A robust correlation analysis framework for imbalanced and dichotomous data with uncertainty

Chun Sing Lai [a,b], Yingshan Tao [a], Fangyuan Xu [a,*], Wing W. Y. Ng [c,*], Youwei Jia [a,d], Haoliang Yuan [a], Chao Huang [a], Loi Lei Lai [a,*], Zhao Xu [d], Giorgio Locatelli [b]

[a] Department of Electrical Engineering, School of Automation, Guangdong University of Technology, Guangzhou 510006, China
[b] School of Civil Engineering, Faculty of Engineering, University of Leeds, Woodhouse Lane, Leeds LS2 9JT, U.K.
[c] Guangdong Provincial Key Lab of Computational Intelligence and Cyberspace Information, School of Computer Science and Engineering, South China University of Technology, Guangzhou 510630, China
[d] Department of Electrical Engineering, The Hong Kong Polytechnic University, Hong Kong SAR, China

Abstract— Correlation analysis is one of the fundamental mathematical tools for identifying dependence between classes. However, the accuracy of the analysis could be jeopardized due to variance error in the data set. This paper provides a mathematical analysis of the impact of imbalanced data concerning Pearson Product Moment Correlation (PPMC) analysis. To alleviate this issue, the novel framework Robust Correlation Analysis Framework (RCAF) is proposed to improve the correlation analysis accuracy. A review of the issues due to imbalanced data and data uncertainty in machine learning is given. The proposed framework is tested with in-depth analysis of real-life solar irradiance and weather condition data from Johannesburg, South Africa. Additionally, comparisons of correlation analysis with prominent sampling techniques, i.e., Synthetic Minority Over-Sampling Technique (SMOTE) and Adaptive Synthetic (ADASYN) sampling techniques are conducted. Finally, K-Means and Wards Agglomerative hierarchical clustering are performed to study the correlation results. Compared to the traditional PPMC, RCAF can reduce the standard deviation of the correlation coefficient under imbalanced data in the range of 32.5% to 93.02%.

Keywords— Pearson product-moment correlation, imbalanced data, clearness index, dichotomous variable.

1. Introduction

With the exponential increase of the amount of data introduced by an increasing number of physical devices, the large-scale advent of incomplete and uncertain data is inevitable, such as those from smart grids (Lai and Lai, 2015; Wu et al., 2014). For sparse data, the number of data points is inadequate for making a reliable judgement. This has been an issue for the successful delivery of megaprojects (Locatelli et al., 2017). In machine learning and data mining applications, redundant data can seriously deteriorate the reliability of models trained from the data.

Data uncertainty is a phenomenon in which each data point is not deterministic but subject to some error distributions and randomness. This is introduced by noise and can be attributed to inaccurate data readings and collections. For example, data produced from GPS equipment are of uncertain nature. The data precision is constrained by the technology limitations of the GPS device. Hence, there is a need to include the mean value and variance in the sampling location to indicate the expected error. A survey of state-of-the-art solutions to imbalanced learning problems is provided in (He and Garcia, 2009). The major opportunities and challenges for learning from imbalanced data are also highlighted in (He and Garcia, 2009).

* Corresponding authors.
E-mail addresses: c.s.lai@leeds.ac.uk (C.S. Lai), yings_tao@foxmail.com (Y. Tao), datuan12345@hotmail.com (F. Xu), wingng@ieee.org (W.W.Y. Ng), corey.jia@connect.polyu.hk (Y. Jia), hunteryuan@126.com (H. Yuan), chao.huang@my.cityu.edu.hk (C. Huang), l.l.lai@ieee.org (L.L. Lai), eezhaoxu@polyu.edu.hk (Z. Xu), g.locatelli@leeds.ac.uk (G. Locatelli)

The number of publications on imbalanced learning has increased by 20 times from 1997 to 2007. Imbalanced data can be classified into two categories, namely, intrinsic and extrinsic imbalanced. Intrinsic imbalance is due to the nature of the data space, whereas extrinsic imbalance is not. Given a dataset sampled from a continuous data stream of balanced data with respect to a specific period of time; if the transmission has irregular disturbances that do not allow the data to be transmitted during this period of time, the missing data in the dataset will result in an extrinsic imbalanced situation obtained from a balanced data space. An example of intrinsic imbalanced could be due to the difference in the number of samples of different weather conditions, i.e., in general, the 'Clear' weather condition has the most occurrences throughout the year, whereas 'Snow' may only have a few occurrences.

There is a growth of interest in class imbalanced problems recently due to the classification difficulty caused by the imbalanced class distributions (Wang and Yao, 2012; Xiao et al., 2017). To solve this problem, several ensemble methods have been proposed to handle such imbalances. Class imbalances degrade the performance of the derived classifier and the effectiveness of selections to enhance classifier performance (Malof et al., 2012).

This paper proposes and validates a new framework for the impact of imbalanced data on correlation analysis. The impact of imbalanced data is described using a mathematical formulation. Additionally, RCAF is proposed for correlation analysis with the aim of reducing the negative effects due to an imbalanced ratio. This will be investigated with a theoretical and real-life case study.

Section 2 provides a literature review on the imbalanced data problem, followed by the correlation analysis of imbalanced data. Section 3 provides an overview of the critical features and the impacts on correlation analysis. Simulations will be conducted to support the findings. Section 4 proposes a new framework for the correlation analysis. Section 5 provides a real-life case study, based on solar irradiance and weather conditions, to evaluate the new framework. Different imbalanced data sampling techniques will be used to compare the correlation analysis performance. Cluster analysis of weather conditions will be given to understand the implications of the correlation results. Future work and conclusions will be given in Section 6.

2. Correlation analysis and imbalanced data

2.1. Imbalanced classification problems

Imbalanced data refers to unequal variable sampling values in a dataset. For example, 90% of sampling data can be in the majority class, with only 10% of the sampling data in the minority class. Therefore, the imbalanced ratio is 9:1. Imbalanced data appears in many research areas. As mentioned in (Krstic and Bjelica, 2015), when TV recommender systems perform well, the number of interactions for users to express positive feedback is anticipated to be greater than the number of negative interactions on the recommended content. This is known as class imbalanced. The misclassification of the unwanted content can be recognized by TV viewers easily, therefore, system performance could decrease.

Commonly, modifying imbalanced datasets to provide a balanced distribution is carried out using sampling methods (Li et al., 2010; Liu et al., 2009; Wang and Yao, 2012). From a broader perspective, over-sampling and under-sampling techniques seem to be functionally equivalent, since they both can provide the same proportion of balance by changing the size of the original dataset. In practice, each technique introduces challenges that can affect learning. The major issue with under-sampling is straightforward, classifiers will miss important information in respect to the majority class, by removing examples from the majority class (Ng et al., 2015). The issues regarding over-sampling are less straightforward. Since over-sampling adds replicated data to the original dataset, multiple instances of certain samples become 'tied', resulting in overfitting. As proposed in (Mease et al., 2007), one solution to the over-sampling problem is to add a small amount of random noise to the predictor so the replicates are not duplicated, which can minimize overfitting. This jittering adds undesirable noise to the dataset but the negative impact of imbalanced datasets has been shown to be reduced. Under-sampling is a favoured technique for class-imbalanced problems; it is very efficient since only a subset of the majority class is used. The main problem with this technique is that many majority class examples are ignored.

Class imbalanced learning is employed to resolve supervised learning problems in which some classes have significantly more samples than others (Xiao et al., 2017). The study of multiclass imbalanced problems and the Dynamic Sampling method (DyS) for multilayer perceptron are provided in (Lin et al., 2013). The authors claim that the DyS method could outperform the pre-sample methods and active learning methods for most datasets. However, a theoretical foundation is necessary to explain the reason a simple method such as DyS could perform so well in practice.

Support Vector Machine (SVM) is a popular machine learning technique that works effectively with balanced datasets (Batuwita and Palade, 2010; Tang et al., 2009). However, with imbalanced datasets, suboptimal classification models are produced with SVMs. Currently, most research efforts in imbalanced learning focus on specific algorithms and/or case studies. Many researchers use machine learning methods such as support vector machines (Batuwita and Palade, 2010), cluster analysis (Diamantini and Potena, 2009), decision tree learning (Mease et al., 2007; Weiss and Provost, 2003), neural networks (Yeung et al., 2016; Zhang and Hu, 2014; Zhou and Liu, 2006), etc., with a mixture of over-sampling and under-sampling techniques to overcome the imbalanced data problems (Liu et al., 2009; Seiffert et al., 2010). A novel machine learning approach to assess the quality of sensor data using an ensemble classification framework is presented in (Rahman et al., 2014), in which a cluster-oriented sampling approach is used to overcome the imbalance issue.

The issues of class imbalanced learning methods and how they can benefit software defect prediction are given in (Wang and Yao, 2013). Different categories of class imbalanced learning techniques, including resampling, threshold moving and ensemble algorithms, have been studied for this purpose. Medical data are typically composed of ‘normal’ samples with only a small proportion of ‘abnormal’ cases, which leads to class imbalanced problems (Li et al., 2010). Constructing a learning model with all the data in class imbalanced problems will normally result in a learning bias towards the majority class.

Imbalanced data can influence the feature selection results. As mentioned in (Zhang et al., 2016), traditional feature selection techniques assume the testing and training datasets follow the same data distribution. This may decrease the performance of the classifier for the application of adversarial attacks in cybersecurity. For real-life applications, the distribution of different datasets and variables may be significantly different and should be thoroughly studied. Feature selection based on methods such as feature similarity measure (Mitra et al., 2002), harmony search (Diao et al., 2014; Diao and Shen, 2012), hybrid genetic algorithms (Oh et al., 2004), dependency margin (Liu et al., 2015b), cluster analysis (Chow et al., 2008) has been developed. The methods have contributed to the quality enhancement of feature selection. However, the fundamental issues of the uncertainty and imbalanced ratio in datasets have not been studied.

2.2. Correlation analysis for imbalanced data problems

Many correlation analyses have been conducted on imbalanced datasets. For example, Community Question Answering (CQA) is a platform for information seeking and sharing. In CQA websites, participants can ask and answer questions. Feedback can be provided in the manner of voting or commenting. (Yao et al., 2015) proposed an early detection method for high-quality CQA questions/answers. Questions of significant importance that would be widely recognized by the participants can be identified. Additionally, helpful answers that would attain a large amount of positive feedback from participants can be discovered. The correlation of questions and answers was performed with Pearson R correlation to test the dependency of the voting score. The classification accuracy with imbalanced data, i.e., the ratio between the number of data for positive and negative feedbacks have not been addressed.

Gamma coefficient is a well-known rank correlation measure that is frequently used to quantify the strength of dependency between two variables in ordinal scale (Ruiz and Hüllermeier, 2012). To increase the robustness of this measure in data with noise, Ruiz et al. (Ruiz and Hüllermeier, 2012) studied the generalization of the gamma coefficient based on fuzzy order relations. The fuzzy gamma has been shown to be advantageous in the presence of noisy data. However, the authors did not consider the imbalanced data issue for correlation analysis.

In clinical studies, the linear correlation coefficient is frequently used to quantify the dependency between two variables, e.g., weight and height. The correlation can indicate if a strong dependency exists. However, in practice, clinical data consists of a latent variable with the addition of an inevitable measurement error component, which affects the reproducibility of the test. The correlation will be less than one even if the underlying physical variables are perfectly correlated. Francis et al. (Francis et al., 1999) studied the reduction in correlation due to limited reproducibility. The implications of experimental design and interpretation were also discussed. It is confirmed that with large measurement errors, the measured correlation for perfectly correlated variables cannot be equal to one but must be less than one (Francis et al., 1999). Francis et al. (Francis et al., 1999) described a method which allows this effect to be quantified once the reproducibility of the individual measurements is known. However, the paper has not resolved the correlation inaccuracy problem and only provides an indication of the effect of noise on the correlation in an imbalanced dataset. The paper concludes that the designers of experiments can relieve the problem of attenuation of correlation in two ways. First, the random component of the error should be minimized, with the aim of improving reproducibility. Technical advances may allow this to occur, but relying on them is not always practical. Random measurement error can also be attenuated statistically but this requires care and logical judgement. Note that some variance errors in the data are inevitable, such as solar irradiance where unexpected phenomenon such as birds flying cannot be avoided.

3. Impact of imbalanced ratio and uncertainty on correlation analysis

Classes exist in various machine learning models and can be in the form of dichotomous variables. The features can be represented by binary classification, i.e., 0 or 1. For example, different weather conditions for solar irradiance prediction can be classified (0 for ‘Clear’ and 1 for ‘Rain’).

3.1. Correlation analysis for imbalanced dichotomous data with uncertainty introduced by noise

In statistical analysis, dependency is defined as the degree of statistical relationship between two sets of data or variables. Dependency can be calculated and represented by correlation analysis. The most commonly used formula is parametric and known as the Pearson Product Moment Correlation (PPMC) coefficient. By definition, the PPMC coefficient has a range from the perfect negative correlation of negative 1.0 to the perfect positive correlation of positive 1.0, with 0 representing no correlation (Mitra et al., 2002).

The following problem is used to describe this research issue.

Assumption: Given two variables X and Y, where $X = \{x_a, x_b\}, Y \in \mathbb{R}_0^+$. In the obtained sampling dataset, the number of samples in x_a is n_a and the number of samples in x_b is n_b, with $n_a + n_b = N$. The noise, i.e., sampling error, occurs in Y. The relationship between each value of Y (y_i) and each value of X (x_i) is $y_i = f(x_i) + Err_i$, $i = \{a, b\}$. Each noise Err_i follows a certain distribution K with mean error μ_{me}. The square of noise error Err_i^2 follows the distribution L with mean square error μ_{mse}.

Fig. 1 presents the PPMC correlation with a variable, i.e., weather being dichotomous. The regression line depicts a negative correlation between Clearness Index (CI) and the two weather conditions. This means the weather transition from ‘Clear’ to ‘Mostly Cloudy’ will reduce the amount of solar resources received.

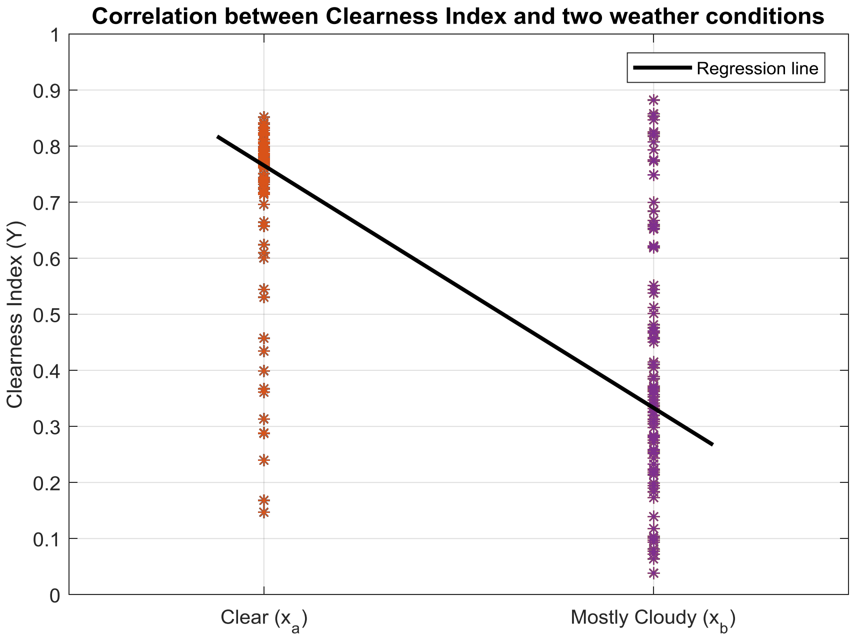


Fig. 1. Correlation analysis with a dichotomous variable.

The PPMC coefficient is given in Equation (1) below:

$$\rho_{XY} = \frac{A-B}{C*D} \tag{1}$$

$$\begin{cases} A = (n_a + n_b) \sum_{i=1}^{n_a+n_b} x_i y_i \\ B = \sum_{i=1}^{n_a+n_b} x_i \, . \sum_{i=1}^{n_a+n_b} y_i \\ C = \sqrt{(n_a + n_b) \sum_{i=1}^{n_a+n_b} x_i^2 - \left(\sum_{i=1}^{n_a+n_b} x_i \right)^2} \\ D = \sqrt{(n_a + n_b) \sum_{i=1}^{n_a+n_b} y_i^2 - \left(\sum_{i=1}^{n_a+n_b} y_i \right)^2} \end{cases}$$

For C to become zero, possible factors include $n_a + n_b = 0$ and all x are zero. Based on Fig. 1, if there is no data, i.e., $n_a + n_b$ and the sample size is zero, it is impossible to conduct the correlation. All x equal to zero signifies there is no value in the variable. Similarly, for D to become zero, possible factors include $n_a + n_b = 0$ and all y are zero. The average value of the sampling set is equal to the expectation of the distribution. Equation (2) depicts this relationship while Equations (3) and (4) are true.

$$\begin{cases} \mu_{me} = \dfrac{\sum_{i=1}^{N} Err_i}{N} \\ \mu_{mse} = \dfrac{\sum_{i=1}^{N} Err_i^2}{N} \end{cases} \quad (2)$$

$$\frac{\sum_{i=1}^{n_a} Err_i}{n_a} = \frac{\sum_{i=1}^{n_b} Err_i}{n_b} \quad (3)$$

$$\frac{\sum_{i=1}^{n_a} Err_i^2}{n_a} = \frac{\sum_{i=1}^{n_b} Err_i^2}{n_b} \quad (4)$$

By considering $y_i = f(x_i) + Err_i$ in Equation (1), further expressions are presented in Equation (5).

$$\begin{cases} A - B = n_a n_b (x_a - x_b)[f(x_a) - f(x_b)] \\ C = \sqrt{n_a n_b (x_a - x_b)^2} \\ D = \sqrt{n_a n_b [f(x_a) - f(x_b)]^2 + (n_a - n_b)^2 . (\mu_{mse} - \mu_{me}^2)} \end{cases} \quad (5)$$

By considering $n_b = \alpha * n_a$, where α is the number ratio between value x_a and value x_b, Equation (5) can be transformed into Equation (6).

$$\begin{cases} \rho_{XY} = \dfrac{A - B}{C * D} = \dfrac{x_a - x_b}{|x_a - x_b|} . \dfrac{f(x_a) - f(x_b)}{|f(x_a) - f(x_b)|} . R \\ R = \dfrac{1}{\sqrt{1 + \dfrac{\mu_{mse} - \mu_{me}^2}{[f(x_a) - f(x_b)]^2} . \left(\dfrac{1}{\alpha} + \alpha + 2\right)}} \end{cases} \quad (6)$$

If $x_a \neq x_b$ and $f(x_a) \neq f(x_b)$, the type of correlation can be expressed by Equation (7).

$$\rho_{XY} \begin{cases} R, \left(x_a < x_b, f(x_a) < f(x_b)\right) \\ -R, \left(x_a < x_b, f(x_a) > f(x_b)\right) \end{cases} \quad (7)$$

Equation (6) shows the correlation may not be +1/-1 given there is an increasing/decreasing linear relationship between X and Y. It is also related to the Momentum Ratio R. For the case $f(x_a) = f(x_b)$, based on Fig. 1, this means the “actual” (excluding error variance) CI for ‘Clear’ is the same as the actual CI for ‘Mostly Cloudy’. Since the variance of Y is zero, the denominator is zero which makes the correlation coefficient undefined.

3.2. Impact of imbalanced ratio

The imbalanced ratio in the dataset is presented by α in Equation (7). Equation (8) extracts the section of R in Equation (7) as given below:

$$coe_\alpha = \frac{1}{\alpha} + \alpha + 2 \tag{8}$$

In Equation (8), the minimum point occurs at α = 1. This indicates R is maximized if the sampling dataset contains an equal number of x_a and x_b. In this section, two functions are employed to study the imbalanced datasets and the correctness of Equation (7). Equation (9) introduces the two functions. The error of each sampling point is assumed to follow a standard normal distribution $N(0,1)$. The first function in Equation (9) establishes a negative relationship while the second function establishes a positive relationship. The correlation can be computed using two methods. Method 1 uses the derived Equation (7) and Method 2 uses the conventional Equation (1).

$$\begin{pmatrix} x_a = 1 \\ x_b = 2 \end{pmatrix} \begin{cases} fun_1 : y = \sin\left(\frac{\pi}{2} x\right) + Err \\ fun_2 : y = \ln(x) + Err \end{cases} \tag{9}$$

Fig. 2 shows the simulation results for the two functions in Equation (9). n_b is fixed at 100 and a sensitivity analysis is conducted for n_a from 1 to 3000. For Function 2, the correlation absolute value increases from 1 to 100 and decreases from 100 to 3000. This shows that Method 1 and Method 2 produce similar results. The simulations in Fig. 2 have proved that Equation (7) is valid. The maximum absolute value of the correlation occurs at n_a = n_b = 100, where α = 1.

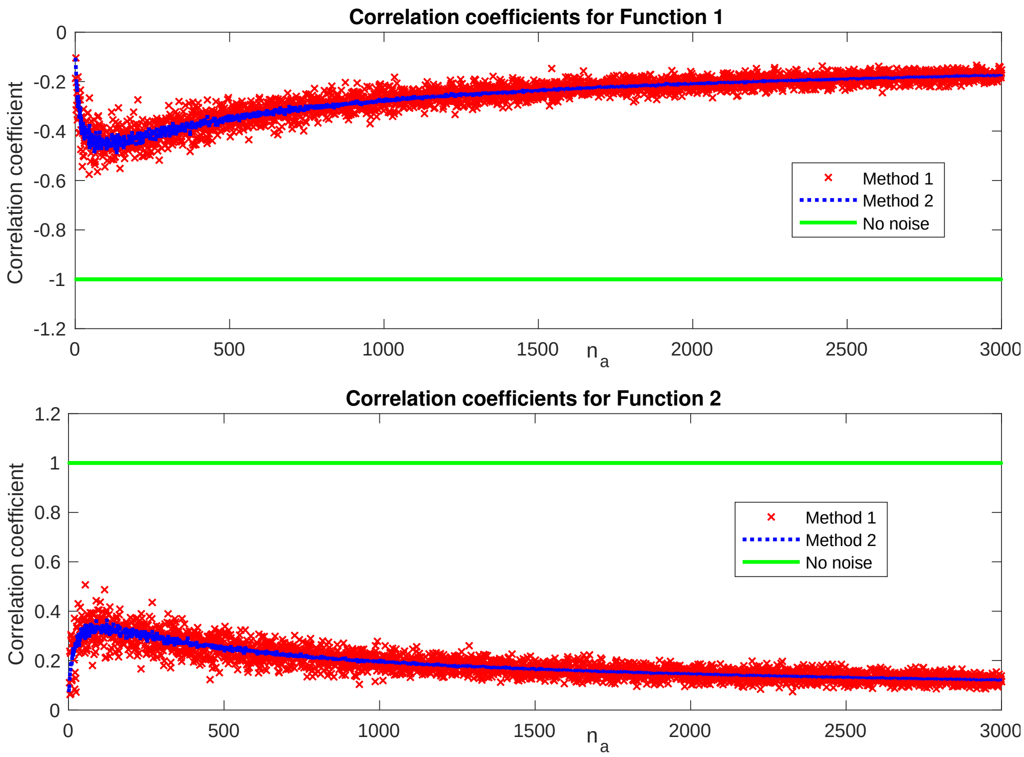


Fig. 2. Correlation for the two functions with imbalanced dataset.

Fig. 2 indicates that although variables X and Y have a confirmed dependence, the correlation may be distorted by imbalanced data. The reason the correlations obtained from Method 1 have more fluctuations than Method 2 is due to the assumption made with Equation (2). A general recognition of correlation with high dependency is usually between 0.7 and 1.0, neutral dependency is between 0.3 and 0.7, and low dependency is between 0 and 0.3. However, for Function 2 in Equation (9), the correlation reaches 0.12 when n_a is 3000 (α = 30), which is far from the maximum value 0.37. This may misinterpret the correlation from 'neutral dependency' to 'low dependency'. The optimal correlation can be realized when the datasets have equal sizes.

3.3. Impact of noise

The contribution of noise to the correlation is presented by Equation (10). Noise represents an unconsidered impact that can cause deviation from the actual value of a variable, which contributes to variance error. It can be recognized as the inaccuracy of measured data.

$$coe_{noise} = \mu_{mse} - \mu_{me}^2 \tag{10}$$

As shown in Equation (7), correlation may be distorted by the imbalanced ratio, with an exceptional condition that coe_{noise} in Equation (10) is equal to zero. If all noise is rejected by a perfect sensor, Equation (7) indicates the correlation will not be influenced by an imbalanced ratio and the resultant Momentum Ratio becomes 1. A simulation is conducted with Equation (9) without noise. The correlation results without noise are presented in Fig. 2. The correlations of the two functions in Equation (9) are shown to be perfectly correlated, i.e., 1 (or -1) when noise does not exist. As n_a increases, the no-noise correlations maintain a value of 1 (or -1). This phenomenon indicates the imbalanced ratio does not influence correlation when noise is removed. Noise is one of the key factors that affect correlation with respect to the imbalanced ratio.

3.4. Impact of output differences

The contribution of the output difference to correlation is presented by Equation (11).

$$coe_{out\_diff} = \frac{1}{[f(x_a) - f(x_b)]^2} \tag{11}$$

In Equation (9), coe_{out_diff} decreases and R in Equation (7) increases if the difference between $f(x_a)$ and $f(x_b)$ increases. This indicates that R can be controlled by the output difference. A larger output difference can counteract the effect of an imbalanced ratio. Similar to Equation (7), for the case $f(x_a) = f(x_b)$, the correlation coefficient is undefined when the variance of Y is zero.

$$\begin{pmatrix} x_a = 1 \\ x_b = 2 \end{pmatrix} \begin{cases} fun_1\colon y = \beta.\sin\left(\frac{\pi}{2}x\right) + Err, \beta = \{1,3,6,9\} \\ fun_2\colon y = \beta.\ln(x) + Err, \beta = \{1,3,6,9\} \end{cases} \tag{12}$$

Fig. 3 presents the simulation results for Equation (12). Note that $[f(x_a) - f(x_b)]^2$ increases as β increases. In addition, the correlation at the same imbalanced ratio is closer to a strong correlation (1 or -1) with an increased β. This indicates that a larger output difference may increase R and counteract the impact of imbalance.

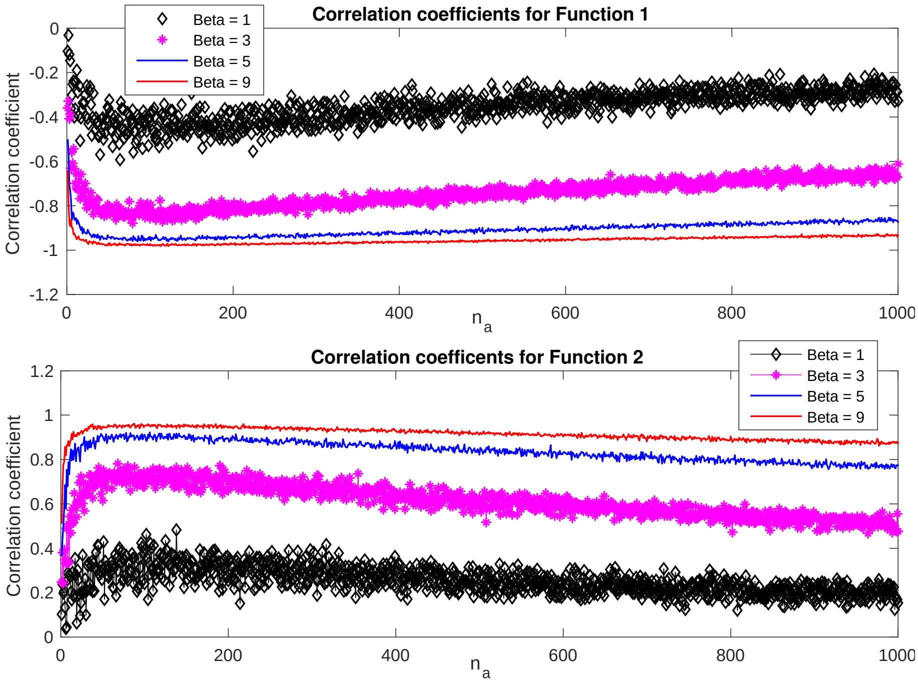


Fig. 3. Correlation on specified function with imbalanced dataset.

4. Robust correlation analysis framework

4.1. Framework

This paper introduces a novel correlation analysis framework to alleviate the negative impact of imbalanced data with noise in correlation analysis. Fig. 4 presents the structure of the framework. In Fig. 4, X has two values (x_a, x_b) in the sampling dataset. The number of data points in x_a and x_b are n_a and n_b, respectively. Each x value and its corresponding y value construct a data pair (x, y). The correlation analysis framework consists of the following two main steps:

- **Step 1: Creating groups of balanced datasets:** The first step is to determine which variable X has the largest amount of data. For example, x_a is selected if $n_a > n_b$, then, select n_b amount of x_a and combine them into pairs with x_b. In this dataset, the number of data points in x_a and x_b is equal to n_b. The procedure is repeated M times to construct a group of balanced sets. To prevent the loss of information from the removal of data and to fully utilize all the data, the method to determine M is shown in Equation (13). In the non-repeated random selector, sampling without replacement is used for sampling purposes to prevent 'tied' data. The ceil function is used to round the value M towards positive infinity.

$$M = ceil\left(\frac{n_a}{n_b}\right) \tag{13}$$

- **Step 2: Correlation integration**: Corr$_i$, which is non-zero, is the correlation of a balance set i calculated with Equation (1). Assume there are M balanced sets, the final correlation can be computed by Equation (14) as below:

$$\frac{1}{Corr_{final}^2} = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{Corr_i^2} \tag{14}$$

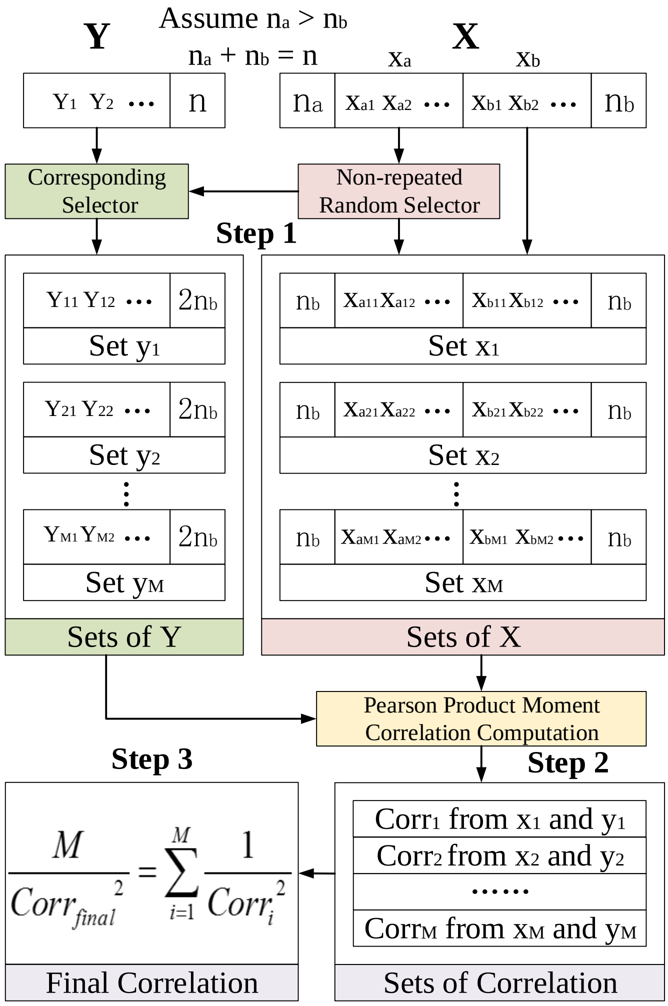
Table 1 presents the detailed algorithm for RCAF. The implementation and pseudocode were developed with MATLAB.

Table 1

Algorithm for RCAF.

Input:
$y_a = (y_{a1}, y_{a2}, y_{a3}, \dots, y_{an})$;
$y_b = (y_{b1}, y_{b2}, y_{b3}, \dots, y_{bn})$;
$n_a = \text{size}(y_a)$;
$n_b = \text{size}(y_b)$;
$x_a = \text{zeros}(n_a, 1) + 1$;
$x_b = \text{zeros}(n_b, 1) + 0$;
Output:
corr_final: PPMC for x and y
Algorithm:
If ρ_{xy} is negative % Use Eq. (1) to determine if the correlation is positive or negative.
  sign = -1;
else
  sign = +1;
end
If $n_a \geq n_b$ **then**
  $M = \text{ceil}(n_a/n_b)$;
  For $\text{counter} = 1{:}M$
    $\text{posi} = \text{randperm}(n_a, n_b)$;
    $xk = x_a(\text{posi})$;
    $yk = y_a(\text{posi})$;
    $x = [xk; x_b]$;
    $y = [yk; y_b]$;
    $\text{cori}(1, \text{counter}) = \text{corr}(x, y)$; % Eq. (1)
    $\text{cori}(1, \text{counter}) = 1./(\text{cori}(1, \text{counter}).\,\hat{}\,2)$;
  end
else
  $M = \text{ceil}(n_b/n_a)$;
  For $\text{counter} = 1{:}M$
    $\text{posi} = \text{randperm}(n_b, n_a)$;
    $xk = x_b(\text{posi})$;
    $yk = y_b(\text{posi})$;
    $x = [xk; x_a]$;
    $y = [yk; y_a]$;
    $\text{cori}(1, \text{counter}) = \text{corr}(x, y)$; % Eq. (1)
    $\text{cori}(1, \text{counter}) = 1./(\text{cori}(1, \text{counter}).\,\hat{}\,2)$;
  end
end
$\text{reg} = \text{mean}(\text{cori})$;
$\text{corr}_{\text{final}} = \text{sign} * (1./(\text{reg}.\,\hat{}\,0.5))$;

As depicted in Table 1, the computational complexity (CC) for RCAF is relatively low. According to Equation (1), the CC for PPMC is linear (Liu et al., 2016) at $O(n)$ with data size n. Since RCAF consists of converting the majority class data into M datasets, with each dataset having the size of the minority class, the CC for RCAF is approximately $O(M(\frac{n}{M}))$ or $O(n)$. Although RCAF has a higher CC due to additional computations, e.g., Equations (13) and (14) and the requirement of more data storage, the improved correlation analysis under imbalanced data can justify the use of RCAF.



$$\frac{M}{Corr_{final}^2} = \sum_{i=1}^{M} \frac{1}{Corr_i^2}$$

Fig. 4. Robust correlation analysis framework.

4.2. Proof of RCAF effectiveness

The Momentum Ratio R should be maximized as explained above. In Step 2 of RCAF, R is calculated with correlations from all balanced sets, as shown in Equation (15). μ_{mse_i} denotes the μ_{mse} of each balanced set. μ_{me_i} denotes the μ_{me} of each balanced set. α_i is α of each balanced set.

$$\frac{1}{R_{final}^2} = \frac{1}{M}\sum_{i=1}^{M}\left[1 + \frac{\mu_{mse\_i} - \mu_{me\_i}^2}{[f(x_a) - f(x_b)]^2} \cdot \left(\frac{1}{\alpha_i} + \alpha_i + 2\right)\right] \tag{15}$$

For each balanced dataset, since the number of data points in x_a and x_b are equal, $a_i = 1$. Equation (15) can be rewritten as Equation (16).

$$\frac{1}{R_{final}^2} = 1 + \frac{4}{M.[f(x_a) - f(x_b)]^2}\left(\sum_{i=1}^{M}\mu_{mse\_i} - \sum_{i=1}^{M}\mu_{me\_i}^2\right) \tag{16}$$

Assuming the sample size, i.e., n_a is large, the noise terms in Equation (16) can be expressed as Equation (17).

$$\begin{cases} \sum_{i=1}^{M} \mu_{mse\_i} = M.\mu_{mse} \\ \sum_{i=1}^{M} \mu_{me\_i}^2 = M.\mu_{me}^2 \end{cases} \tag{17}$$

By considering Equations (7), (16), and (17); Equation (18) gives the equations of R for the original correlation and the new correlation. Note that the term α disappears in the Momentum Ratio under RCAF.

$$\begin{cases} Original: \dfrac{1}{R^2} = 1 + \dfrac{\mu_{mse} - \mu_{me}^2}{[f(x_a) - f(x_b)]^2}.\left(\dfrac{1}{\alpha} + \alpha + 2\right) \\ New: \dfrac{1}{R_{final}^2} = 1 + \dfrac{\mu_{mse} - \mu_{me}^2}{[f(x_a) - f(x_b)]^2}.4 \end{cases}$$

$$\because 4 < \frac{1}{\alpha} + \alpha + 2 \tag{18}$$

$$\therefore \frac{1}{R_{final}^2} < \frac{1}{R^2}$$

$$\therefore R_{final} > R$$

4.3. Theoretical study stimulations

Base on Equation (9), the correlations under RCAF are much more stable and slanting does not occur with respect to the increase of the imbalanced ratio. Fig. 5 shows the simulation results. The imbalanced ratio increases as n_a increases. However, the correlations under RCAF do not have a large variation and the optimal value is maintained.

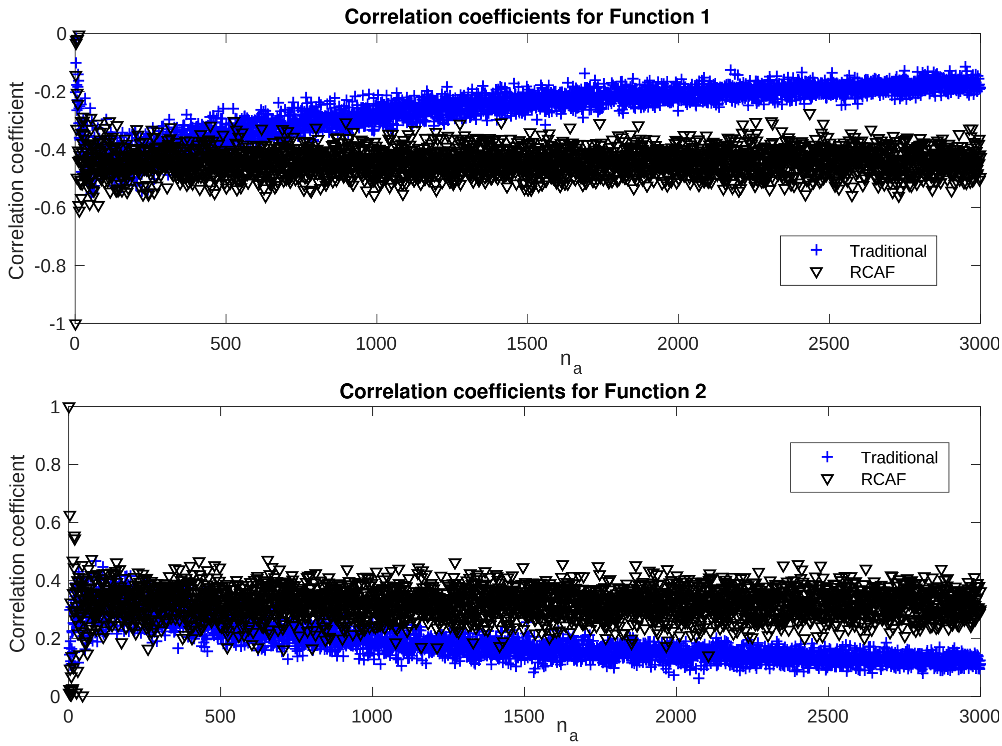


Fig. 5. Correlation comparison between traditional approach and RCAF.

5. Real-life case study: correlation for weather conditions and clearness index

5.1. Problem context and correlation analysis

Weather condition is one of the major factors affecting the amount of solar irradiance reaching earth. As a consequence, one of the most important applications affected by solar irradiance due to weather perturbation is Photovoltaic (PV) system. Weather condition changes affect the electrical power generated by a PV system with respect to time.

Using CI in Equation (19) is one method to evaluate the influence of weather conditions with respect to solar irradiance (Lai et al., 2017a). The analysis of these fluctuations with regard to solar energy applications should focus on the instantaneous CI (Kheradmanda et al., 2016; Liu et al., 2015a; Woyte et al., 2007; Woyte et al., 2006). CI can effectively characterize the attenuating impact of the atmosphere on solar irradiance by specifying the proportion of extra-terrestrial solar radiation that reaches the surface of the earth. In Equation (19) for each time of the year, $I_{pyranometer}$ is the irradiance on the surface of the earth measured with a pyranometer device and I_{model} is the clear-sky solar irradiance (Lai et al., 2017a). The CI value will be between 0 and 1, where 0 and 1 indicate no solar irradiance and the maximum amount of solar irradiance will arrive on the surface of earth, respectively. This index can be used to quantify the amount of atmospheric fluctuation based on different weather conditions.

$$CI = \frac{I_{pyranometer}}{I_{model}} \quad (19)$$

The commercial weather service website 'Weather Underground' (Weatherunderground.com, 2017) represents the weather condition using String, which is the most typically used data type. Due to the nature of climate and the hemisphere of the earth, the number of samples for each weather condition, e.g., 'Overcast' and 'Heavy Rain', is expected to be disproportional for a given location.

The data structure for the correlation analysis is presented in Table 2. The data pairs in each row represent an observation. Column 1 represents the type of weather condition, i.e., 0 and 1 for weather conditions 1 and 2, respectively. Column 2 is the CI value.

Solar irradiance data between 2009 to 2012 in Johannesburg, South Africa was collected with a SKS 1110 pyranometer sensor for the real-life case study. The solar data adopted in this work has been studied and used for solar energy system research in (Lai et al., 2017a; Lai et al., 2017b; Lai and McCulloch, 2017). The corresponding weather condition information for the solar irradiance data in Johannesburg was obtained from Weather Underground. There are 41 types of weather conditions in Johannesburg from 2009 to 2012. The sampling size of all weather conditions in Johannesburg is listed in Table 5 in the appendix. The same weather conditions can results in different CI values due to other perturbation effects that are factored

Table 2
Typical representation of a dataset for the correlation analysis.

Weather type (binary) X = 0 for weather type 1 X = 1 for weather type 2	Y = CI
1	0.71
1	0.69
0	0.43
1	0.61
0	0.32
1	0.54

out by the weather. The solar altitude angle range studied is between 0.8 and 1. The correlation results under the traditional approach and the novel correlation framework are provided in Fig. 6 and Fig. 7, respectively. The entire correlation matrix is a 41x41 square matrix.

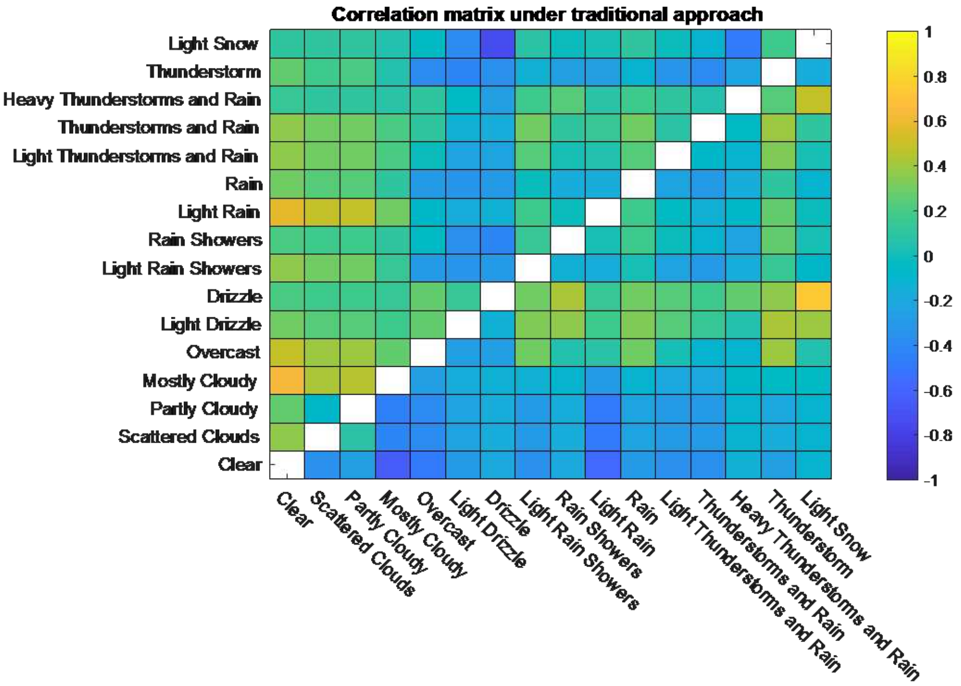


Fig. 6. Correlation matrix under traditional PPMC.

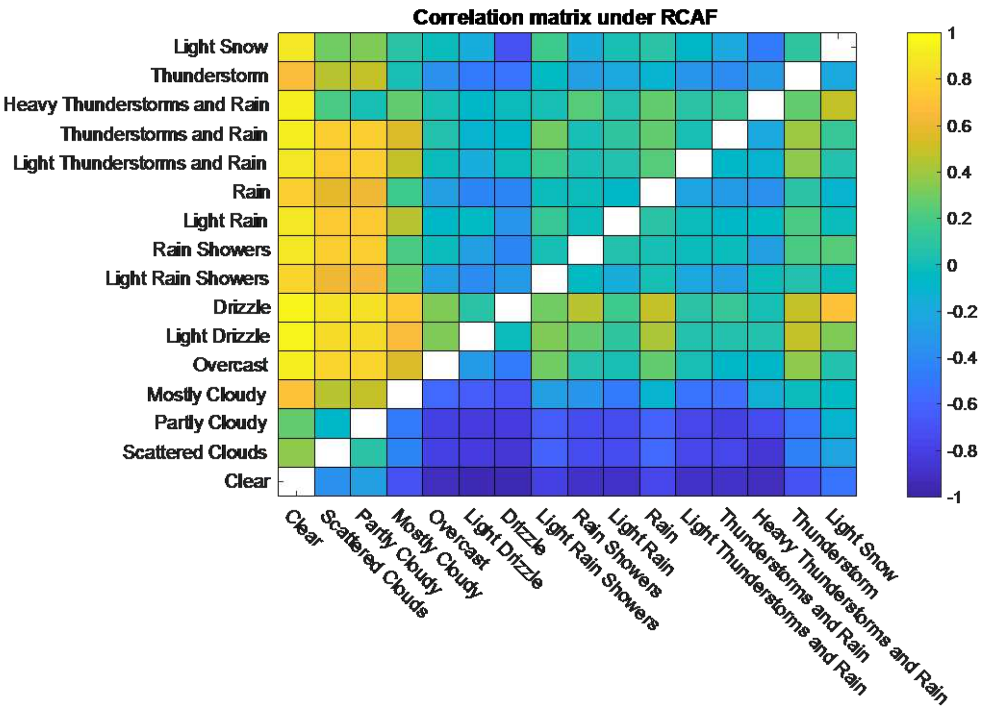


Fig. 7. Correlation matrix under RCAF.

The correlation between X and Y represents the variation of CI for the two weather transitions. A high correlation absolute value means the CI changes significantly with weather condition transitions. In contrast, if the absolute value of the correlation is low, CI changes slightly when the weather condition changes.

5.2. Clearness index and weather conditions statistical analysis

The following section of this paper examines the correlation results in Fig. 6 and Fig. 7. To understand the uncertainty and stochastic properties of CI with respect to weather conditions, it is crucial to provide statistical measures and a mathematical description of the random phenomenon for the variables.

The mean and standard deviation with error bars are presented in Fig. 8 for the weather conditions and CI for a solar altitude angle between 0.8 and 1.0. Bootstrapping is used to quantify the error in the statistics. The bootstrapped 95% confidence intervals for the population mean and standard deviation are calculated. Eight weather conditions selected from the correlation matrix are studied. The mean and standard deviation are calculated using Equations (20) and (21), respectively, for the weather conditions. s is the sample size of the weather condition. To compute the 95% bootstrap confidence interval of the mean and standard deviation, 2000 bootstrap samples are used.

$$w_{mean} = \frac{1}{s}\sum_{i=1}^{s} CI_i \quad (20)$$

$$w_{sd} = \sqrt{\frac{1}{s}\sum_{i=1}^{s}(CI_i - w_{mean})^2} \quad (21)$$

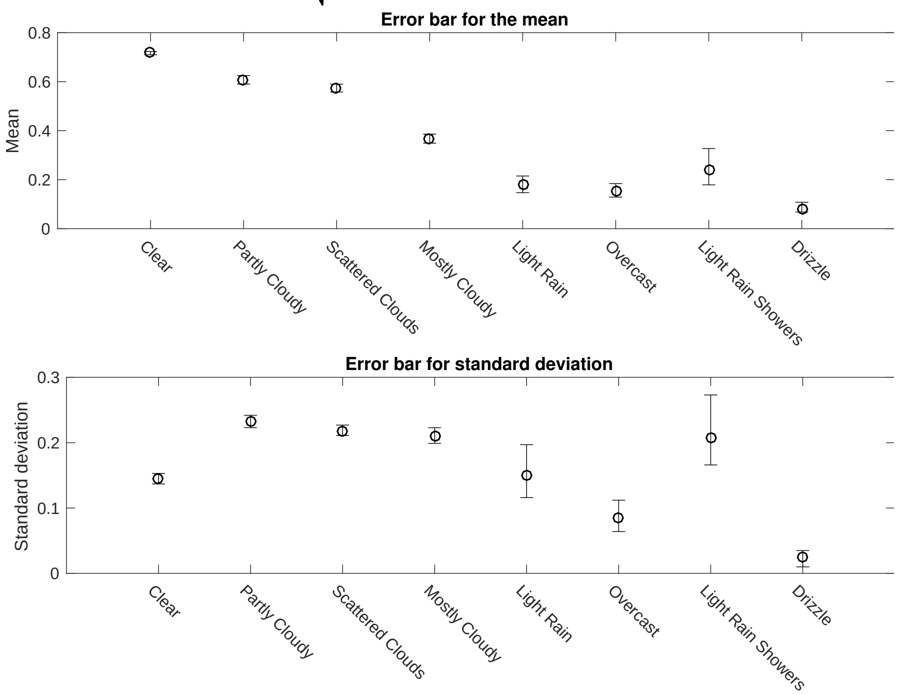


Fig. 8. Error bars for mean and standard deviation with eight types of weather conditions.

A graphical representation of the distribution of variables is presented in the histograms in Fig. 9. This effectively displays the probability distribution of CI for the weather conditions. The histogram shows that different weather conditions result in different distributions. The 'Clear' case is a monomodal distribution with a peak at 0.8 CI, whereas 'Mostly cloudy' has a peak at 0.3 CI. CIs are generally high for the 'Clear' weather condition due to the frequency of high CI occurrences. In contrast, 'Mostly Cloudy' has a high frequency of lower CI value occurrences.

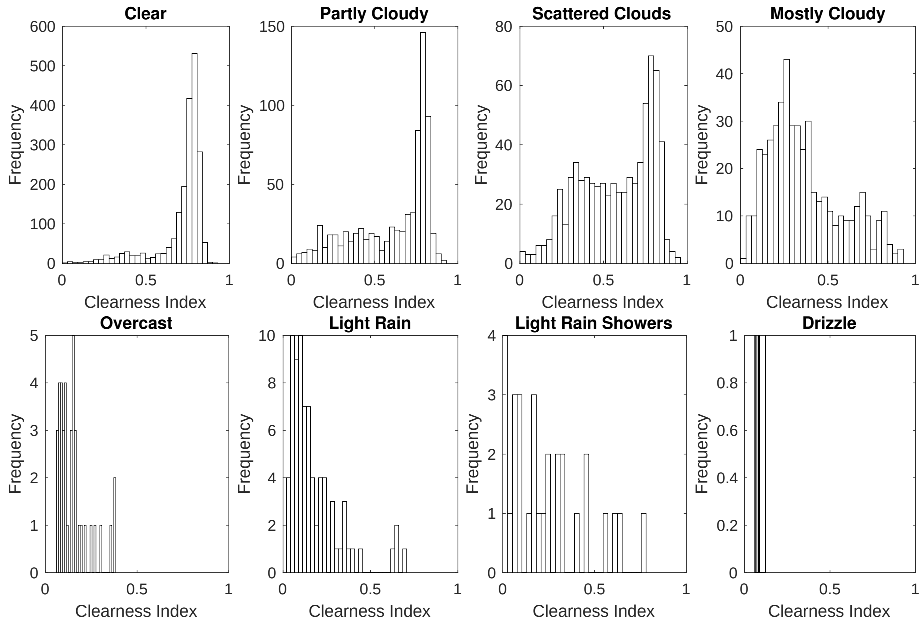


Fig. 9. Histograms of CI with respect to different weather conditions.

Due to the highly stochastic nature of CI, as shown in the histogram, it is impossible to use a parametric method where an assumption of the data distribution is made. Kernel Density Estimation (KDE) is a non-parametric method to estimate the probability density function (pdf) of a random variable. KDE is a data smoothing problem where inferences about the population are made, based on a finite data sample. Let $(x_1, x_2, \dots, x_n)$ be a sample drawn from distributions with an unknown density f. The kernel density estimator is:

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} G_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} G\left(\frac{x - x_i}{h}\right) \tag{22}$$

where n is the sample size. $G(\bullet)$ is the kernel function, a non-negative function that integrates to one and has a mean of zero. h is a smoothing parameter called the bandwidth and has the properties of h > 0.

The kernel smoothing function defines the shape of the curve used to generate the pdf. KDE constructs a continuous pdf with the actual sample data by calculating the summation of the component smoothing functions.

The Gaussian kernel is:

$$G(u) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}u^2} \tag{23}$$

Therefore, the kernel density estimator with a Gaussian kernel is:

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{j \neq i}^{n} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\left(\frac{x_j - x_i}{h}\right)^2} \tag{24}$$

The aim is to minimize the bandwidth, h. However, there is a trade-off between the bias of the estimator and its variance. In this paper, the bandwidth is estimated by completing an analytical and cross-validation procedure. The bandwidth estimation consists of two steps:

1. Use an analytical approach to determine the near-optimal bandwidth;
2. Adopt log-likelihood cross-validation method to determine the optimal bandwidth.

This adopted method has the advantage of avoiding use of the expectation maximization iterative approach to estimate the optimal bandwidth. The near-optimal bandwidth can be calculated with the analytical approach and could be further improved by using the maximum likelihood cross-validation method. This simplifies the estimation process and could potentially reduce the computational effort as this method is not an iterative approach.

a) Analytical method

For a kernel density estimator with a Gaussian kernel, the bandwidth can be estimated with Equation (25), the Silverman's rule of thumb (Silverman, 1986).

$$h = \left(\frac{4\sigma^5}{3n}\right)^{\frac{1}{5}} \approx 1.06\sigma n^{-\frac{1}{5}} \tag{25}$$

where σ is the standard deviation of the dataset. The rule of thumb should be used with care as the estimated bandwidth may produce an over-smooth pdf if the population is multimodal. An inaccurate pdf may be produced when the sample population is far from normal distribution.

b) Maximum likelihood 10-fold cross-validation method

The maximum likelihood cross-validation method was proposed by Habbema (Habbema, 1974) and Duin (Duin, 1976). In essence, the method uses the likelihood to evaluate the usefulness of a statistical model. The aim is to choose h to maximize pseudo-likelihood $\prod_{i=1}^{n} \widehat{f}_h(x_i)$.

A number of observations $x_K = \{x_1, x_2, \dots, x_k\}$ from the complete set of original observations x can be retained to evaluate the statistical model. This would provide the log-likelihood $\log\left(\hat{f}_{-k}(x_i)\right)$. The density estimate constructed from the training data is defined in Equation (26).

$$\hat{f}_{-k}(x_i) = \frac{1}{n_t h} \sum_{t \neq i}^{n_t} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\left(\frac{x_i - x_t}{h}\right)^2} \tag{26}$$

where $n_t = n - n_k$. Let n_t and n_k be the number of sample data for training and testing, respectively. The number of training data will be the number of the entire sample dataset minus the number of testing data. Since there is no preference for which observation is omitted, the log-likelihood is averaged over the choice of each omitted data sample, x_K, to give the score function. The maximum log-likelihood cross-validation (MLCV) function is given as follows:

$$MLCV(h) = \left(\frac{1}{n_k} \sum_{i=1}^{n_k} \log\left[\sum_{t \neq i}^{n_k} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\left(\frac{x_i - x_t}{h}\right)^2}\right] - \log(n_k h)\right) \tag{27}$$

The bandwidth is chosen to maximize the function $MLCV(h)$ for the given data as shown in Equation (28).

$$h_{mlcv} = \underset{h>0}{\operatorname{argmax}}\, MLCV(h) \tag{28}$$

KDE has been applied to compute the continuous pdf of CI for different weather conditions. Fig. 10 shows the density estimation with the maximum log-likelihood cross-validation method for the ‘Clear’ weather condition. The top figure shows the histogram and the density function fitted on the histogram. The bottom left figure shows the shape variation of kernel density with various bandwidths shaded in grey. The best bandwidth is highlighted in red. The bottom right figure shows the log-likelihood plot with respect to the bandwidth. The red circle identifies the bandwidth with the highest log-likelihood. The cross-validated pdf has a good fit with the histogram and has been confirmed with the log-likelihood. The optimal bandwidth estimation approach is shown to be effective and the density function gives a good representation of the histogram. The optimal bandwidth for the weather conditions can be found in Table 3.

Table 3
Optimal bandwidth for PDFs.

Weather condition	Optimal bandwidth h
‘Clear’	0.0124
‘Partly Cloudy’	0.0132
‘Scattered Clouds’	0.0224
‘Mostly Cloudy’	0.0313
‘Light Rain’	0.0316
‘Overcast’	0.0291
‘Light Rain Showers’	0.1023
‘Drizzle’	0.0260

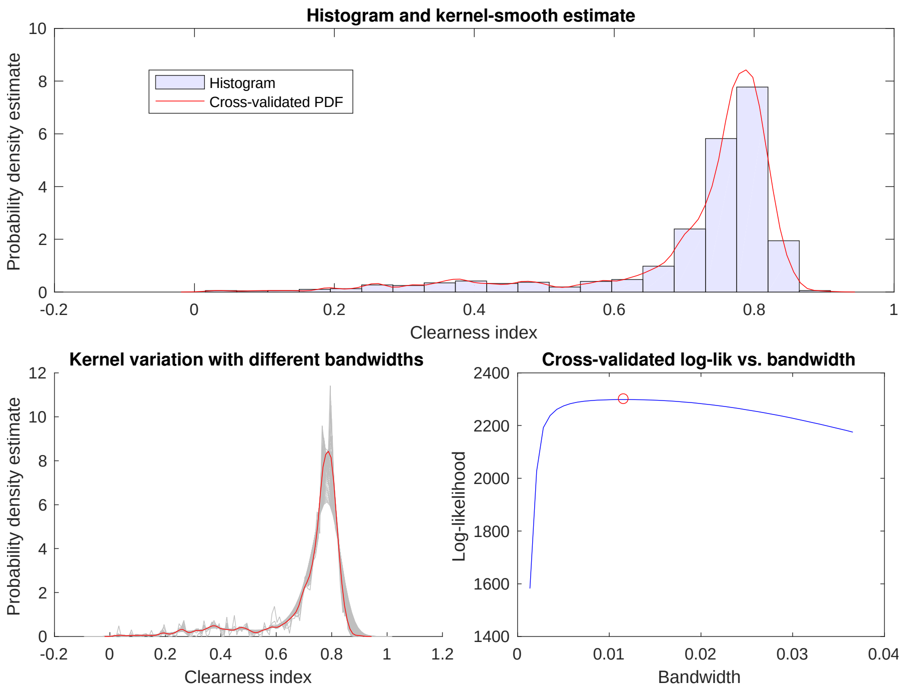


Fig. 10. Kernel density estimation for ‘Clear’.

The pdfs produced using KDE for the eight weather conditions are given in Fig. 11. Note that the pdf (such as for ‘Light rain’) could be in the range of negative CI due to the nature of a fitted function. In practice, CI cannot be negative as this means the irradiance will have a negative value. This will give a negative value for solar power estimation. Hence, negative CI values should not be considered.

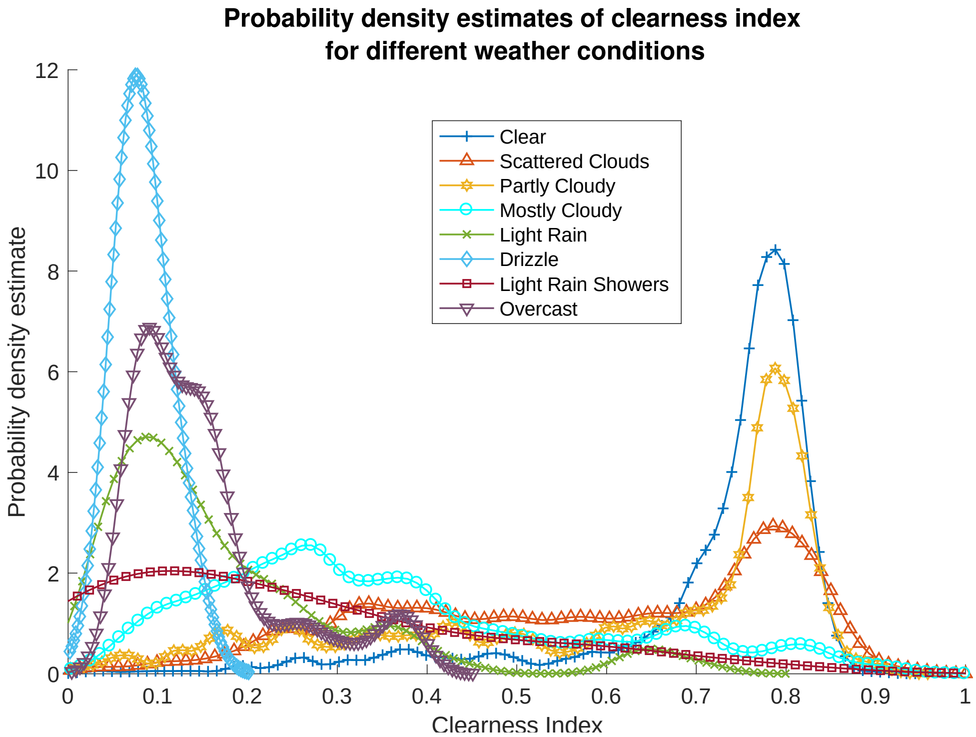


Fig. 11. PDF for various weather conditions.

5.3. Comparison of sampling techniques in correlation analysis

To compare the proposed framework with previous sampling methods for correlation analysis, the prominent sampling techniques: Synthetic Minority Over-Sampling Technique (SMOTE) and Adaptive Synthetic (ADASYN) sampling are employed in this study. SMOTE (Chawla et al., 2002) was introduced in 2002 and is an over-sampling technique with K-Nearest Neighbours (KNN). First, the KNN is considered for a sample of the minority class. To create an additional synthetic data point, the difference between the sample and the nearest neighbour is calculated and multiplied with a random number between zero and one. The randomly generated synthetic data point will be within the two specific samples. In 2008, He et al. (He et al., 2008) introduced ADASYN for over-sampling of the minority class. ADASYN is an improved technique that uses a weighted distribution for individual minority class samples depending on their level of learning difficulty. As such, additional synthetic samples are generated for minority class samples that are more difficult to learn. SMOTE generates an equal number of synthetic data points for each minority sample.

In this study, the number of nearest neighbours for SMOTE is produced according to the imbalanced ratio, as this suggests the number of data points needs to be generated. If the number of nearest neighbours for over-sampling is greater than five, under-sampling by randomly removing samples in the majority class will be similar; as the number of nearest neighbours would be too large for effective sampling (Chawla et al., 2002). In this work, the K-Nearest Neighbours for both ADASYN and SMOTE are considered to be five, which is the value used in the original work.

The constructed pdfs in Fig. 11 are useful for studying PPMC with different sampling methods. A sensitivity analysis is conducted to provide comparisons of the traditional approach and the RCAF approach. Data are generated from the pdf with random sampling. The aim of this analysis is to understand the influence of the variation of dataset size on correlation results. The size of the dataset for each weather condition, at a solar altitude angle between 0.8 and 1.0, is given in Table 5 in the appendix. The dataset size for ‘Clear’ is determined to be 1993 data points. A range of samples from 1 to 1993 is generated from the 'Clear' pdf to study the impact of imbalanced data on correlation. Seven weather conditions are studied for this purpose. The dataset size for the seven weather conditions is fixed throughout the analysis. As shown in Fig. 12, the correlation calculated with one data point for RCAF, SMOTE-under sampling, and under sampling is at perfect correlation, i.e., 1. This can be explained by the fact that the correlation between two data points at two different classes (except for the case where the two data points are equal) will be a perfect positive or perfect negative correlation.

As expected, the traditional PPMC and RCAF correlation at the end of the sensitivity analysis given in Fig. 12 can refer to the correlation of the correlation matrices in Fig. 6 and Fig. 7. The deviation between the correlation for all methods increases as the imbalanced ratio increases. This is also shown in Table 4. Additionally, the high standard deviation and mean error in Fig. 8 can result in a larger sampling range, and consequently will result in increased correlation inaccuracy.

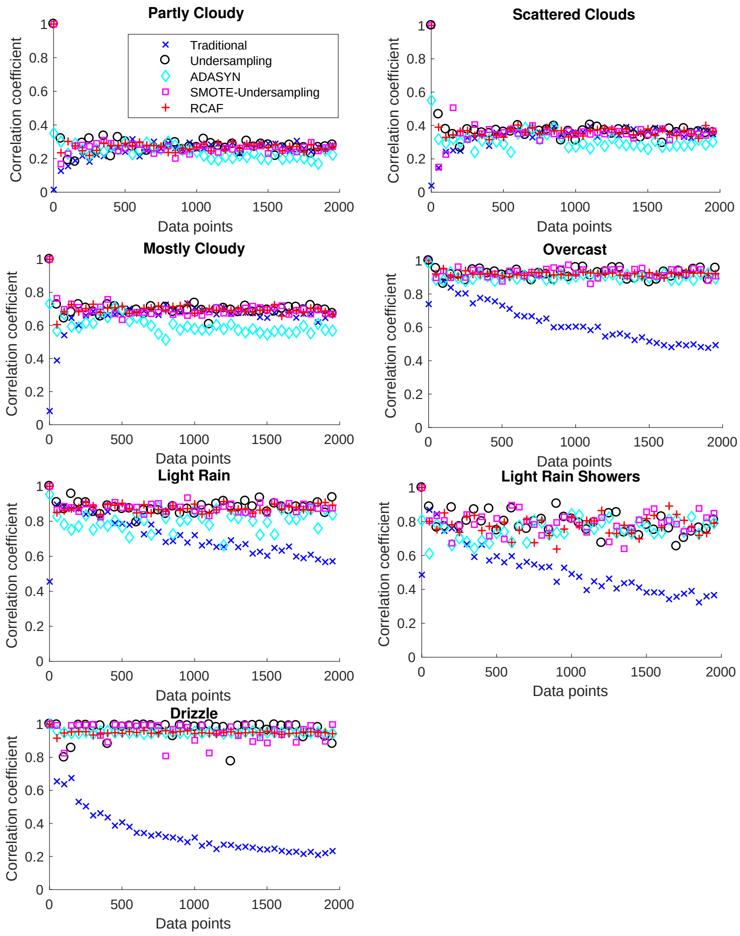


Fig. 12. Sensitivity analysis of correlation with no sampling (traditional) and different sampling methods.

The correlation reaches a steady state as the imbalanced ratio decreases, where the imbalanced ratio will have an insignificant effect on correlation in the traditional approach. The SMOTE-Under-sampling and ADASYN sampling methods are competitive with the proposed RCAF. However, SMOTE may generate data between the inliers and outliers. ADASYN focuses on generating more synthetic data points for difficult trained samples, and may focus on generating from the outlier samples and deteriorate the correlation. (Amin et al.,

2016) suggests the previous sampling techniques should investigate outliers for optimal performance.

To quantify the variation in correlation with imbalanced data, Table 4 presents the standard deviation of the correlations with respect to different methods, as presented in Fig. 12. The correlation with one sample data is excluded in the standard deviation calculation, since it can be considered an outlier as explained above.

Table 4
Standard deviation of correlation coefficients with imbalanced data.

	Traditional	Under-sampling	ADASYN	SMOTE-Under-sampling	RCAF	Percentage difference between Traditional and RCAF (%)
'Partly Cloudy'	0.040	**0.026**	0.049	0.036	0.027	32.50
'Scattered Clouds'	0.047	0.030	0.035	0.035	**0.023**	51.06
'Mostly Cloudy'	0.057	0.025	0.041	0.030	**0.018**	68.42
'Overcast'	0.129	0.029	0.016	0.024	**0.012**	90.70
'Light Rain'	0.095	0.029	0.051	0.026	**0.020**	78.95
'Light Rain Showers'	0.122	0.066	0.069	0.050	**0.048**	60.66
'Drizzle'	0.129	0.069	**0.008**	0.044	0.009	93.02

5.4. Cluster analysis of weather conditions

Classes with high correlation should be separated and in contrast, classes with weak correlation should be clustered together. According to the rule of thumb, a correlation less than 0.3 (Ratner, 2009) is considered a weak correlation. As shown in Fig. 6 and considering the case for 'Clear', i.e., column for 'Clear', most of the correlations under the traditional approach are in the range 0 - 0.3. This signifies they can be clustered as one weather group. However, the correlations computed with RCAF, as shown in Fig. 7, signify that only two other weather conditions, i.e., 'Partly Cloudy' and 'Scattered Clouds', are weakly correlated with 'Clear'. The following section of the paper employs two clustering approaches, K-Means and Ward's Agglomerative hierarchical clustering, to cluster weather conditions and understand the implications of the correlation results. However, since the number of data points is different for the weather conditions, the mean calculated with Equation (20) is used to duplicate an equal amount of data points to match the majority class, i.e., 'Clear', for cluster analysis.

K-Means is an iterative unsupervised learning algorithm for clustering problems. The basis of the algorithm is to allocate the data point to the nearest centroid. The centroid is calculated as the mean value; based on the data in the cluster at the current iteration. The K-Means algorithm with Euclidean distance for time-series clustering can be referred to (Lai et al., 2017a). The K-Means clustering results for weather conditions with K=2 is shown in Fig. 13. As shown, the CIs are generally higher for 'Clear', 'Partly Cloudy' and 'Scattered Clouds' conditions. Due to the insufficient amount of data in minority classes, e.g., 'Partly Cloudy', the values after the 740th data point will be denoted with the mean value of its dataset. The mean value will not deteriorate the clustering results since the K-Means algorithm calculates the centroid as the mean value.

Fig. 13. K-Means clustering results for weather conditions.

In Ward's Agglomerative hierarchical clustering (Murtagh and Legendre, 2014), the clustering objective is to minimize the error sum of squares, where the total within-cluster variance is minimized. At each iteration, pairs of clusters are merged which leads to a minimum increase in total within-cluster variance. The results for the hierarchical clustering of weather conditions are depicted in Fig. 13. The weather conditions can be separated into two major branches with 'Scattered Clouds', 'Partly Cloudy', and 'Clear' as one cluster. The results are consistent with the correlation results from RCAF.

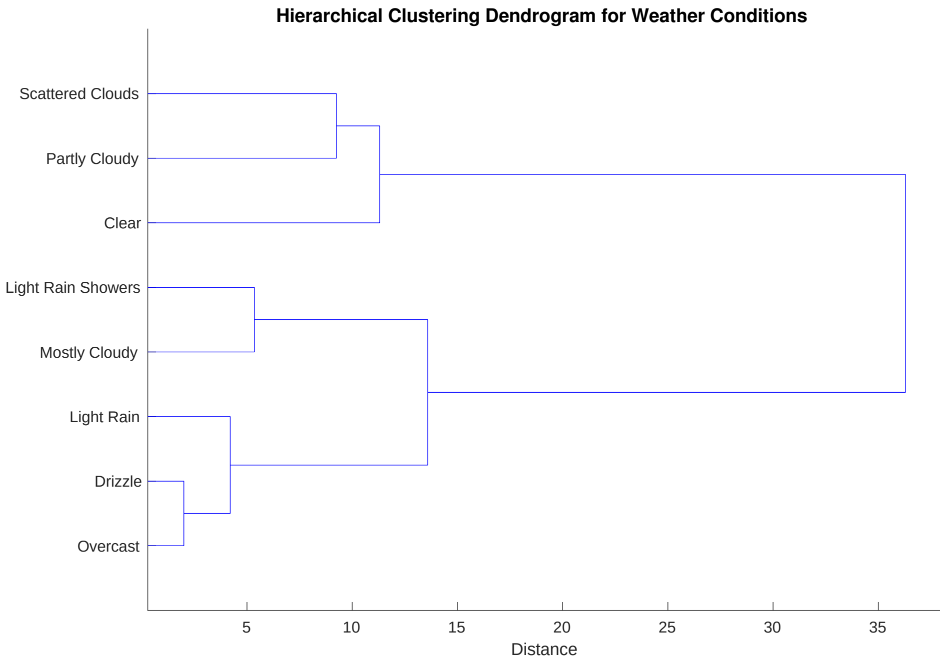


Fig. 14. Ward's Agglomerative hierarchical clustering results for weather conditions.

6. Future work and conclusions

6.1. Future work

The absolute value of the correlation may be very high if the sample size is extremely low, such as the case for 'Heavy drizzle' in which only one data point is available. The correlation of 'Heavy drizzle' under RCAF becomes 1 while the coefficient is less than 0.1 using the traditional approach. Numerous small sample balanced datasets are created in RCAF. A challenging research question that remains is that a severe lack of data points can be an issue for the correlation analysis. The limitations of RCAF and methods to overcome such issues need to be investigated.

The theoretical study of the imbalanced data effect on PPMC for continuous variables should be a focus in future work. This may provide a broader application in PPMC analysis and the method may be generalized.

The study of imbalanced data and noise in rank-order correlations will greatly benefit exploring relationships involving ordinal variables. PPMC measures the linear relationship between two continuous variables (it is also possible for one variable to be dichotomous as studied in this research) and Spearman-Rank measures the monotonic relationship between continuous or ordinal variables. Additionally, rank correlations such as Kendall's τ, Spearman's ρ, and Goodman's γ will be explored. Since a dichotomous variable is a special form of continuous variable, i.e., by treating the continuous data as binary values, providing a mathematical deduction for the correlation measures with continuous variable is challenging and will be future work.

6.2. Conclusions

Uncertainty and imbalanced data can adversely affect correlation results. This paper presents a study on the effects of imbalanced data with variance error in Pearson Product Moment Correlation analysis for dichotomous variables. A novel Robust Correlation Analysis Framework (RCAF) is proposed and tested to minimize correlation inaccuracy. A detailed theoretical study is provided with simulation results to determine whether RCAF is a feasible solution for real correlation problems. Based on the current study with seven weather conditions under imbalanced data, the proposed correlation methodology can reduce the standard deviation in a range from 32.5% to 93% when compared to the traditional approach. Solar irradiance data were collected with a pyranometer, and the respective weather conditions were obtained from the weather station database to examine the correlation analyses. Comparison with prominent sampling techniques were made. RCAF is a generalized technique and can be applied to other dichotomous variables for Pearson product moment correlation. This will be useful for understanding the dependency of dichotomous variables and subsequently improve the course of pattern analysis and decision making. The practical case study conducted in this paper will be useful for solar energy system operation and planning, by learning the dependency between different weather conditions in the context of clearness index.

Acknowledgements

This research work was supported by the Guangdong University of Technology, Guangzhou, China under Grant from the Financial and Education Department of Guangdong Province 2016[202]: Key Discipline Construction Programme; the Education Department of Guangdong Province: New and Integrated Energy System Theory and Technology Research Group, Project

Number 2016KCXTD022 and National Natural Science Foundation of China under Grant Number 61572201.

Appendix

Table 5
Complete list of weather conditions and number of samples (bad data rejection included).

Weather condition	Number of data points	
	Full	Solar altitude angle between 0.8 and 1
Clear	32626	1993
Partly Cloudy	5947	740
Scattered Clouds	5373	716
Mostly Cloudy	4631	470
Haze	2350	0
Unknown	1982	0
Light Rain	1097	76
Light Rain Showers	550	30
Smoke	534	0
Overcast	516	39
Light Thunderstorms and Rain	476	21
Mist	460	0
Thunderstorms and Rain	335	19
Rain	209	20
Thunderstorm	181	18
Fog	178	0
Light Drizzle	169	10
Rain Showers	120	6
Drizzle	64	5
Patches of Fog	56	0
Light Thunderstorm	47	0
Heavy Thunderstorms and Rain	20	2
Heavy Fog	18	0
Heavy Rain Showers	16	0
Light Snow	15	2
Partial Fog	12	0
Shallow Fog	10	0
Light Fog	8	0
Heavy Drizzle	5	0
Heavy Rain	4	0
Blowing Sand	3	0
Widespread Dust	3	0
Thunderstorm with Small Hail	2	0
Thunderstorms with Hail	2	0
Heavy Thunderstorms with Small Hail	1	0
Light Small Hail Showers	1	0
Light Hail Showers	1	0
Heavy Hail Showers	1	0
Small Hail	1	0
Light Ice Pellets	1	0
Snow	1	0
Light Snow Showers	1	0

References

Amin, A., S. Anwar, A. Adnan, M. Nawaz, N. Howard, J. Qadir, A. Hawalah, and A. Hussain. 2016. Comparing oversampling techniques to handle the class imbalance problem: a customer churn prediction case study. IEEE Access. 4:7940-7957.

Batuwita, R., and V. Palade. 2010. FSVM-CIL: fuzzy support vector machines for class imbalance learning. IEEE Transactions on Fuzzy Systems. 18:558-571.

Chawla, N.V., K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research. 16:321-357.

Chow, T.W., P. Wang, and E.W. Ma. 2008. A new feature selection scheme using a data distribution factor for unsupervised nominal data. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). 38:499-509.

Diamantini, C., and D. Potena. 2009. Bayes vector quantizer for class-imbalance problem. IEEE Transactions on Knowledge and Data Engineering. 21:638-651.

Diao, R., F. Chao, T. Peng, N. Snooke, and Q. Shen. 2014. Feature selection inspired classifier ensemble reduction. IEEE Transactions on Cybernetics. 44:1259-1268.

Diao, R., and Q. Shen. 2012. Feature selection with harmony search. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). 42:1509-1523.

Duin, R.P.W. 1976. On the choice of smoothing parameters for Parzen estimators of probability density functions. IEEE Transactions on Computers. C-25:1175-1179.

Francis, D.P., A.J. Coats, and D.G. Gibson. 1999. How high can a correlation coefficient be? Effects of limited reproducibility of common cardiological measures. International Journal of Cardiology. 69:185-189.

Habbema, J. 1974. A stepwise discriminant analysis program using density estimetion. In Compstat. Physica-Verlag. 101-110.

He, H., Y. Bai, E.A. Garcia, and S. Li. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on. IEEE. 1322-1328.

He, H., and E.A. Garcia. 2009. Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering. 21:1263-1284.

Kheradmanda, S., O. Nematollahi, and A.R. Ayoobia. 2016. Clearness index predicting using an integrated artificial neural network (ANN) approach. Renewable and Sustainable Energy Reviews. 58:1357-1365.

Krstic, M., and M. Bjelica. 2015. Impact of class imbalance on personalized program guide performance. IEEE Transactions on Consumer Electronics. 61:90-95.

Lai, C.S., Y. Jia, M. McCulloch, and Z. Xu. 2017a. Daily clearness index profiles cluster analysis for photovoltaic system. IEEE Transactions on Industrial Informatics. 13:2322-2332.

Lai, C.S., and L.L. Lai. 2015. Application of big data in smart grid. In Systems, Man, and Cybernetics (SMC), 2015 IEEE International Conference on. IEEE. 665-670.

Lai, C.S., X. Li, L.L. Lai, and M.D. McCulloch. 2017b. Daily clearness index profiles and weather conditions studies for photovoltaic systems. Energy Procedia. 142:77-82.

Lai, C.S., and M.D. McCulloch. 2017. Sizing of stand-alone solar PV and storage system with anaerobic digestion biogas power plants. IEEE Transactions on Industrial Electronics. 64:2112-2121.

Li, D.-C., C.-W. Liu, and S.C. Hu. 2010. A learning method for the class imbalance problem with medical data sets. Computers in Biology and Medicine. 40:509-518.

Lin, M., K. Tang, and X. Yao. 2013. Dynamic sampling approach to training neural networks for multiclass imbalance classification. IEEE Transactions on Neural Networks and Learning Systems. 24:647-660.

Liu, J., W. Fang, X. Zhang, and C. Yang. 2015a. An improved photovoltaic power forecasting model with the assistance of aerosol index data. IEEE Transactions on Sustainable Energy. 6:434-442.

Liu, X.-Y., J. Wu, and Z.-H. Zhou. 2009. Exploratory undersampling for class-imbalance learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). 39:539-550.

Liu, Y., T. Pan, and S. Aluru. 2016. Parallel pairwise correlation computation on intel xeon phi clusters. In Computer Architecture and High Performance Computing (SBAC-PAD), 2016 28th International Symposium on. IEEE. 141-149.

Liu, Y., F. Tang, and Z. Zeng. 2015b. Feature selection based on dependency margin. IEEE Transactions on Cybernetics. 45:1209-1221.

Locatelli, G., M. Mikic, M. Kovacevic, N.J. Brookes, and N. Ivanišević. 2017. The successful delivery of megaprojects: a novel research method. Project Management Journal. 48:78-94.

Malof, J.M., M.A. Mazurowski, and G.D. Tourassi. 2012. The effect of class imbalance on case selection for case-based classifiers: An empirical study in the context of medical decision support. Neural Networks. 25:141-145.

Mease, D., A.J. Wyner, and A. Buja. 2007. Boosted classification trees and class probability/quantile estimation. Journal of Machine Learning Research. 8:409-439.

Mitra, P., C. Murthy, and S.K. Pal. 2002. Unsupervised feature selection using feature similarity. IEEE Transactions on Pattern Analysis and Machine Intelligence. 24:301-312.

Murtagh, F., and P. Legendre. 2014. Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? Journal of Classification. 31:274-295.

Ng, W.W., J. Hu, D.S. Yeung, S. Yin, and F. Roli. 2015. Diversified sensitivity-based undersampling for imbalance classification problems. IEEE Transactions on Cybernetics. 45:2402-2412.

Oh, I.-S., J.-S. Lee, and B.-R. Moon. 2004. Hybrid genetic algorithms for feature selection. IEEE Transactions on Pattern Analysis and Machine Intelligence. 26:1424-1437.

Rahman, A., D.V. Smith, and G. Timms. 2014. A novel machine learning approach toward quality assessment of sensor data. IEEE Sensors Journal. 14:1035-1047.

Ratner, B. 2009. The correlation coefficient: Its values range between +1/−1, or do they? Journal of Targeting, Measurement and Analysis for Marketing. 17:139-142.

Ruiz, M.D., and E. Hüllermeier. 2012. A formal and empirical analysis of the fuzzy gamma rank correlation coefficient. Information Sciences. 206:1-17.

Seiffert, C., T.M. Khoshgoftaar, J. Van Hulse, and A. Napolitano. 2010. RUSBoost: A hybrid approach to alleviating class imbalance. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans. 40:185-197.

Silverman, B.W. 1986. Density estimation for statistics and data analysis. CRC press.

Tang, Y., Y.-Q. Zhang, N.V. Chawla, and S. Krasser. 2009. SVMs modeling for highly imbalanced classification. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). 39:281-288.

Wang, S., and X. Yao. 2012. Multiclass imbalance problems: Analysis and potential solutions. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). 42:1119-1130.

Wang, S., and X. Yao. 2013. Using class imbalance learning for software defect prediction. IEEE Transactions on Reliability. 62:434-443.

Weatherunderground.com, Historical data. [Online]. Available: https://www.wunderground.com/history/. [Accessed on 5th Nov. 2017].

Weiss, G.M., and F. Provost. 2003. Learning when training data are costly: the effect of class distribution on tree induction. Journal of Artificial Intelligence Research. 19:315-354.

Woyte, A., R. Belmans, and J. Nijs. 2007. Fluctuations in instantaneous clearness index: Analysis and statistics. Solar Energy. 81:195-206.

Woyte, A., V. Van Thong, R. Belmans, and J. Nijs. 2006. Voltage fluctuations on distribution level introduced by photovoltaic systems. IEEE Transactions on Energy Conversion. 21:202-209.

Wu, X., X. Zhu, G.-Q. Wu, and W. Ding. 2014. Data mining with big data. IEEE Transactions on Knowledge and Data Engineering. 26:97-107.

Xiao, Y., B. Liu, and Z. Hao. 2017. A sphere-description-based approach for multiple-instance learning. IEEE Transactions on Pattern Analysis and Machine Intelligence. 39:242-257.

Yao, Y., H. Tong, T. Xie, L. Akoglu, F. Xu, and J. Lu. 2015. Detecting high-quality posts in community question answering sites. Information Sciences. 302:70-82.

Yeung, D.S., J.-C. Li, W.W. Ng, and P.P. Chan. 2016. MLPNN training via a multiobjective optimization of training error and stochastic sensitivity. IEEE Transactions on Neural Networks and Learning Systems. 27:978-992.

Zhang, F., P.P. Chan, B. Biggio, D.S. Yeung, and F. Roli. 2016. Adversarial feature selection against evasion attacks. IEEE Transactions on Cybernetics. 46:766-777.

Zhang, X., and B.-G. Hu. 2014. A new strategy of cost-free learning in the class imbalance problem. IEEE Transactions on Knowledge and Data Engineering. 26:2872-2885.

Zhou, Z.-H., and X.-Y. Liu. 2006. Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Transactions on Knowledge and Data Engineering. 18:63-77.