Journal of Energy in Southern Africa

Volume 35 Number 1
2024

An evaluation of variable selection methods using Southern Africa solar irradiation data

Daniel Maposa[1*], **Amon Masache** [2], **Precious Mdlongwa**[2], **Caston Sigauke**[3]

1 Department of Statistics and Operations Research, University of Limpopo, Sovenga, South Africa
2 Department of Statistics and Operations Research, National University of Science and Technology, Ascot, Bulawayo, Zimbabwe
3 Department of Mathematical and Computational Sciences, University of Venda, Thohoyandou, South Africa

Abstract
Dimensionality poses a challenge in developing quality predictive models. Often when modelling solar irradiance (SI), many covariates are considered. Training such data has several disadvantages. This study sought to identify the best variable embedded selection method for different location and time horizon combinations from Southern Africa solar irradiance data. It introduced new variable selection methods into solar irradiation studies, namely penalised quantile regression (PQR), regularised random forests (RRF), and quantile regression forest (QRF). Stability analysis, performance and accuracy metric evaluations were used to compare them with the common lasso, elastic and ridge regression methods. The QRF model performed best in all locations followed by the shrinkage methods on hourly data. However, it was found that QRF is not sensitive to associations through correlations, thereby ignoring the relevance of variables while focusing on importance. Among the shrinkage methods, the lasso performed best in only one location. On the 24-hour horizon, elastic net dominated the performances among the shrinkage methods, but QRF was best in three locations of the six considered. Results confirmed that variable selection methods performed differently on different situational data sets. Depending on the strengths of the methods, results were combined to identify the most paramount variables. Day, total rainfall, and wind direction were superfluous features in all situations. The study concluded that shrinkage methods are best in cases of extreme multicollinearity, while QRF is best on data sets with outliers or/and heavy tails.

Keywords: *multicollinearity; quantile regression; relevance; random forests; shrinkage methods; variable selection*

Journal of Energy in Southern Africa 35(1), 1–23
DOI: https://dx.doi.org/10.17159/2413-3051/2024/v35i1a16336
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: Daniel.maposa@ul.ac.za; danmaposa@gmail.com

Nomenclature

12V	12V battery average
12VMax	12V battery maximum
12Min	12V battery minimum
24V	24V battery average
24VMax	24 battery maximum
24VMin	24V battery minimum
BPAvg	Barometric pressure average
BPMax	Barometric pressure maximum
BPMin	Barometric pressure minimum
CAA	Calculated azimuth angle
CNR	Clustering, nested modelling and regression
CTA	Calculated tilt angle
DHITot	Total diffuse horizontal irradiance
DNIAvg	Direct normal irradiance
DNICal	Calculated direct normal irradiance
F	Cumulative distribution function
GHI	Global horizontal irradiance
MAE	Mean absolute error
MASE	Mean absolute scaled error
NUST	Namibia University of Science and Technology
OOB	out-of-bag
PE	Prediction error
PQR	Penalised quantile regression
QR	Quantile regression
QRF	Qauntile regression forest
R^2	Coefficient of determination
RF	Random forest
RHAvg	Averaged relative humidity
RHMax	Relative humidity maximum
RMSE	Root mean square error
SAURAN	Southern Africa Universities Radiometric Association Network
S_i	impurity decrease
SI	Solar irradiation
TAvg	Temperature average
TMax	Temperature maximum
TR	Total rainfall
VIF	Variance inflation factor
WD	Wind direction
WDAvg	Wind direction average
WDStD	Wind direction standard deviation
WSAvg	Wind speed
WSMax	Wind speed maximum
WVM	Wind vector magnitude
X	vector of covariates
x_i	the ith covariate
Y	response variable
y_i	the ith value of the response variable
β	vector of regression coefficients
β_0	the regression intercept
λ	tunning parameter
α	shrinkage parameter
τ	quamtile level
Q_τ	the τ^{th} quantile
σ	population standard deviation
n	sample size
N	large sample size

1. Introduction

It is inevitable that when studying solar irradiance (SI) one has to consider a lot of variables that may influence the radiation of the sun's energy on the earth's surface. Considering all variables in forecasting models introduces the challenge of the curse of dimensionality. This is a phenomenon where models are negatively affected by the increase in the number of covariates. Some of these covariates may be giving redundant information which does not have any influence on prediction processes. These superfluous variables must be excluded when training the data. The question becomes, which variables are paramount to consider when building a model? It is necessary to find ways of identifying those insignificant features and exclude them in model development before training the model. In addition, a high dimensional training data set can negatively affect a predictive model in several ways: (1) Prediction accuracy is reduced; (2) models do not learn well a large number of irrelevant variables; (3) some important variables may not be picked due to interference from irrelevant variables; (4) it makes the model complex to interpret; (5) the algorithm processing time is increased; (6) too many resources are used in the prediction process; (7) maintenance is difficult. As proved by Hossain et al. (2013), including an optimal feature subset provides better prediction accuracy in forecasting solar power, and the selection of a small (possibly minimal) feature set giving the best possible classification results is desirable for practical reasons. The subset fits well the data because it contains the most paramount variables. An optimal subset reduces overfitting in the model-building process. Different variable selection methods have been used to find this optimal subset of features, but the methods were developed to suit different data conditions. As a result, they focused on different aspects of data sets that have been developed thereby introducing different variable selection performances. That is, on different situational data sets the methods would give different optimal subsets. Therefore, this study is motivated by the need to establish situational SI data sets when different embedded variable selection methods are best applicable. A comparative investigation of existing variable selection methods in SI studies is made here, and new variable selection methods are introduced to achieve this objective.

1.1 Rationale and contribution of the study

Several methods are applicable to solve the curse of dimensionality in different situational data sets. However, according to the best of our knowledge, all studies that included variable selection on Southern Africa SI data applied lasso and/or its extensions (shrinkage methods only) without any comparative analysis. The methods might not have been the best approaches for those different situational data sets. A comparison of the variable selection methods is necessary before application.

The main contribution of the current study is to demonstrate to the solar energy industry, meteorologists and the body of knowledge at large that different variable selection approaches perform differently on different situational data sets. The study showed that, although it may appear that lasso is the ideal variable selection method for Southern Africa SI data sets, it depends on location, time horizon and nature of the data. To complement regularisation in its sensitivity to outliers we introduced penalised quantile regression (PQR). That is, adding the robustness to outliers and/or heavy-tailed data property of quantile regression (QR) to shrinkage methods. Since shrinkage methods only measure the relevance of a variable, random forests (RFs) which measure feature importance were also proposed. RFs were developed to improve learning performance by use of a voting system which enables them to measure the importance of variables as well as predict the response variable. That is, the current study proposed the inclusion of evaluating the importance of variables in addition to relevance when finding an optimal feature subset. It suggested that a subset without corresponding variable importance measures is a local minimal. A global minimal subset of variables should consider variables of both best relevance and importance. In addition to proposing RFs, the study also checked whether regularising them improves their performance. It further proposed hybridising the RF method with QR modelling. Apart from being robust to outliers and/or heavy-tailed data, QR gives unique insights into the predictor-response variable relationship through percentiles. Though this concept of hybridising models has become popular in machine learning methods, the study investigated whether hybridisation improves the embedded variable selection methods used to solve the curse of dimensionality when modelling SI. Hybridisation may not be necessary in some situations.

1.2 Research highlights

This study shows that, although lasso has been popular in Southern Africa SI modelling, it is not always the best among shrinkage methods as a variable selection technique. The root mean squared error (RMSE) was used to measure the goodness-of-fit of the shrinkage methods. However, shrinkage methods concentrate on relevance, so the evaluation study introduced separate RFs. RMSEs were also used to evaluate the goodness-of-fit of RFs and the PQR model. The PQR model was included to check if the weakness of shrinkages in failure to handle data with outliers and/or noise can be improved by hybridising with QR. The adjusted R^2 and mean absolute scaled error (MASE) were used as measures of performance and accuracy respectively. RFs would not perform better than shrinkage methods amid multicollinearity. As a result, a regularised RF (RRF) was also considered. Noting that SI data is heavy-tailed and sometimes contains outliers, it was checked whether hybridising separately both shrinkage methods and RFs with QR would improve their performances. Analysis of variables selected by the different models was done and results (together with the R^2) were used to check the stability of the methods through sensitivity analysis.

2. Related literature

Amongst the several studies done on solar irradiation in the Southern Africa region so far, only five considered the selection of variables, to the best of our knowledge. Four of them applied lasso, but none did a comparison of the variable selection methods to check which one would best apply to their different situational data sets. However, outside Southern Africa, the latest variable selection method comparison in SI studies was done by Muller (2021). Their developed clustering, nested modelling and regression (CNR) model performed the best under high sensitivity when compared to lasso, lasso least angle regression, and elastic net. The study restricted the number of features, and CNR identified relevant information better than any other. In contrast, El Motaki and El Fengour (2021) used meteorological and geographical data with no correlated features as conditional variants in comparing different filter, wrapper and embedded variable selection methods. However, conclusions were made from the reduced number of features, stability and regression accuracy comparatives. Instead, lasso was considered by Tang et al. (2017) as a solar power generation forecasting tool. Comparison analysis was focused on forecasting accuracy rather than optimal variable selection and found lasso's capability to optimise feature selection as a trade-off between complexity and model forecasting accuracy. Outside the solar energy industry, Yilmaz and Kuvat (2023) used the R-square metric to investigate the effect of nine feature selection methods (lasso and elastic net included) on the success of overall equipment effectiveness prediction. Omoruyi et al. (2019) applied mean rank to optimise model selection among direct search, forward selection, backward and stepwise on gross domestic product data. Sanchez-Pinto et al. (2018) concluded that variable section method performance is associated with sample size after comparing regression-based and tree-based algorithms. The elastic net was found to be the most superior among the six algorithms compared by Williams et al. (2015) when applied to frequency and severity models of homeowner insurance claims. Simulation has been found useful in variable selection method comparative studies (Kipruto and Sauerbrei, 2021; Mehmood et al., 2020; Celeux et al., 2015). Surveys and reviews alike can be approaches for investigating the strengths and weaknesses of variable selection methods (Li et al., 2017; Wang, Wang and Chang, 2016; Khalid et al., 2014). All these previous studies were focused on the properties of variable selection methods as the basis for comparison, except El Motaki and El Fengour (2021), who considered variants in the data properties. Therefore, the current study extends variable methods comparison investigation to data sets with separate and combined existence of multicollinearity, heavy tails and outliers. It also adds time horizon variants to meteorological and geographical variants, which has not been done before. Among the existing embedded variable selection methods used in previous studies, the current research introduces PQR, RRF and QRF models.

Lasso being the most common variable selection method in SI studies using Southern Africa data, Mpfumali et al. (2019) and Chandiwana et al. (2021) extended the implementation via hierarchical interactions between predictor variables. They claim that consideration of interactions greatly expands the understanding of the relationships among variables. Mpfumali et al. (2019) highlighted that the lasso is useful especially when dealing with highly correlated predictors but they mixed up relevance and importance in interpreting their results. Mutavhatsindi et al. (2020) presented a bar chart of variable coefficient values and also misinterpreted them as important features. A feature coefficient may be shrunk to zero by a regularisation process, but it does not necessarily mean that it is not important. Ratshilengo et al. (2021) agreed with so many other researchers who used lasso that the method tackles issues of model overfitting. The many advantages of lasso over other regularisation methods have made the method so popular. However, the question is whether the method would be the best in all SI cases.

Leng et al. (2006) highlighted that when superfluous variables exist in a regression model and the design matrix is correlated then the probability of a regularisation method on identifying the true set of important variables is less than a constant, not depending on the sample size. Thus, prediction-accuracy-based criteria are not suitable in such a situation. As a result, some researchers have suggested some adjustments to the lasso, including Alhamzawi and Ali (2018), who extended the adjustment to the Bayesian adaptive lasso. We agree with Belloni and Chernozhukov (2011) that running the selection procedure at quantile levels may help. They developed a PQR, which is a hybrid feature selection of regularisation and quantile regression. Randa et al. (2022) stipulated that penalisation removes at least nearly all covariates whose population coefficients are zero in a QR process. Since penalised least squares regression is not robust to outliers or heavy-tailed error (Su and Wang, 2021), combining the technique with QR (a useful method on heteroscedasticity) has been found to improve regularisation. QR estimates the response feature at different quantile levels which gives precise insight into the relationship at the upper and lower tails of the distribution. Gu et al. (2017) added that the asymptotic behaviour of quantiles can be directly investigated amid increasing data dimensions. PQR is very competitive, accurate and efficient. However, QR is generally inconsistent in high dimensional settings while shrinkage deals with high multicollinearity among covariates. The method has not been applied to SI data from Southern Africa.

The other weakness of shrinkage methods is that they only focus on associations of covariates with the response through correlation. That is, they lack importance measurement of a variable on the response, and yet lack of correlation does not necessarily mean no importance. RFs classify features as important or rejected. Leo Breiman in 2001 advocated that RFs are the best classifiers for high dimensional data. They form an ensemble of weak unbiased classifiers which combine their results during final classification. No tuning is necessary since trees are grown until each leaf contains just a few elements. Munshi and Moharil (2022) added that RFs can handle missing values with no overfitting. They are less affected by noise in the data, robust to outliers, and stable. Villegas-Mier et al. (2022) found that RFs gave a robust performance with similar results in two different scenarios. RFs delivered accurate and precise results when mapping SI at high latitudes (Babar et al., 2020). Lee et al. (2020) also found that lagged solar irradiance features contribute significantly to the ensemble model. Their RF model produced SI at one-hour lag, relative humidity and showed to have high importance scores on USA data from six stations. Zeng et al. (2020) concluded that the RF model had high performance under different climates and geographic conditions. The importance scores computed by Ibrahim and Khatib (2017) showed that sunshine, hour and temperature were the most important features. We appreciate that assessing importance scores guides the feature selection process. However, the concept of ignoring correlations makes RFs insensitive to interaction effects. Therefore, Deng and Runger (2012) proposed a tree regularisation framework based on the random forest method called regularised random forest (RRF). The model was developed to improve the performance of RFs amid data sets with significant multicollinearity.

Now, the quantile regression forest (QRF) model, which includes percentiles, adds a no-parametric property of QR to give valuable information about the dispersion of observations. Freeman et al. (2023) claimed that it is possible to map the upper and lower bounds of predictions with QRF because predicted median and quantiles can be mapped. These predictions from individual trees in the model can follow any probability distribution. Vaysse and Lagaricherie (2017) used the extended ensemble method on soil properties, and it performed better than the common regression kriging. Another strong property of QRF is that it does not assume any prior distribution or stationarity of the response variable, so it is a better methodology to describe variability in the real world than linear QR. This is probably one of the reasons why Vantas et al. (2020) compared QRF and the state-of-the-art kriging method. QRF compared very well with rainfall erosivity data in Greece. Asnaghi et al. (2017) used QRF as a novel approach for coastal management of harmful algal blooms. They found the methodology flexible in such a way that it could be extended to other ecological phenomena that are dependent on meteorological features. Maxwell et al. (2021) also addressed weaknesses in geostatistical methods to model coal properties by proposing a QRF algorithm. The algorithm performed better than the most popular regression kriging method in the field of geostatistics. However, they found out that the algorithm is less intuitive and computationally demanding. The novel approach has not yet been applied to Southern Africa SI data and any SI variable selection study in the globe according to the best of our knowledge.

3. Materials and methods

3.1 Data and variables

Radiometric stations in the Southern Africa region are geographically located as shown in Table 1.

Data uploaded from the radiometric stations by the Southern Africa Universities Radiometric Association Network (SAURAN) into their database can be accessed at their website (https://sauran.ac.za). The Meteorological Services Department in Zimbabwe supplied daily averaged insolation data from their Goetz observatory in Bulawayo. There were at least nineteen features considered from each SAURAN station on the hourly recorded SI datasets, with Windhoek having the highest number of twenty-one variables. The daily recorded data sets had more variables considered with Windhoek having the highest number of thirty-three variables. Solar irradiance from the SAURAN database was measured as hourly averaged global horizontal irradiation (GHI) in watts per square metre. However, at the Goetz observatory, solar irradiation was measured as daily total insolation also in watts per square metre. Features considered from each location are shown in Table 2.

Table 1: Periods of data considered from radiometric stations.

Station	*Latitude*	*Longitude*	*Elevation*	*Location*	*Period*
Goetz	-20.1418223	28.6125335	1346 m	Bulawayo	Jan 2018–Sept 2022
NUST	-22.5650005	17.07500076	1683m	Windhoek	Jul 2017–Jun 2021
SUN	-33.9281006	18.86540031	119m	Cape Town	Jul 2017–Jun 2021
UGB	-24.6609993	25.93400002	1014m	Gaborone	Jan 2015–Nov 2020
UKZNH	-29.8709793	30.97694969	150m	Durban	Dec 2015–Jun 2021
UP	-25.7530804	28.22859001	1410m	Pretoria	Jul 2017–Jun 2021
UV	-23.1310005	30.42399979	628m	Venda	Jul 2017–Jun 2021

Table 2: Variables considered from the SAURAN stations.

Variable	*Units*	*Durban*	*Gaborone*	*Windhoek*	*Cape Town*	*Pretoria*	*Venda*
Diffuse horizontal irradiance	w/m^2	x	x	x	x	x	x
Direct normal irradiance	w/m^2	x	x	x	x	x	-
Calculated direct normal 1rradiance	w/m^2	-	-	x	x	x	x
Temperature	° C	x	x	x	x	x	x
Relative humidity	%	x	x	x	x	x	x
Total rainfall	mm	x	x	x	-	x	x
Wind speed	m/s	x	x	x	x	x	x
Maximum wind speed	m/s	-	-	x	-	-	x
Wind direction	degrees	-	x	x	x	x	x
Wind direction standard deviation	degrees	-	x	x	x	x	x
Wind vector magnitude	degrees		x	x	-	x	x
Barometric pressure	mbar	x	x	x	x	x	x
12V battery average	volts	x	x	x	-	x	x
12V battery minimum	volts	-	x	-	x	x	x
24V battery average	volts		-	x	-	x	x
24V battery minimum	volts	-	-	-	-	x	x
24V-105Ah battery average	volts	x	x	-	-	-	-
12V-105Ah battery average	volts	-	x	-	-	-	-
Logger temperature	° C	-	x	x	-	x	x
Calculated azimuth angle	degrees	-	-	x	-	-	-
Calculated tilt angle	degrees	-	-	x	-	-	-

3.2 Main assumption

It is assumed that all radiometric stations in the Southern Africa region experience similar climatic conditions. As a result, though variable selection methods may select different variables in different locations, conclusions would apply to any other location within the Southern Africa region. That is, a change of location within the same climatic region does not significantly affect features that influence the amount of solar energy received on Earth.

3.3 Shrinkage algorithms

3.3.1 Regularisation methods

A shrinkage method is defined as a general procedure used to improve a least squares estimator and comprises reducing variance by adding constraints on the value of the coefficients. They remove irrelevant variables by reducing the fitted coefficients to zero.

Theorem 1. *A variable X_i is irrelevant to Y concerning X (notwithstanding that $X_i \in X$), for which, for all $A \subseteq X$, the conditional mutual information, I (X_i; Y |A) of X_i and Y given the variable in A is equal to zero.*

Thus, lasso and elastic net (ridge regression also being a special case of the elastic net) were developed to identify such variables in Theorem 1 by shrinking the coefficients of irrelevant variables to zero. The algorithm developed by Friedman et al. (2010) generalises naturally to the unstandardised cases of observations such that

$$\frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - x_i^T)^2 + \lambda P_\alpha(\beta), \tag{1}$$

is minimised, where

$$P_\alpha(\beta) = \sum_{j=1}^{p}\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha\,|\,\beta_j\,|, \tag{2}$$

is the elastic net penalty.

The lasso penalty is also a compromise of the elastic net penalty when $\alpha = 1$ in Equation 2. When $\alpha = 1-\varepsilon, (\varepsilon > 0)$, elastic net performs much like a lasso, but removes any degeneracies and wild behaviour caused by extreme correlations. When $\alpha = 0$ the penalty becomes a compromise to the ridge regression penalty. Ridge regression was proposed to minimise the residual sum of squares subject to the constraint (Equation 3).

$$\sum_{j=1}^{p}\|\,\beta_j\,\|^2 \le t(l_2 \text{ norm}). \tag{3}$$

The solution to the problem in Equation 1 is unique given the condition in Theorem 2, which is fulfilled when $p \le n$. That is, assuming that X has a full rank then:

Theorem 2. β^* *is a unique solution to the Lagrangian form in Equation 5 if*

$$C(\beta^*) \cap N(X) = \{O\},$$

where O is a p-dimensional null vector and $c(\beta^*) = nx \in C(\beta^*) = \{x \in R^p : x^T e \le 0, \quad \forall e\}$.

However, it is noted that although the solution is not necessarily unique it is still convex and neighbourhood search results can be used. If cross-validation is applied when solving the problem in Equation 1 for a good value of a regularisation parameter, then λ is chosen to minimise the prediction error (PE) in Equation 4:

$$PE = E(Y - \beta X)^2, \tag{4}$$

where $Y \in R^{N\times 1}$ is a vector of the response variable observations and $X \in R^{N\times p}$ is the matrix of predictor variables. Different structure models are obtained from several choices of λ's that give the same PE. However, the choice of the parameter using an *n*-fold cross-validation is not stable in many cases (Park and Casella, 2008).

The main objective is to achieve better prediction in the face of multicollinearity while preventing overfitting. Lasso shrinks the coefficients of correlated variables towards each other which allows them to borrow strength from each other. However, the method shrinks all regression coefficients to zero and yet some variables may be important. On the other hand, lasso ensures that only relevant variables are selected. However, lasso tends to pick one and ignore the rest of a group of highly correlated variables. The selected coefficient has a high variance because this collinearity causes the corresponding coefficient standard error to become unstable. Regularisation is achieved by continuously shrinking the coefficients such that if λ is sufficiently large then some coefficients are shrunk to zero. However, a little bias to reduce the variance of the predicted values is compromised at the overall benefit of improving the prediction accuracy. It becomes a relevant technique for situations where regularised methods are applied as the inclusion of statistical modelling for the feature selection process and the results used for further analyses. Feature selection and regularisation performed by lasso enhance model prediction accuracy by tackling the problem of overfitting. It does so on finite samples and performs well in cases where *p* may grow faster than *n*. It is a novel approach to problems of high dimensional non-linear modelling where the structure of the model has to be detected. Most other methods are not adequate, in that they do not deal well with large numbers of irrelevant explanatory variables. This makes lasso to be an advantageous variable selection procedure in difficult forecasting problems. However, lasso in particular selects at most *n* variables before saturating in small *n* but large number of variables data sets. It is a challenge to apply shrinkage methods when some variables are recorded through some calculations, because there would be no common referral point for the features to choose which ones should be selected.

3.3.2 Penalised quantile regression

When cross-validation is repeated, shrinkage methods tend to be unstable. In addition, these regularised least square regression methods are not robust to outliers or heavy-tailed error distribution. Yet QR is robust, and sparse and gives unique insights into the relationship between features and response variables. Penalisation in QR removes at least nearly all features whose population coefficients are shrunk to zero whilst (in particular comparison to lasso) the method of percentiles deals with its coefficients' instability under repeated cross-validations. So, penalised QR offers sparse solutions as well as performing automatic feature selection. The *l*l_1 PQR estimator $\hat{\beta}(\tau)$ is a solution to the optimisation problem, as shown in Equation 5.

$$\underbrace{\min}_{\beta\in R^p}\left\{\hat{Q}_\tau(\beta)+\frac{\lambda\sqrt{\tau(1-\tau)}}{n}\sum_{j=1}^{p}\hat{\sigma}_j^2\,|\,\beta_j\,|\right\}, \qquad (5)$$

where $\hat{\sigma}_j^2 = E(x_{ij}^2)$ and T ⊂ (0, 1) such that

$$F^{-1}_{y_i|x_i}(\tau\,|\,x_i)=x'\beta(\tau), \quad \beta(\tau)\in R^p, \quad \forall\tau\in T. \qquad (6)$$

The overall penalty level $\lambda\sqrt{\tau(1-\tau)}$ depends on each quantile τ, whilst λ depends on T.

Table 3 gives a summarised comparison of the lasso, elastic net, ridge and PQR as a group of shrinkage variable selection methods.

3.4 Tree-based algorithms

3.4.1 Random forests

RFs are popular learning models for solving a variety of classification and regression problems. They are based on a multitude of decision trees which are independently developed on different sample bags taken from the training set. RFs are an ensemble method in which classification is performed by voting of multiple unbiased decision trees. Different subsets of attributes are randomly selected at each step of tree construction. These subsets are different bootstrap samples of the training set. Each bootstrap sample is a result of the replacement selection of the same objects as in the original set. Trees that are trained on different parts of the same training set are averaged to reduce variance, but this increases bias and the models may lose interpretability. Consequently, this will pull together efforts of the tree algorithms. Therefore, the performance of a single random tree is improved by this teamwork of many trees. However, the performance of the final model is greatly boosted. Trees that are grown very deep

Table 3: Comparison of shrinkage methods.

Method	*Strengths*	*Weaknesses*
Lasso	Ensures that only relevant variables are selected. Both a variable selector and forecasting model. Adequate on too high-dimensional data and non-linear modelling.	Pick one and ignore the rest of a group of highly correlated variables. Selects at most *n* variables before saturating in small sample sizes but high dimensional cases.
Elastic net	Removes degeneracies and wild behaviour caused by extreme correlations. Encourages a grouping effect. Provides a compromise between lasso and ridge regression.	It is hard to tune two hyperparameters simultaneously. It may not be interpretable or explainable.
Ridge	Achieves good prediction when covariates are correlated while preventing overfitting.	The coefficients of some important variables may be reduced to zero. The penalty term cannot force the coefficients to be exactly zero. Coefficient estimates can change substantially when multiplying a given predictor by a constant.
PQR	Robust to outliers. Gives unique insights into the relationship between features and response variables at all quantile levels of the distribution. Deals with its coefficients' instability under repeated cross-validation.	Lacks the ability to reveal grouping information. Computationally complex. The check loss function is not smooth.

can learn highly irregular patterns in the data and have a low bias at the expense of overfitting the training sets. The original dataset is extended by adding the so-called shadow features whose values are randomly permuted among the training cases to remove their correlations with a decision variable. Then importance estimation of a feature is calculated as the loss of classification accuracy caused by a random permutation of feature values of cases. That is, the importance of any variable is evaluated as the mean decrease impurity importance (MDI), shown in Equation 7.

$$\mathrm{Imp}(X_i) = \frac{1}{n_J}\sum_{J}\sum_{j\in J: v(s_j)=X_i} p(j)\delta_i(s_i, j), \tag{7}$$

where $p(j)$ is the proportion n_j/n of samples reaching j nodes, $p(j)\delta_i(s_i, j)$ is the weighted impurity decrease and $v(s_j)$ is the variable used in split s_j. Now, if X^{-i} denotes the subset $X\{X_i\}$ and $P_k(X^{-i})$ is the set of subsets of X^{-i} of cardinality k then we can use Theorem 3 to compute MDI.

Theorem 3. *MDI importance of Xi ∈ X for Y as computed with an infinite ensemble of fully developed randomised trees and an infinitely large training sample is as shown in Equation 8.*

$$\mathrm{Imp}(X_i) = \sum_{k=0}^{p-1}\frac{1}{{}^{k}C_p}\frac{1}{p-1}\sum_{A\in P_k(X^{-i})} I(X_i;Y\,|\,A). \tag{8}$$

We deduce from Theorem 4 that an irrelevant variable has no importance, but among relevant variables we can have strongly relevant ones, which are classified as confirmed important. Then weakly relevant variables are classified as tentative.

Theorem 4. *X_i ∈ X is irrelevant to Y concerning A if and only if its infinite sample size importance as computed with an infinite ensemble of fully developed randomized trees built on X for Y is 0.*

RFs improve learning performance through a voting system given a set number of decision trees. As new objects come in, all trees in the forest classify them and the final decision on the new objects is made through this voting system. Trees vote for the classification of objects which were not involved in their classification. The votes for a correct class are recorded for each tree. Then values of variables are randomly permuted across objects and the classification is repeated. In summary, RFs exhibit characteristics of random feature selection, bootstrap sampling, out-of-bag (OOB) error estimation and full-depth decision tree growing. That is, an RF model first extracts some of the samples by bootstrap sampling and then randomly selects the features of these samples, as shown in Figure 1.

These two steps of random sampling make RF more tolerant to the noise in the data and reduce the possibility of overfitting. However, when data has random correlations in a large number of variables it is difficult for RFs to distiguish truly important variables from those that gain importance. An RRF may be considered on such data.

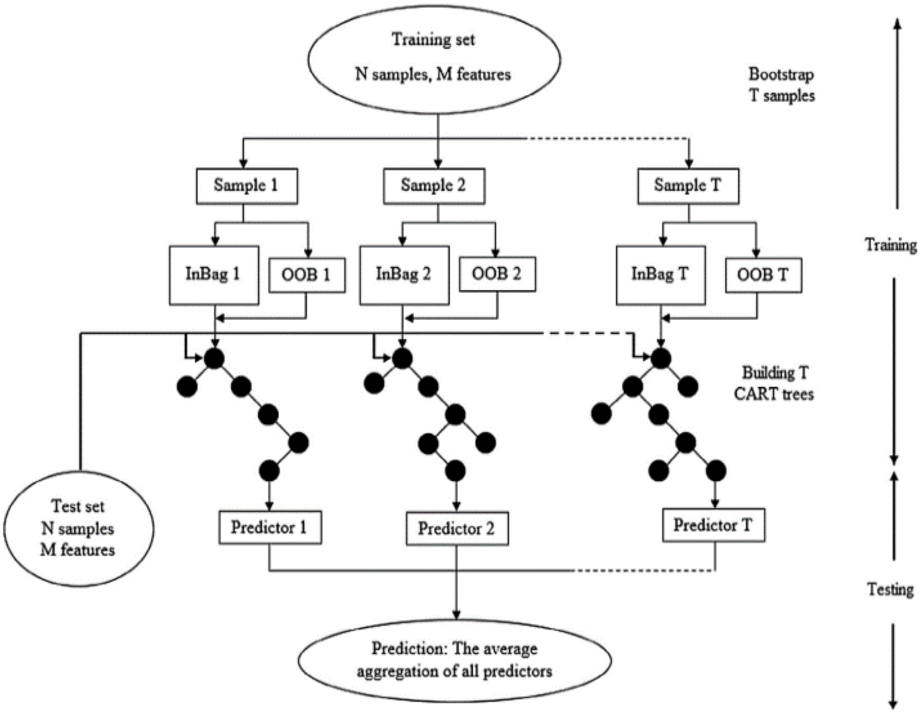

Figure 1: RFs algorithm (Source: Ibrahim and Khatib, 2017).

3.4.2 Regularised random forests

The RRF framework is generally for feature selection where the process is greedy and features are chosen on a sub-sample from each node. That is, in each of those nodes all observations are analysed. Feature selection is done by avoiding features not belonging to *F* (a feature set used in previous splits in a tree model) unless a regularised information gain is defined by Equation 9, where

$$\text{gain}\,(X_i, v) = \begin{cases} \lambda.\text{gain}\,(X_i, v), & \text{for} \quad X_i \notin F \\ \text{gain}\,(X_i, v), & \text{for} \quad X_i \in F \end{cases} \qquad (9)$$

is significantly larger than the maximum gain. $\lambda \in [0, 1]$ is the penalty charged on gain (X_i , v) for a feature that does not belong to *F*. Therefore, the assumption is that maximising gain (X_i , v) selects the splitting feature at any tree node.

3.4.3 Quantile random forests

QRFs assess the conditional distribution of the response i.e. in each tree all of the leaves keep all of the relevant observations. If the observations are unequally weighted, then a good approximation of the full conditional distribution can be delivered. If the weights $W_i(x)$ are calculated, then the conditional distribution of the response given by Equation 10.

$$F(y \mid X = x) = P(Y \leq y \mid X = x) \qquad (10)$$

can be estimated as

$$\hat{F}(y \mid X = x) = \sum_{i=1}^{n} W_i(x) 1_{\{Y_i \leq y\}}. \qquad (11)$$

That is, for any τ -quantile level defined by Equation 12,

$$\tau = P(Y < Q_\tau(x) \mid X = x), \quad \tau \in (0,1), \qquad (12)$$

we can consider the spread of the response in the form of a quantile, as in Equation 13.

$$Q_\tau(x) = \inf \left\{ y : \hat{F}(y \mid X = x) \geq \tau \right\}. \qquad (13)$$

Now, this conditional distribution model can be solved as an optimisation problem, by Equation 14.

$$\hat{F}(y \mid X = x) = \sum_{i=1}^{n} W_i(x) 1_{\{Y_i \leq y\}}, \quad (14)$$

where λ is the tuning parameter.

3.5 Methodology

The evaluation of variable selection methods considered in this study was focused on a comparative investigation approach in different locations and time horizons. The study focused on performance comparison of embedded variable selection methods. The methods determine the best feature subset while building the statistical learning model itself which is ideal for solar irradiance modelling. Embedded methods are very relevant algorithms to our study because they select variable subsets in the course of training the data (El Motaki and El Fengour, 2021) and, further, use their results to determine how solar irradiation forecasting can be improved. It was deduced from the literature that solar irradiation data contain some outliers and have significant multicollinearity among covariates. The presence of outliers and correlated covariates can significantly influence the performance of variable selection algorithms. Thus, the variance inflation factor (VIF) was used to assess the collinearity of variables considered from each data set. Then multicollinearity was measured as a proportion of variables with VIFs greater than 1. A Grubb's test was also conducted on each data set to check the presence of at least one outlier. The normality of the response has a bearing on the properties of variable selection algorithms, so the distribution of solar irradiance was explored through the skewness statistic computation, box plot construction and interpretation. As a result, there was a need to apply robust and sparse variable selection algorithms like embedded. Apart from giving diverse samples that are good enough to come up with significant associations between radiometric stations, the multi-site forecasting approach enhances the statistical power of an algorithm (Sigauke et al., 2023). A cross-validation process was then implemented on each multi-site data set, that is, splitting the data set: 80% going into training and 20% into test samples. Models were fitted on the training data frames and then variable selection capabilities were analysed using the testing data frames. The partitioning strategy avoids the overfitting problem, improves prediction based on bias and/or variance, assesses how effectively the model will perform in real-world scenarios and allows for predicting how well a model will perform on data that it has not seen before (Yilmaz and Kuvat, 2023).

The embedded methods were also classified into two groups, namely the shrinkage methods (lasso, elastic net, ridge, PQR) and tree-based methods (RF, RRF, QRF). Shrinkage algorithms were developed for better performance while avoiding overfitting on multicollinear covariates through the assessment of variable relevance. On the other hand, tree-based algorithms may overfit the data but can learn highly irregular patterns in solar irradiation data with low bias. The ensemble of decision trees selects variables by classifying them as important or not. QR hybrids were also introduced among the algorithms because QR is robust to outliers, whose presence in solar irradiation data has been indicated by the literature. All tree-based algorithms are non-parametric models, which are best adapted for modelling response variables like solar irradiance, which has no known relationship structures with its covariates as yet. The RMSE metric was used to find the best among lasso, elastic net and ridge. These three shrinkage methods are called regularisation algorithms. The three regularisation algorithms have the same model structure, i.e. Equation 1. As a result, we presumed that, for any sample data, if the best algorithm (in terms of the RMSE) among these three regularisation methods is inferior to any other model compared in this study then the inferior models among the regularisation methods can never outperform that other model. For example, suppose that for the Venda data set, the elastic net has the lowest RMSE among the regularisation algorithms but it is outperformed by the RF algorithm. We would not expect lasso or ridge to outperform the RF algorithm. That is, the best regularisation algorithm is the one we compared with PQR and the rest of the tree-based algorithms. In that context, we combined the regularisation algorithms through the 'glmnet' R programming software package developed by Hastie et al. (2023). The package uses one penalty, λ but adds another parameter $\alpha \in [0, 1]$ such that the general structure of the regularisation model can be written as Equation 15:

$$L(\hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n} w_i l(y_i, \beta_0 - \beta^T x_i) + \lambda\left[(1-\alpha \parallel \beta \parallel_2^2 / 2 + \alpha \parallel \beta \parallel_1)\right], \qquad (15)$$

where $L(\hat{\beta})$ is minimised. The parameter λ controls how much of the tuning needs to be done on the penalised least squares regression process. The software has an inbuilt k-fold cross-validation (CV) algorithm where k (model size) is automatically selected on which the resulting test has the smallest CV error. While the three regularisation algorithms used multiple linear regression to learn the data, PQR uses multiple linear QR and the method of quantiles to estimate the regression coefficients. The Boruta algorithm was used to train the RF and RRF models while the QRF was trained using non-parametric quantile regression.

To prevent the researchers' judgment from biasing the results, a forward selection technique was implemented on each multi-site data set. That is, all variables in each data set were fed into the model at once and then assessed according to the extent each algorithm provides an optimal subset that results in a highly predictive model. The assessments were done by analysing the experimental error according to the RMSE, adjusted-R^2 and MASE evaluation metrics. The RMSE was used as a basis for comparing the goodness-of-fit of the models, and with adjusted-R^2 their predictive performances were compared. Since solar irradiation data include zero values of GHI, MASE was used as a basis metric for predictive accuracy comparisons. Performance scores on each metric evaluation were generated and a system of ranking the models was introduced to finalise the comparison investigation. Other algorithm comparative considerations were the ease of setting up the model in software and the speed of processing results when running the coded algorithm.

Through listing, common variables with coefficients shrunk to zero and rejected were identified by inspection. They were determined from the relevant and importance scores calculated using the respective fitted models. The flow chart in Figure 2 summarises the comparative investigation approach adopted in this study. The stability of the models was checked through a sensitivity analysis, where it was observed how the model performances changed when sample sizes were changed. That is, we checked for consistencies in R^2 values as we varied sample sizes and general performances across different locations. We also checked whether the algorithms were selecting the same variables in these different locations.

3.5.1 Goodness-of-fit evaluation

The goodness-of-fit of the models was evaluated by measuring the deviations of the fitted from the actuals using computing the RMSE, as given in Equation 17.

$$RMSE = \sqrt{\frac{1}{n}(y_i - \hat{y}_i)^2}, \qquad (17)$$

where y_i is the actual observed SI and $\hat{y}_i$ is the corresponding model fitted value. That is, it is the standard deviation of the residuals. The smaller the RMSE the better the model fits the data.

3.5.2 Performance evaluation

Performance is the universal metric for evaluating a learning model. Performance was measured by calculating the proportion of the total variation in solar irradiation that could be explained by the covariates. That proportion was then expressed as a percentage through the coefficient of multiple determination, R^2. Difficulties in interpreting it are avoided by considering the adjusted R^2. For a p dimensional data set the adjusted R^2 can be computed as in Equation 18.

$$AdjR^2 = 1 - \left(\frac{n-1}{n-p}\right)(1 - R_p^2). \qquad (18)$$

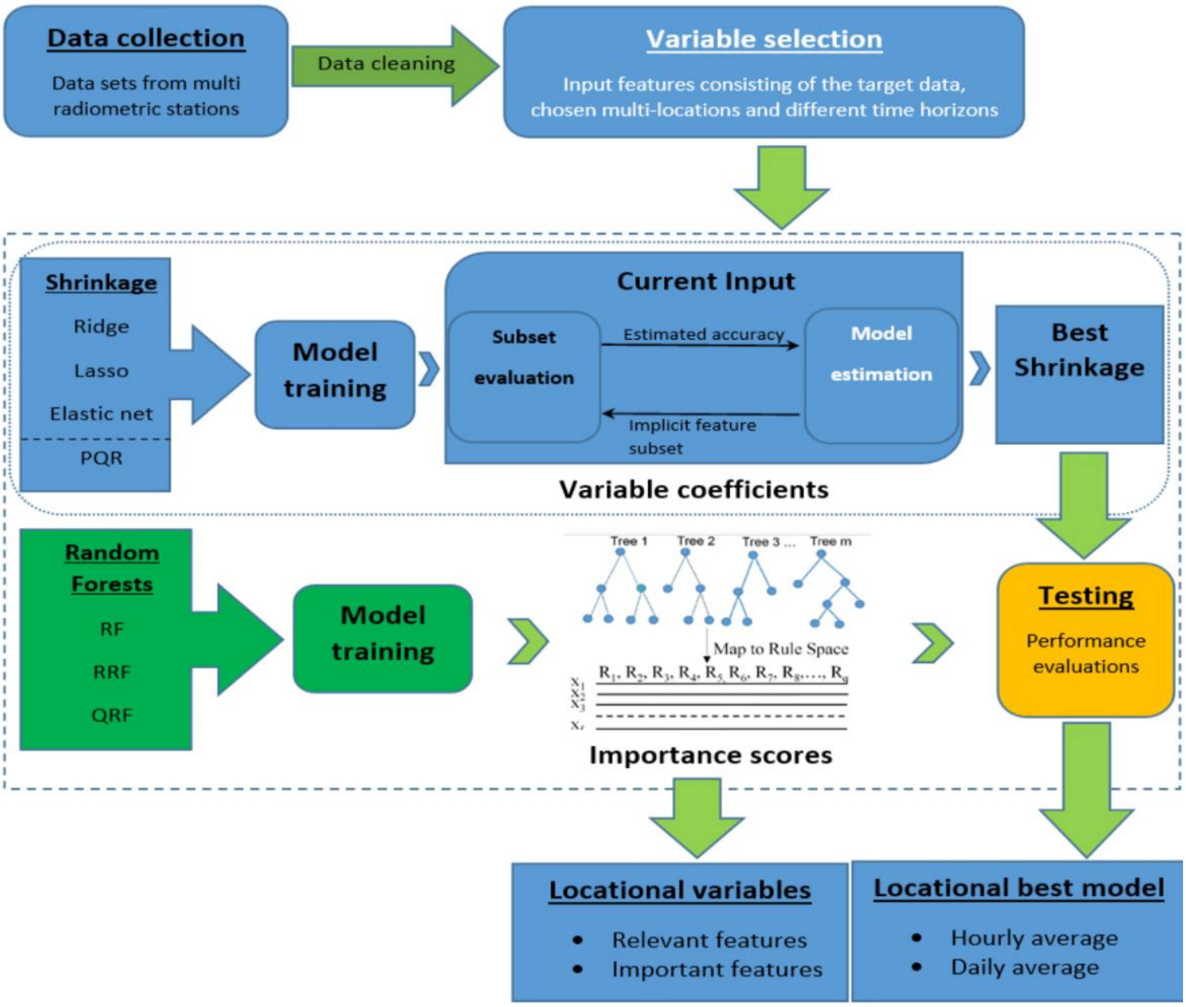

Figure 2: Methodology flow chart.

The adjusted R^2 does not necessarily change as more and more covariates are introduced and the best model is the one that gives the maximum value of the metric.

3.5.3 Accuracy evaluation

The MASE provides an interpretable measure of accuracy in predictive modelling. It is a good metric to use when comparing models trained on different datasets. It is one of the most appropriate metrics when the response has zero or near zero values. The metric is computed by dividing the mean absolute error (MAE) of the trained model by the MAE of the corresponding naïve mode. The naïve model predicts the value at a time point as the previous historical value. That is, MASE indicates the effectiveness of a forecasting model concerning a naïve model. As a result, a MASE greater than 1 means that the forecasting model is performing worse than the naïve benchmark otherwise it is better. A forecasting model with a lower MASE is a better model than the one compared to.

3.5.4 Sensitivity analysis

A sensitivity analysis was done to compare conclusions between the analysis carried out and another analysis in which some aspect of the approach is changed – for example, changing parameters or assumptions of the modelling process, such as support, confidence, or lift, to observe how the model changes and find the optimal values. The sensitivity analysis attempts to assess the appropriateness of a particular model specification and to appreciate the strength of the conclusions being drawn from such a model. The process involves a series of methods to quantify how the uncertainty in the output of a model is related to the uncertainty in its inputs. In this way it assesses how "sensitive" the model is to fluctuations in the parameters and/or data on which it is built.

4. Results and discussions

4.1 Data exploration

Multicollinearity was quite high on the hourly data sets from all except Cape Town, which had 24%; the rest had more than 30%, as shown in Table 4. Windhoek had an extreme multicollinearity of 72% on daily recorded variables but Bulawayo had a very low multicollinearity of 13%. A Grubb's test for outliers shows that Cape Town and Durban hourly SI had outliers because they had p-values less than 0.05. Bulawayo is the only one that had outliers on the 24-hour horizon (p-value $=1.07\times10^{-6}$). Hourly SI is positively skewed because all of the skewness values were more than 1.0, while 24-hour SI is not skewed. The skewness values of 24-hour SI can be approximated to zero, as shown in Table 4. Hourly SI is also heavily right-tailed on all locations, as shown by all box plots in Figure 3, whilst daily SI is symmetrically distributed (Figure 4). All of the box plots in Figure 3 have right-hand side whiskers and no left-hand side whiskers.

4.2 Evaluation of regularisation algorithms

The first model comparisons were done amongst the regularisation algorithms because they use the same model structure shown in Equation 1. To determine the best regularisation algorithm, RMSEs from trained models in Equation 14 for each value of α were calculated and are given in Table 5. The best regularisation algorithm for a particular location is the one with an RSME in bold. The results show that lasso was the best shrinkage method in only one location, Windhoek on hourly SI data. The data set had a high multicollinearity percentage. Either elastic-net or ridge regression was the best for the other locational data sets. It can also be observed that ridge regression was the best in Venda, where the data set had the highest multicollinearity percentage.

Table 4: Multicollinearity, outliers' tests and skewness results.

	Multicollinearity percentage		*Grubb's test p-value*		*Skewness*	
Location	*Hourly*	*Daily*	*Hourly*	*Daily*	*Hourly*	*Daily*
Bulawayo	-	13	-	1.07×10^{-6}	-	-0.060
Cape Town	24	26	1.34×10^{-7}	158.41021	5.906	0.066
Durban	32	-	4.90×10^{-10}	-	1.856	-
Gaborone	42	42	1	1	1.111	0.017
Pretoria	35	26	1	1	1.126	-0.006
Windhoek	43	72	1	0.21	1.004	0.283
Venda	50	57	1	1	1.296	0.000

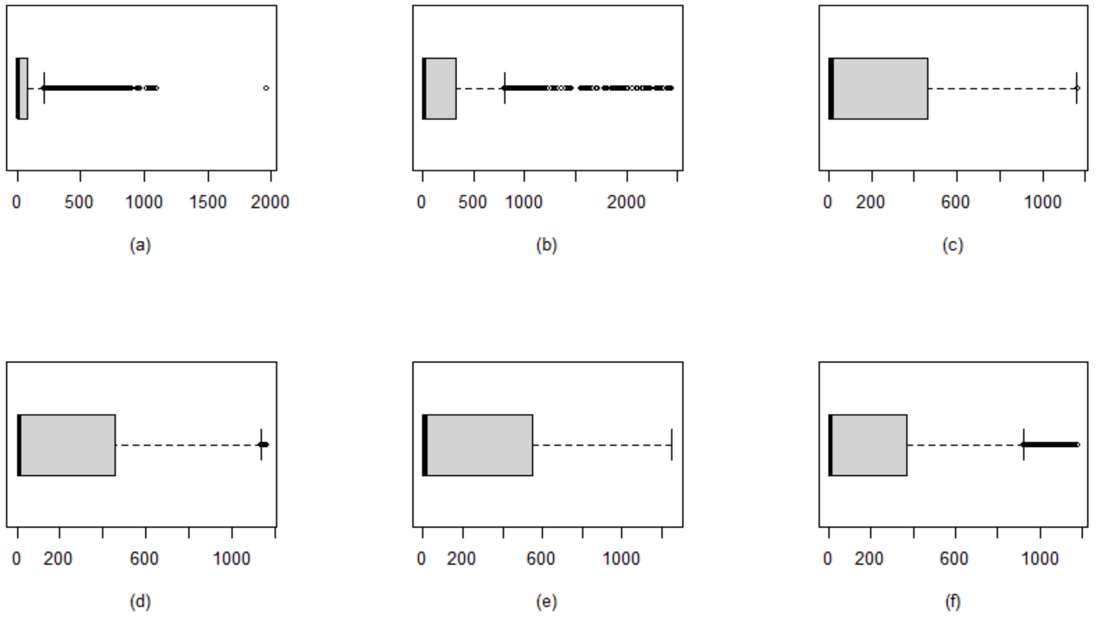

Figure 3: Hourly averaged SI box plot for (a) Cape Town, (b) Durban, (c) Gaborone, (d) Pretoria, (e) Windhoek, and (f) Venda.

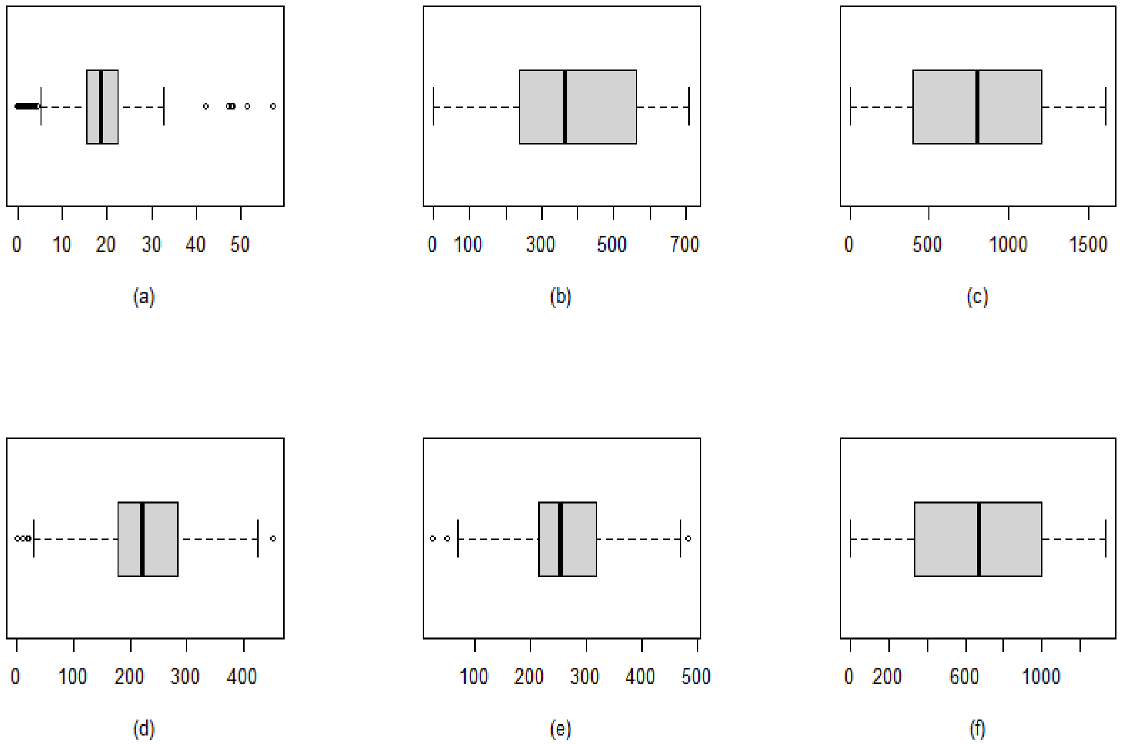

Figure 4: Daily averaged SI box plot for (a) Bulawayo, (b) Cape Town, (c) Gaborone, (d) Pretoria, (e) Windhoek, and (f) Venda.

Table 5: Comparison of regularisation algorithms using the RMSE for hourly SI.

Alpha	*Venda*	*Pretoria*	*Windhoek*	*Cape Town*	*Gaborone*	*Durban*
0.0	0.158	0.743	0.869	0.775	0.249	1.169
0.1	0.841	0.734	0.740	0.827	0.653	0.927
0.2	0.878	0.733	0.722	0.828	0.655	0.929
0.3	0.885	0.737	0.725	0.829	0.657	0.931
0.4	0.888	0.740	0.727	0.830	0.650	0.933
0.5	0.888	0.742	0.722	0.831	0.648	0.936
0.6	0.881	0.744	0.722	0.831	0.648	0.936
0.7	0.884	0.747	0.722	0.832	0.645	0.939
0.8	0.881	0.749	0.719	0.833	0.643	0.941
0.9	0.877	0.750	0.719	0.833	0.644	0.941
1.0	0.878	0.752	0.718	0.834	0.642	0.941
Best	Ridge	Elastic net	Lasso	Ridge	Ridge	Elastic net

A 24-hour time horizon was also considered, in order to check how the time horizon influences the variable selection methods. Results show that lasso was not the best in any of the locations considered, as shown in Table 6. The elastic net was the dominant 24-hour SI in the variable selection context, instead. Bulawayo was the only location where the elastic net was inferior to ridge regression. We observe that Bulawayo 24-hour data had the lowest multicollinearity percentage of 13, but outliers were existent in the data set. The data set was one of only two that was negatively skewed.

Table 6: Comparison of regularisation algorithms using the RMSE for 24-hour SI.

Alpha	*Venda*	*Pretoria*	*Windhoek*	*Cape Town*	*Gaborone*	*Bulawayo*
0.0	7.530	2.500	2.002	0.948	15.956	0.2568
0.1	4.026	2.663	1.261	0.310	16.001	0.2571
0.2	3.808	2.636	1.307	0.313	16.003	0.2571
0.3	3.943	2.618	1.319	0.270	16.043	0.2572
0.4	3.475	2.614	1.153	0.228	15.915	0.2572
0.5	3.437	2.496	1.160	0.181	15.701	0.2571
0.6	3.273	2.562	1.140	0.226	16.024	0.2573
0.7	3.340	2.597	1.148	0.167	16.168	0.2571
0.8	3.431	2.607	1.121	0.137	16.187	0.2573
0.9	3.599	2.617	1.148	0.240	16.268	0.2571
1.0	3.513	2.575	1.154	0.217	16.519	0.2573
Best	Elastic net	Elastic net	Elastic net	Elastic net	Elastic net	Ridge

Table 7: RMSEs for hourly SI.

Location	*Shrinkage*	*PQR*	*RF*	*RRF*	*QRF*
Cape Town	45.772	84.708(τ =0.9)	434.678	434.148	11.707(τ =0.5)
Durban	65.532	48.020(τ =0.6)	406.293	406.368	38.410(τ =0.6)
Gaborone	106.431	109.273(τ =0.6)	455.393	455.461	28.199(τ =0.5)
Pretoria	82.969	87.486(τ =0.5)	438.869	438.855	18.069(τ =0.5)
Windhoek	94.351	109.965(τ =0.5)	505.733	505.728	13.094(τ =0.5)
Venda	69.757	71.562(τ =0.5)	416.607	416.524	30.659(τ =0.5)

4.3 Comparisons of variable selection methods

The best regularisation algorithm on each location was compared with the PQR, RF, RRF and QRF variable selection methods. All algorithms were coded and run in R programming software where the 'quantreg' (Koenker, 2018) and 'rqPen' packages were used to fit the PQR model, the 'Boruta' (Kursa and Rudnicki, 2022) package fitted the RF model, the 'randomForest' and 'RRF' (Deng et al, 2022) packages were used to fit the RRF model and 'quantregForest' (Meinshausen, 2022) package fitted the QRF model.

4.3.1 Goodness-of-fit evaluations

The RMSE for the best regularisation algorithm on each location was also compared against PQR, RF, RRF and QRF. QRF had the lowest RMSEs on all locations for hourly SI as shown in Table 5 (RMSE in bold shows the smallest RSME for that particular location). Results show that PQR had larger RSMEs than regularisation algorithms in all locations except for Durban where there was an improvement from elastic net. We suspect that the reason is that more than 90% of the features considered in all locations were important as evidenced by importance scores from RF, RRF and QRF models. That is, data sets considered did not have superfluous features. Therefore, hybridising a regularisation model with quantile regression would not improve the model. Thus, the results demonstrate that shrinkage methods have been developed to overcome the challenges of multicollinearity in modelling because they could handle the high multicollinearity in the data sets considered. On all locations shrinkage methods performed markedly better than RFs. Even regularising the RF did not improve the selection process except for Gaborone 24-hour data as shown in Table 7. The RMSEs in both Table 7 and Table 8 of all corresponding RRFs were higher than those of RFs, worsening the performance of the RFs. Surprisingly, adding the non-parametric property of QR to the RF model significantly improved the feature selection performance of an RF on all locations except for 24-hour SI in Gaborone. QRF hybrid model became the best method for all hourly SI. The non-parametric property of QR and its other several advantages in regression modelling could be attributed to this significant improvement.

Table 8: Selection methods RMSEs for 24-hour SI.

Location	*Shrinkage*	*PQR*	*RF*	*RRF*	*QRF*
Bulawayo	5.598	5.702(τ =0.5)	7.599	7.599	5.343(τ =0.5)
Cape Town	36.477	40.652(τ =0.8)	252.090	252.285	32.636(τ =0.6)
Gaborone	336.317	351.345(τ =0.6)	564.911	202.070	212.501(τ =0.6)
Pretoria	18.823	14.939(τ =0.6)	103.403	103.569	21.227(τ =0.5)
Windhoek	12.560	12.676(τ =0.6)	91.268	91.302	20.224(τ =0.6)
Venda	321.936	353.215(τ =0.7)	485.782	486.607	260.412(τ =0.6)

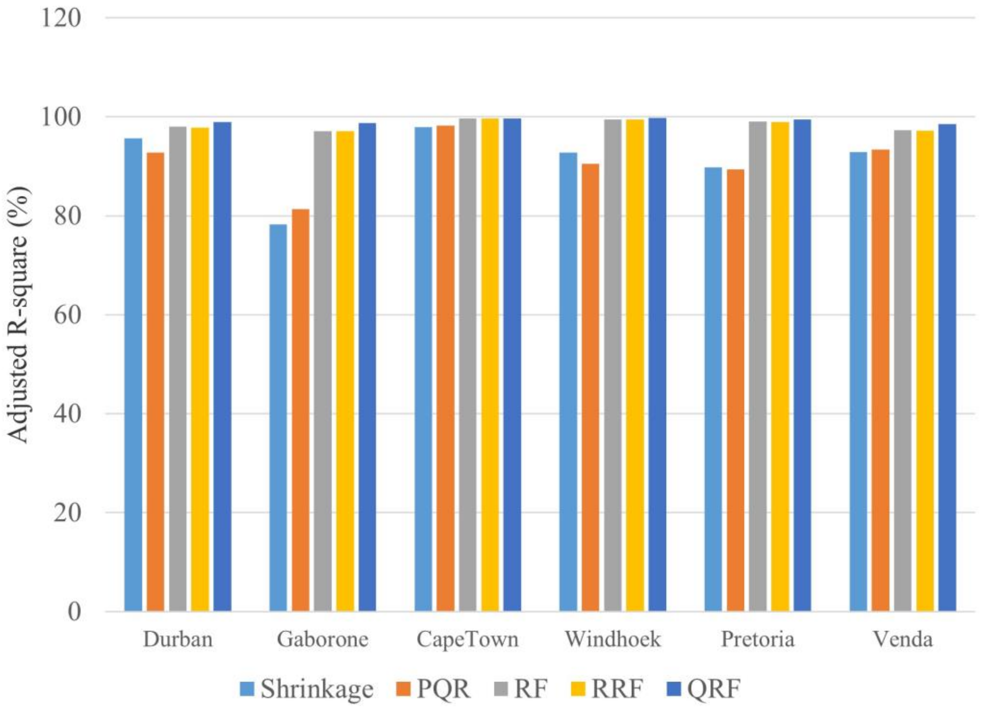

Figure 5: Bar chart exhibiting the R-squared scores.

It is noted that both QR and RFs are robust to outliers. RFs are also tolerant of outliers. In locations where there were outliers (Cape Town for hourly time horizon and Bulawayo for daily time horizon) the performance of QRF was the best among the locations. Median conditional distribution (i.e. at τ = 0.5) gave the best description of solar irradiation in almost all locations. However, the model cannot measure the association of features with the response through correlations. Though QRF was better than any of the RFs, it performed best in only three locations and elastic net was the best in Windhoek for the daily time horizon (it is noted that Windhoek has extreme multicollinearity). Likewise, hybridising a shrinkage method with QR on hourly time horizon data sets did not improve the selection method on daily recorded SI except for Pretoria, where a PQR was the best selection method.

4.3.2 Adjusted R-square comparisons

The adjusted R-square was used as a performance indicator in this study. To analyse the performance, Figure 5 shows the R-squared scores. All of the adjusted R-squared scores were at least 90% except for shrinkage and PQR methods on the Gaborone data set. The shrinkage methods had a 78.25% score while PQR had 81.33%. The data set had a relatively high multicollinearity percentage of 42, no outliers and was comparatively skewed like any other data set used in this study. Further investigations may be required on the data set to find out the reason for comparatively low adjusted R-squared scores. The new QRF model had the highest adjusted R-squared scores from all data sets, followed by RRF and then RF. It showed that the hybridisation of an RF with QR performed better than the RF on its own. Results show that RF-based algorithms had better predictive performances than regularisation methods. This means that RF-based algorithms could explain the total variation in solar irradiation caused by the covariates considered in this study better than shrinkages and PQR. Results also show that introducing QR on regularisation methods did not improve their performance, because the PQR model had smaller or almost equal adjusted R-squared scores than shrinkage on all locations.

4.3.3 Accuracy evaluations

Results in Table 9 show that all of the models performed better than their corresponding naïve models when trained to the data from all locations under study. All of the MASE values were less than one. The results also show that QRF had the smallest MASE values in each location (the locational MASE values in bold) meaning that it was the most accurate variable selection method for each location. It is also observable that it is the same method that had the smallest (MASE=0.026) among all computed MASE values. Therefore it can be deduced that QRF was the most accurate variable selection method.

4.3.4 Model rankings

The average rankings of the variable selection methods over all of the locations are shown in Table 10. Results show that QRF was ranked first on all of the metrics. That is, QRF is the overall superior variable selection method among the methods considered in this study. The shrinkage method was ranked second on RMSE but last on both adjusted R-square and MASE. Though RFs were ranked better on performance and accuracy, shrinkages were better fitting the data.

4.4 Feature selection evaluation

Features with coefficients shrunk to zero or rejected were different in the selection method, location and time horizon. Table 11 shows that total rain, wind speed, maximum wind speed, wind direction, wind direction standard deviation, 12V battery, 12V battery minimum and 24V battery can be excluded from hourly SI prediction modelling. The features were either rejected or had coefficients that were shrunk to zero in at least two locations or two methods.

Table 9: MASEs for hourly SI.

Location	*Shrinkage*	*PQR*	*RF*	*RRF*	*QRF*
Cape Town	0.110	0.070	0.038	0.038	0.032
Durban	0.119	0.115	0.051	0.051	0.034
Gaborone	0.305	0.286	0.115	0.114	0.064
Pretoria	0.250	0.246	0.069	0.071	0.045
Windhoek	0.209	0.202	0.041	0.040	0.026
Venda	0.209	0.187	0.106	0.107	0.069

Table 10: Model comparison by average rankings.

Metric	*Shrinkage*	*PQR*	*RF*	*RRF*	*QRF*
RMSE	2	3	5	4	1
Adjusted R^2	5	4	2	3	1
MASE	5	4	2	3	1

Table 11: Features not selected on hourly SI.

Location	*Shrinkage*	*PQR*	*RF*	*RRF*	*QRF*
Cape Town	None	Month,12V	None	None	None
Durban	TR,WSAvg	WSAvg	None	None	None
Gaborone	None	Hour,WSAvg	None	TR	TR
Pretoria	12V,12VMin, 24VMin	WSAvg,WVM,24Min, WDStD,BPAvg,12V,TR	None	TR	TR
Windhoek	12V,Year,Day	WSAvg,WVM,WDStD, RHAvg,WSMax,BPAvg,12V	None	TR,WDAvg, WSAvg,WSMax	None
Venda	None	WSAvg,12V,12Min	None	TR	TR

Table 12: Features not selected on daily time horizon.

Location	*Shrinkage*	*PQR*	*RF*	*RRF*	*QRF*
Bulawayo	None	None	None	None	Day, WSAvg
Cape Town	Year,RHMax, TMax,BPMax	DHITot,BPMax	None	None	Year,Day, WSAvg
Gaborone	WSAvg,12VMin	DNICalc,WSAvg	None	None	None
Pretoria	Year,WVM, WDAvg,Day,12V 12VMin	WDAvg,WVM	None	None	Day
Windhoek	DNIAvg,TMax WVM,24VMin	DNIAvg, TAvg, WDAvg, 24VMin	None	None	Year,24V
Venda	None	WDStD,WSMax, 12VMax, 24VMax, DNICalc	Day	Day	Day,12V

Table 13: Features with coefficients shrunk to zero.

Time	*Bulawayo*	*Cape Town*	*Durban*	*Gaborone*	*Pretoria*	*Windhoek*	*Venda*
Hourly	-	Day,CAA, WDAvg	TR, Day WDAvg	WDAvg	Day	WSAvg, WDAvg	WDAvg
Daily	None	Year, Day, TMax, RHMax, BPMax, DHISTot	-	WDAvg, 12VMin	WDAvg, WVM	TMax, WVM, BPMin, 24VMin	None

Table 14: Rejected or less important features.

Time	*Bulawayo*	*Cape Town*	*Durban*	*Gaborone*	*Pretoria*	*Windhoek*	*Venda*
Hourly	-	None	TR, WDAvg WDStD	Day, CTA,CAA	TR	WSAvg, WDAvg	TR
Daily	WSAvg, Day	None	-	None	Day	Day, Year, 12V, WDStD, WVM, 24VMin	Day, 12V

On 24-hour SI, features that were not selected on at least two locations were day, maximum barometric pressure, wind speed, wind direction, wind velocity magnitude, 12V battery, 24V battery minimum and averaged DNI (see Table 12). Table 13 shows features with coefficients that were shrunk to zero when applying a better method between shrinkage and PQR. The coefficient of wind direction was shrunk to zero from all locations except Pretoria on hourly SI. The day coefficient was also shrunk to zero from Pretoria and the other two locations. On a 24-hour SI, day, wind speed, 12V battery average, wind vector magnitude wind direction can be removed. Rejected variables or those with less than 1.5% importance scores were extracted from the best selection method among RF, RRF and QRF. Table 13 shows that wind direction, wind standard deviation and total rainfall are not important for hourly SI. The day is also not important on 24-hour SI.

Year, month, temperature, DHI, wind speed, 12V battery and 24V battery were found to have the most significant relevance on hourly SI, whilst month and minimum temperature were the most relevant variables on 24-hour SI. Hour, DHI, DNI, temperature, relative humidity and barometric pressure were the most important features of hourly SI. Month and DHI were the most important features of 24-hour SI.

4.5 Sensitivity analysis

Results from Section 4.4 show that hour, DNI, DHI, temperature, relative humidity, barometric pressure and wind speed should be always considered covariates when modelling SI. These results agree with previous studies that included variable selection when modelling SI. Since RFs are non-parametric models, the stability of the models was checked through variations in sample sizes. Results given in Figure 6 show that the QRF algorithm was not sensitive to any sample size changes. It was demonstrated to be the most stable random variable selection method. The regularisation, PQR, RF and

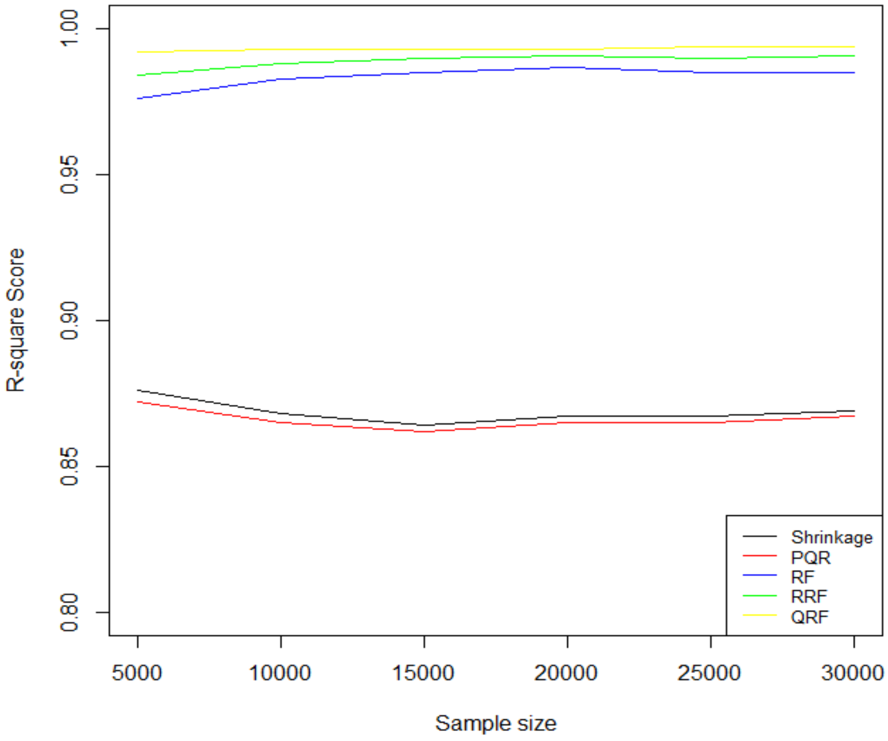

Figure 6: Sensitivity analysis results.

Table 15: Variables not selected from different sample sizes.

N	*Shrinkage*	*PQR*	*RF*	*RRF*	*QRF*
5000	WD	WD	None	None	None
10000	WD	WD	None	None	None
15000	WD	WD	None	None	None
20000	WD	WD	None	None	None
25000	WD	WD	None	None	None
30000	WD	WD	None	None	None

RRF models were sensitive to smaller sample size changes. However, they were not sensitive to large sample size changes. That is, the algorithms became more stable as the sample size increased. All of the algorithms selected the same variables from all of the different sample sizes considered, as shown in Table 15. However, it has to be noted that all shrinkage methods shrunk the coefficient of wind direction to zero, while all tree-based algorithms selected all of the variables (considered for sensitivity analysis) as important features when modelling SI.

5. Discussion

Although lasso is the most common variable selection method among previous SI studies in Southern Africa, the results from this study show that, among the regularisation algorithms considered, ridge was the best in most locations. The focus of this study was on comparing the variable selection capabilities of different algorithms. Literature confirms that ridge is good for only selecting relevant variables in the presence of multicollinearity but lasso is good when the selection method is further applied as a forecasting model. However, here lasso was the best in only one location, Windhoek. Ridge prevents overfitting when covariates are correlated and that is why it became the best in Venda and Gaborone where there was the highest multicollinearity. Ridge regression is a very good algorithm when focusing on dimension reduction only, as in this study. The combination of ridge and lasso, the elastic net (Sigauke et al., 2023), which is expected to overcome their limitations, was the best in Durban, where there were outliers, and in Pretoria. Pretoria data did not have outliers or multicollinearity. Though outliers do not cause serious problems with lasso and ridge, the two algorithms do not perform well in the presence of many outliers. The performance of the elastic net in the presence of several outliers is suspected to be attributable to its ability to remove degeneracies and wild behaviours in the data. Since the elastic net provides a compromise between lasso and ridge regression it can be expected that the algorithm will perform better than both lasso and ridge in a data set like Pretoria, with no outliers and low multicollinearity.

The new QRF algorithm outperformed all variable selection methods considered in this study on all metric evaluations when training all different SI data in Table 2. This excellent superiority can be attributed to the ability of QRF to learn any pattern of any response for any provided data (Dega et al, 2023). SI data is non-Gaussian and this study shows that the algorithm can be applied to both deterministic correction and probabilistic calibration of a skewed distribution of meteorological features, as claimed by Evin et al (2021). Apart from exhibiting strengths of both RFs and QR modelling, the hybrid algorithm demonstrated that it handled well the weaknesses in both QR and RF modelling separately. Though it is difficult to discern truly important variables from those that gain importance when applying a tree-based algorithm on high dimensional data that has random correlations, the hybrid algorithm handled excellently very high multicollinearity in Gaborone, Venda and Windhoek. QRF was robust to outliers in Cape Town and Durban SI data, although QR depends on the completeness of the meteorological data (Ayodele et al., 2016), and its prediction errors of the next value in the series are often large on short-term forecasting. As a result, the noise in the data did not affect the algorithm as much as it affected other algorithms. Those are the powerful properties the hybrid algorithm inherited from QR, being robust to outliers and tolerant to noise in the data (Diez-Olivan et al., 2018; Vantas et al., 2020). In addition, QRF infers conditional quantiles and gives a non-parametric and accurate way of estimating conditional quantiles in high-dimensional cases (Gostkowski and Gajowniczek, 2020).

Although shrinkage algorithms were outright inferior to the proposed QRF, they had better RMSE values than the rest of the other algorithms. This showed that the inbuilt k-fold cross-validation and regularisation in shrinkage algorithms tackled the problem of overfitting quite well. RMSE measures deviations of the fitted from the actuals as a goodness-of-fit metric and shrinkages were developed to achieve both feature selection and regularisation in the presence of multicollinearity. Though variance reduction done by shrinkages introduces little bias, the continued shrinking of the coefficients improves the minimisation of the prediction error, thus enhancing prediction accuracy by tackling overfitting (Fonti and Belitser, 2017). However, the shrinkage algorithms work well in small sample sizes and high-dimensional data, which is why lasso in particular performs well in finite sample cases where p may grow faster than n (Brink-Jensen and Ekstrom, 2021). The smallest sample size in this study was 5000 data points and the smallest variable dimension was 10. In addition, the process of shrinking coefficients focuses on feature relevance while ignoring importance. The relevance of a feature is determined by measuring associations through correlation. Thus, indirectly applying Gaussian and parametric modelling assumptions, the data exploration results here showed that SI data is non-Gaussian, as well as the relationship structures between SI and its covariates not yet being known. When dealing with SI data it is always best to work with non-linearity assumptions. Non-parametric assumptions are even better. Consequently, these results show the superiority of tree-based algorithms against regularisation algorithms when using the R^2 and MASE metrics. Tree-based algorithms are non-parametric models that have a very good ability to learn highly irregular patterns in non-linear data (Ibrahim and Khatib, 2017). That is why both the RF and RRF algorithms had better R^2 and adjusted- R^2 values than shrinkages. By pulling together a forest of trees, the performance of single trees is greatly boosted in RFs. Better MASE values on tree-based algorithms than shrinkages to the OOB estimation they do can be accounted for through randomly selecting features from bootstrap samples. The RF and RRF algorithms are noted as having similar feature selection processes, but RRF selects features by avoiding features not belonging to a feature set used in previous splits in the tree model. This avoidance did not lead to any superiority of RRF against RF in our SI study, because all metric values of the two algorithms were approximately equal. However, the present results agree with with those of Deng and Runger (2012), that the RRF was developed to improve the performance of RFs amid data sets with significant multicollinearity. Adjusted R-squared scores and MASE values from the RRF algorithm were slightly higher than those from the RF algorithm in Gaborone, Venda and Windhoek. Literature highlights the possibilities of overfitting in tree-based algorithms and the packages used in this study to fit them did not have inbuilt k-fold cross-validation. So, we attribute the reduction of overfitting possibility in tree-based algorithms to the partitioning strategy employed on all data sets. The very high adjusted R-squared scores and very low MASE values from all algorithms compared in this study indicate that our results agree with the finding of Ludwig et al. (2015) that both shrinkages and tree-based algorithms can select the right variables. Literature also specifies that shrinkage and tree-based algorithms are stable variable selection methods, and the results in this study have demonstrated that when the algorithms selected the same variables on different situational data sets they have approximately equal R-squared scores on different sample sizes.

6. Conclusion

This study introduced PQR to feature subset selection in SI studies and compared it with other shrinkage methods. Though lasso is the most popular model among regularisation algorithms it was not the best in some locations. Therefore, a comparison of the regularised methods should be made before the application of a selected variable selection method. Tree-based variable selection methods were also included in the study and were compared with shrinkage methods. It emerged that there are data situations when one or other of the techniques is best. The study focused on multicollinearity, outlier existence, skewness and heavily tail-distribution data situations. The hybrid model between QR and RFs, that is, the QRF model performed the best in most of these situations. The conclusion was drawn that the QRF model is the best method for SI data sets on which there is often existence of outliers. However, the RF in the hybrid model is not sensitive to associations through correlations, while QR offered a way of exploring sources of heterogeneity in covariates. As a result, it would not be advised to conclude the feature selection exercise using results from the QRF model only. As features are classified as important or not, it is also paramount to measure their relevance. Therefore, it would be prudent to run the feature selection process in two stages, starting with the relevance of features to associations through correlations and then classifying their importance. We conclude that a variable coefficient is shrunk to zero if and only if it is not important, but a relevant one may not be important. That is, relevant variables can be classified as strongly relevant or weakly relevant through importance scores. Further studies can be done to find how regularisation can be done on QRF modelling; that is, developing a model that is stable and accurate in multicollinearity, outlier existence and heavily-tailed data situations. If there exist groups of highly correlated features, then group regularisation should be considered. Further studies can include interactions in the features as well. It can also be concluded that hourly or monthly time and temperature are paramount variables in SI modelling in Southern Africa. Time can be recorded in hourly or monthly units depending on the study. Day recorded as a variable is neither relevant nor important when modelling SI. Apart from location and time horizon, This study also leads to the conclusion that covariates paramount for predicting SI may vary, depending on the context of the study or application.

Acknowledgements

The authors acknowledge the Research and Innovation Office at the National University of Science and Technology for funding (Grant No: RDB/57/22) the purchase of data from Goetz station. They are also grateful for the services and help offered by the Meteorological Services Department staff.

Author contributions

Conceptualisation, A.M., D.M. and P.M.; methodology, A.M., D.M. and P.M.; software, A.M.; validation, A.M., D.M., P.M. and C.S.; formal analysis, A.M., D.M., P.M. and C.S.; investigation, A.M., D.M. and P.M.; resources, A.M.; data curation, A.M.; writing—original draft preparation, A.M.; writing—review and editing, A.M., D.M., P.M, and C.S.; visualisation, A.M., D.M., P.M. and C.S; supervision, D.M., P.M. and C.S.; project administration, A.M. All authors have read and agreed to the published version of the manuscript.

Data availability statement

Most of the data used in this study are from the SAURAN website (https://sauran.ac.za, accessed on 12 June 2022).

References

Alhamzawi, R. and Ali, H.T.M. 2018. The Bayesian adaptive lasso regression. *Mathematical Biosciences* 303: 75-82. .

Asnaghi, V., Pecorino, D., Ottaviani, E., Pedroncini, A., Bertolotto, R.M. and Chiantore, M. 2017. A novel application of an adaptable modelling approach to the management of toxic microalgal bloom events in coastal areas. *Harmful Algae* 63: 184-192.

Ayodele, T. R., Ogunjuyigbe, A. S. O., and Monyei, C. G. 2016. On the global solar radiation prediction methods. *Journal of Renewable and Sustainable Energy* 8: 023702-1; . http://dx.doi.org/10.1063/1.4944968.

Babar, B., Luppino, L.T., Bostrom, T. and Anfinsen, S.N. 2020. Random forest regression for improved mapping of solar irradiance at high latitudes. *Solar Energy* 198:81-92.

Belloni, A. and Chernozhukov, V. 2011. l_1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics* 39(1): 82-130; DOI: 10.1214/10-AOS827.

Brink-Jensen, K., and Ekstrom, C. T. 2021. Inference for feature selection using the Lasso with high-dimensional data. arXiv:1403.4296v1 [stat.ME]; https://doi.org/10.48550/arXiv.1403.4296.

Celeux, G., Martin-Magniette, M-L., Maugis-Rabusseau, C. and Raftery, A. E. 2015. Comparing model selection and regularization approaches to variable selection in model-based clustering. *Journal de la Société française de Statistique* 155(2): 57–71.

Chandiwana, E., Sigauke, C., and Bere, A, 2021. Twenty-four-hour ahead probabilistic global horizontal irradiation forecasting using Gaussian process regression. *Algorithms* 14: 177. .

Deng, H., and Runger, G. 2012. Feature selection via regularized trees, *WCCI 2012. Proceedings of the IEEE World Congress on Computational Intelligence, Brisbane, Australia, 10-15 June 2012*. .

Deng, H., Guan, X., Liaw, A., Breiman, L., and Cutler, A. 2022. Package 'RRF', *CRAN*. .

Diez-Olivan, A., Averos, X., Sanz, R., Sierra, B., and Estvez, I. 2018. Quantile regression forests-based modelling and environmental indicators for decision support in broiler farming. *Computers and Electronics in Agriculture* 161: 141-150; https://doi.org/10.1016/j.compag.2018.03.025.

El Motaki, S., and El Fengour, A. 2021. A statistical comparison of feature selection techniques for solar energy forecasting based on geographical data. CAMES 28(2): 105–118; DOI: 10.24423/cames.324.

Evin, G., Lafaysse, M., Taillardat, M., and Zamo, M. 2021. Calibrated ensemble forecasts of the height of new snow using quantile regression forests and ensemble model output statistics. *Nonlinear. Processes in Geophysics* 28: 467–480; https://doi.org/10.5194/npg-28-467-2021.

Fonti, V., and Belitser, E. 2017. Feature selection using lasso. *VU Amsterdam research paper in business analytics*, *30*, 1-25.

Friedman, J., Hastie, T. and Tibshirani, R. 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* 33(1): 1-22.

Freeman, E.A., Frescino, T.S. and Moisen, G.G. 2023. Pick your flavour of random forest. *CRAN.*

Gostkowski, M., and Gajowniczek, K. 2020. Weighted Quantile Regression Forests for Bimodal Distribution Modeling: A Loss Given Default Case. *Entropy* 22: 545; ; DOI:10.3390/e22050545.

Gu, Y., Fan, J., Kong, L., Ma, S. and Zou, H. 2017. ADMM for High-Dimensional Sparse Penalized Quantile Regression. *Technometrics*; DOI: 10.1080/00401706.2017.1345703 .

Hastie, T., Qian, J. and Tay, K. 2023. An Introduction to glmnet. CRAN.

Hossain, M. R., Than Oo, A. M. and Shawkat Ali, A. B. M. 2013. The Effectiveness of feature selection method in solar power prediction. *Journal of Renewable Energy* (2013): http://dx.doi.org/10.1155/2013/952613.

Ibrahim, I.A. and Khatib, T. 2017. A novel hybrid model for hourly global solar radiation prediction using random forests technique and firefly algorithm. *Energy Conversion and Management* 138 (2017): 413-425.

Khalid, S., Khalil, T. and Nasreen, S. 2014. A survey of feature selection and feature extraction techniques in machine learning. *Science and Information Conference Proceeding. London, UK, August 27-29, 2014*. .

Kipruto, E. and Sauerbrei, W. 2022. Comparison of variable selection procedures and investigation of the role of shrinkage in linear regression protocol of a simulation study in low-dimensional data. *PLOS ONE* 17(10): e0271240; https://doi.org/10.1371/journal.pone.0271240.

Koenker, R. 2018. Quantile regression in R: A vignette. *CRAN*.

Kursa, M. B., and Rudnicki, W. R. 2022. Package 'Boruta'. *CRAN*.

Lee, j., Wang, W., Harrou, F. and Sun, Y. 2020. Reliable solar irradiance prediction using ensemble learning-based models: A comparative study. *Energy conversion and management* 208(2020): 112-582. .

Leng, C., Lin Y. and Wahba, G. 2006. A note on the lasso and related procedures in model selection. *Statistica Sinica* 16(4): 1273-1284.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J. and Liu, H. 2017. Feature selection: A data perspective. *ACM Computing Surveys* 50(6) Article 94: 45 pages; ttps://doi.org/10.1145/3136625.

Ludwig, N., Feuerriegel, S. and Neumann, D. 2015. Putting Big Data analytics to work: Feature selection for forecasting electricity prices using the LASSO and random forests. *Journal of Decision Systems* 24(1):19-36; http://dx.doi.org/10.1080/12460125.2015.994290.

Maxwell, K., Rajabi, M. and Esterle, J. 2021. Spatial interpolation of coal properties using geographic quantile regression forest. *International Journal of Coal Geology* 248: 103869 .

Mehmood, T., Sæbø, S. and Liland, K. H. 2020. Comparison of variable selection methods in partial least squares regression. *Journal of Chemometrics* 34:e3226: Doi.org/10.1002/cem.3226.

Meinshausen, N. 2022. Package 'quantregForest'. *CRAN*.

Mpfumali, P., Sigauke, C., Bere, A. and Mlaudzi, S. 2019. Day Ahead Hourly Global Horizontal Irradiance Forecasting-Application to South African Data. *Energies* 12: 1-28. .

Muller, I. M. 2021. Feature selection for energy system modelling: Identification of relevant time series information. *Energy and AI* 4(2021): 100057; https://doi.org/10.1016/j.egyai.2021.100057.

Munshi, A. and Moharil, R.M. 2022. Solar radiation forecasting using random forest. *AIP Conference Proceedings 2424*, 050003 (2022); DOI.org/10.1063/5.0076827 .

Mutavhatsindi, T., Sigauke, C. and Mbuvha, R. 2020. Forecasting Hourly Global Horizontal Solar Irradiance in South Africa, *IEEE Access* 8: 198873.

Omoruyi, F. A., Obubu, M., Omeje, I. L., Echebiri, U., Onyekwere, K. C., Lilian, N. O. and Hamzat K. I. 2019. Comparison of some variable selection techniques in regression analysis. *American Journal of Biomedical Science and Research* 6(4): 281-293; DOI: 10.34297/AJBSR.2019.06.001044.

Park, T. and Casella, G. 2008. The Bayesian Lasso, *Journal of the American Statistical Association* 103(482): 681-686.

Randa, T.M., Tinungki, G.M. and Sunusi, N. 2022. Application of lasso and lasso quantile regression in the identification of factors affecting poverty levels in Central Java. *International Journal of Academic and Applied Research* 6(4):350-353.

Ratshilengo, M., Sigauke, C. and Bere, A. 2021. Short-Term Solar Power Forecasting Using Genetic Algorithms: An Application Using South African Data. *Applied Sciences* 11: 4214 .

Sanchez-Pinto, L. N., Venable, L. R., Fahrenbach, J. and Churpek, M. M. 2018. Comparison of variable selection methods for clinical predictive modelling. *International Journal of Medical Information* 116: 10–17; doi:10.1016/j.ijmedinf.2018.05.006.

Su, M. and Wang, W. 2021. Elastic net penalized quantile regression model. *Journal of Computational and Applied Mathematics* 392 (2021): 113462. .

Vantas, K., Sidiropoulos, E. and Loukas, A. 2020. Estimating Current and Future Rainfall Erosivity in Greece Using Regional Climate Models and Spatial Quantile Regression Forests. *Water* 12 (2020): 687; DOI:10.3390/w12030687 .

Vaysse, K. and Lagacherie, P. 2017. Using quantile regression forest to estimate the uncertainty of digital soil mapping products. *Geodema* 291 (2017):55-64. .

Villegas-Mier, C. G., Rodriguez-Resendiz, J., Alvarez-Alvarado, J.M., Jimenez-Hernandez, H. and Odry, A. 2022. Optimized Random Forest for Solar Radiation Prediction Using Sunshine Hours. *Micromachines* 13: 1406 .

Wang, L., Wang, Y. and Chang, Q. 2016. Feature selection methods for big data bioinformatics: A survey from the search perspective. *Methods* (2016); doi: http://dx.doi.org/10.1016/j.ymeth.2016.08.014.

Williams, B., Hansen, G., Baraban, A. and Santoni, A. 2015. A practical approach to variable selection comparison of various techniques. *Casualty Actuarial Society E-Forum*, Summer 2015.

Yilmaz, U. and Kuvat, O. 2023. Investigating the effect of feature selection methods on the success of overall equipment effectiveness prediction. *Uludağ University Journal of The Faculty of Engineering* 28(2): 437-452; DOI: 10.17482/uumfd.1296479.

Zeng, Z., Wang, Z., Gui, K., Yan, X., Gao, M., Luo, M., Geng, H., Liao, T., Li, X., An, J., Liu, H., He, C., Ning, G. and Yang, Y. 2020. Daily global solar radiation in China estimated from high density meteorological observations: A random forest model framework. *Earth and Space Science* 7: e2019EA001058; DOI. org/10.1029/2019EA001058.

Zhang, L. and Wen, J. 2019. A systematic feature selection procedure for short-term data-driven building energy forecasting model development. *Energy & Buildings* 183: 428–442; https://doi.org/10.1016/j.enbuild.2018.11.010