Volume 36 Number 1 2025

Journal of Energy in Southern Africa

Outlier detection in ground-measured solar resource data using statistical classification models

Chantelle Clohessy *, Warren Brettenny, Waldo Abrahams

Department of Statistics, Nelson Mandela University, Gqeberha (Port Elizabeth), South Africa

Abstract

Ground-based solar resource measurements are known to be preferred to synthetic or simulated data for a given location, but outliers present in this data can significantly impact the accuracy of predictions used in viability assessments. For solar energy installations to be self-sustaining and viable, accurate ground-based solar resource data for the location of these installations are essential for decision-making and planning. Conventional outlier detection techniques used for solar resources, including graphical plots to complex numerical approaches, often have difficulty identifying these outliers to a satisfactory degree. This study proposes the use of simulated outliers added to synthetic data to train and compare the effectiveness of traditional outlier detection methods and several statistical learning methods, including *k*NN, naïve Bayes, support vector machines and advanced tree-based models for the purpose of outlier detection in this field. The results indicate that the advanced tree-based models provide accurate identification of outliers in the simulation step and are demonstrated to be effective on a ground-based real world data set collected in Gqeberha, South Africa. The use of the proposed approach can aid in reducing the uncertainty in measured solar resource data and, as a result, help to promote the use of solar energy solutions in areas with unreliable solar resource data.

Keywords: outlier detection, solar resource assessment, statistical learning

Highlights:

- Simulation method proposed to leverage the use of supervised classification algorithms for outlier detection.
- Tree-based models achieved >99% cross validated accuracy in simulated data.
- Case study for South African solar resource showed promising result.
- Improved accuracy for outlier detection compared to current methods.

Journal of Energy in Southern Africa 36(24),1–14

DOI: https://dx.doi.org/10.17159/2413-3051/2025/v36i1a20742

Published by the University of Cape Town ISSN: 2413-3051 https://journals.assaf.org.za/jesa

Sponsored by the Department of Science and Innovation

Corresponding author: E-mail: chantelle.clohessy@mandela.ac.za

Abbreviations used

GHI	Global horizontal irradiance
DHI	Diffuse horizontal irradiance
DNI	Direct normal irradiance
SAURAN	South African Universities Radiometric Network
LOF	Local outlier factor
BCT	bootstrap aggregated (or bagged) classification tree
kNN	k-nearest-neighbours
SVC	Support vector classification
XGBoost	Extreme gradient boosted
AdaBoost	Adaptive boosting model
TP	True positive
FP	False positive
FN	False negative
TN	True negative
MCC	Matthews correlation coefficient
NPV	Negative predictive value

1. Introduction

The growing interest in renewable energy research is motivated by the detrimental impacts that burning fossil fuels has on the environment, the unpredictability of fossil fuel prices, and the increased demand for electricity due to population growth (Clohessy, 2017). Countries are promoting the move to renewable energy sources by providing incentives and funding for such projects, including tax reductions and grants (Shahbaz et al., 2020). As a result, the renewable energy market is rapidly increasing across the globe, with energy being harnessed from a variety of abundant natural resources such as solar, wind, geothermal, hydro and modern biomass. This study focuses on solar energy.

Accurate solar resource data are essential for viability assessments, as this data is used to forecast energy yield (Jensen et al., 2023). Energy yield, in turn, is required by the net present value and levelised costs of electricity calculations for the economic assessment of solar energy system installations (Clohessy, 2017). Therefore, accurate forecasts are essential to ensure the long-term economic viability of a solar installation. To create reliable forecasts for such an assessment, accurate historical solar resource data is required. Data for the assessment of solar resource at a given location can be grouped into three categories, namely ground-based measures, remote-sensing (e.g. satellite based) measures, and weather prediction models (Yang et al., 2022). Of these, ground-based measures are known to be most accurate when properly calibrated (Yang et al., 2022), and are often preferred when making financial viability judgements. Outliers in such data can indicate the presence of uncalibrated equipment and also lead to incorrect assessments. The identification and removal of outliers in such data is thus a necessary and important step in the viability assessment process.

The present study investigates the use of statistical classification algorithms to identify outliers in solar resource data – including global horizontal irradiance (GHI), diffuse horizontal irradiance (DHI), direct normal irradiance (DNI), and temperature. These classification algorithms are proposed as an alternative approach to the existing outlier detection and quality control methods currently used in the field.

2. Outliers in solar resource data

Outliers are defined as observations that are far removed from the other observations in the dataset (Cousineau and Chartier, 2010). Outliers, therefore, may have an adverse effect on any analysis performed on that data. Outliers can also diminish the power of statistical tests and increase the error variance (Osborne and Overbay, 2004). Divya and Babu (2016) describe three categories of outliers: point outliers, contextual outliers, and collective outliers. A point outlier is an observation that is far removed from the rest of the data. These are the most commonly understood outlier category. Contextual outliers consist of observations that might only be regarded as outliers within a specific context. For example, a GHI measurement of 900 W/m^2 is quite possible at 11h00; however at 06h00 such a measurement is unlikely and a potential outlier. The last category of outlier described by Divya and Babu (2016) consists of a collection or group of observations within a dataset that may be considered as outliers when compared to the rest of the data.

Outliers in solar resource data can result from faulty equipment, incorrectly calibrated measurement systems, obstructed or damaged sensors, maintenance work, and incorrectly labelled mea-

surements (Younes et al., 2005). Outliers as a result of these causes can be expected to be observed as point, contextual or collective outliers.

Data-driven models are typically used to predict the energy yield of a solar energy installation. The model's accuracy relies heavily on the quality of the measured input data. Moradi (2009) highlights that solar radiation is one of the more difficult variables to measure as it is prone to complexities during the recording process, such as technical and operational-related problems. Previous studies have made use of various techniques to assess the quality of measured solar resource data.

Younes et al. (2005) reviewed and proposed several quality control measures to assess solar resource data. The most popular technique used involves user-defined and calculated threshold values which serve as a basis for the removal of a potential outlying data point (Sheng et al., 2017). Such methods include the use of the Page clear-sky and overcast models to calculate upper and lower bounds for the assessment of irradiance measurements. Additionally, the NREL SERI QC programme, the CIE automatic quality control, and the Muneer and Fairooz quality control procedure (Muneer and Fairooz, 2002) were mentioned. Younes et al. (2005) proposed a four-step procedure for the assessment of solar resource data. The first three steps include physical limit tests, while the last is based on a statistical method where the mean, weighted mean and standard deviation were required to construct a quality control envelope.

There also exist web-based quality control algorithms, such as the Helioclim algorithm, which checks the plausibility of the data by comparing observations with their predicted values based on the extraterrestrial irradiance and a simulation of irradiance from clear skies. Similarly, Molineaux and Ineichen (1994) developed web-based tools which compare measured and predicted values to assess measurement errors and to determine potential data quality anomalies.

The temporally varying nature of irradiance data led to the use of threshold limits which themselves change according to the time of day. Journée and Bertrand (2011) and Lee et al. (2013) are some of the studies which investigate this aspect of outlier detections and quality control assessments. The National Renewable Energy Labor-atory (Wilcox and McCormack, 2011) summarised the best practices for data quality assessments by including the assessment of long-term trends and redundant measurements on time series plots.

Other statistical methods used for outlier and anomaly detection in solar resource and environmental data include box-plots, quartiles intervals, Gaussian weighted regression (Sheng et al., 2017), support vector machines (Zhang et al., 2009), standard deviation tests and Bayesian tolerance intervals (Clohessy, 2017).

The South African Universities Radiometric Network (SAURAN) (Brooks et al., 2015) is a South African database containing solar resource data at several sites in the country and neighbouring countries. The quality control mechanisms in place on SAURAN are based on those proposed by Jacovides et al. (2006). Using this approach, a measurement is identified as problematic if it falls outside of predetermined expected ranges. Each time this happens, the measurement is flagged. Brooks et al. (2015), however, emphasise that the SAURAN quality control flags method is not designed, in principle, to detect outliers, but rather to provide caution to observations that are erroneous. However, since this method is the only one in place to assess the quality of the SAURAN data, the methods proposed in this paper are assessed in comparison to these flags - which represent an 'as-is' representation of the current system in place.

3. Data

This study made use of Meteonorm software to collect data for the resource variables used to train the methods proposed in this study. The data for these variables are generated by proprietary statistical models using data at a location close to the site of interest. The data are simulated for an expected year at the location specified by the user (Remund et al., 2012). Solar resource data for the GHI (W/m^2), DHI (W/m^2), DNI (W/m^2), and temperature (°C) were collected from Meteonorm for the Gqeberha region in South Africa. Ten years of typical data were simulated for each of these variables at this location.

As the data from Meteonorm are simulated from prevailing conditions in the region and have gone through various quality checks, these data are assumed to be clean and contain no outliers (Clohessy, 2017). Furthermore, Clohessy (2017) conducted a study assessing the quality of the data generated by Meteonorm for the Gqeberha region and found the data to satisfy this assumption.

3.1 Data collation and outlier simulation

The ten years of data, simulated from Meteonorm, were sorted into seasons, as Summer (December, January, February), Autumn (March, April, May), Winter (June, July, August) and Spring (September, October, November). Splitting the data into seasons was considered an effective and intuitive mechanism to account for the variability of the data points between each seasons and aligns with method used by Clohessy (2019). Owing to the nature of the study, only daylight hours were

considered, since solar irradiance does not occur outside of daylight hours.

As the data generated by Meteonorm is considered to be clean and contain no outliers, it is necessary to simulate appropriate outliers and add these to the seasonal data sets. Once these outliers are added, statistical learning algorithms and other outlier detection method can be trained, validated, and implemented on the data set. Adding outliers to such solar resource data can, however, be complex, due to the temporal nature of the data. As such, these aspects of the data were considered when outliers were incorporated.

Specifically, two types of noise, generated using zero-centred Gaussian and Cauchy distributions, were created for each observation. Gaussian noise was added by using a zero-centred Gaussian distribution with standard deviation chosen empirically such that the added noise generated outliers which were sufficiently distinct from the Meteonorm data to allow for model training and implementation. Cauchy noise was similarly created from a zero-centred Cauchy distribution. The Cauchy distribution was selected along with the traditional normal distribution as it has heavier tails and thus will be able to generate a higher proportion of noise further from the origin. The simulated noise was added to the Meteonorm generated data separately, resulting in two noise-augmented data sets.

Since the random generation of noise originates from zero-centred distributions, it is common for the generated noise to result in values which are similar to the Meteonorm observations. For example, an observation might have a temperature value of 20°C and the noise-augmented value could be 20.5°C, which would not be considered an outlier. To establish bounds for the determination of outliers, Chebyshev's inequality is used. For any random variable X (i.e. the solar resource variables in this study), Equation 1 (Chebyshev's inequality) states that

$$P(|X - \mu_X| \geq k\sigma_X) \leq \frac{1}{k^2}\,, (1)$$

where μ_X and σ_X are the mean and standard deviation of X, respectively. For this study a value of $k = 4$ is chosen as this ensures that only the most extreme 6.25% of the augmented data would be considered outliers (Abrahams, 2021) .

Since there is a significant temporal component to the data, that is, the data vary by both time of day and season, this needs to be taken into account when determining outliers in each context. For example, a GHI value of 1000 W/m^2 at 06h00 can be regarded as an outlier, owing to the time of the day at which it occurred, but this same observed value would not be considered an outlier if it were observed at 12h00 on the same day. To identify these so-called temporal outliers, a moving window approach was used to localise the assessment and outlier identification process to specific time intervals.

Once the outliers were identified in the outlier-augmented data sets, they were added to Meteonorm data. This data was then standardised, by centring and scaling in order to remove undesired constant offsets, and also to remove magnitude (unit measurement) differences between the different variables.

Outliers were added to the Meteonorm data in different proportions to test the applicability of the methods in varying conditions. To this end, outliers were added as 5%, 10%, 20%, 30%, and 50%. These values were decided upon as they range from typical proportions to the more extreme proportions that are usually indicative of malfunctioning equipment.

3.2 Training, validation and testing

The seasonal hourly transformed data, as constructed above, were used to create training, validation and testing datasets. The data were split according to an 80:20 split, where 80% of the data was used for training and validation and the remaining 20% was used as unseen data for testing. This split criterion was used based on its simplicity and popularity when performing cross-validation (Gholamy et al., 2018). Model validation, prior to testing, was performed using 10-fold cross-validation on the training data. The cross-validation was applied to tune the model parameters and estimate the true prediction error of models. These methods were then evaluated on the created testing data.

4. Methodology

4.1 Conventional outlier detection methods

Outlier detection and identification is an important part of any data analysis. Methods to perform such tasks range from simple graphical identification to complex numerical approaches. The conventional methods covered and investigated in this study are the use of SAURAN flags, which is the method currently in use at the site of interest, and the local outlier factor (LOF), a common density-based method for outlier detection.

4.1.1 SAURAN flags

As discussed, the current method that is employed to assess the quality of solar resource data in Southern Africa is the quality control flags that are offered by Brooks et al. (2015). The exact flags used at SAURAN are provided in Table 1.

Table 1: SAURAN flags used for measurement assessments.

Flag	*Limits*	*Descriptions*
2	$\|E - E_d\| < 5, E_{dn} < 1.5$ and $E > 600$	Warns of problems with the tracking system.
3	$E_d > 1.1E$	Identifies values that are impossible to attain.
4	$E > 1.2E_o$	Identifies values that are impossible to attain.
5	$E_d > 0.8E_o$	Identifies values that are impossible to attain.
6	$E_d < 5$	Removes all night-time/sunrise/sunset observations.
7	$E < 5$	Removes all night-time/sunrise/sunset observations.
8	$E - E_d > E_o$	Identifies values that are impossible to attain.
12	$E_{dn_{calc}} > 1367$	Upper limit of the solar constant for DNI.

Key: E_o denotes the extra-terrestrial GHI, E_d is the diffuse irradiance, E and E_{dn} the global and direct irradiance, respectively and $E_{dn_{calc}} = \frac{E-E_o}{\cos(Z)}$, where Z denotes the zenith angle Brooks et al. (2015).

Table 1 also provides an indication of the variable which is assessed for each flag. Flags 2, 4, 7 and 8 are used to assess the GHI data, 3, 5 and 6 are used for DHI data and 2 and 12 to assess DNI data. Flags 6 and 7 are controls which to remove low-light and night-time observations.

If an observation satisfies any of the flags mentioned in Table 1, this observation would be flagged for further investigation. A limitation with this method is that is not specifically designed to detect outliers and is very limited in its application; rather than aiming to detect outliers, these flags are only intended to warn users of any potential problems with the data (Brooks et al., 2015). Notwithstanding this intention, they flags have been actively used by SAURAN for this purpose. It it with this in mind that these quality control flags were considered in this study as an approach to identify outliers as a comparison point.

4.1.2 Local outlier factor

The LOF is an unsupervised outlier detection method which falls under the umbrella of density-based approaches. Such approaches are known to be more robust than traditional distance based measures in instances when clustering is present in the data (Ha et al., 2015). Density based approaches compare the local density of an observation to that of its immediate neighbourhood (Smiti, 2020). The LOF method is chosen for consideration in this study as it is one of the most popular methods used for outlier detection in the related literature (Ha et al., 2015).

The LOF was developed and proposed by Breunig et al. (2000) as a score which measures the ratio of the local density for an observation x to that of it's nearest neighbours (Smiti, 2020). Specifically, the LOF score for an observation a is calculated by first determining the k-nearest neighbours to a. This is determined using a pre-specified distance measure such as the Euclidean distance or Manhattan distance.

The reachability distance between observations a and x, $rd_k(a, x)$, is defined by Equation 2 as

$$rd_k(a, x) = \max\{d(x, a), d_k(a)\}, \quad (2)$$

where $d(x, a)$ is the distance between points a and x, and $d_k(a)$ is the k-distance for an observation a, defined as $d_k(a) = d(a, k_a^*)$ where k_a^* is the k^{th}-nearest-neighbour of the observation a. Let $N_k(a)$ denote the set that contains all observations which are within $d_k(a)$ of a. Then the local reachability density of an observation a, $LRD_k(a)$, is calculated using Equation 3:

$$LRD_k(a) = \frac{| N_k(a) |}{\sum_{x \in N_k(a)} d_k(a, x)}, \quad (3)$$

where $| N_k(a) |$ refers to the cardinality of $| N_k(a) |$ (Breunig et al., 2000). The LOF of an observation a is then defined by Equation 4 (Breunig et al., 2000)

$$LOF_k(x) = \frac{\sum_{x \in N_k(a)} \frac{LRD_k(x)}{LRD_k(a)}}{| N_k(x) |}. \quad (4)$$

Once the LOF scores were calculated, the outliers were identified as those LOF scores which were greater than three standard deviations above the mean of the LOF values. Since the data set used in this study is highly variable, this is a conservative approach which will help to ensure that legitimate observations are not incorrectly identified as outliers.

4.2 Classification models

Consider a data set represented as (X, y), where X represents the predictor variables (in this case the solar resource variables) and y represents a categorical outcome (in this case the binary outcome of outlier or not). Using this terminology, the kNN, naïve Bayes and tree-based classification methods can be outlined and described.

4.2.1 k-nearest-neighbours classification

The kNN classification, possibly the simplest and most intuitive classification method, is also one of the most widely used (Ali et al., 2019). The concept was introduced by Fix (1985) and later formalised and expanded upon by Cover and Hart (1967).

The process followed to determine a kNN estimate for at some point in the predictor space, x, starts by creating a 'neighbourhood', $K_x \subseteq (X, y)$, which consists of the k observations which are nearest to x. Several distance metrics can be used to determine the proximity of each observation to x, with the most common being the traditional Euclidean distance (Ali et al., 2019).

The prediction of the outcome, y, at the point x is then simply determined as the outcome value which occurs most frequently in K_x, i.e. the neighbourhood of x. To ensure that an outright majority is always reached when the outcome is binary, odd values of k are typically chosen.

4.2.2 Naïve Bayes

The naïve Bayes model is based on the well-known Bayes theorem that states that for any category of the outcome y_c $c \in 1, \dots, C.$, the posterior conditional probability $P(y_c|X)$ can be determined by Equation 5 as

$$= \frac{P(y_c)P(X|y_c)}{P(X)} . (5)$$

Expanding this for each of the q predictors, Equation 6 can show that

$$P(y_c|X) \propto P(y_c)P(X|y_c) = P(y_c, X_1, \dots, X_q). (6)$$

Also referred to as idiot's or independence Bayes (Hand and Yu, 2001), the naïve Bayes method is so named owing to the assumptions which are made when it is used. Namely, it is assumed that there is independence between all predictors, $X_1, \dots, X_q$, within the model, with dependence only accounted for between each predictor and the response. This allows for a simplification of the posterior probability representation to Equation 7:

$$P(y_c|\underline{x}) \propto P(y_c) \prod_{j=1}^{q} P\left(X_j|y_c\right). (7)$$

Using this, the naïve Bayes estimator, Equation 8 can be defined as

$$\underset{c}{\text{argmax}} \left\{ P(y_c) \prod_{j=1}^{q} P\left(X_j|y_c\right) \right\}. (8)$$

That is, the estimator returns the category c for which the posterior probability of occurrence (conditional on the observed values of the predictor variables) is the highest.

4.2.3 Support vector classification

The support vector classification (SVC) is a popular statistical learning approach that classifies data points by determining the hyperplance $b_1X_1 + b_2X_2 + \cdots + b_qX_q = w$, which best separates the data points from the different classes in the q-dimensional space spanned by the predictors $X_1, \dots, X_q$. The selected hyperplane is the one which maximises the distance between the hyperplane and the nearest point in each class. These points are known as the support vectors.

In many cases a simple linear hyperplance is unable to perfectly separate the data as described. In this case, instead of adapting the shape of the hyperplane, the original data are transformed (using kernels) in such a way that linear separation using a simple hyperplane is possible – this is known as the kernel trick. The most popular and widely used kernel in practice is the Gaussian or radial basis function kernel,

$$K(\underline{x}_j, \underline{x}_j) = \exp\left(-\gamma|\underline{x}_j - \underline{x}_j|^2\right), (9)$$

and is the one which is utilised in this study.

If perfect linear separation is still not possible after the transformation is applied, then data are permitted to fall on the wrong side of the hyperplance, but such instances are penalised and the hyperplance is chosen to minimise the penalty incurred and thus optimise the separation of classes.

4.2.4 Classification trees

A classification tree is a predictive method which recursively splits the predictor space into distinct regions, with each region being allocated predicted value or outcome for the response, $\hat{y}_c$. The splits are made by selecting the variable and splitting value combination at each step, which optimises some loss function, $L(y_c, \hat{y}_c)$. In particular, a loss function commonly used in this application is the Gini impurity, defined by Equation 10:

$$\sum_{c=1}^{C} 1 - p_c^2, (10)$$

where p_c is the relative frequency of the outcome category $c \in 1, \dots, C$.

A bootstrap aggregated (or bagged) classification tree, BCT, is an algorithm that incorporates a bootstrapping method to boost the performance of the results obtained from a single classification tree. This algorithm was first introduced by Breiman (1996) and is considered one of the simplest techniques that can be used to boost a model's performance (Lemmens and Croux, 2006).

Consider a classification tree $f(X)$, which can be considered a process by which values of the predictor variables X are converted to a predicted outcome $\hat{y}$. For each data set (X, y), a single classification tree can be constructed for prediction purposes. In order to augment this, B bootstrap samples of the (X, y), denoted $\left(X^*_{(b)}, y^*_{(b)}\right)$, $b = 1, \dots, B$ can be created through random sampling, with replacement, from (X, y). In so doing, B distinct classification trees, $f_{(b)}(X)$ $b = 1, \dots, B$, can be created, each resulting in a prediction for the response variable, $\hat{y}_{(b)}$ $b = 1, \dots, B$.

The bagging approach then simply creates an estimator for the response variable by aggregating the individual predictions $\hat{y}_{(b)}$, $b = 1, \dots, B$ into a single predictor as the most prevalent outcome within these predictions.

Specific shortcomings of this approach are addressed through the use of random forests, and boosting methods, which control for the tree complexity, such as the extreme gradient boosted (XGBoost) tree classification algorithm developed by Chen and Guestrin (2016) and the adaptive boosting model (AdaBoost) developed by Freund and Schapire (1997). Depending on the data set and specific problem being addressed, these boosting methods may not provide significant improvements over the standard bagging approach.

4.3 Assessment metrics

Assessments of classification algorithms, like those indicated in the previous sections, are often based on a confusion matrix which helps to visualise and summarise the predictions (Singh et al., 2021). The specific confusion matrix for the present study is provided in Table 2.

Table 2: Confusion matrix representation for the classification of outliers.

Prediction	*Outlier*	*Normal*
Outlier	TP (true positive)	FP (false positive)
Normal	FN (false negative)	TN (true negative)

From Table 2 it can be seen that observations which fall into the TP and FP cells represent correct classifications of outliers and normal observations, respectively. Similarly, the FN and FP cells represent incorrect classifications of actual outlying and normal observations, respectively. From these outcomes traditional assessment metrics such as the accuracy, sensitivity, specificity, precision, balanced accuracy and Matthews correlation coefficient (MCC) can be defined.

The accuracy is defined as the proportion of all predictions which were correct. This is calculated by Equation 11.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}. \quad (11)$$

The sensitivity indicates the proportion of true outliers that were correctly classified. This is calculated by Equation 12.

$$\text{Sensitivity} = \frac{TP}{TP + FN}. \quad (12)$$

The specificity indicates the proportion of true normal observations that were correctly classified. This is calculated by Equation 13.

$$\text{Specificity} = \frac{TN}{TN + FP}. \quad (13)$$

The precision indicates the proportion of predicted outliers that were truly outliers. This is calculated by Equation 14.

$$\text{Precision} = \frac{TP}{TP + FP}. \quad (14)$$

The negative predictive value (NPV) indicates the proportion of predicted normal observations that were correct. This is calculated by Equation 15.

$$NPV = \frac{TN}{TN + FN}. \quad (15)$$

The balanced accuracy, an extension of the accuracy, allows for the assessment of accuracy in unbalanced datasets. These are datasets for which one class makes up the (usually vast) majority of the observations. The balanced accuracy is calculated as the arithmetic mean of sensitivity and specificity by Equation 16.

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2}. \quad (16)$$

All of the above metrics range from 0 to 1, with values closer to 1 indicating better performance.

The MCC is also a preferred method in the case of imbalanced data (Chicco and Jurman, 2020) and is defined by Equation 17.

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}}. \quad (17)$$

Similar to the traditional correlation coefficient, the MCC ranges between -1 and 1. Values closer to 1 indicate that both classes are being predicted well.

These measures will be used to assess the suitability of the various methods for the classification of outliers in the current application and make a suggestion regarding the preferred method for use in the field.

All analysis for this study was performed using R 4.0.3 (R Core Team, 2020) using the *ruta* (Charte et al., 2019), *dbscan* (Hahsler et al., 2019), and *caret* (Kuhn, 2020) packages.

5. Results

This section summarises the results for the LOF, SAURAN flags, kNN, naïve Bayes and classification tree methods for the purpose of outlier detection in solar resource data. As discussed, simulated data sets containing a different proportion of outliers (5%, 10%, 20%, 30%, 50%) were used for training and testing the methods. Based on the comparisons of these proposed methods, as well as other computational considerations, the identification and suggestion of the most appropriate method for this application is proposed. The section concludes with a case study for solar resource data collected in the Gqeberha region, South Africa, making use of the identified and proposed method from the simulation study.

5.1 Simulation study

To assess the effect on the classification ability each of the methods caused by the proportion of outliers present in the simulated data, the MCC metric is used. The MCC values were calculated and aggregated for each of the the four seasons and split according to the outlier prevalence. These cross-validated results are provided in Table 3.

From Table 3 the MCC values suggest that there is little change in prediction capability across the different prevalences of outliers in the simulated data. This suggests the quantity of outliers in the data does not have a meaningful impact on performance of any of the models.

The traditional methods, LOF and SAURAN flags, perform poorly at the classification task. These methods are outperformed, according to this metric, by all but one of the statistical learning methods. The SVC approach performed particularly poorly at this task compared to the remaining statistical learning methods. While it does appear to improve with an increase in outlier prevalence, it still does not outperform the SAURAN flags approach.

Most prominent, however, is the performance of the tree-based methods, namely BCT, XGBoost, and AdaBoost. These methods achieve MCC near 1 (indicating perfect classification results) in all the simulated scenarios. It is clear that the tree based methods are far superior to the other approaches for this task.

To explore the results further, the investigation now focuses on the individual performances of the methods by considering the the balanced accuracy, sensitivity, specificity, precision, and NPV measures averaged over all the seasons. These results are provided in Table 4.

Table 3: MCC values for each method for each level of outlier prevalence

Method	*5%*	*10%*	*20%*	*30%*	*50%*
LOF	0.2242	0.1499	0.1194	0.1593	0.1603
SAURAN flags	0.4279	0.4182	0.4225	0.4062	0.4295
kNN	0.599	0.6216	0.6743	0.6533	0.7466
Naïve Bayes	0.819	0.7985	0.8045	0.7984	0.8032
SVC	0.1671	0.2243	0.2861	0.2901	0.4223
BCT	0.9821	0.9865	0.9903	0.9884	0.9946
XGBoost	0.9848	0.9854	0.9892	0.9908	0.9962
AdaBoost	0.9839	0.9868	0.9896	0.9916	0.9971

Table 4: Assessment of the classification ability and suitability of all algorithms and methods

Method	*Balanced accuracy*	*Sensitivity*	*Specificity*	*Precision*	*NPV*
LOF	0.5233	0.0519	0.9942	0.9407	0.7904
SAURAN flags	0.6118	0.2236	1.0000	1.0000	0.8218
kNN	0.7922	0.6003	0.9842	0.8496	0.9016
Naïve Bayes	0.8523	0.7047	1.0000	1.0000	0.9230
SVC	0.6006	0.2390	0.9622	0.5776	0.8229
BCT	0.9945	0.9902	0.9987	0.9901	0.9975
XGBoost	0.9947	0.9907	0.9988	0.9913	0.9978
AdaBoost	0.9954	0.9921	0.9987	0.9907	0.9982

Table 5: Run times for training the BCT and XGBoost methods

	Summer	*Autumn*	*Winter*	*Spring*	*Average*
BCT	0.6529	0.6193	0.5629	0.6449	0.6200
XGBoost	5.4876	5.0117	4.7966	5.4310	5.1817
AdaBoost	43.6135	35.3874	28.7824	40.8657	37.1623

The results presented in Table 4 reinforce the findings shown in the MCC assessment.

The LOF method achieves a balanced accuracy of only 0.5233. This is largely due to a poor sensitivity score of 0.0519. This indicates that the LOF has difficulty identifying outliers, and tends to classify observations as normal in most circumstances. When a outlier is identified, however, it tends to be correct, but not enough of these predictions are made to consider this method to be appropriate for the task. The SAURAN flags and SVC method exhibit similar findings and performances and display a slightly better ability to identify outliers than the LOF. The SAURAN network currently utilises these flags to identify errors and correct quality in ground-measured data. These results show that around 78% of simulated outliers are not identified using this approach and highlights the need for a more robust and accurate method.

The results show that out of the six statistical learning based classification approaches tested, the kNN method performs the worst, while still achieving a balanced accuracy of almost 80%. As with the traditional methods, the kNN method performs worst in its ability to identify outliers. A sensitivity of 0.6003 indicates that approximately 40% of the simulated outliers are not identified. Also, the precision vcalue indicates that if an outlier is identified by the kNN method it is only correct around 85% of the time. The naïve Bayes algorithm outperforms the kNN method in all metrics and, unlike the kNN ethods, when a outlier is identified by the method it is 100% accurate. The drawback is in that this method only identifies around 70% of the outliers.

The tree-based methods, BCT, XGBoost and AdaBoost, outperform all other methods by a considerable margin in every metric. There is little to distinguish between the methods, with all achieving values greater than 0.99 for all metrics. The AdaBoost method tends to slightly outperform the other methods, albeit not in all metrics. These results reiterate the findings shown using the MCC metric, and show that these methods have little problem distinguishing between outliers and normal observations.

In order to propose a preferred method between the BCT, XGBoost, and AdaBoost algorithms for this application, the computational aspects of the methods are considered. The XGBoost and AdaBoost methods are, by design, significantly more complex than the BCT algorithm, and this may have an impact on the ability to implement these techniques in practice. As such, the training runtimes (in minutes) for these models is provided, for each season, in Table 5.

Taking the computation times into consideration, it is evident that the BCT algorithm exhibits considerably faster training times than the XGBoost and AdaBoost methods, while achieving very similar results. It is for this reason that that the BCT algorithm trained using this process be used in a practical setting.

5.2 Case study

The case study used the trained BCT classification model to identify potential outliers on a data set collected at Nelson Mandela University. For this assessment, hourly observations on the variables GHI, DNI, DHI and temperature were obtained for the period of 10 December 2015 to 27 April 2021.

The BCT models trained for each season in the simulation study above, were used to predict outliers for the real data collected at the site. The results of the BCT classification for the summer season are presented for the DNI, DHI and GHI variables plotted against temperature in Figures 1, 2, and 3 respectively.

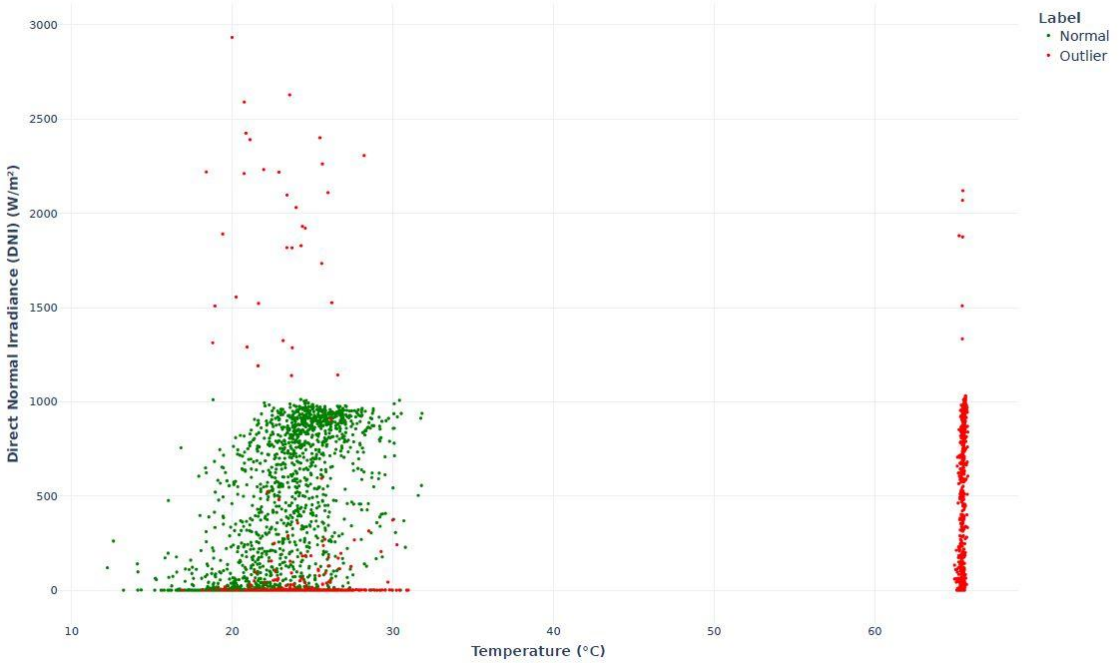

Figure 1: Plot of DNI against temperature with BCT outlier predictions for the Gqeberha region, South Africa (Summer)

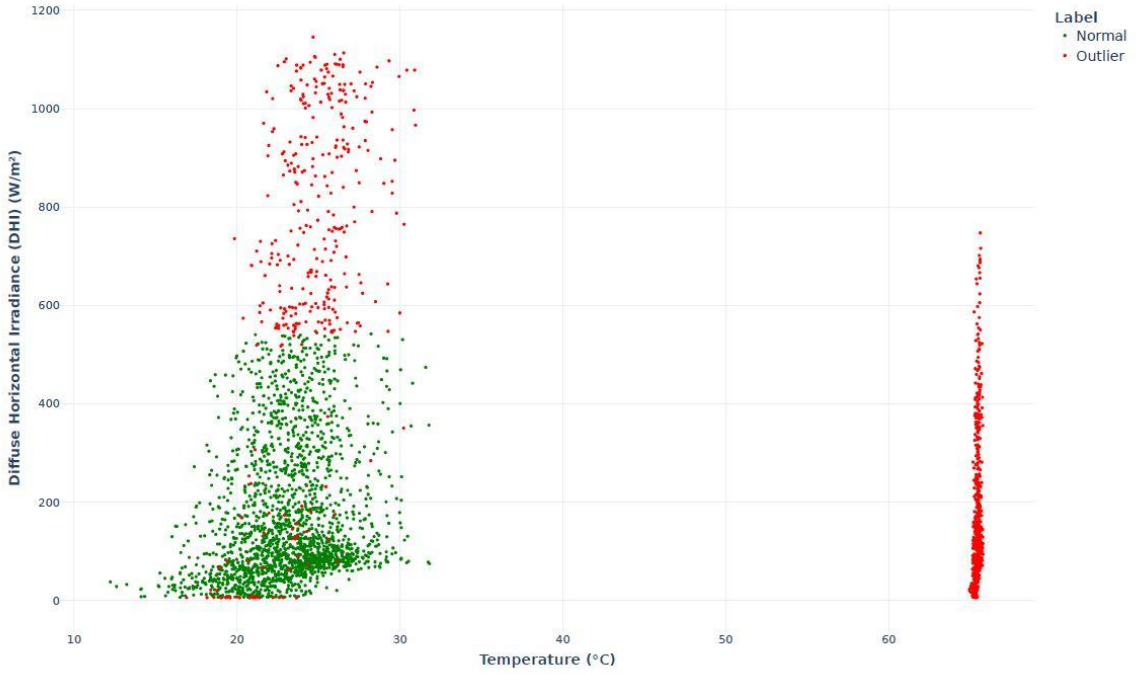

Figure 2: Plot of DHI against temperature with BCT outlier predictions for the Gqeberha region, South Africa (Summer)

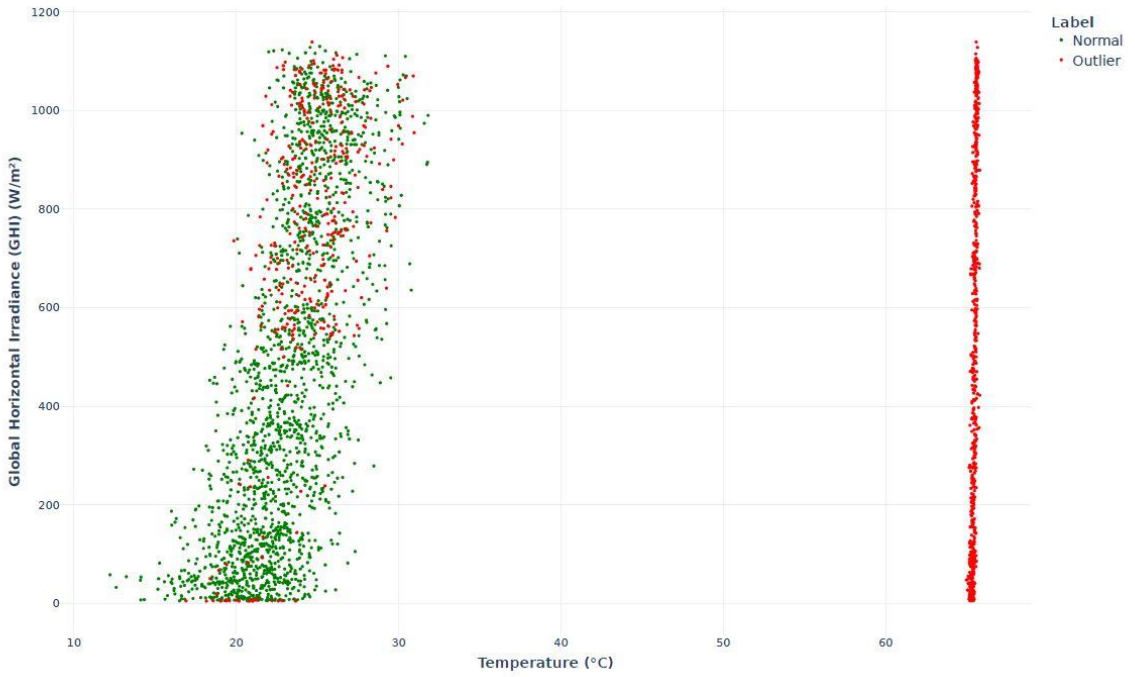

Figure 3: Plot of GHI against temperature with BCT outlier predictions for the Gqeberha region, South Africa (Summer)

Table 6: Approximate accuracy of BCT predictions on case study data

	Summer	*Autumn*	*Winter*	*Spring*
BCT predicted outlier prevalence (Gqeberha)	34%	32%	19%	31%
Comparable simulated prevalence	30%	30%	20%	30%
Approximate balanced accuracy of predictions	0.9959	0.9965	0.9944	0.9944

Figures 1, 2 and 3 indicate that the trained BCT classifier consistently and correctly classifies extreme temperature observations as outliers. This is evidenced by observing the classification of observations with temperature values greater than 60°C. This temperature value has never been recorded in the Gqeberha region, and demonstrates that the BCT classifier is able to identify obvious outliers.

Also evident in the figures is that the trained BCT classifier also identifies outliers within the cluster of normal observations. These should not be considered as misclassifications, but are more likely correctly identified as outliers owing to the time of day at which they were observed. As mentioned, a GHI value of 1000 W/m^2 will likely be an outlier at 06h00 but not necessarily at 12h00. Since time is not indicated on these plots, this is not immediately evident.

From Figure 1 and Figure 2 it is observed that the observations which have extreme irradiance values are identified as being outliers. This is especially apparent when considering Figure 1 where several extreme DNI values are labelled as outliers. This is expected since these observations far exceed 1000 W/m^2 an approximate upper limit of this measure.

As the case study data is unlabelled, it is impossible to identify the true accuracy of the trained BCT model in predicting outliers in this case. However, approximations of the accuracies can be established through appropriate comparisons to the simulations study. Table 6 summarises the prevalence of predicted outliers returned by the trained BCT model for the case study. The percentages are compared to the simulated test set percentages to allow for an approximation of the balanced accuracy of the classifications. From these results it is apparent that a balanced accuracy exceeding 99% is expected from the predictions.

6. Discussion

Identifying outliers in real world ground-measured data can be difficult. The main reason for this is that outliers are unlabelled and unknown, which

preclude them from typical supervised learning approaches. That is, in typical supervised classification tasks the outcome is known – and can be verified – but outliers in ground-measured solar resource data are not. Outside of values which are physically impossible (e.g. those that are identified using the SAURAN flags), outliers within otherwise normal data are often missed. This study introduces simulated outlier data into acquired clean resource data, specific to a given location, and fits a supervised classification model to detect these outliers. The trained model can now be used to identify outliers in unlabelled real world data.

The results of the case study demonstrate that, although the models are fitted to artificial simulated data, they remain effective in detecting outliers in real world situations. Even in situations where the outlying data are unlike the simulated data (e.g. in the case of the temperature measurements being above 60 degrees Celsius or DNI values above 1000 W/m^2) the model was able to accurately detect these. As such, the fitted classification model accounts for situational and physical outliers with a high degree of accuracy.

Such a fitted model would be able to be used for many solar installations with similar solar resource and geographical characteristics. In cases where significant disparity exists between the physical and geographical characteristics of the modelled and measured data, the two-step approach of outlier simulation and model fitting using tree-based models, as proposed in this study, is shown to be an effective means to accurately detect outliers in real world use cases.

A barrier to implementing this approach on a large scale is that synthetic resource data is required (eg: Meteonorm) and is often not open sourced. However, practioners and researchers in this field typically have access to this type of data and would be able to use the methodology proposed.

7. Conclusion

Collection of ground-based solar resource data at an installation site is important for the maintenance, administration and potential further development of a site. Unlike synthetic data generated from a software-based solution source, the ground-based data provides a true reflection of the solar potential of the site. This data however, is not without its own limitations. The data is collected directly from sensors and passed through systems that all can introduce erroneous readings from time to time. This can be from malfunctions of the sensors, or even errors in the reading and recording systems. These outliers can have a considerable effect on the assessment of the resource available at a solar installation. The accurate detection of these outliers is, therefore, an important step to take in the assessment of the data that is recorded.

This study proposed and validated a two step approach which allows for the development of an statistical classification of outliers through first simulating outlier data and adding it to synthetic software-based data for a site and then fitting a classification model to the resulting outlier-infused synthetic data. Such a model could easily be setup within the data pipeline at a solar installation site and trigger an alarm when outliers are detected.

The study found that the most typical and currently used measures to flag outlying data, that is the LOF and SAURAN flags, have been shown to miss many of the simulated outliers. The LOF performs particularly poorly, while the SAURAN flags identify data that are physically impossible, but fail in more nuanced situations where the readings were, for example, the time of day would preclude the possibility of a particular measurement.

Of the statistical classification methods that were investigated, the tree-based algorithms performed very well for the identification of outliers, achieving accuracies in excess of 99%. It is believed that any additional methods considered would only introduce marginal gains, typically at the expense of speed, and incur considerably more computational resources. It is for this reason that BCT method, as used in this study, is recommended as the most suitable for the identification of outliers in the proposed two-step solution. Using the proposed method, it has been shown that over 99% of outliers in ground-measured solar resource data can be detected in close to real time, should this approach be implemented as part of a data validation step in the data flow at a solar installation.

For future studies, alternative outlier simulation approaches could be introduced along with the utilisation of other statistical methods as kernal ridge regression or machine learning approaches such as neural networks for comparison with the current results. In addition, a three-stage approach utilising both unsupervised and supervised learning methods could be used to better assess the underlying resource data. For example, a clustering step could be introduced prior to the classification step in the proposed method. The clustered data may provide more detail for assessing the solar resource data and thus provide a practioner with additional insight into the data.

Acknowledgements

The financial assistance of the South African National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at are those of the authors and are not

necessarily to be attributed to the NRF. The workstation used for this research is funded through NRF grants SFH180517331201 and TTK190408428135. We also acknowledge the NRF (Academic Statistics Grant – 127889), the South African Statistical Association and Nelson Mandela University PGRS (Postgraduate Research Scholarships) for their financial support.

References

Abrahams, W. (2021). Classification and clustering based methods for outlier detection of solar resouce data. Nelson Mandela University. https://vital.seals.ac.za/

Ali, N., Neagu, D., & Trundle, P. (2019). Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets. SN Applied Sciences, 1, 1-15. https://link.springer.com/ content/pdf/10.1007/s42452-019-1356-9.pdf.

Breiman, L. (1996). Bagging predictors. Machine learning, 24, 123-140.

Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000). LOF: identifying density-based local outliers. In Proceedings of the 2000 ACM SIGMOD international conference on Management of data, 93-104. https://dl.acm.org/doi/pdf/10.1145/342009.335388

Brooks, M. J., Du Clou, S., Van Niekerk, W. L., Gauché, P., Leonard, C., Mouzouris, M. J., & Vorster, F. J. (2015). SAURAN: A new resource for solar radiometric data in Southern Africa. Journal of energy in Southern Africa, 26(1), 2-10. https://www.scielo.org.za/pdf/jesa/v26n1/01.pdf

Charte, D., Herrera, F., & Charte, F. (2019). Ruta: implementations of neural autoencoders in R. Knowledge-Based Systems, 174, 4-8. https://fcharte.com/assets/pdfs/2019-KBS-Ruta.pdf

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 785-794. https://dl.acm.org/doi/pdf/10.1145/2939672.2939785

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC genomics, 21, 1-13. https://link.springer.com/content/pdf/10.1186/s12864-019-6413-7.pdf

Clohessy, C.M., (2017). Statistical viability assessment of a photovoltaic system in the presence of data uncertainty (PhD thesis). Nelson Mandela Metropolitan University. https://vital.seals.ac.za/

Clohessy, C. M., Sharp, G., Hugo, J., & Van Dyk, E. (2019). Inferential based statistical indicators for the assessment of solar resource data. Journal of Energy in Southern Africa, 30(1), 21-33. https://www.scielo.org.za/scielo.php?pid=S1021-447X2019000100003&script=sci_arttext.

Cousineau, D., & Chartier, S. (2010). Outliers detection and treatment: a review. International journal of psychological research, 3(1), 58-67. https://www.redalyc.org/pdf/2990/299023509004.pdf

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE transactions on information theory, 13(1), 21-27.

Divya, D., & Babu, S. S. (2016). Methods to detect different types of outliers. In 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE), 23-28. https://ieeexplore.ieee.org/abstract/document/7684114/

Fix, E. (1985). Discriminatory analysis: nonparametric discrimination, consistency properties. USAF school of Aviation Medicine.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences, 55(1), 119-139.

Gholamy, A., Kreinovich, V., & Kosheleva, O. (2018). Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation. International Journal of Intelligent Technologies and Applied Statistics, 11(2), 105-111. https://scholarworks.utep.edu/cgi/viewcontent.cgi?article=2202&context=cs_techrep

Ha, J., Seok, S., & Lee, J. S. (2015). A precise ranking method for outlier detection. Information Sciences, 324, 88-107.

Hahsler, M., Piekenbrock, M., & Doran, D. (2019). dbscan: Fast density-based clustering with R. Journal of Statistical Software, 91, 1-30.

Hand, D. J., & Yu, K. (2001). Idiot's Bayes—not so stupid after all?. International statistical review, 69(3), 385-398.

Jacovides, C. P., Tymvios, F. S., Assimakopoulos, V. D., & Kaltsounides, N. A. (2006). Comparative study of various correlations in estimating hourly diffuse fraction of global solar radiation. Renewable energy, 31(15), 2492-2504. https://ideas.repec.org/a/eee/renene/v31y2006i15p2492-2504.html

Jensen, A. R., Anderson, K. S., Holmgren, W. F., Mikofski, M. A., Hansen, C. W., Boeman, L. J., & Loonen, R. (2023). pvlib iotools—Open-source Python functions for seamless access to solar irradiance data. Solar Energy, 266, 112092. https://www.sciencedirect.com/science/article/pii/S0038092X23007260

Journée, M., & Bertrand, C. (2011). Quality control of solar radiation data within the RMIB solar measurements network. Solar Energy, 85(1), 72-86.

Kuhn, M. (2020). Caret: classification and regression training. Astrophysics Source Code Library, ascl-1505.

Lee, K., Yoo, H., & Levermore, G. J. (2013). Quality control and estimation hourly solar irradiation on inclined surfaces in South Korea. Renewable energy, 57, 190-199. https://www.sciencedirect.com/science/article/abs/pii/S0960148113000669

Lemmens, A., & Croux, C. (2006). Bagging and boosting classification trees to predict churn. Journal of Marketing Research, 43(2), 276-286. https://pure.uvt.nl/ws/portalfiles/portal/1425373/lemmens_bagging.pdf

Molineaux, B., & Ineichen, P. (1994). Automatic quality control of daylight measurements: software for IDMP stations.

Moradi, I. (2009). Quality control of global solar radiation using sunshine duration hours. Energy, 34(1), 1-6. https://www.sciencedirect.com/science/article/abs/pii/S0360544208002466

Muneer, T., & Fairooz, F. (2002). Quality control of solar radiation and sunshine measurements–lessons learnt from processing worldwide databases. Building Services Engineering Research and Technology, 23(3), 151-166. https://journals.sagepub.com/doi/abs/10.1191/0143624402bt038oa

Osborne, J. W., & Overbay, A. (2004). The power of outliers (and why researchers should always check for them). Practical Assessment, Research, and Evaluation, 9(1). https://scholarworks.umass.edu/ bitstreams/-af7a80b9-e2b9-4809-b353-b6271b1e5314/download

R Core Team, (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Remund, J., Mueller, S., Kunz, S., & Schilter, C. (2012). Meteonorm handbook, part II: theory. Bern, Switzerland, Meteotest.

Shahbaz, M., Raghutla, C., Song, M., Zameer, H., & Jiao, Z. (2020). Public-private partnerships investment in energy as new determinant of CO2 emissions: the role of technological innovations in China. Energy Economics, 86, 104664. https://mpra.ub.uni-muenchen.de/97909/1/MPRA_paper-_97909.pdf

Sheng, H., Xiao, J., Cheng, Y., Ni, Q., & Wang, S. (2017). Short-term solar power forecasting based on weighted Gaussian process regression. IEEE Transactions on Industrial Electronics, 65(1), 300-308. https://ieeexplore.ieee.org/abstract/document/7945510

Singh, P., Singh, N., Singh, K. K., & Singh, A. (2021). Diagnosing of disease using machine learning. In Machine learning and the internet of medical things in healthcare, Academic Press, 89-111.

Smiti, A. (2020). A critical overview of outlier detection methods. Computer Science Review, 38, 100306.

Wilcox, S. M., & McCormack, P. (2011). Implementing Best Practices for Data Quality Assessment of the National Renewable Energy Laboratory's Solar Resource and Meteorological Assessment Project (No. NREL/CP-5500-50897). National Renewable Energy Lab.(NREL), Golden, CO (United States). https://www.nrel.gov/docs/fy11osti/50897.pdf

Yang, D., Wang, W., & Xia, X. A. (2022). A concise overview on solar resource assessment and forecasting. Advances in Atmospheric Sciences, 39(8), 1239-1251. https://link.springer.com/content/pdf/- 10.1007/s00376-021-1372-8.pdf

Younes, S., Claywell, R., & Muneer, T. (2005). Quality control of solar radiation data: Present status and proposed new approaches. Energy, 30(9), 1533-1549. https://www.sciencedirect.com/science- /article/abs/pii/S0360544204002233

Zhang, Y., Meratnia, N., & Havinga, P. (2009). Hyperellipsoidal svm-based outlier detection technique for geosensor networks. In *International conference on GeoSensor Networks*, 31-41.