
	36.51	4.08	0.68	0	15.65	34.24	0	2.95	5.9	100
5	580	59	4	0	52	673	0	13	20	1401
	41.4	4.21	0.29	0	3.71	48.04	0	0.93	1.43	100
6	40	0	1	0	1	23	0	0	0	65
	61.54	0	1.54	0	1.54	35.38	0	0	0	100
7	438	34	5	0	77	272	0	20	22	868
	50.46	3.92	0.58	0	8.87	31.34	0	2.3	2.53	100
8	152	4	2	0	18	66	0	1	25	268
	56.72	1.49	0.75	0	6.72	24.63	0	0.37	9.33	100
Total	4217	332	46	0	334	2036	5	98	176	7244
	58.21	4.58	0.64	0	4.61	28.11	0.07	1.35	2.43	100
Priors	0.33	0.08	0.13	0.04	0.06	0.19	0.01	0.12	0.04	

Error Count Estimates for FIREC										
0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Total**	
Rate	0.2975	0.7483	0.9902	1	0.8435	0.5196	1	0.977	0.9067	0.6396
Priors	0.33	0.08	0.13	0.04	0.06	0.19	0.01	0.12	0.04	

Table 11. Classification of "Undetermined" Fire Cause by Quadratic Discriminant Analysis

Fire Cause	0	1	2	3	4	5	6	7	8
Classification	1	0	487	53	2589	3	0	15	412

Table 12. Optimal number of clusters

Method	'Ward'	'complete'
Kl (Krzanowski & Lai 1988)	3	8
Silhoutte (Rousseeuw 1987)	2	2
Cindex (Hubert & Levin 1976)	19	2
Ch (Calinski & HArabasz 1974)	3	6
Beale (Beale 1969)	3	2
Ratkowsky (Ratkowsky & Lance1978)	4	6
Hartigon (Hartigan 1975)	3	6
Ball (Ball & Hall 1965)	3	3
Scott (Scott & Symons 1971)	11	5
Rubin (Friedman & Rubin 1967)	3	6
Dunn (Dunn 1974)	2	2
Friedman (Friedman & Rubin 1967)	11	5

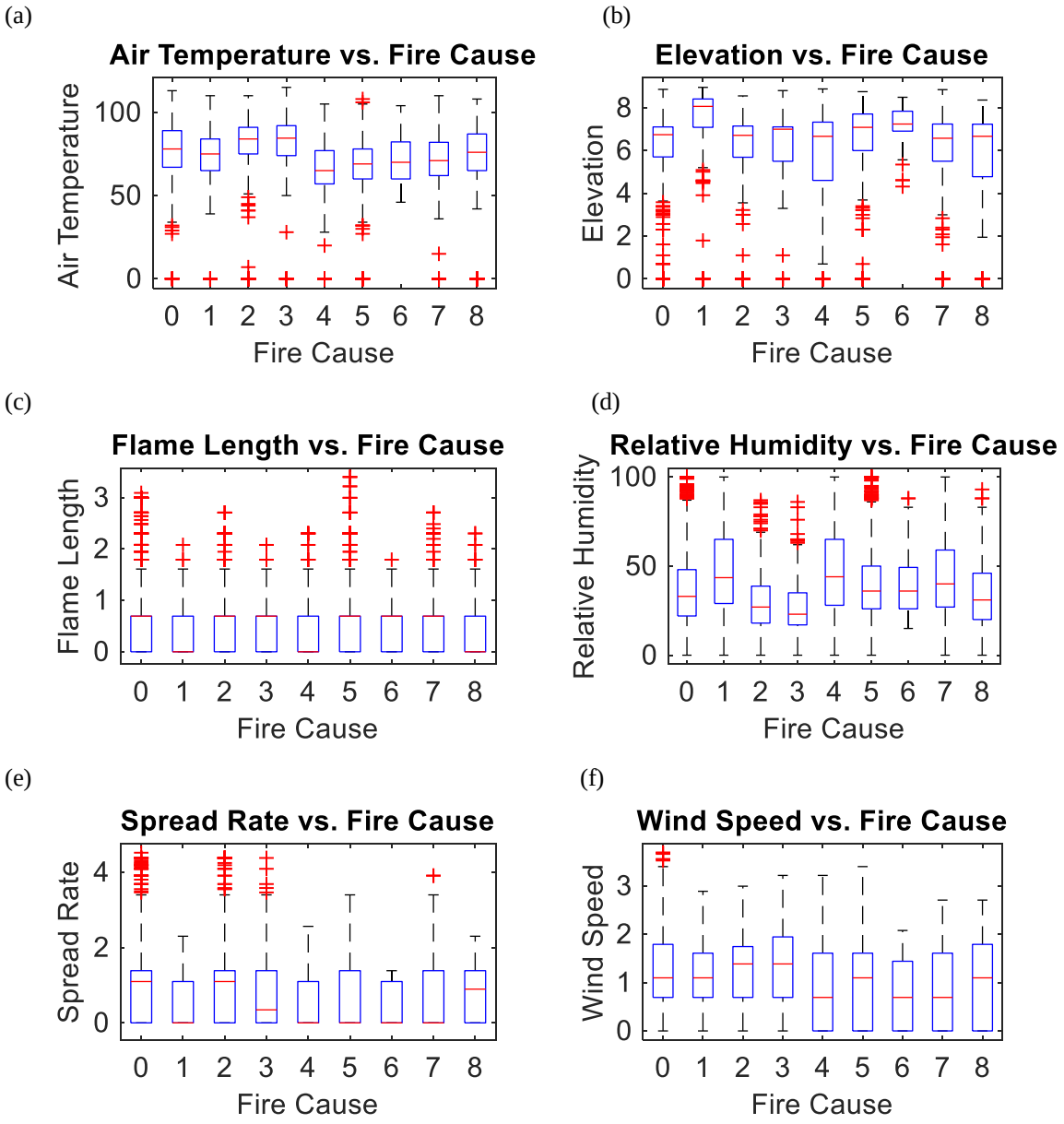

Figure 1. Box plots with respect to various fire causes for (a) air temperature, (b) elevation, (c) flame length, (d) relative humidity, (e) spread rate and (f) wind speed.

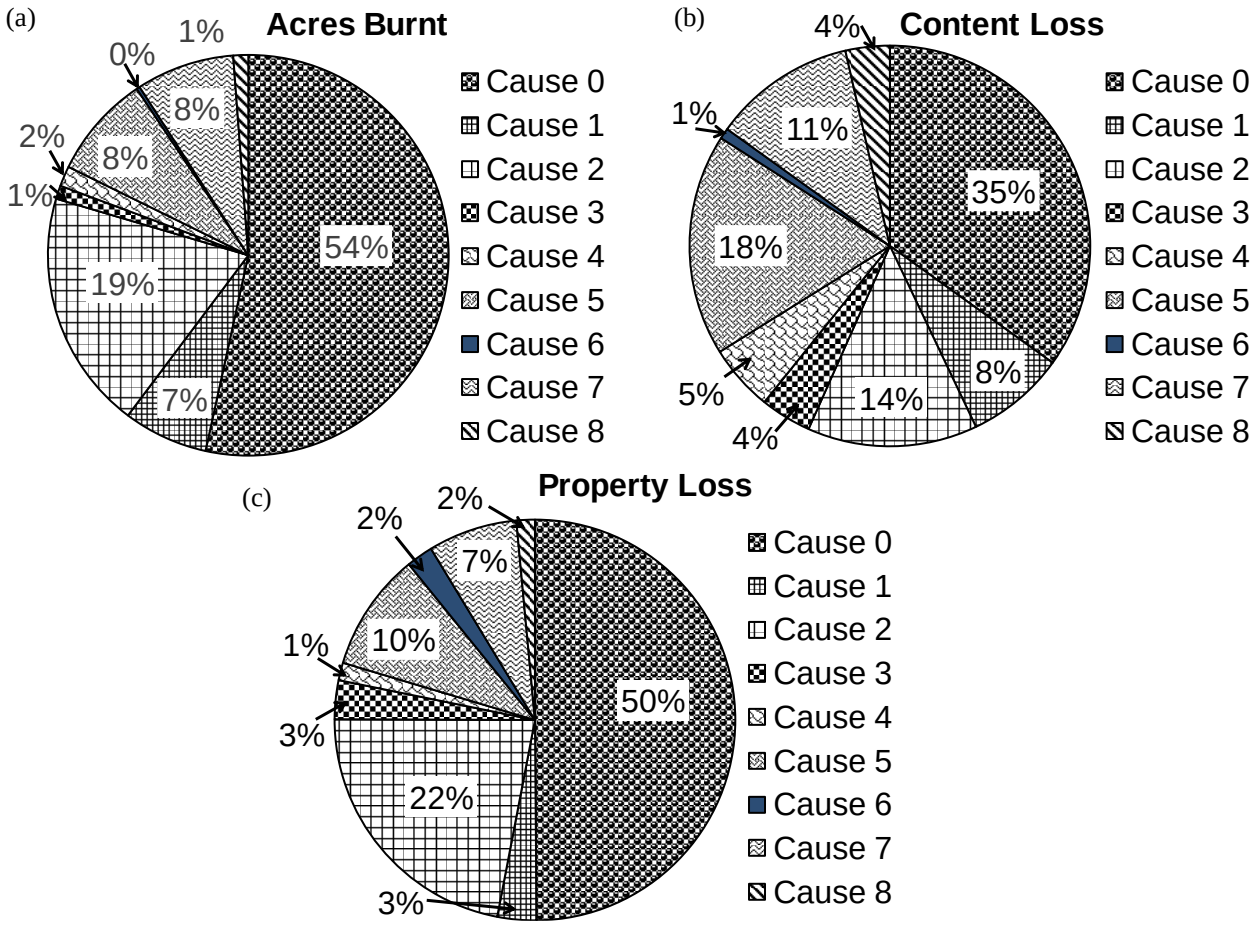

Figure2. Pie charts demonstrating the contribution of each fire cause towards (a) acres burnt (b) content loss and (c) property loss

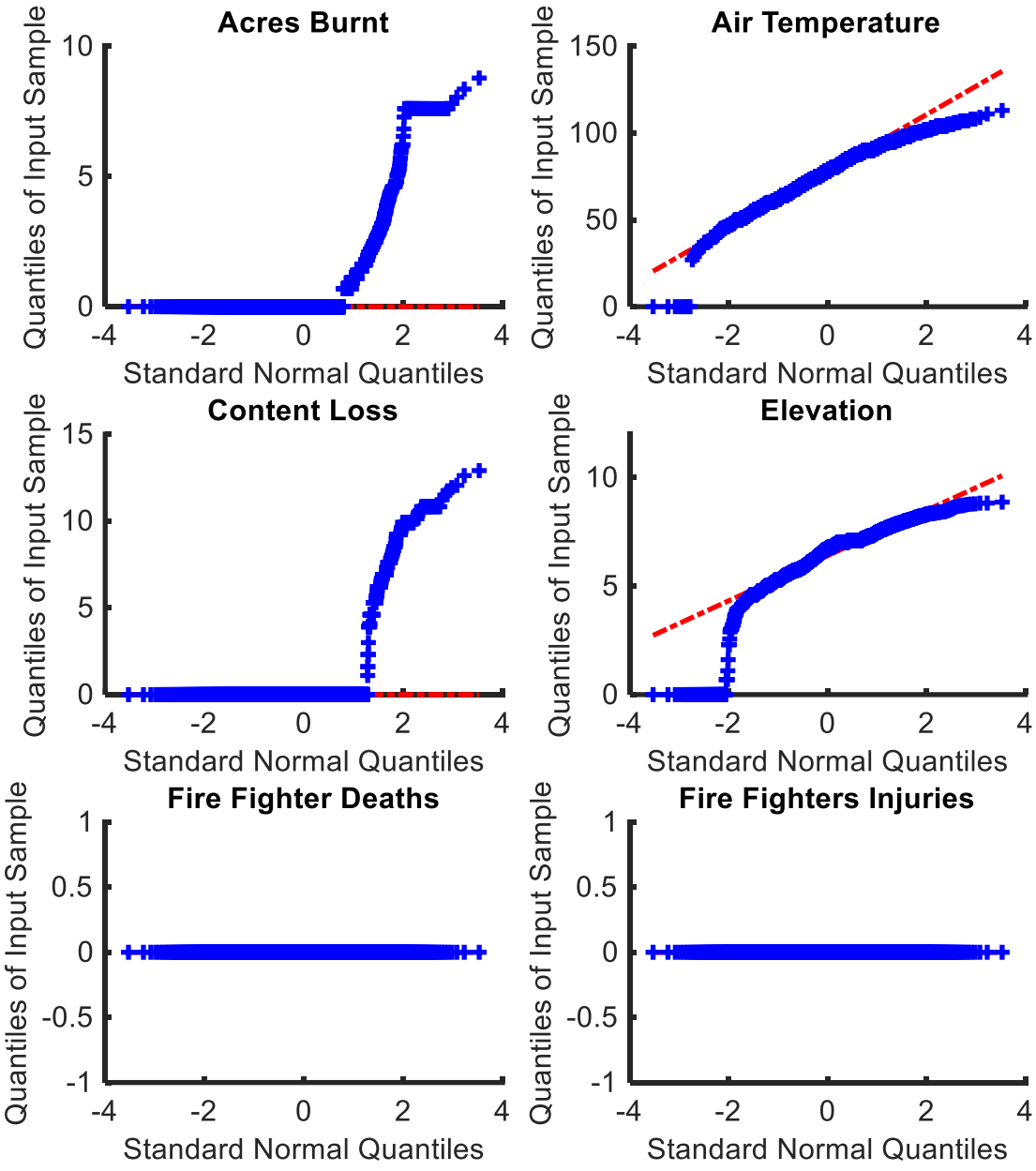

Figure 3. Q-Q plots of all variables corresponding to fire cause '0' (Contd.)

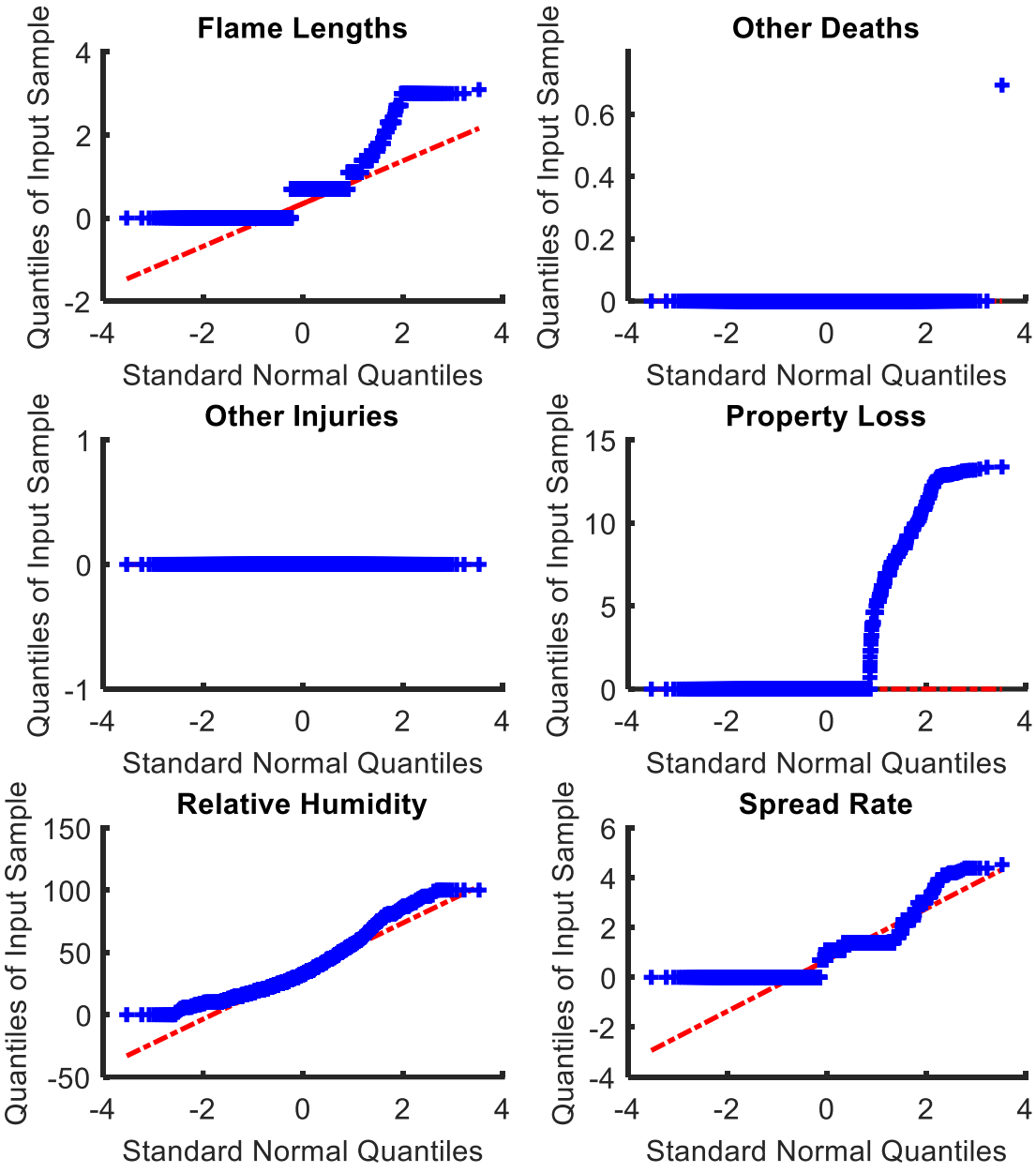

Figure 3. Q-Q plots of all variables corresponding to fire cause '0' (Contd.)

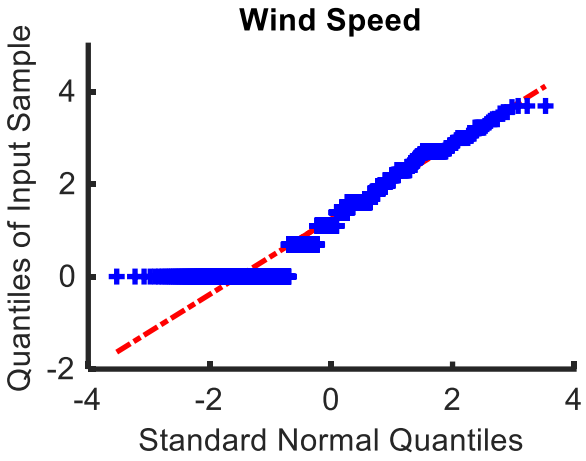

Figure 3. Q-Q plots of all variables corresponding to fire cause ‘0’.

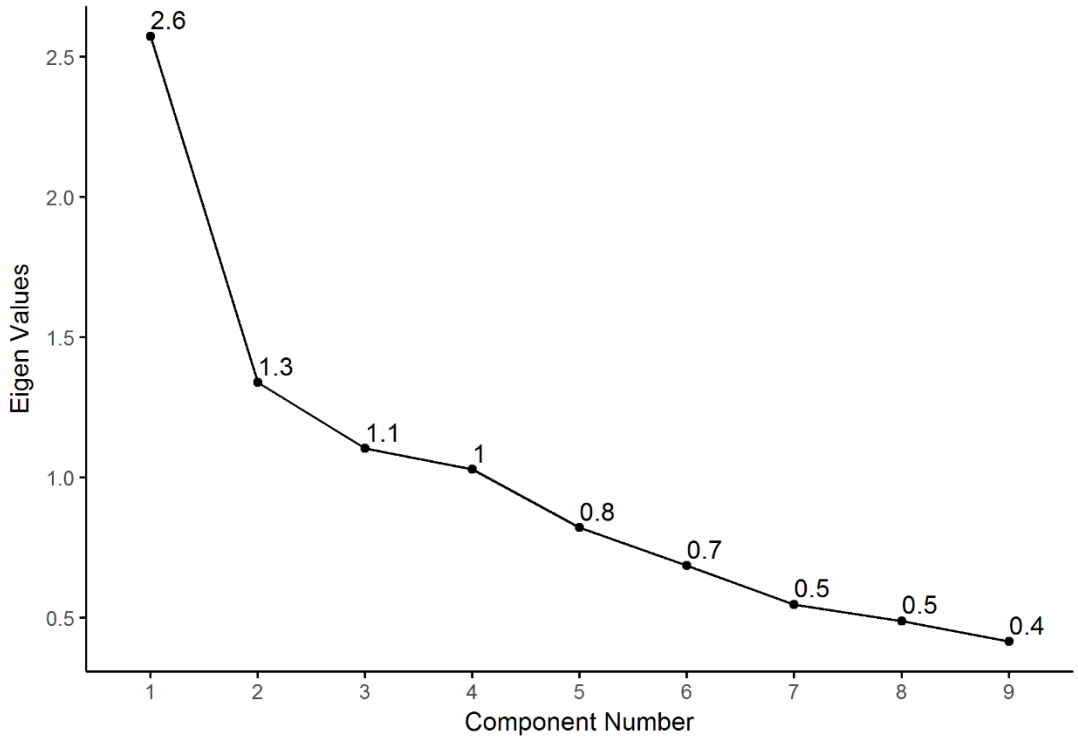

Figure 4. Scree plot for eigen values

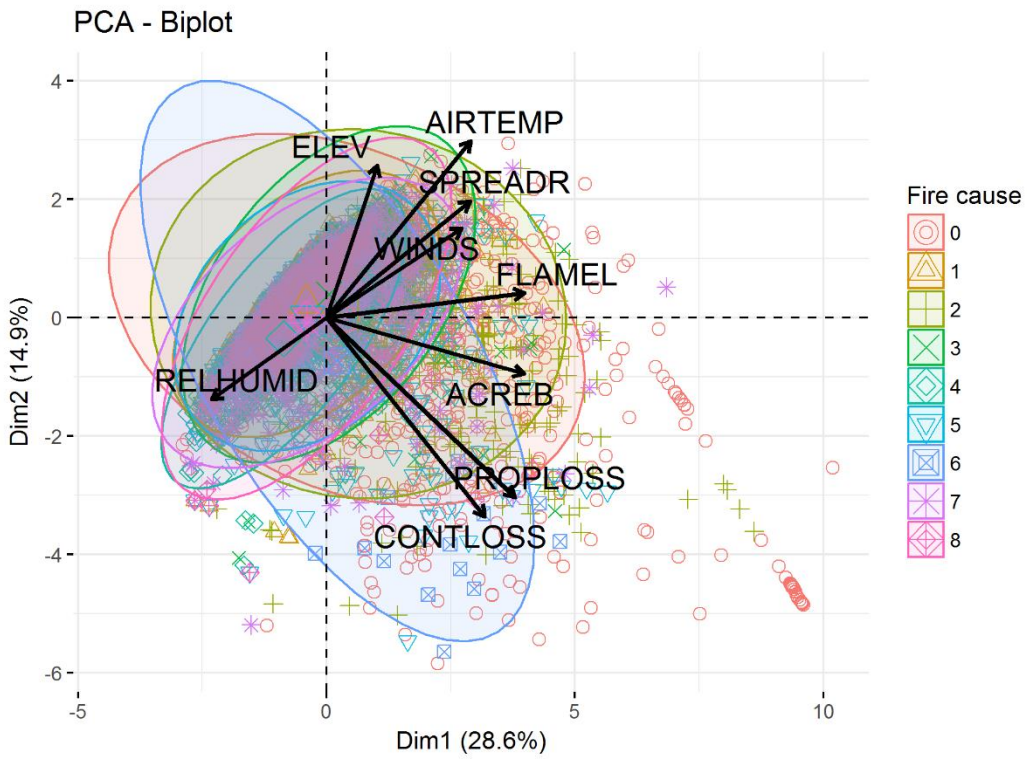

Figure 5. Biplot of the wildland fire accident data with respect to PC1 and PC2.

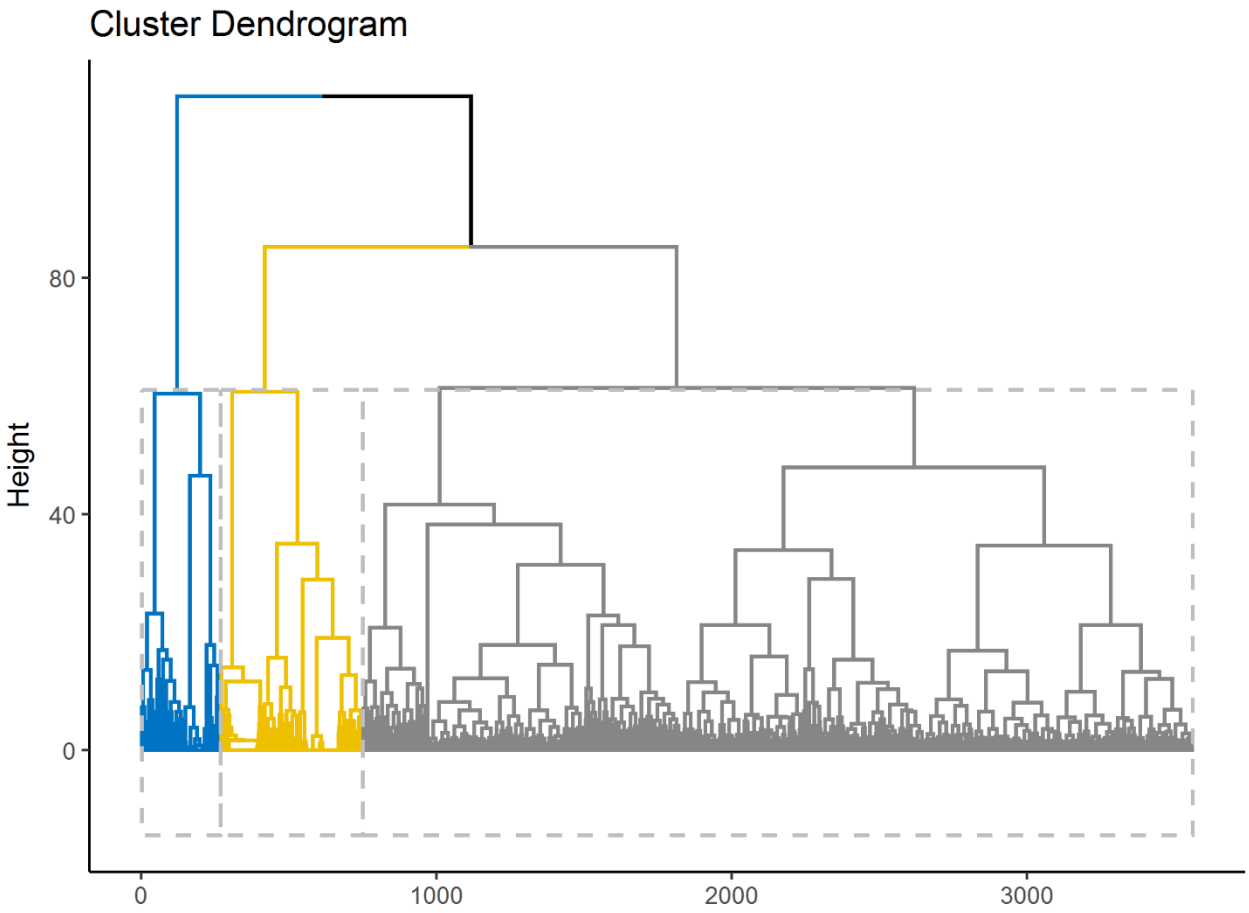

Figure 6. Dendrogram for Ward's method

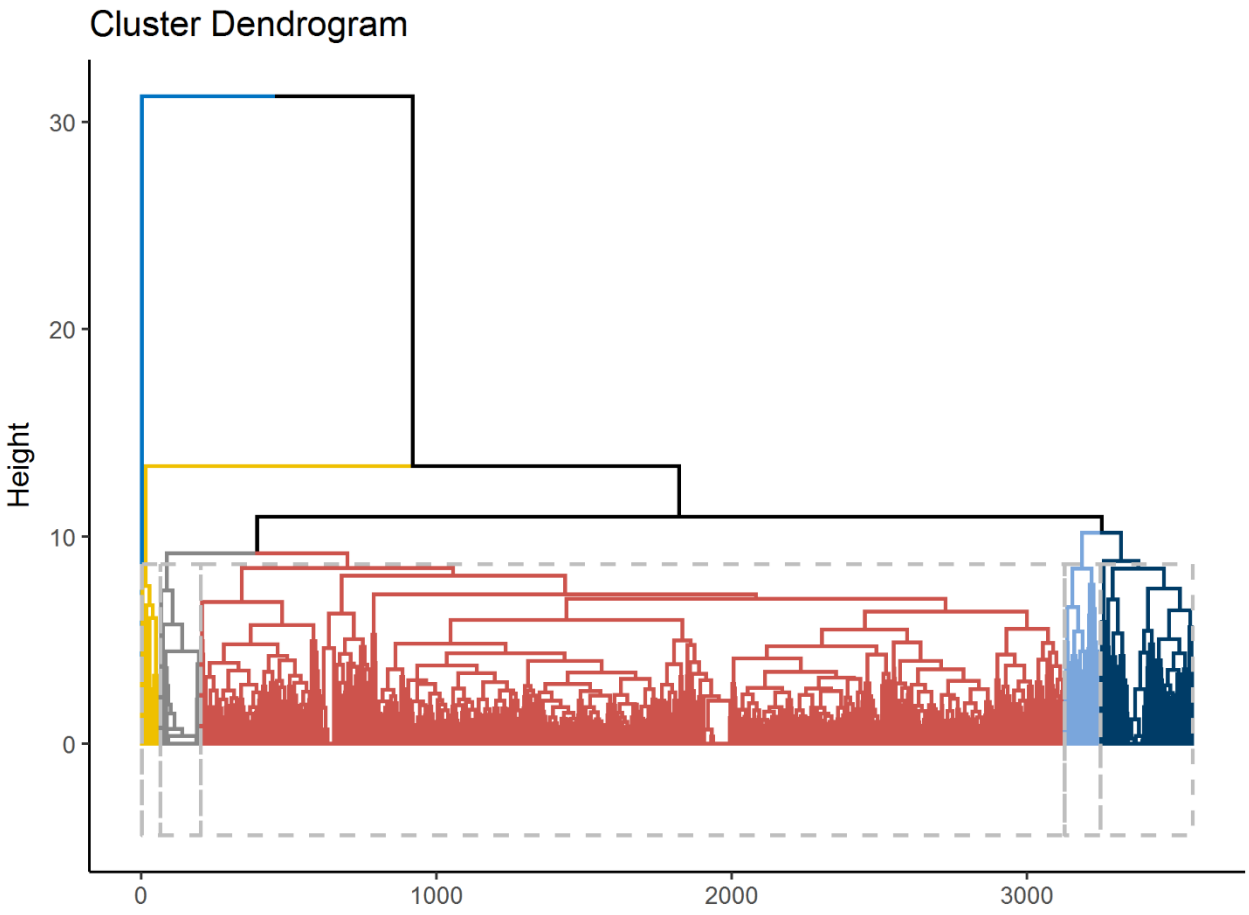

Figure 7. Dendrogram for complete linkage method

A.Appendix

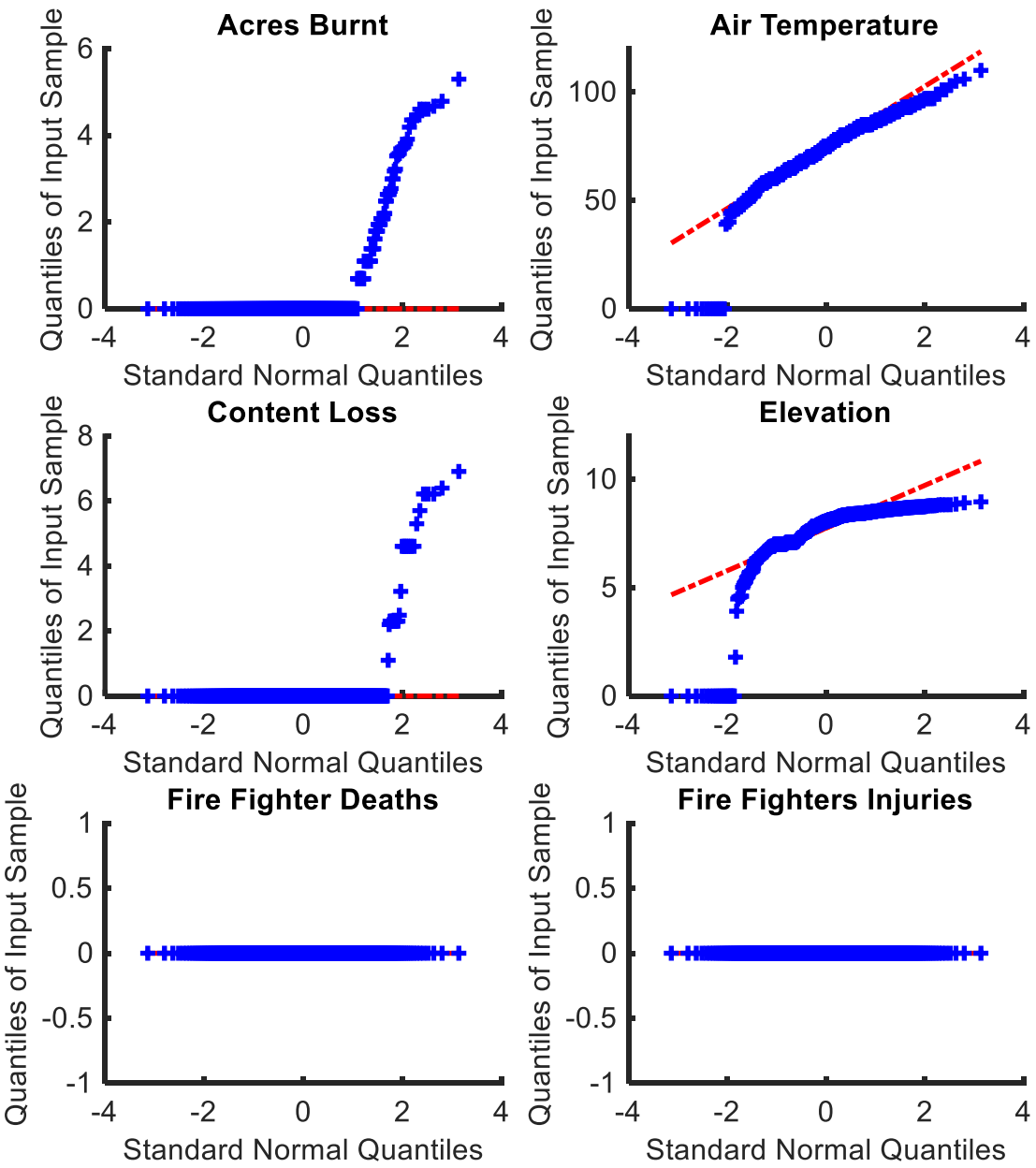

Figure A.1. Q-Q plots of all variables corresponding to fire cause '1' (Contd.)

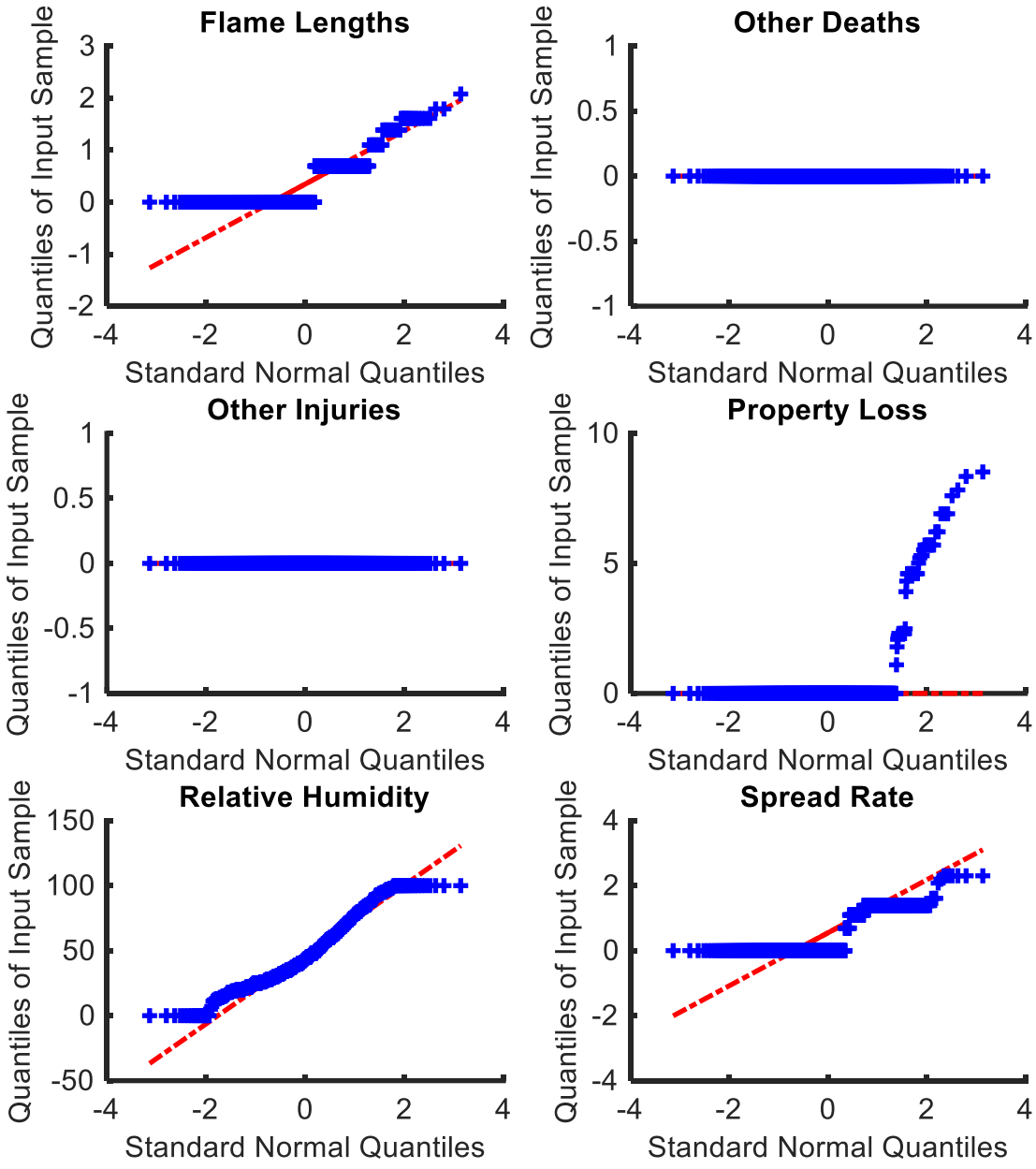

Figure A.1. Q-Q plots of all variables corresponding to fire cause '1' (Contd.)

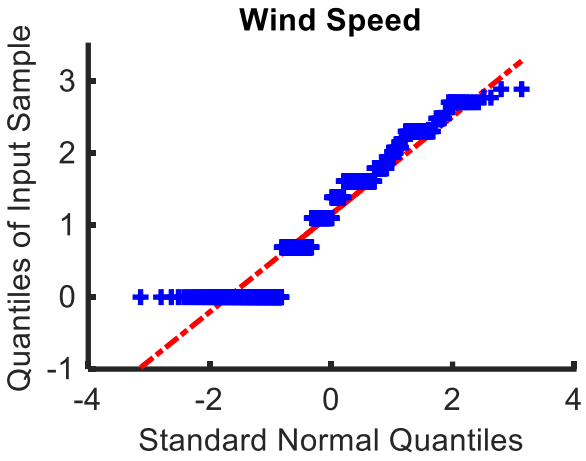

Figure A.1.Q-Q plots of all variables corresponding to fire cause ‘0’.

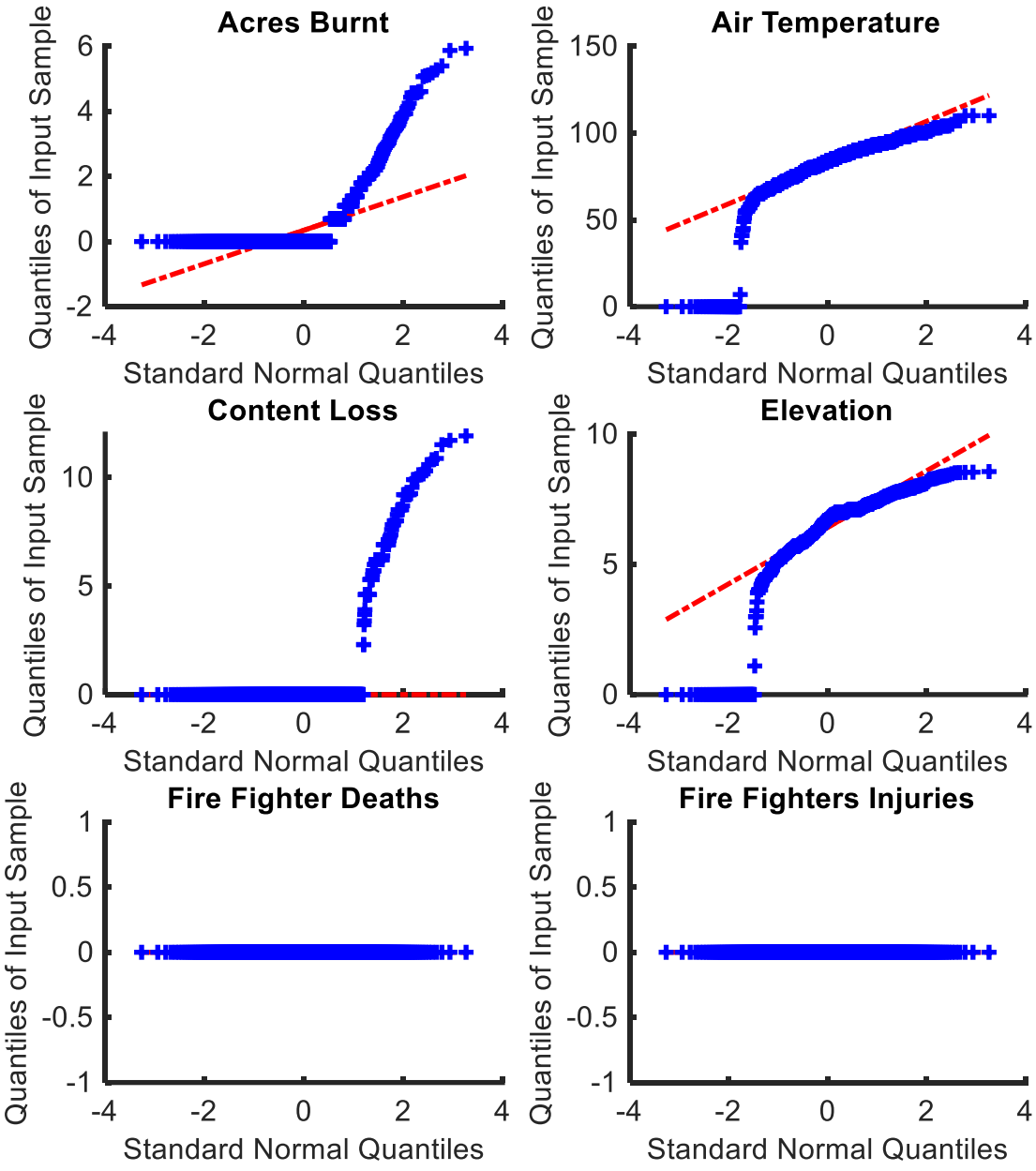

Figure A.2. Q-Q plots of all variables corresponding to fire cause '2' (Contd.)

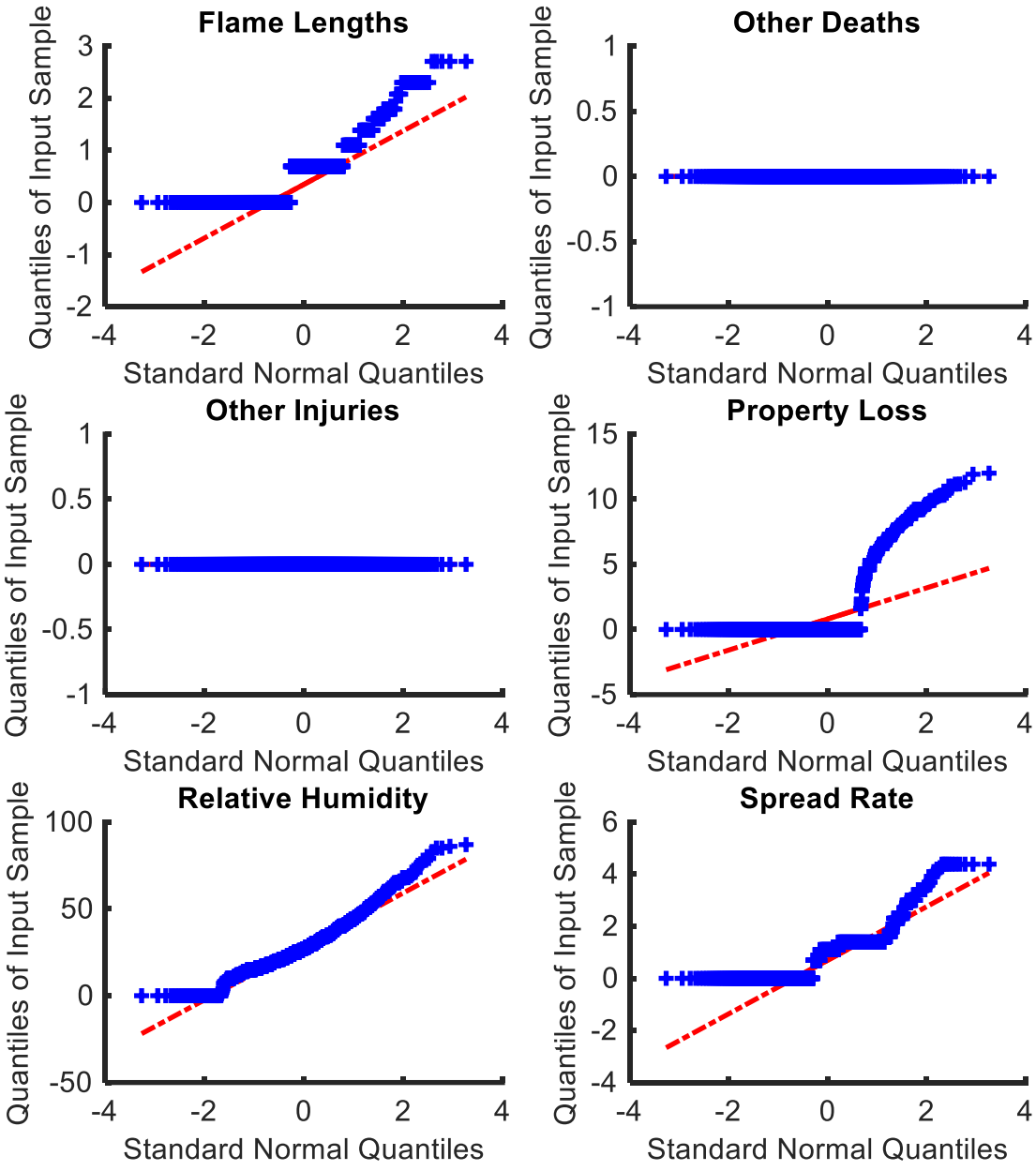

Figure A.2. Q-Q plots of all variables corresponding to fire cause '2' (Contd.)

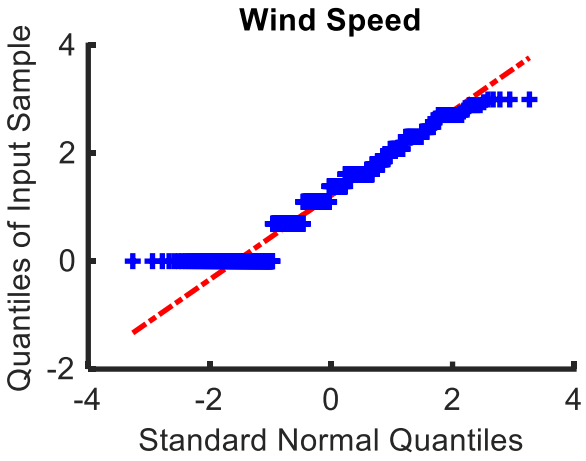

Figure A.2.Q-Q plots of all variables corresponding to fire cause ‘2’.

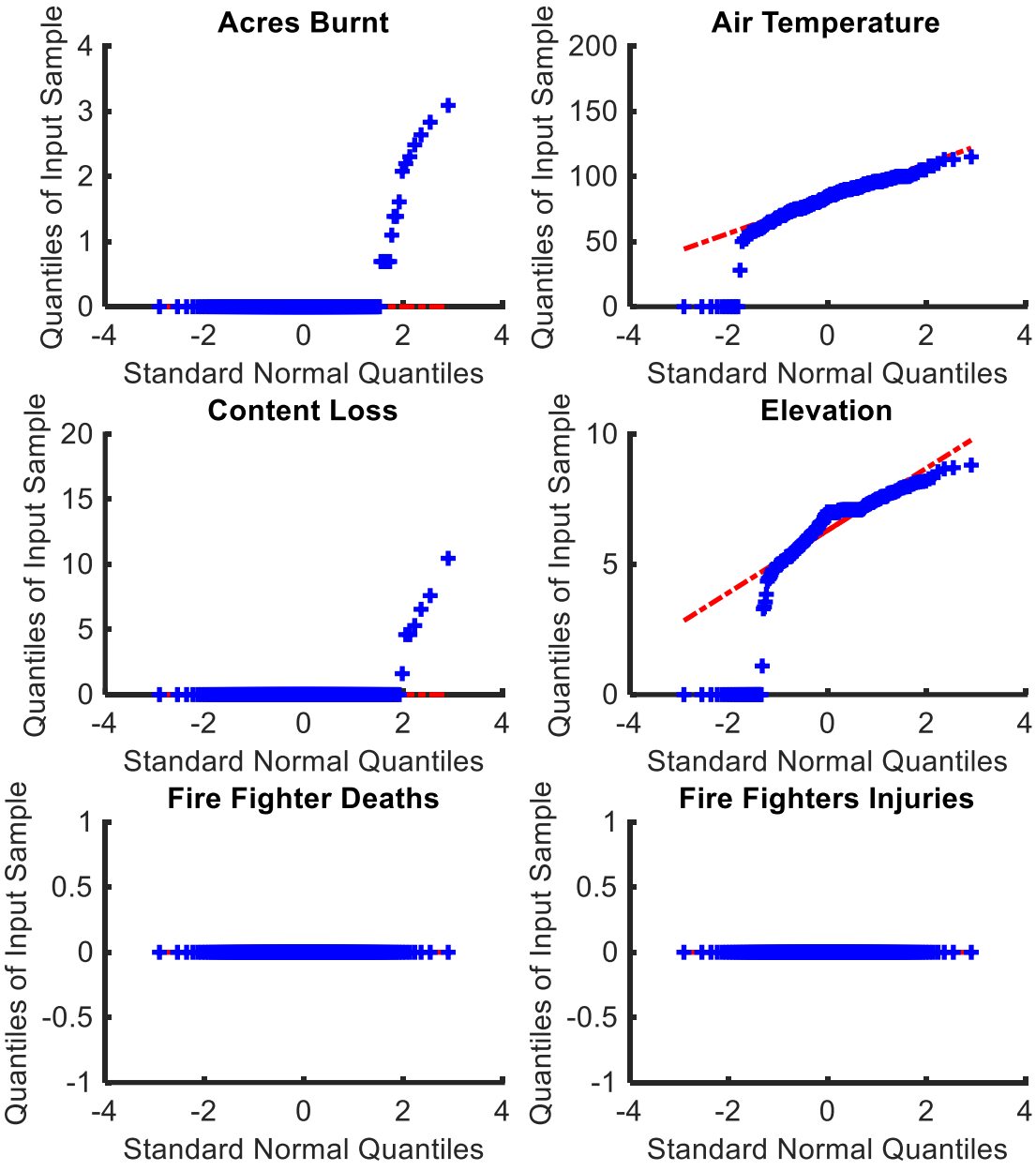

Figure A.3. Q-Q plots of all variables corresponding to fire cause '3' (Contd.)

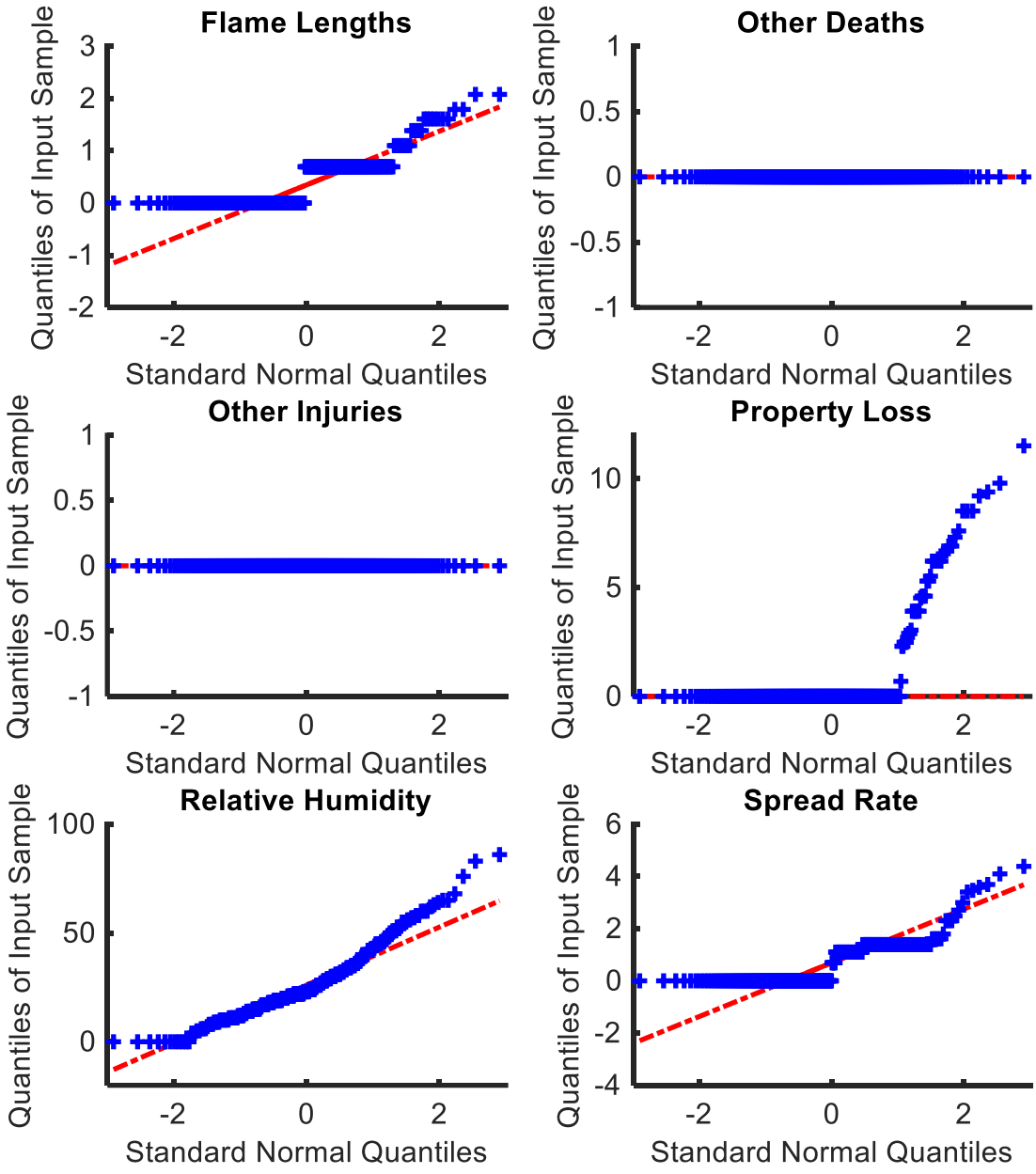

Figure A.3. Q-Q plots of all variables corresponding to fire cause '3' (Contd.)

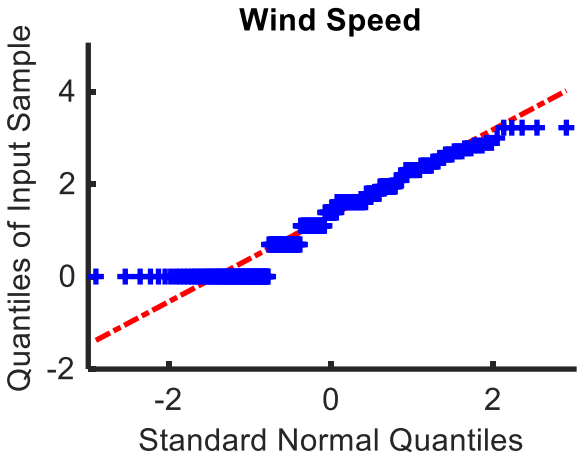

Figure A.3. Q-Q plots of all variables corresponding to fire cause ‘3’.

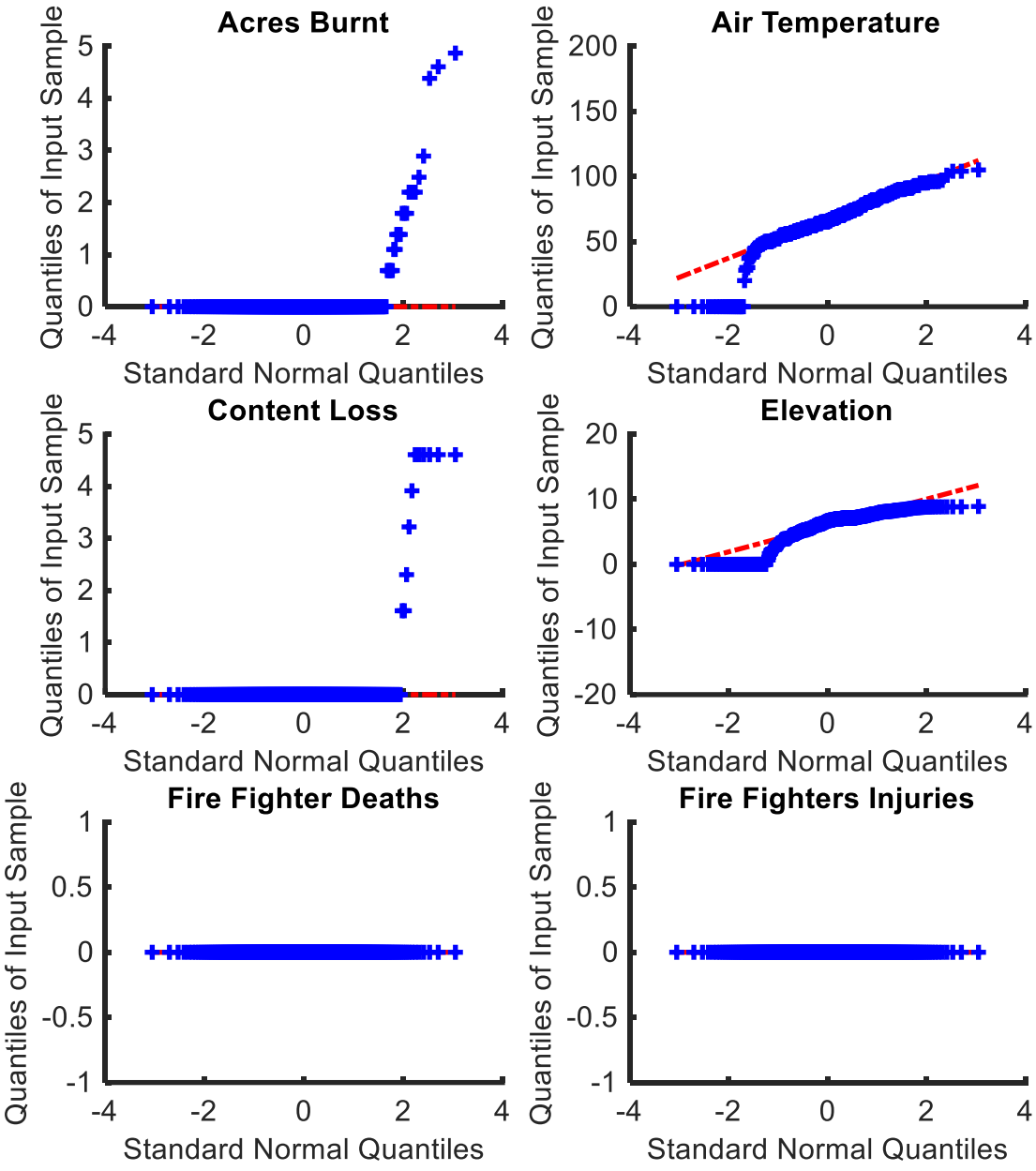

Figure A.4. Q-Q plots of all variables corresponding to fire cause ‘4’ (Contd.)

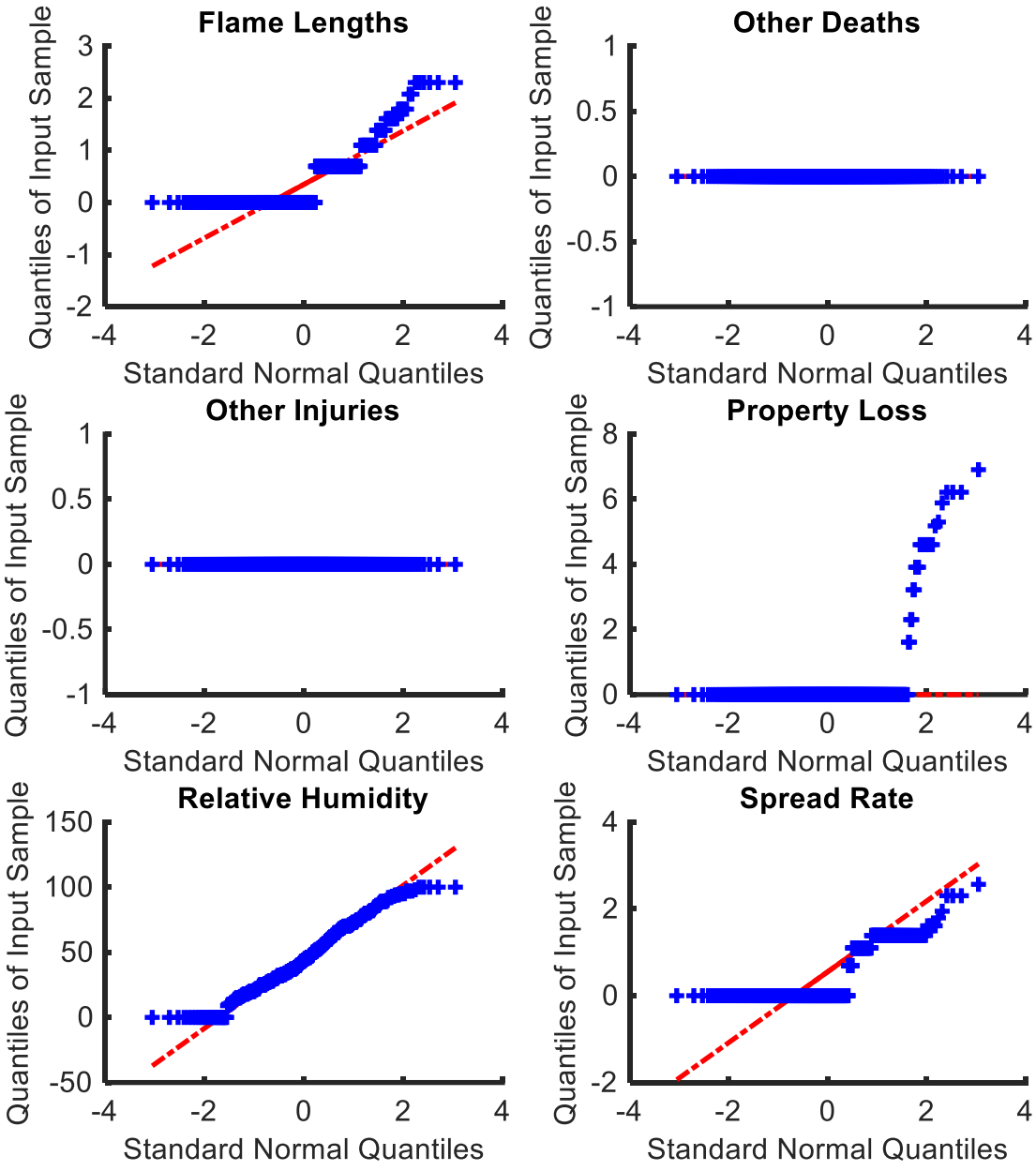

Figure A.4. Q-Q plots of all variables corresponding to fire cause ‘4’ (Contd.)

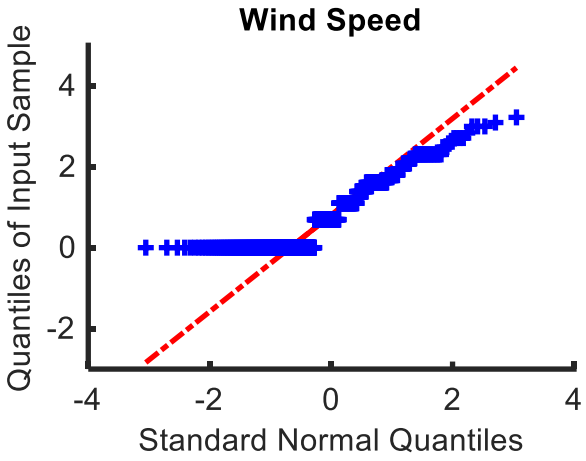

Figure A..4. Q-Q plots of all variables corresponding to fire cause ‘4’.

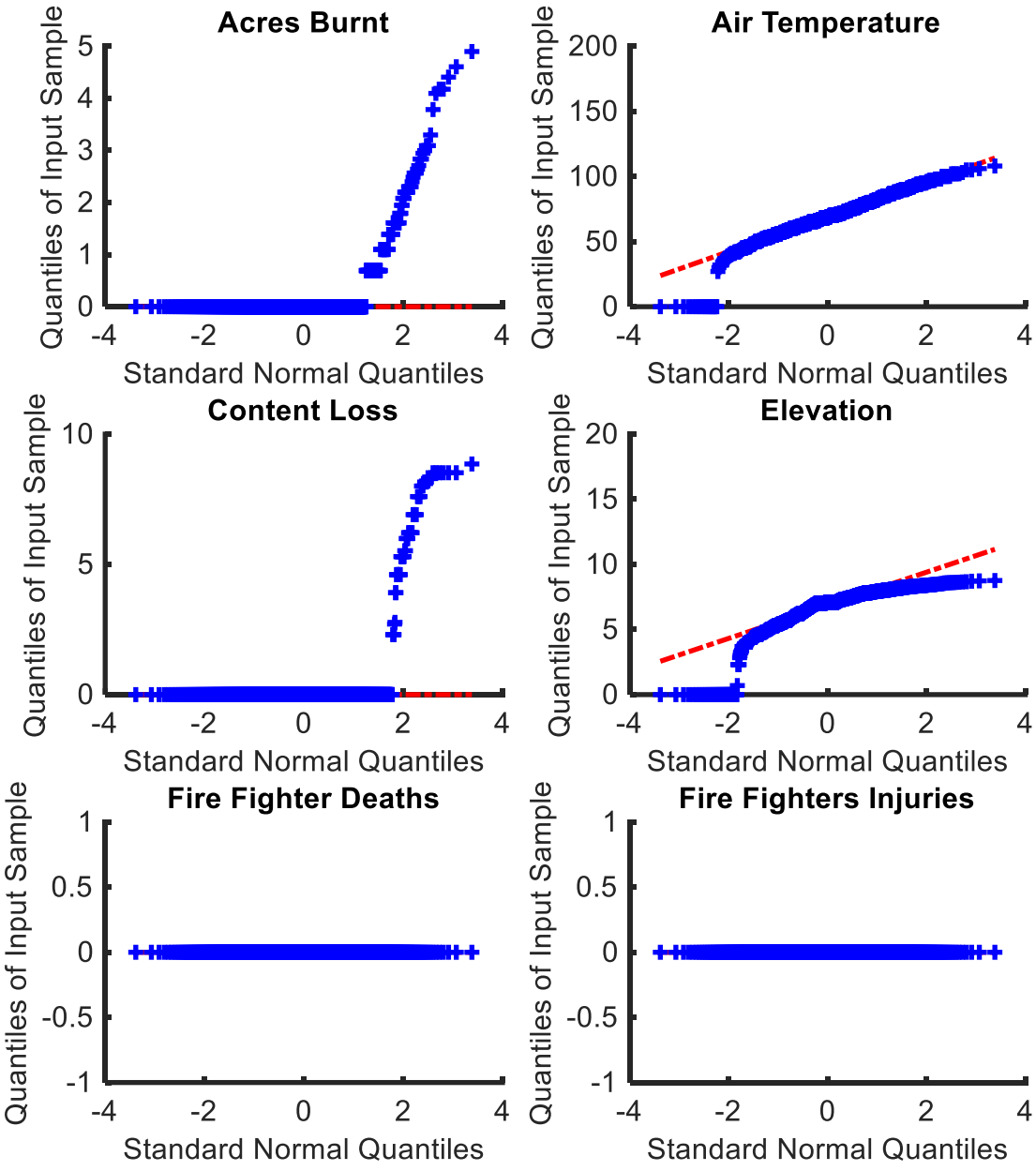

Figure A.5. Q-Q plots of all variables corresponding to fire cause ‘5’ (Contd.)

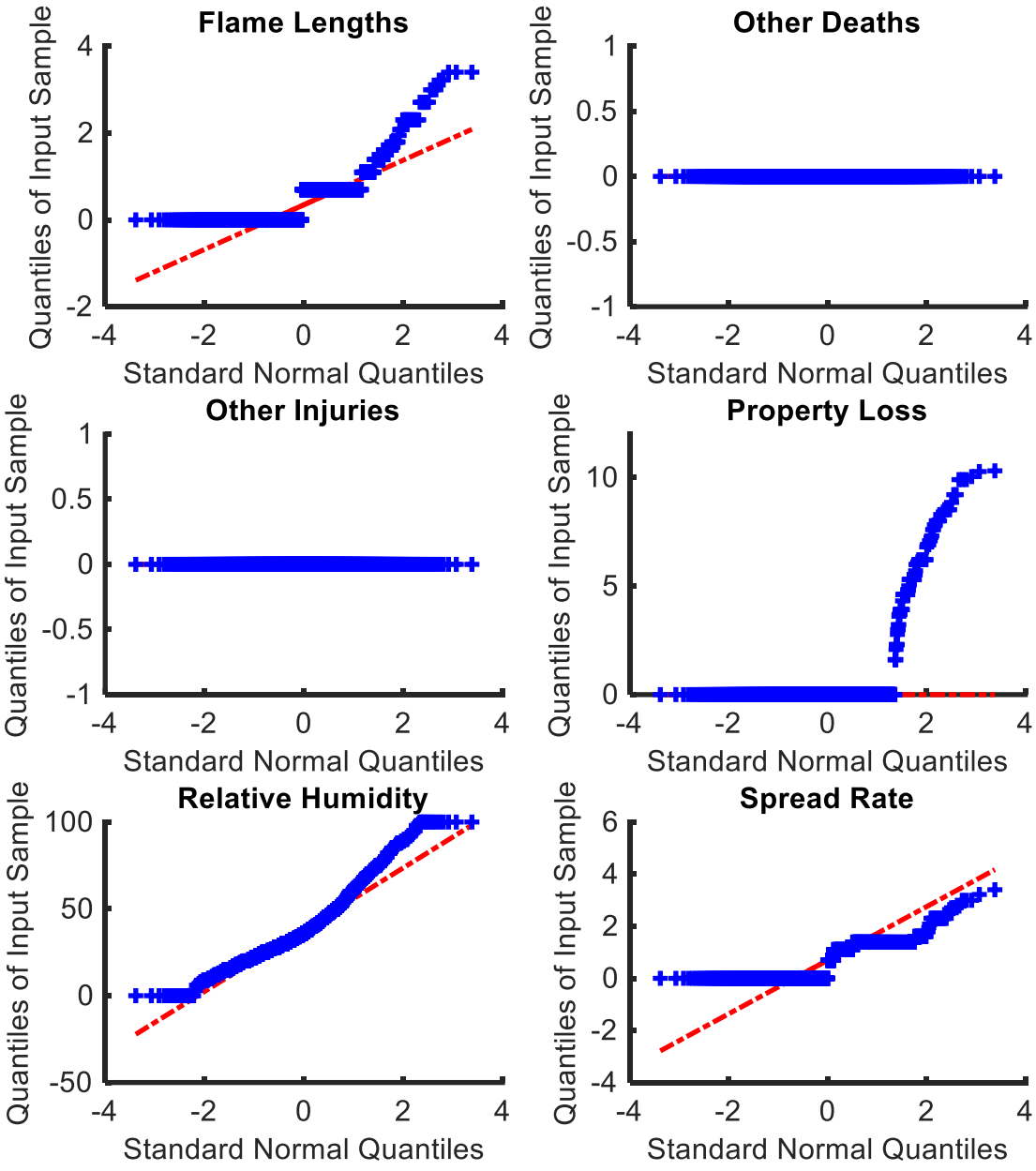

Figure A.5. Q-Q plots of all variables corresponding to fire cause ‘5’ (Contd.)

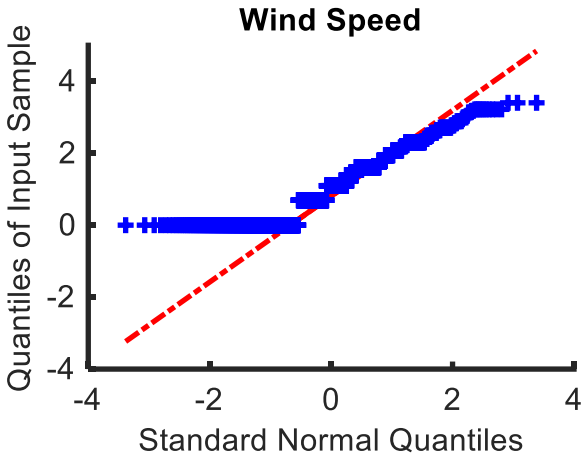

Figure A.5. Q-Q plots of all variables corresponding to fire cause ‘5’.

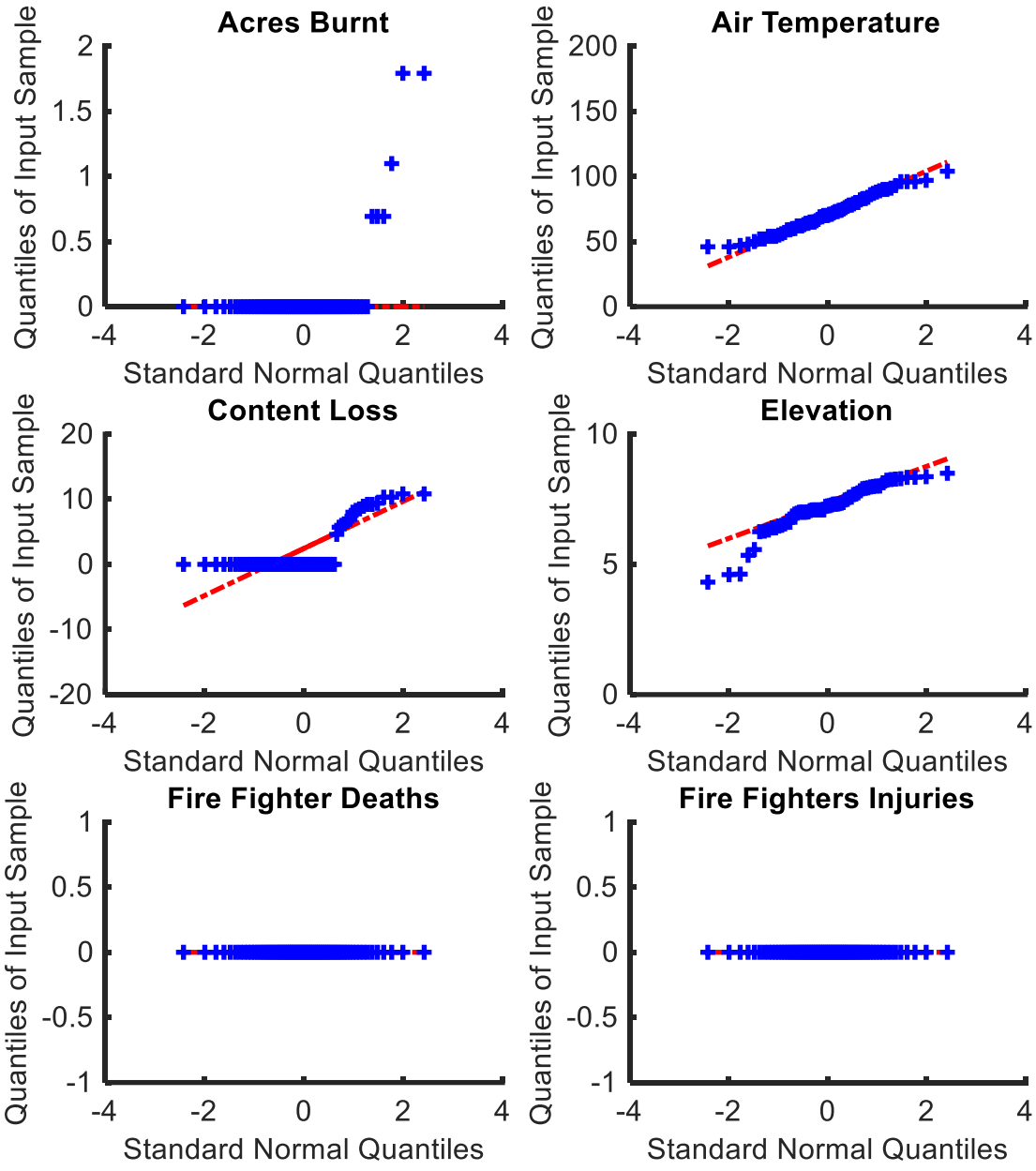

Figure A.6. Q-Q plots of all variables corresponding to fire cause ‘6’ (Contd.)

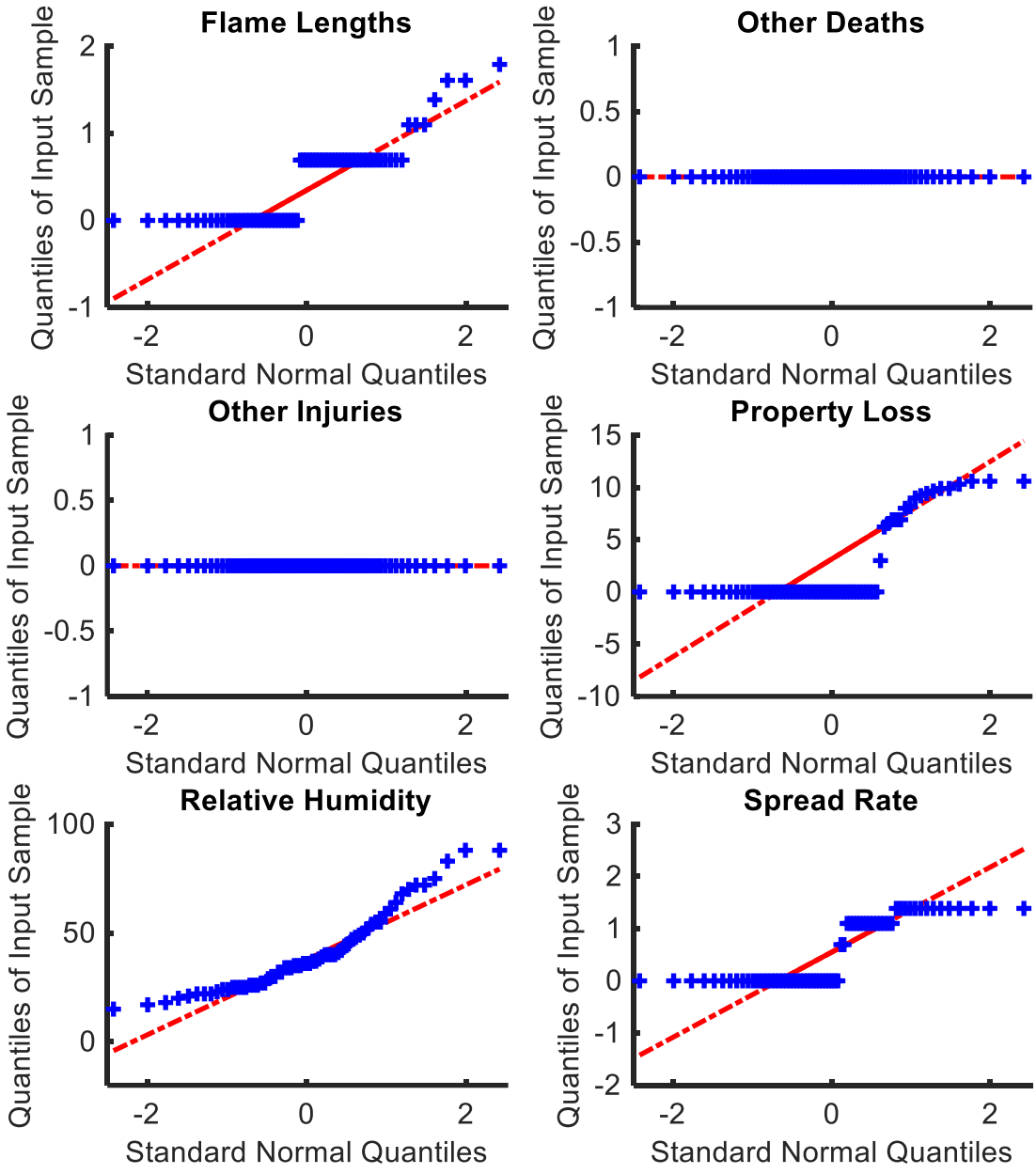

Figure A.6. Q-Q plots of all variables corresponding to fire cause ‘6’ (Contd.)

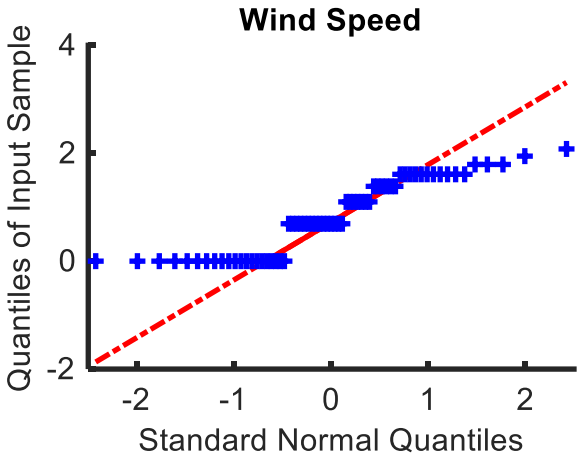

Figure A.6. Q-Q plots of all variables corresponding to fire cause ‘6’.

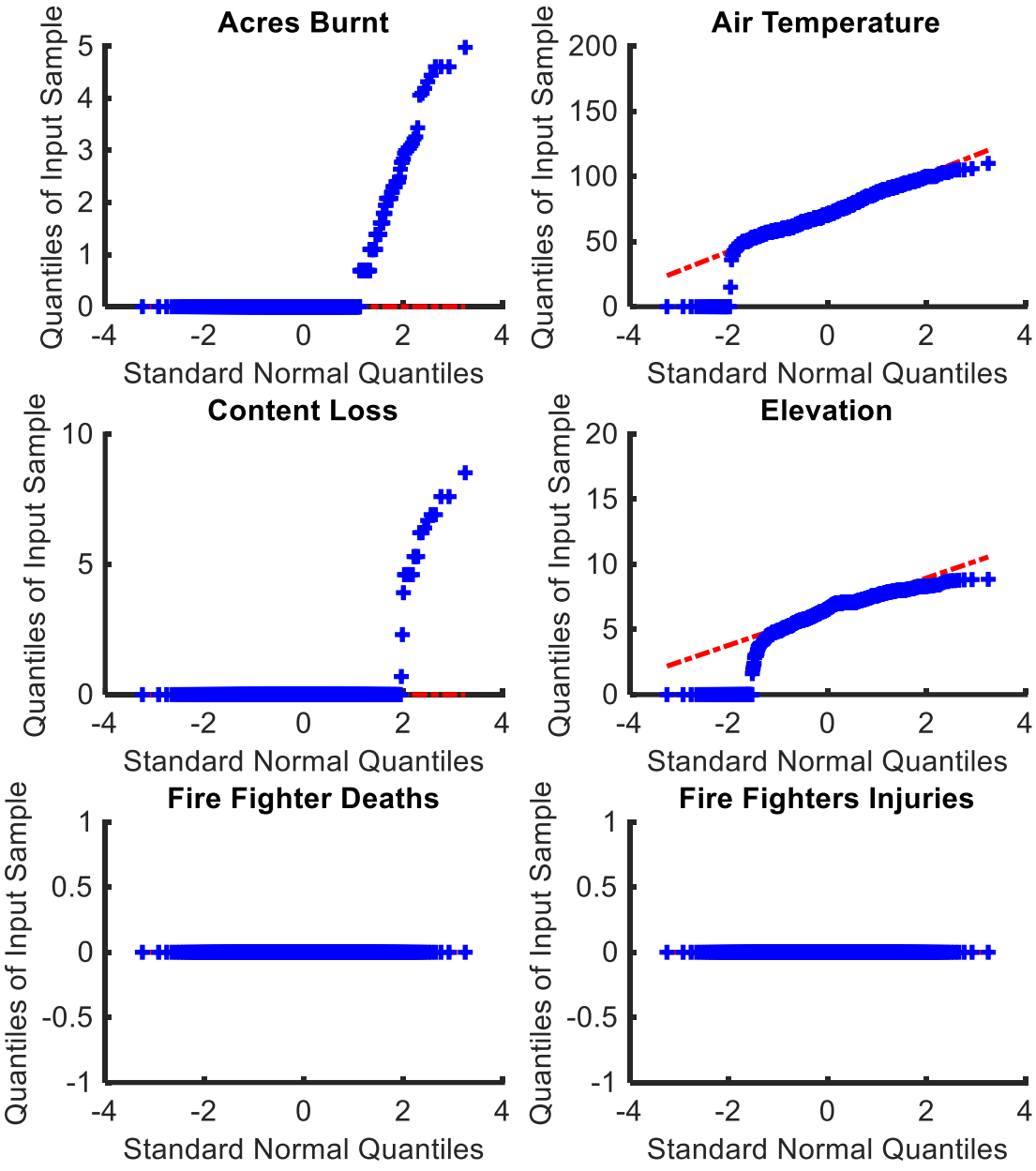

Figure A.7. Q-Q plots of all variables corresponding to fire cause ‘7’ (Contd.)

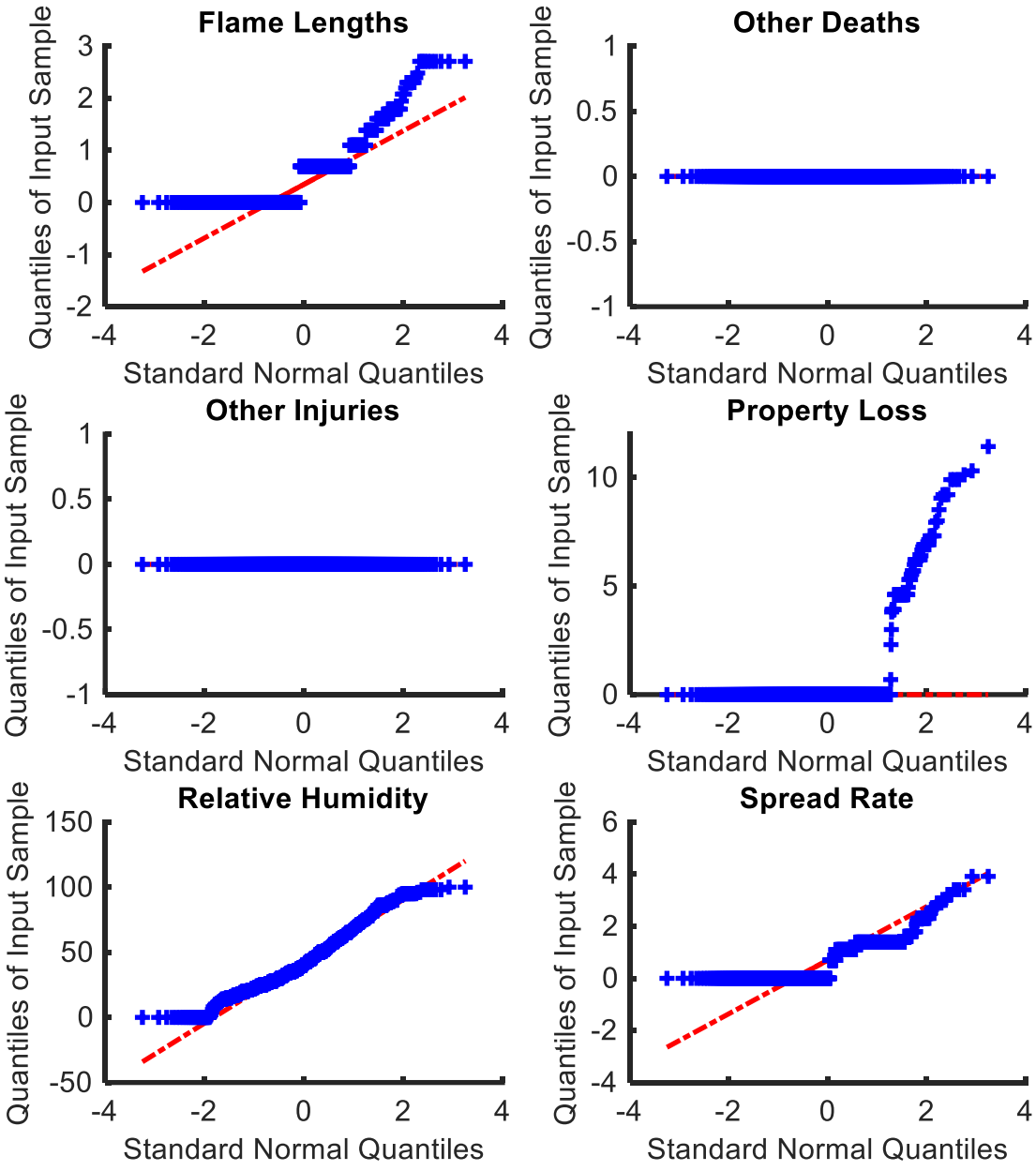

Figure A.7. Q-Q plots of all variables corresponding to fire cause ‘7’ (Contd.)

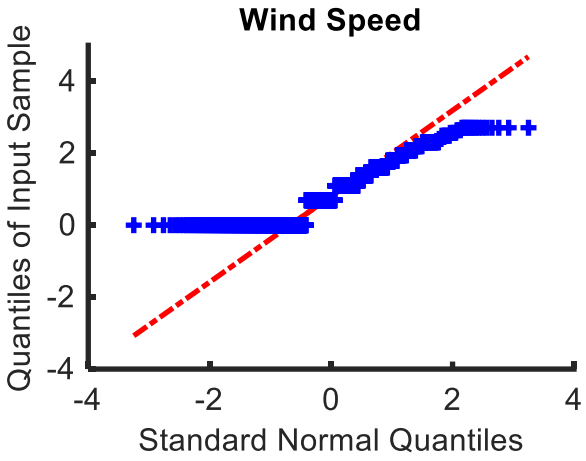

Figure A.7. Q-Q plots of all variables corresponding to fire cause ‘7’.

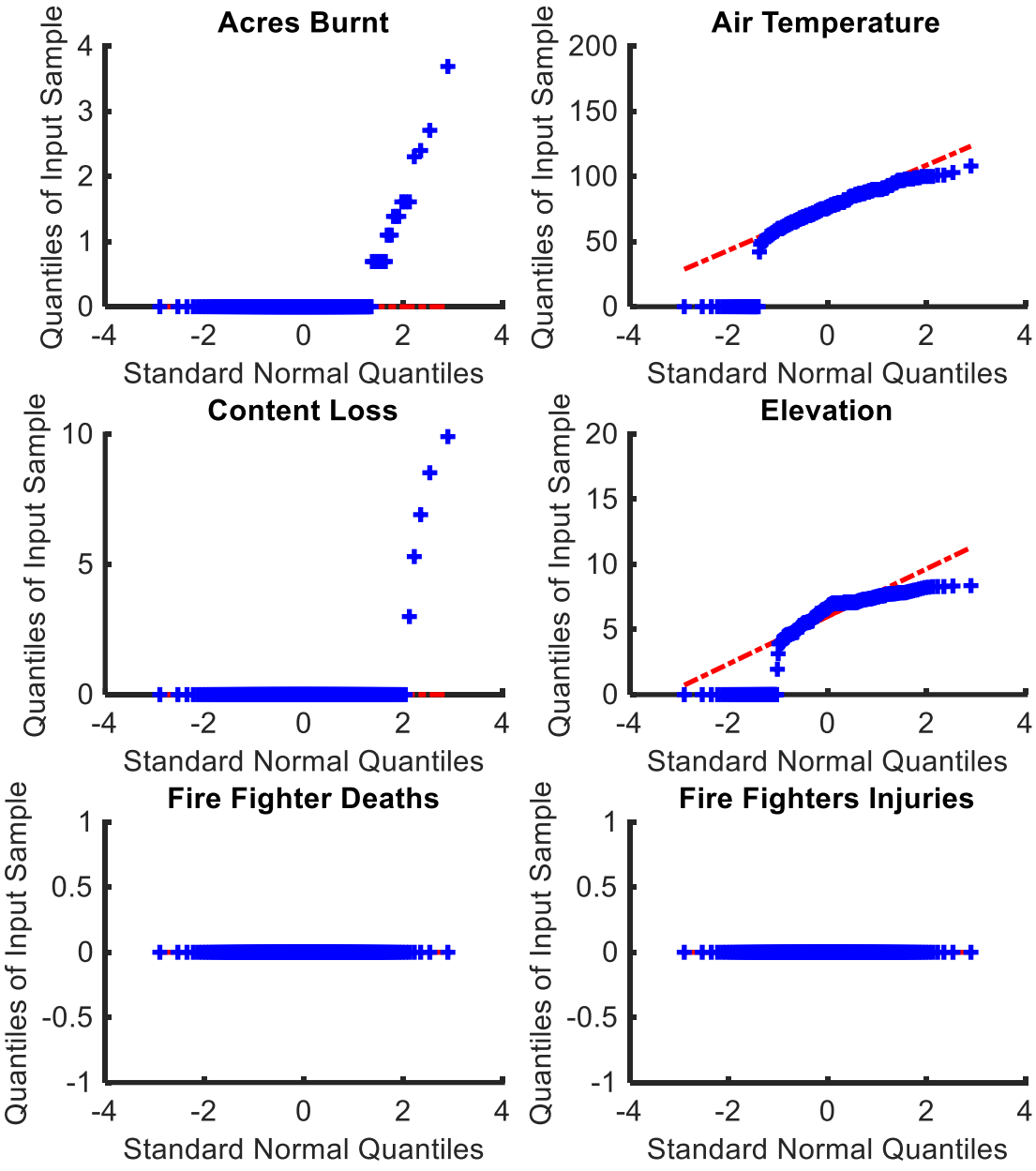

Figure A.8. Q-Q plots of all variables corresponding to fire cause ‘8’ (Contd.)

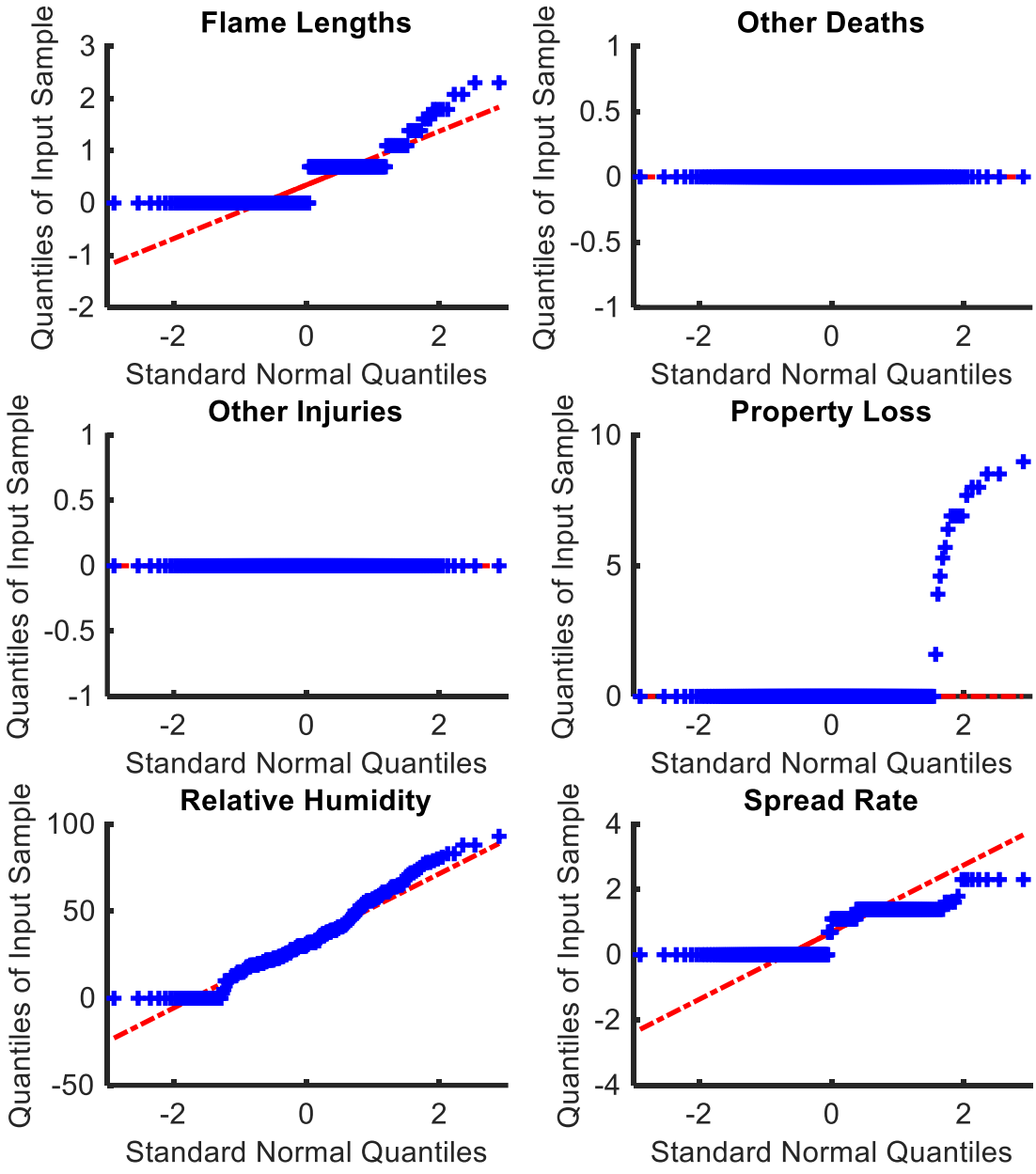

Figure A.8. Q-Q plots of all variables corresponding to fire cause '8' (Contd.)

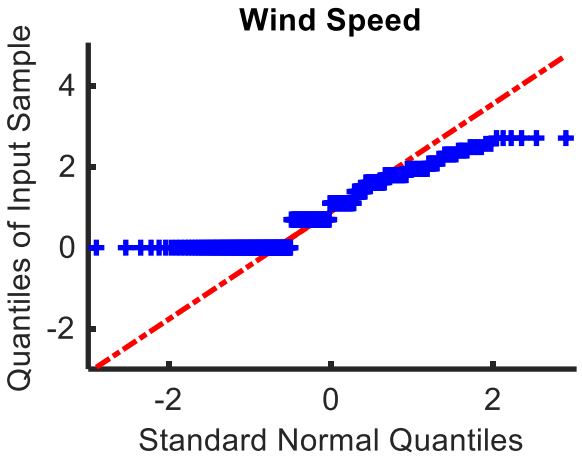

Figure A.8. Q-Q plots of all variables corresponding to fire cause ‘8’.

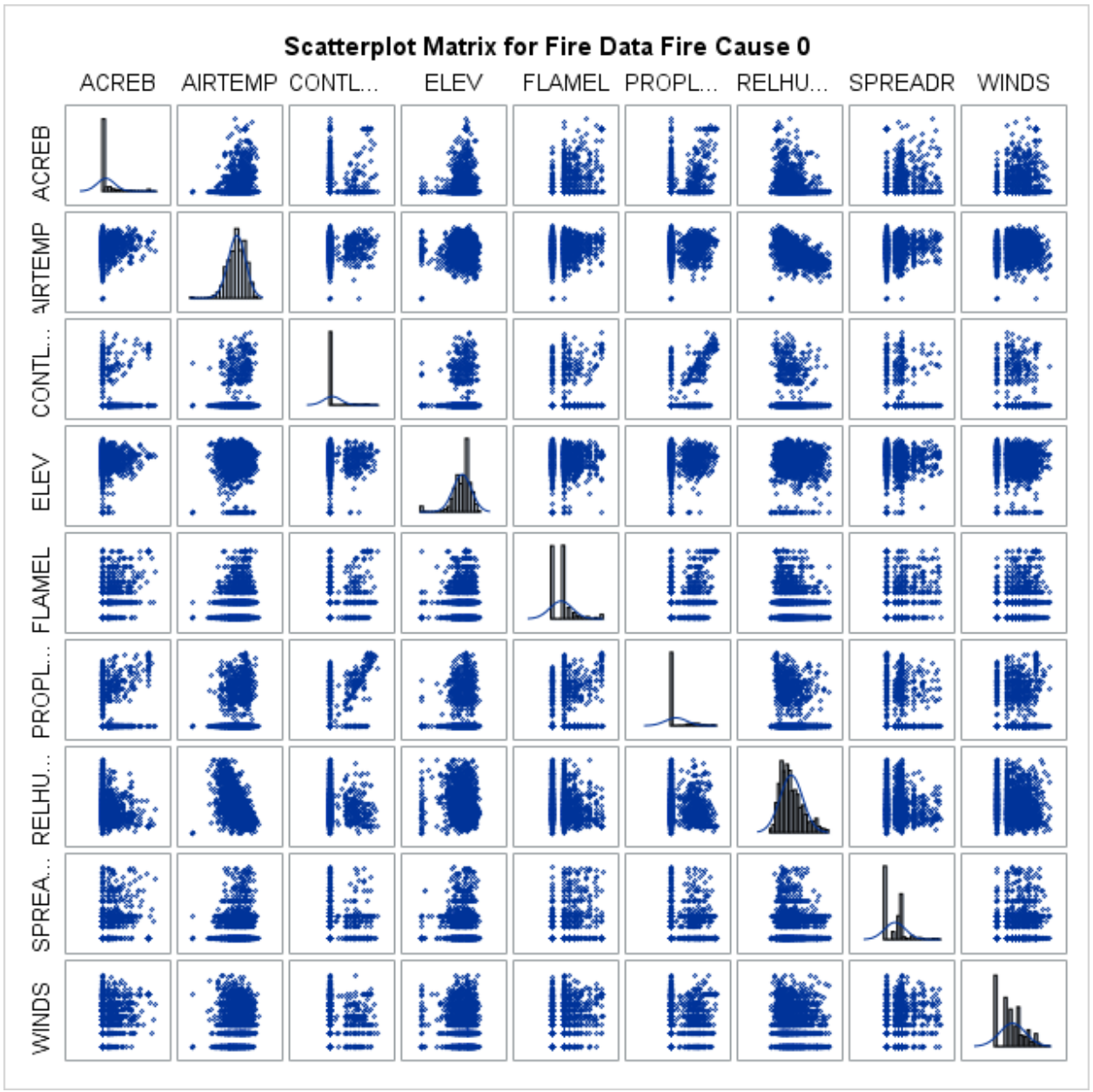

Figure A.9. Scatterplot Matrix of 9 variables corresponding to fire cause ‘0’.

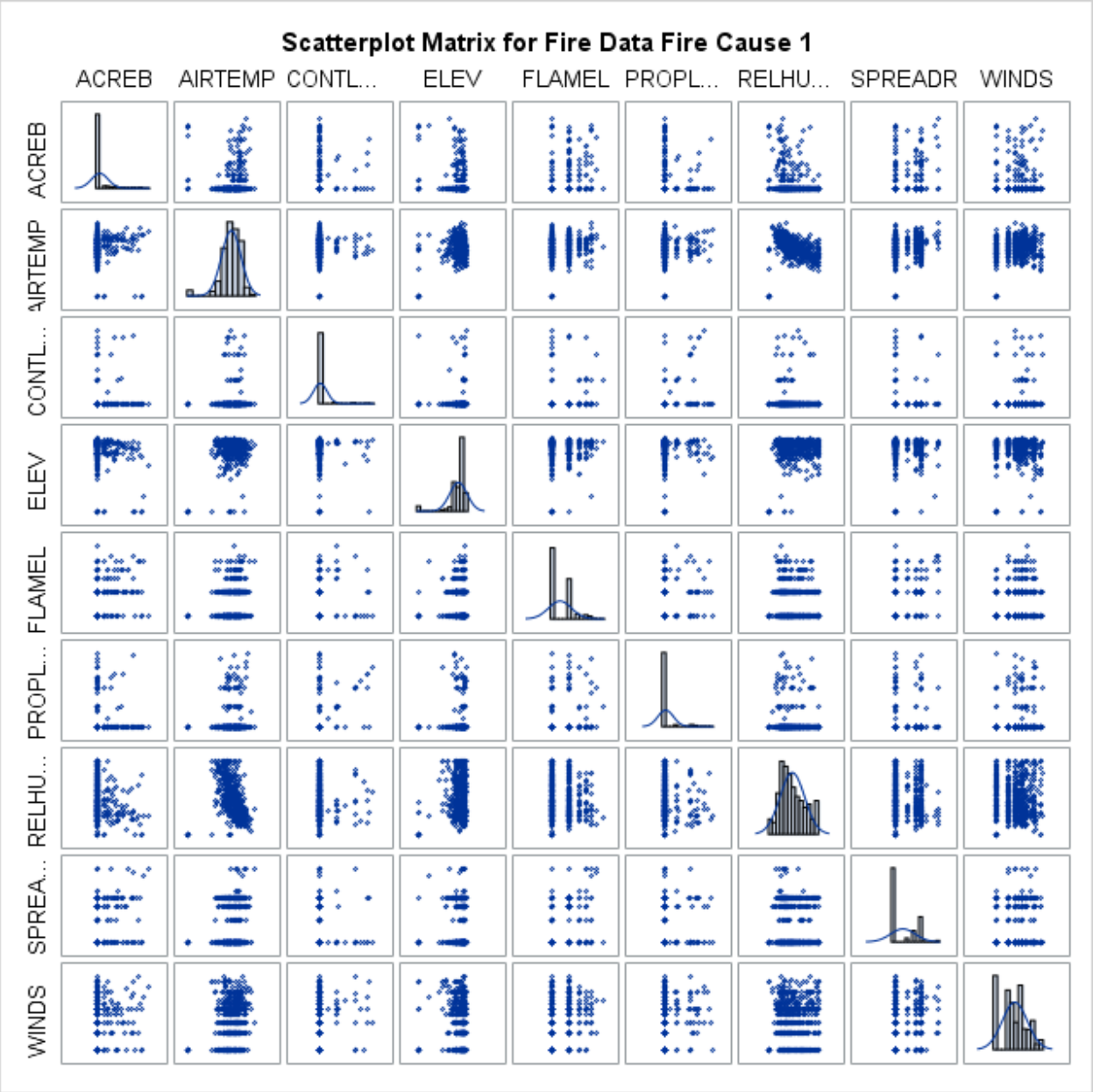

Figure A.10. Scatterplot Matrix of 9 variables corresponding to fire cause ‘1’.

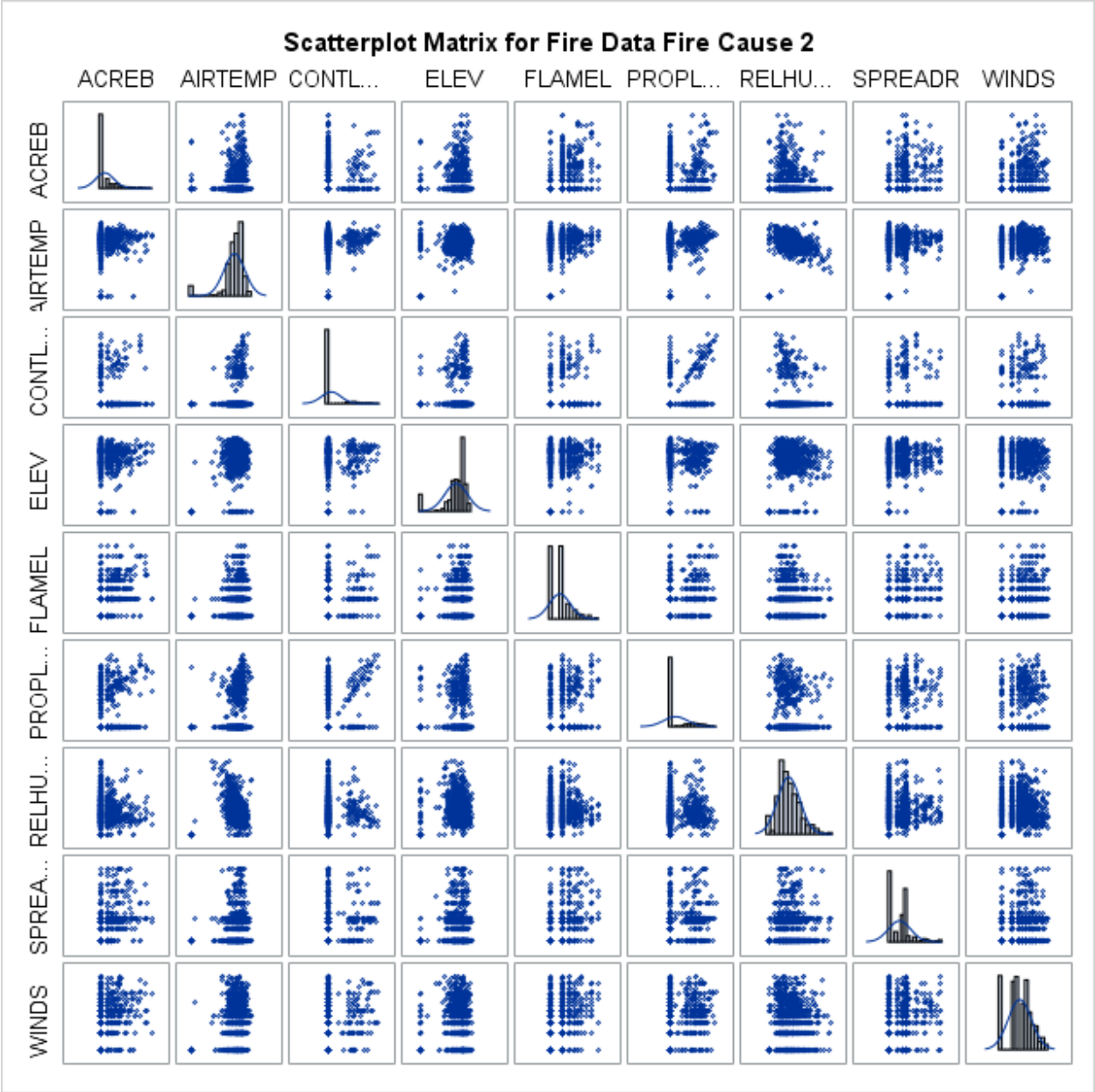

Figure A.11. Scatterplot Matrix of 9 variables corresponding to fire cause '2'.

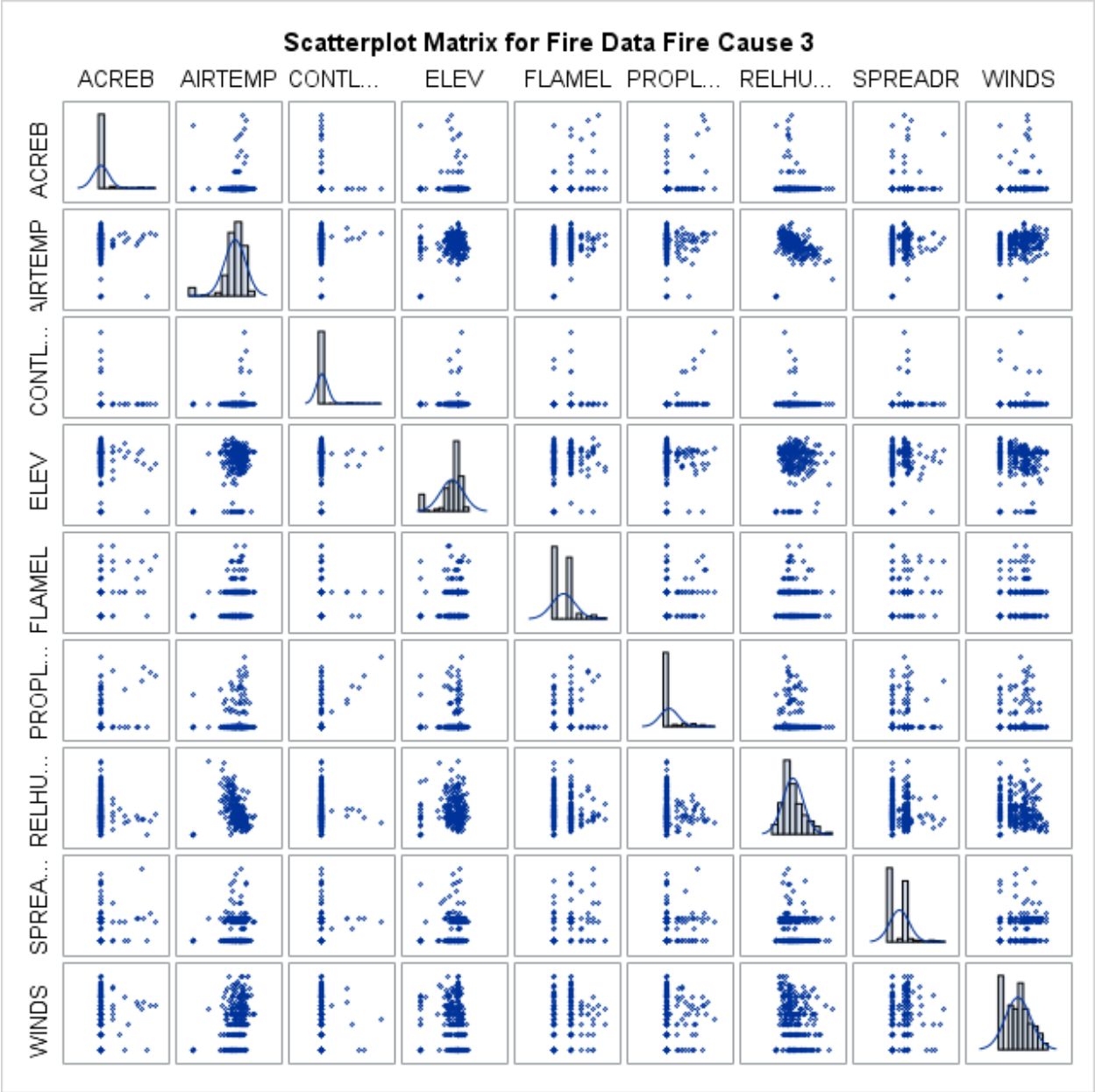

Figure A.12. Scatterplot Matrix of 9 variables corresponding to fire cause '3'.

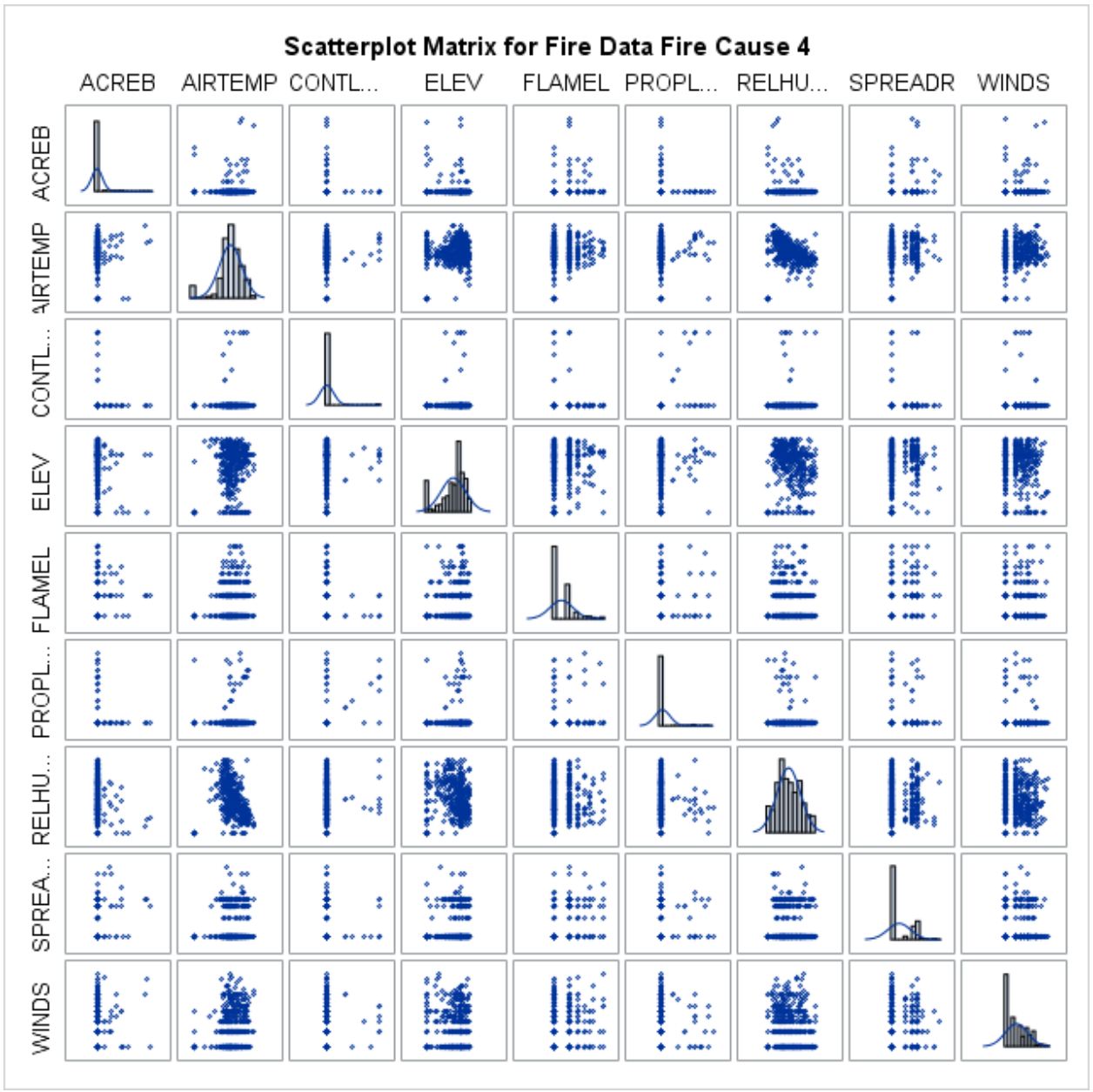

Figure A.13. Scatterplot Matrix of 9 variables corresponding to fire cause ‘4’.

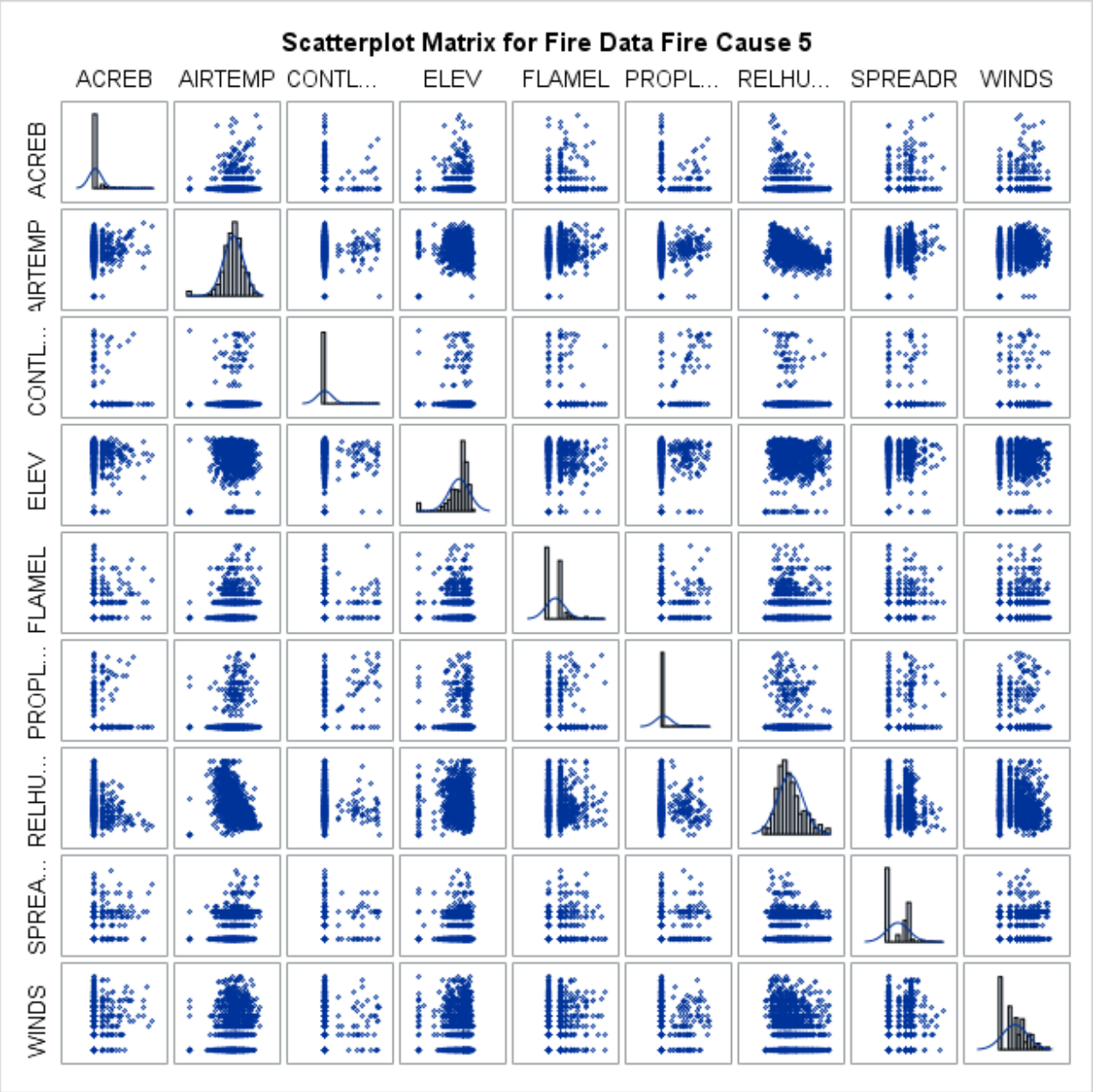

Figure A.14. Scatterplot Matrix of 9 variables corresponding to fire cause ‘5’.

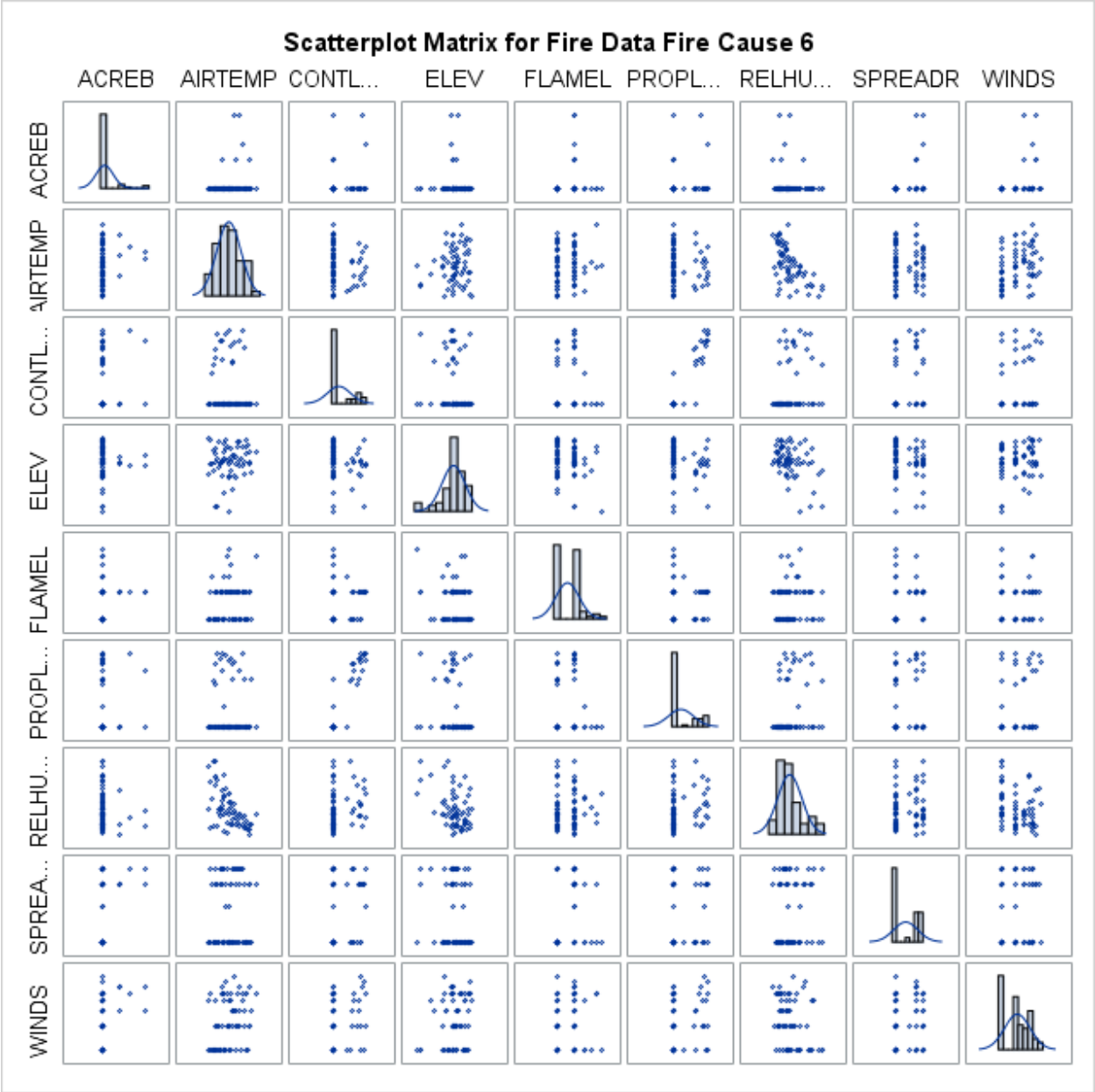

Figure A.15. Scatterplot Matrix of 9 variables corresponding to fire cause ‘6’.

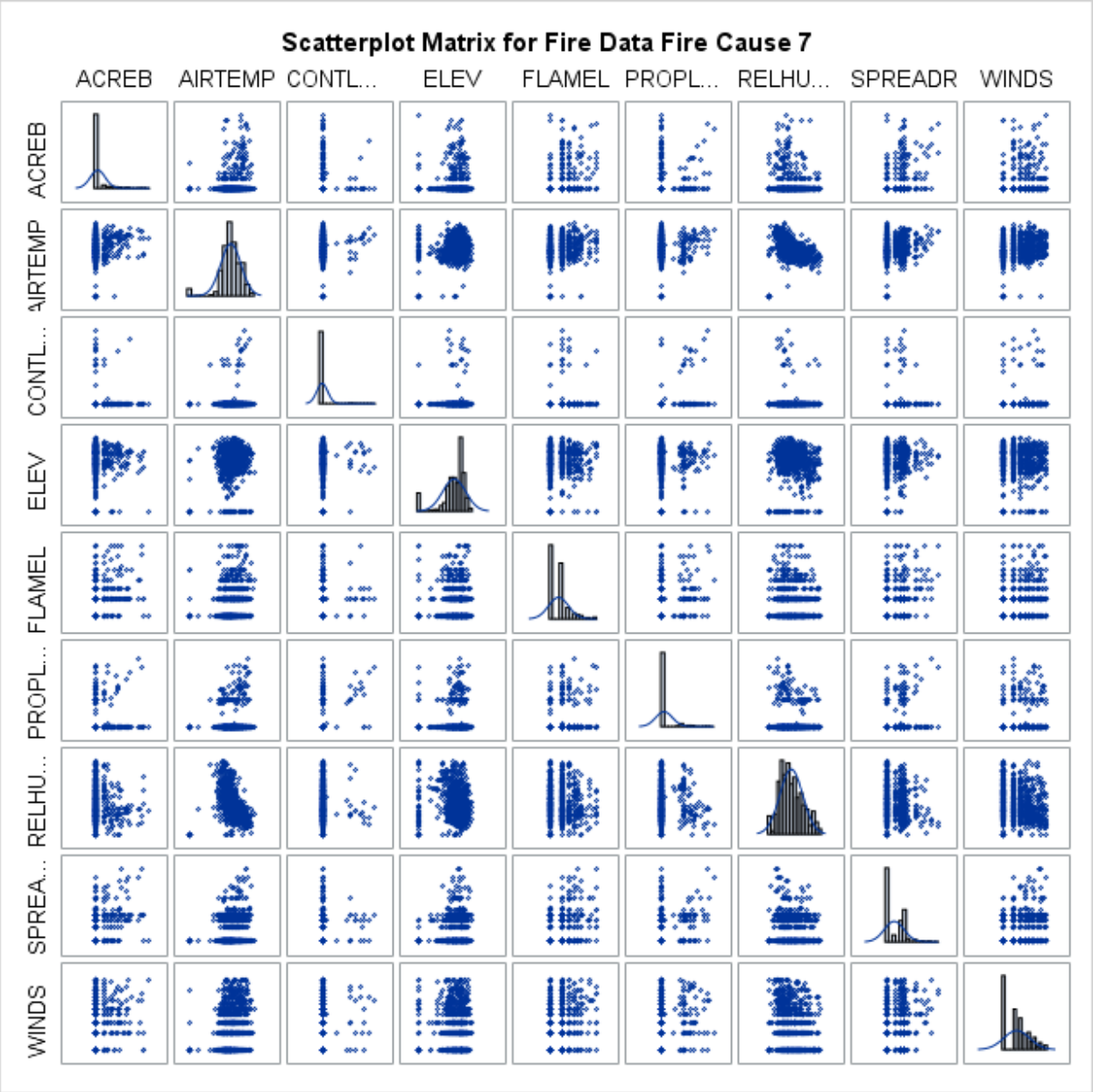

Figure A.16. Scatterplot Matrix of 9 variables corresponding to fire cause '7'.

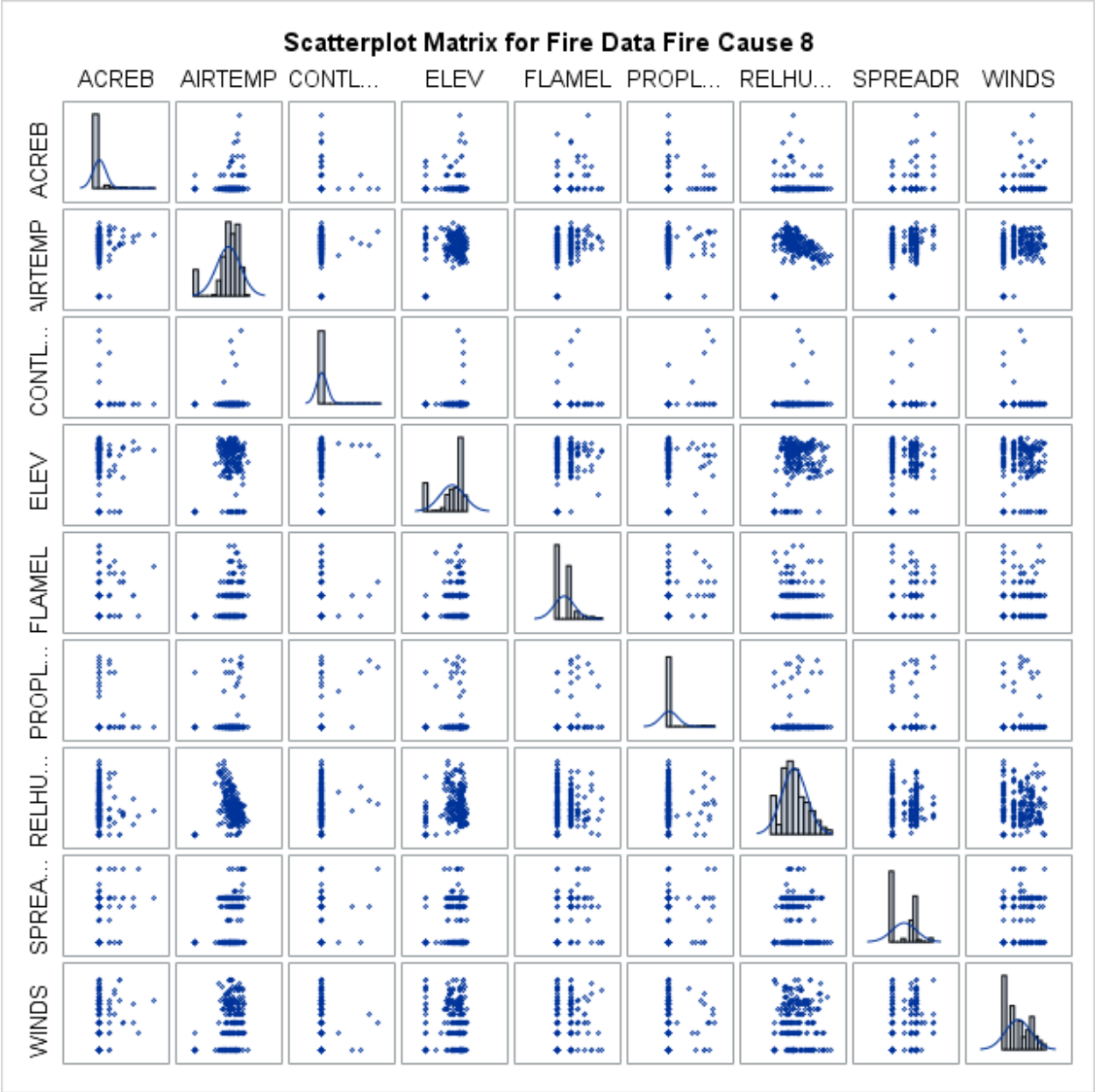

Figure A.17. Scatterplot Matrix of 9 variables corresponding to fire cause ‘8’.

Appendix B

B.1 SAS code to generate initial Q-Qplots

```
OPTIONSLS=84;
DATA FIRE;
INFILE'E:\Fall 2017 courses\STAT 764 Methods of Multivariate Methods\Project\dataset\master_dataset_all_years_testing.txt';
INPUT FIREC ACREB AIRTEMP CONTLOSS ELEV FFDEATH FFINJ FLAMEL OTHDEATH OTHINJ PROPLOSS RELHUMID
SPREADR WINDS;
RUN;
odspdffile="E:\Fall 2017 courses\STAT 764 Methods of Multivariate
Methods\Project\code\qqplots_master_data_allyears_testing.pdf";
PROCCAPABILITYDATA=FIRE NORMALTEST;
        BY FIREC;
        QQPLOT ACREB AIRTEMP CONTLOSS ELEV FFDEATH FFINJ FLAMEL OTHDEATH OTHINJ PROPLOSS
RELHUMID SPREADR WINDS;
RUN;
odspdfclose;
```

B.2 SAS code to generate bi-scatter plots

```
OPTIONSLS=84;
DATA FIRE;
INFILE'E:\Fall 2017 courses\STAT 764 Methods of Multivariate
Methods\Project\dataset\standardized box cox transformed data all years for pca.txt';
INPUT FIREC ACREB AIRTEMP CONTLOSS ELEV FLAMEL PROPLOSS RELHUMID SPREADR WINDS;
RUN;
DATA FIRE0;
set FIRE;
where FIREC=0;
run;
DATA FIRE1;
set FIRE;
where FIREC=1;
run;
DATA FIRE2;
set FIRE;
where FIREC=2;
run;
DATA FIRE3;
set FIRE;
where FIREC=3;
run;
DATA FIRE4;
set FIRE;
where FIREC=4;
run;
DATA FIRE5;
set FIRE;
where FIREC=5;
run;
DATA FIRE6;
set FIRE;
where FIREC=6;
run;
DATA FIRE7;
set FIRE;
where FIREC=7;
run;
DATA FIRE8;
set FIRE;
where FIREC=8;
run;
odspdffile="E:\Fall 2017 courses\STAT 764 Methods of Multivariate Methods\Project\code\scatterplot_transformed.pdf";
procsgscatterdata=FIRE0;
```

```
title"Scatterplot Matrix for Fire Data Fire Cause 0";
matrix ACREB AIRTEMP CONTLOSS ELEV FLAMEL PROPLOSS RELHUMID SPREADR WINDS
      / diagonal=(histogram normal);
run;
procsgscatterdata=FIRE1;
title"Scatterplot Matrix for Fire Data Fire Cause 1";
matrix ACREB AIRTEMP CONTLOSS ELEV FLAMEL PROPLOSS RELHUMID SPREADR WINDS
      / diagonal=(histogram normal);
run;
procsgscatterdata=FIRE2;
title"Scatterplot Matrix for Fire Data Fire Cause 2";
matrix ACREB AIRTEMP CONTLOSS ELEV FLAMEL PROPLOSS RELHUMID SPREADR WINDS
      / diagonal=(histogram normal);
run;
procsgscatterdata=FIRE3;
title"Scatterplot Matrix for Fire Data Fire Cause 3";
matrix ACREB AIRTEMP CONTLOSS ELEV FLAMEL PROPLOSS RELHUMID SPREADR WINDS
      / diagonal=(histogram normal);
run;
procsgscatterdata=FIRE4;
title"Scatterplot Matrix for Fire Data Fire Cause 4";
matrix ACREB AIRTEMP CONTLOSS ELEV FLAMEL PROPLOSS RELHUMID SPREADR WINDS
      / diagonal=(histogram normal);
run;
procsgscatterdata=FIRE5;
title"Scatterplot Matrix for Fire Data Fire Cause 5";
matrix ACREB AIRTEMP CONTLOSS ELEV FLAMEL PROPLOSS RELHUMID SPREADR WINDS
      / diagonal=(histogram normal);
run;
procsgscatterdata=FIRE6;
title"Scatterplot Matrix for Fire Data Fire Cause 6";
matrix ACREB AIRTEMP CONTLOSS ELEV FLAMEL PROPLOSS RELHUMID SPREADR WINDS
      / diagonal=(histogram normal);
run;
procsgscatterdata=FIRE7;
title"Scatterplot Matrix for Fire Data Fire Cause 7";
matrix ACREB AIRTEMP CONTLOSS ELEV FLAMEL PROPLOSS RELHUMID SPREADR WINDS
      / diagonal=(histogram normal);
run;
procsgscatterdata=FIRE8;
title"Scatterplot Matrix for Fire Data Fire Cause 8";
matrix ACREB AIRTEMP CONTLOSS ELEV FLAMEL PROPLOSS RELHUMID SPREADR WINDS
      / diagonal=(histogram normal);
run;
title;
odspdfclose;
```

B.3 SAS code to standardize dataset

```
DATA FIRE;
INFILE'E:\Fall 2017 courses\STAT 764 Methods of Multivariate
Methods\Project\dataset\master_dataset_box_cox_transformed_all_years_testing.txt';
INPUT FIREC ACREB AIRTEMP CONTLOSS ELEV FFDEATH FFINJ FLAMEL OTHDEATH OTHINJ PROPLOSS RELHUMID
SPREADR WINDS;
RUN;
odspdffile="E:\Fall 2017 courses\STAT 764 Methods of Multivariate Methods\Project\code\standardized.pdf";
procsql;
createtable STATS_QUERY_FORM as
            select
            Min(ACREB) as Min_ACREB,Max(ACREB) as Max_ACREB,            avg(ACREB) as Means_ACREB,std(ACREB) as
Std_Deviation_ACREB,
            Min(AIRTEMP) as Minimum_AIRTEMP,Max(AIRTEMP) as Max_AIRTEMP,avg(AIRTEMP) as
Means_AIRTEMP,std(AIRTEMP) as Std_Deviation_AIRTEMP,
            Min(CONTLOSS) as Min_CONTLOSS,Max(CONTLOSS) as Max_CONTLOSS,avg(CONTLOSS) as
Means_CONTLOSS,std(CONTLOSS) as Std_Deviation_CONTLOSS,
            Min(ELEV) as Min_ELEV,Max(ELEV) as Max_ELEV,avg(ELEV) as Means_ELEV,std(ELEV) as Std_Deviation_ELEV,
            Min(FFDEATH) as Min_FFDEATH,Max(FFDEATH) as Max_FFDEATH,avg(FFDEATH) as
Means_FFDEATH,std(FFDEATH) as Std_Deviation_FFDEATH,
            Min(FFINJ) as Min_FFINJ,Max(FFINJ) as Max_FFINJ,avg(FFINJ) as Means_FFINJ,std(FFINJ) as Std_Deviation_FFINJ,
```

```
        Min(FLAMEL) as Min_FLAMEL,Max(FLAMEL) as Max_FLAMEL,avg(FLAMEL) as Means_FLAMEL,std(FLAMEL) as
Std_Deviation_FLAMEL,
        Min(OTHDEATH) as Min_OTHDEATH,Max(OTHDEATH) as Max_OTHDEATH,avg(OTHDEATH) as
Means_OTHDEATH,std(OTHDEATH) as Std_Deviation_OTHDEATH,
        Min(OTHINJ) as Min_OTHINJ,Max(OTHINJ) as Max_OTHINJ,avg(OTHINJ) as Means_OTHINJ,std(OTHINJ) as
Std_Deviation_OTHINJ,
        Min(PROPLOSS) as Min_PROPLOSS,Max(PROPLOSS) as Max_PROPLOSS,avg(PROPLOSS) as
Means_PROPLOSS,std(PROPLOSS) as Std_Deviation_PROPLOSS,
        Min(RELHUMID) as Min__RELHUMID,Max(RELHUMID) as Max_RELHUMID,avg(RELHUMID) as
Means_RELHUMID,std(RELHUMID) as Std_Deviation_RELHUMID,
        Min(SPREADR) as Min_SPREADR,Max(SPREADR) as Max_SPREADR,avg(SPREADR) as
Means_SPREADR,std(SPREADR) as Std_Deviation_SPREADR,
        Min(WINDS) as Min_WINDS,Max(WINDS) as Max_WINDS,avg(WINDS) as Means_WINDS,std(WINDS) as
Std_Deviation_WINDS
from FIRE;
quit;
procsql;
createtable STANDARDIZED_DATA as
        select FIREC,(ACREB-Means  ACREB)/Std  Deviation_ACREB AS ACREB,(AIRTEMP-
Means_AIRTEMP)/Std_Deviation_AIRTEMP AS AIRTEMP,
        (CONTLOSS-Means_CONTLOSS)/Std_Deviation_CONTLOSS AS CONTLOSS,(ELEV-
Means_ELEV)/Std_Deviation_ELEV AS ELEV,
        (FFDEATH-Means_FFDEATH)/Std_Deviation_FFDEATH AS FFDEATH,(FFINJ-Means_FFINJ)/Std_Deviation_FFINJ AS
FFINJ,
        (FLAMEL-Means_FLAMEL)/Std_Deviation_FLAMEL AS FLAMEL,(OTHDEATH-
Means_OTHDEATH)/Std_Deviation_OTHDEATH AS OTHDEATH,
        (OTHINJ-Means_OTHINJ)/Std_Deviation_OTHINJ AS OTHINJ,(PROPLOSS-
Means  PROPLOSS)/Std  Deviation  PROPLOSS AS PROPLOSS,
        (RELHUMID-Means_RELHUMID)/Std_Deviation_RELHUMID AS RELHUMID,(SPREADR-
Means_SPREADR)/Std_Deviation_SPREADR AS SPREADR,
        (WINDS-Means_WINDS)/Std_Deviation_WINDS AS WINDS
        from FIRE,STATS_QUERY_FORM;
quit;
procEXPORTdata=STANDARDIZED_DATA
outfile='E:\Fall 2017 courses\STAT 764 Methods of Multivariate
Methods\Project\dataset\standardized_box_cox_transformed_data_all_years_testing_exported.txt'
dbms=dlm;
delimiter='&';
run;
odspdfclose;
```

B.4 SAS code to perform Discriminant Analysis

```
DATA FIRE;
INFILE'E:\Fall 2017 courses\STAT 764 Methods of Multivariate
Methods\Project\dataset\standardized_box_cox_transformed_data_all_years_exported.txt';
INPUT FIREC ACREB AIRTEMP CONTLOSS ELEV FFDEATH FFINJ FLAMEL OTHDEATH OTHINJ PROPLOSS RELHUMID
SPREADR WINDS;
RUN;
odspdffile="E:\Fall 2017 courses\STAT 764 Methods of Multivariate Methods\Project\code\discriminantresult_standard.pdf";
PROCSTEPDISC STEPWISE SIMPLESTDMEANTCORRWCORR;
CLASS FIREC;
TITLE'STEPWISE';
RUN;
odspdfclose;
```

B.5 R code to generate summary statistics

```
library(factoextra)
library(FactoMineR)
#taking input
setwd("E:/Fall 2017 courses/STAT 764 Methods of Multivariate Methods/Project/dataset")
datafile='standardized_box_cox_transformed_data_all_years_for_pca.txt'
training=as.matrix(read.table(datafile))
r=dim(training)[1]
c=dim(training)[2]
Species=training[,1]
```

```
train1=training[,2:c]
colnames(train1)<-c("ACREB","AIRTEMP","CONTLOSS","ELEV","FLAMEL","PROPLOSS","RELHUMID","SPREADR","WINDS")

#calculating summary stats
summaryResult=summary(train1)
write.xlsx(summaryResult,"E:/Fall 2017 courses/STAT 764 Methods of Multivariate Methods/Project/code/summaryResult.xlsx")
covtrain1=cov(train1)
write.xlsx(covtrain1,"E:/Fall 2017 courses/STAT 764 Methods of Multivariate Methods/Project/code/covarianceResult.xlsx")
cortrain1=cor(train1)
write.xlsx(cortrain1,"E:/Fall 2017 courses/STAT 764 Methods of Multivariate Methods/Project/code/correlationResult.xlsx")
```

B.6 SAS code to perform Prinicipal Component Analysis

```
DATA FIRE;
INFILE'E:\Fall 2017 courses\STAT 764 Methods of Multivariate
Methods\Project\dataset\standardized_box_cox_transformed_data_all_years_for_pca.txt';
INPUT FIREC ACREB AIRTEMP CONTLOSS ELEV FLAMEL PROPLOSS RELHUMID SPREADR WINDS;
RUN;
odspdffile="E:\Fall 2017 courses\STAT 764 Methods of Multivariate Methods\Project\code\pcaresult_standard.pdf";
odsgraphicson;
PROCPRINCOMPOUT=RESULTS1 plots=score(ellipse);
        VAR ACREB AIRTEMP ELEV PROPLOSS RELHUMID SPREADR;
procprint;
run;
odsgraphicsclose;
odspdfclose;
```

B.7 R code to perform Principal Component Analysis

```
library(factoextra)
library(FactoMineR)

#taking input
setwd("E:/Fall 2017 courses/STAT 764 Methods of Multivariate Methods/Project/dataset")
datafile='standardized_box_cox_transformed_data_all_years_for_pca.txt'
training=as.matrix(read.table(datafile))
r=dim(training)[1]
c=dim(training)[2]
Species=training[,1]
train1=training[,2:c]
colnames(train1)<-c("ACREB","AIRTEMP","CONTLOSS","ELEV","FLAMEL","PROPLOSS","RELHUMID","SPREADR","WINDS")

#calculating prinicipal components
res.pca <- prcomp(train1, scale = TRUE)

#to plot screeplot
tiff("E:/Fall 2017 courses/STAT 764 Methods of Multivariate Methods/Project/dataset/screeplotnew1.tiff", width = 6.6, height = 4.8,
units = 'in', res = 300)
fviz_eig(res.pca, geom="line",choice = "eigenvalue",addlabels=TRUE, main = " ",xlab ="Component Number",ylab = "Eigen
Values",ggtheme = theme_classic() )

#to plot biplot
tiff("E:/Fall 2017 courses/STAT 764 Methods of Multivariate Methods/Project/dataset/biplotnew1.tiff", width = 6.6, height = 4.8, units
= 'in', res = 300)
fviz_pca_biplot(res.pca, label="var",  labelsize = 5,pointsize = 2.5,arrowsize = 1,habillage=Species,
        col.var = "black",addEllipses=TRUE, ellipse.level=0.95,repel=TRUE,legend.title="Fire cause ")

dev.off()
```

B.8 SAS code to perform M test

```
DATA FIRE;
INFILE'E:\Fall 2017 courses\STAT 764 Methods of Multivariate
Methods\Project\dataset\standardized_box_cox_transformed_data_all_years_for_pca.txt';
INPUT FIREC ACREB AIRTEMP CONTLOSS ELEV FLAMEL PROPLOSS RELHUMID SPREADR WINDS;
RUN;
```

```
odspdffile="E:\Fall 2017 courses\STAT 764 Methods of Multivariate Methods\Project\code\mtestResult.pdf";
DATA SCORES;
INFILE PSYCH;
INPUT SEX TEST1 TEST2 TEST3 TEST4;
PROCIML;
USE FIRE;
READ ALL VAR {ACREB AIRTEMP CONTLOSS ELEV FLAMEL PROPLOSS RELHUMID SPREADR WINDS} INTO Y;
  n = nrow(Y);
  p = ncol(Y);

  Y0 = Y[1:2420,];
  Y1 = Y[2421:3008,];
  Y2 = Y[3009:3927,];
  Y3 = Y[3928:4201,];
  Y4 = Y[4202:4642,];
  Y5 = Y[4643:6043,];
  Y6 = Y[6044:6108,];
  Y7 = Y[6109:6976,];
  Y8 = Y[6977:7244,];

  n0 = nrow(Y0);
  y0b = 1/n0 * t(Y0)*J(n0,1);
  S0 = 1/(n0-1) * t(Y0) * (I(n0) - 1/n0*j(n0)) * Y0;

  n1 = nrow(Y1);
  y1b = 1/n1 * t(Y1)*J(n1,1);
  S1 = 1/(n1-1) * t(Y1) * (I(n1) - 1/n1*j(n1)) * Y1;

  n2 = nrow(Y2);
  y2b = 1/n2 * t(Y2)*J(n2,1);
  S2 = 1/(n2-1) * t(Y2) * (I(n2) - 1/n2*j(n2)) * Y2;

  n3 = nrow(Y3);
  y3b = 1/n3 * t(Y3)*J(n3,1);
  S3 = 1/(n3-1) * t(Y3) * (I(n3) - 1/n3*j(n3)) * Y3;

  n4 = nrow(Y4);
  y4b = 1/n4 * t(Y4)*J(n4,1);
  S4 = 1/(n4-1) * t(Y4) * (I(n4) - 1/n4*j(n4)) * Y4;
  n5 = nrow(Y5);
  y5b = 1/n5 * t(Y5)*J(n5,1);
  S5 = 1/(n5-1) * t(Y5) * (I(n5) - 1/n5*j(n5)) * Y5;
  n6 = nrow(Y6);
  y6b = 1/n6 * t(Y6)*J(n6,1);
  S6 = 1/(n6-1) * t(Y6) * (I(n6) - 1/n6*j(n6)) * Y6;
  n7 = nrow(Y7);
  y7b = 1/n7 * t(Y7)*J(n7,1);
  S7 = 1/(n7-1) * t(Y7) * (I(n7) - 1/n7*j(n7)) * Y7;
  n8 = nrow(Y8);
  y8b = 1/n8 * t(Y8)*J(n8,1);
  S8 = 1/(n8-1) * t(Y8) * (I(n8) - 1/n8*j(n8)) * Y8;

  Spl = 1/(n0+n1+n2+n3+n4+n5+n6+n7+n8-9) * ((n0-1)*S0+(n1-1)*S1 + (n2-1)*S2+(n3-1)*S3+(n4-1)*S4+(n5-1)*S5+(n6-1)*S6+(n7-
1)*S7+(n8-1)*S8);

  v0 = n0-1;
  v1 = n1-1;
  v2 = n2-1;
  v3 = n3-1;
  v4 = n4-1;
  v5 = n5-1;
  v6 = n6-1;
  v7 = n7-1;
  v8 = n8-1;

  v = v0+v1+v2+v3+v4+v5+v6+v7+v8;

  detS0 = det(S0);
  detS1 = det(S1);
  detS2 = det(S2);
```

```
  detS3 = det(S3);
  detS4 = det(S4);
  detS5 = det(S5);
  detS6 = det(S6);
  detS7 = det(S7);
  detS8 = det(S8);

  detSpl = det(Spl);

  logM =(1/2) * (v0*log(detS0) + v1*log(detS1) + v2*log(detS2) + v3*log(detS3) + v4*log(detS4) + v5*log(detS5) + v6*log(detS6) +
v7*log(detS7) + v8*log(detS8) - v*log(detSpl));

  exactu = -2 * logM;

  k = 9;
  c1 = (1/v0 + 1/v1 + 1/v2 + 1/v3 + 1/v4 + 1/v5 + 1/v6 + 1/v7 + 1/v8 - 1/v) * (2*p**2 + 3*p - 1) / (6 * (p+1) * (k-1));
  u = -2*(1-c1)*logM;
  dfX2 = (1/2) * (k-1) * p * (p+1);
  X2crit = quantile('CHISQUARE', 0.95, dfX2);
  X2pval = 1 - cdf('CHISQUARE', u, dfX2);


  c2 = (1/v0**2 + 1/v1**2 + 1/v2**2 + 1/v3**2 + 1/v4**2 + 1/v5**2 + 1/v6**2 + 1/v7**2 + 1/v8**2 - 1/v**2) * (p-1) * (p+2) / (6*(k-1));
  a1 = (1/2) * (k-1) * p * (p+1);
  a2 = (a1+2) / abs(c2 - c1**2);
  b1 = (1 - c1 - a1/a2) / a1;
  b2 = (1 - c1 - 2/a2) / a2;

START FVALUE;
IF c2>c1**2THEN DO;
      F = -2*b1*logM;
PRINT F;
END;
ELSE DO;
      F = -a2*b2*logM/(a1*(1+2*b2*logM));
PRINT F;
END;
FINISH;
RUN FVALUE;

   Fcrit = quantile('F', 0.95, a1, a2);
   Fpval = 1 - cdf('F', F, a1, a2);

PRINT v0 v1 v2 v3 v4 v5 v6 v7 v8 detS0 detS1 detS2 detS3 detS4 detS5 detS6 detS7 detS8 detSpl, logM;

PRINT'EXACT TEST STATISTIC';
PRINT exactu;

PRINT'CHI-SQUARE APPROXIMATION';
PRINT c1 u X2crit X2pval;

PRINT'F APPROXIMATION';
PRINT c2 a1 a2 b1 b2 F Fcrit Fpval;

odspdfclose;
```

B.9 SAS code to perform Quadratic Discriminant Classification

```
DATA FIRE;
INFILE'E:\Fall 2017 courses\STAT 764 Methods of Multivariate
Methods\Project\dataset\standardized_box_cox_transformed_data_all_years_for_pca.txt';
INPUT FIREC ACREB AIRTEMP CONTLOSS ELEV FLAMEL PROPLOSS RELHUMID SPREADR WINDS;
RUN;
odspdffile="E:\Fall 2017 courses\STAT 764 Methods of Multivariate Methods\Project\code\classificationResult3.pdf";
PROCDISCRIMLISTTCORRWCORRCROSSVALIDATEPOOL=NO;
CLASS FIREC;
VAR AIRTEMP ELEV PROPLOSS RELHUMID SPREADR;
PRIORS'0'=0.33'1'=0.08'2'=0.13'3'=0.04'4'=0.06'5'=0.19'6'=0.01'7'=0.12'8'=0.04;
RUN;
```

```
odspdfclose;
```

B.10 R code to perform Quadratic Discriminant Classification

```
library(knitr)
library(ISLR)
library(MASS)

#taking input
setwd("E:/Fall 2017 courses/STAT 764 Methods of Multivariate Methods/Project/dataset")
testfile='standardized_box_cox_transformed_data_all_years_testing_exported.txt'
trainfile='standardized_box_cox_transformed_data_all_years_for_pca.txt'
training=read.table(trainfile)
testing=read.table(testfile)

#assigning column names
names(training)<-
c("FIREC","ACREB","AIRTEMP","CONTLOSS","ELEV","FLAMEL","PROPLOSS","RELHUMID","SPREADR","WINDS")
names(testing)<-
c("FIREC","ACREB","AIRTEMP","CONTLOSS","ELEV","FLAMEL","PROPLOSS","RELHUMID","SPREADR","WINDS")

#applying QDA to training data
qda.fit<-qda(FIREC~.,data = training)
qda.fit

#applying prediction to testing data
qda.class <- predict(qda.fit, testing)$class
table(qda.class)
```

B.11 R code to perform Clustering

```
library(factoextra)
library(FactoMineR)
library(NbClust)
library(fpc)

#taking input
setwd("E:/Fall 2017 courses/STAT 764 Methods of Multivariate Methods/Project/dataset")
datafile='standardized_box_cox_transformed_data_all_years_testing_exported.txt'
training=as.matrix(read.table(datafile))
r=dim(training)[1]
c=dim(training)[2]
train1=training[,2:c]

#clustering using ward method
nb <- NbClust(train1, distance = "euclidean", min.nc = 2,
        max.nc = 20, method = "ward.D",index=('kl'))
fviz_nbclust(nb)

nb <- NbClust(train1, distance = "euclidean", min.nc = 2,
        max.nc = 20, method = "ward.D",index=('Silhoutte'))
fviz_nbclust(nb)

nb <- NbClust(train1, distance = "euclidean", min.nc = 2,
        max.nc = 20, method = "ward.D",index=('Cindex'))
fviz_nbclust(nb)

nb <- NbClust(train1, distance = "euclidean", min.nc = 2,
        max.nc = 20, method = "ward.D",index=('ch'))
fviz_nbclust(nb)

nb <- NbClust(train1, distance = "euclidean", min.nc = 2,
        max.nc = 20, method = "ward.D",index=('beale'))
fviz_nbclust(nb)

nb <- NbClust(train1, distance = "euclidean", min.nc = 2,
        max.nc = 20, method = "ward.D",index=('Ratkowsky'))
```

```
fviz_nbclust(nb)

nb <- NbClust(train1, distance = "euclidean", min.nc = 2,
        max.nc = 20, method = "ward.D",index=('Hartigon'))
fviz_nbclust(nb)

nb <- NbClust(train1, distance = "euclidean", min.nc = 2,
        max.nc = 20, method = "ward.D",index=('ball'))
fviz_nbclust(nb)

nb <- NbClust(train1, distance = "euclidean", min.nc = 2,
        max.nc = 20, method = "ward.D",index=('scott'))
fviz_nbclust(nb)

nb <- NbClust(train1, distance = "euclidean", min.nc = 2,
        max.nc = 20, method = "ward.D",index=('rubin'))
fviz_nbclust(nb)

nb <- NbClust(train1, distance = "euclidean", min.nc = 2,
        max.nc = 20, method = "ward.D",index=('dunn'))
fviz_nbclust(nb)

nb <- NbClust(train1, distance = "euclidean", min.nc = 2,
        max.nc = 20, method = "ward.D",index=('friedman'))
fviz_nbclust(nb)

#clustering using complete linkage method
nb <- NbClust(train1, distance = "euclidean", min.nc = 2,
        max.nc = 20, method = "complete",index=('kl'))
fviz_nbclust(nb)

nb <- NbClust(train1, distance = "euclidean", min.nc = 2,
        max.nc = 20, method = "complete",index=('Silhoutte'))
fviz_nbclust(nb)

nb <- NbClust(train1, distance = "euclidean", min.nc = 2,
        max.nc = 20, method = "complete",index=('cindex'))
fviz_nbclust(nb)

nb <- NbClust(train1, distance = "euclidean", min.nc = 2,
        max.nc = 20, method = "complete",index=('ch'))
fviz_nbclust(nb)

nb <- NbClust(train1, distance = "euclidean", min.nc = 2,
        max.nc = 20, method = "complete",index=('beale'))
fviz_nbclust(nb)

nb <- NbClust(train1, distance = "euclidean", min.nc = 2,
        max.nc = 20, method = "complete",index=('Ratkowsky'))
fviz_nbclust(nb)

nb <- NbClust(train1, distance = "euclidean", min.nc = 2,
        max.nc = 20, method = "complete",index=('Hartigon'))
fviz_nbclust(nb)

nb <- NbClust(train1, distance = "euclidean", min.nc = 2,
        max.nc = 20, method = "complete",index=('ball'))
fviz_nbclust(nb)

nb <- NbClust(train1, distance = "euclidean", min.nc = 2,
        max.nc = 20, method = "complete",index=('scott'))
fviz_nbclust(nb)

nb <- NbClust(train1, distance = "euclidean", min.nc = 2,
        max.nc = 20, method = "complete",index=('rubin'))
fviz_nbclust(nb)

nb <- NbClust(train1, distance = "euclidean", min.nc = 2,
        max.nc = 20, method = "complete",index=('dunn'))
fviz_nbclust(nb)
```

```
nb <- NbClust(train1, distance = "euclidean", min.nc = 2,
              max.nc = 20, method = "complete",index=('friedman'))
fviz_nbclust(nb)

#plotting dendogram using ward method
hc.res <- eclust(train1, "hclust", k = 3, hc_metric = "euclidean",
                 hc_method = "ward.D2", graph = FALSE)
tiff("E:/Fall 2017 courses/STAT 764 Methods of Multivariate Methods/Project/dataset/dendo1.tiff", width = 6.6, height = 4.8, units =
'in', res = 300)

fviz_dend(hc.res, show_labels = FALSE,
          palette = "jco", as.ggplot = TRUE)

dev.off()
```