scientific reports

Check for updates

OPEN

Hybrid models of sparse and robust regression to solve heterogeneity problem in black pepper big data

Paavithashnee Ravi Kumar[1], Olayemi Joshua Ibidoja[1,2], Majid Khan Majahar Ali[1]✉ & Wan Rosli Wan Ishak[3]✉

Data analytics is increasingly important in agriculture, particularly in smart farming, enhancing decision-making and sustainability. Research on factors affecting moisture content removal in black pepper drying using solar dryers is crucial for cost reduction and improving product quality and quantity. This drying process involves numerous parameters, resulting in big data complexity. Heterogeneity among these parameters can introduce bias, leading to incorrect inferences, while multicollinearity and outliers impact model validation and interpretation. This study proposes hybrid models of sparse and robust regression to solve the heterogeneity problem using black pepper big data. Sparse regression techniques such as elastic net, ridge and LASSO are used to identify the 25, 35, 45, 55 and 100 highest-ranking variables for black pepper moisture content removal. These models are hybridized with robust regression estimators (M Bi Square, M Hampel, M Huber, S and MM) for handling outliers. Results indicate that before heterogeneity, the hybrid Ridge model with M Bi-Square performs best under both 2-sigma and 3-sigma limits. After heterogeneity removal, LASSO model with S estimator proves to be the most effective across both limits.

Keywords Heterogeneity, Outliers, Robust estimation, Robust prediction, Variable selection

Data analytics in statistics is an effective method for analysing complex data to gain meaningful insights, supported by visualisation tools to summarise data, detect trends, and aid decision-making across various fields[1]. In agriculture, it plays a vital role in smart farming by enhancing efficiency, sustainability, and crop yields. Managing complex and varied agricultural data requires efficient systems for fast and reliable processing[2]. By integrating advanced systems and innovations such as machine learning, the Internet of Things (IoT), Artificial Intelligence (AI), sensors and automation, smart farming empowers farmers with real-time insights and optimised decision-making[3]. This innovative approach addresses global issues like food security, climate change, and increasing costs while improving resource efficiency and promoting sustainable agricultural practices[4]. Notably, these technologies significantly enhance various stages of herb and spice cultivation and processing, preserving and improving their quality. A notable example is black pepper.

Black pepper, renowned as the "King of Spices" enjoys widespread recognition and is extensively consumed as a spice worldwide. Peppercorns that are almost ripe or have a greenish appearance are sun-dried until they turn brownish black to produce black pepper. The main processing steps involved in black pepper production include harvesting, blanching, drying, cleaning, grading, packaging and storage[5]. Piperine, a significant bioactive compound in black pepper, exhibits diverse physiological and drug-like effects. Black pepper, the most traded spice globally, particularly in Asia, witnessed Vietnam, Brazil, Indonesia, India, and Malaysia as the leading producers in 2022[6].

The primary method for preserving black pepper is through the drying process, crucial for moisture removal and prevention of microbial decay. Water activity is a crucial determinant of the stability, safety, and quality of agricultural products. It represents the amount of free water available for microbial and biochemical reactions, which directly affects product shelf life and nutritional properties[7]. Controlling water activity is therefore essential in the drying of black pepper, as inadequate or uneven moisture removal can increase the available water for microbial growth and oxidative reactions[8]. Such reactions accelerate the degradation of essential oils and bioactive compounds like piperine, which are responsible for black pepper's distinctive aroma, flavour and

[1]School of Mathematical Sciences, Universiti Sains Malaysia USM, 11800 Penang, Malaysia. [2]Department of Mathematics, Federal University Gusau, Gusau, Nigeria. [3]School of Health Sciences, Universiti Sains Malaysia, Health Campus, 16150 Kubang Kerian, Kelantan, Malaysia. ✉email: majidkhanmajaharali@usm.my; wrosli@usm.my

therapeutic properties[9]. Therefore, maintaining optimal water activity through effective drying techniques is essential to ensure microbial safety, chemical stability and overall product quality.

Drying also minimizes pest infestations and reduces the pepper volume and weight of the pepper for better storage and transportation suitability[10]. Harvested peppercorns typically contain around 65 to 70% d.b. moisture and require drying to achieve a moisture content of 11% to 12% for storage. Sun- drying black pepper, which is a conventional method, presents challenges like a significantly prolonged drying period and contamination from dust, dirt, insects and other pollutants, thus negatively affecting the quality of products. Meanwhile, mechanical drying systems like tunnel dryers and fluidized bed dryers also exhibit drawbacks, such as their inefficiency, high usage of fossil fuels, and a requirement for substantial labour[11].

Solar dryers present a promising alternative to conventional methods in drying technology, as they can effectively control moisture loss during drying to preserve product quality and nutritional value. They offer numerous advantages, including requiring less space, producing clean and high-quality commodities, avoiding insect and animal threats and providing a controlled drying process[12–14]. However, assessing solar dryer efficiency requires consideration of various factors impacting the dried product quality. Several research studies have explored solar dryers for black pepper drying, with[15] achieving a final moisture content of 9.4% over a—12 h drying period using an indirect-type solar-biomass hybrid dryer. These findings demonstrated the enhanced product quality achievable with solar dryers compared to conventional methods.

However, drying black pepper using a solar dryer with smart IoT monitoring systems involves numerous parameters, resulting in complex big data processing in the cloud database. These complexities come with many challenges, particularly in agricultural data analysis, where issues such as heterogeneity, multicollinearity and outliers are common[16,17]. Addressing these issues through advanced data analytics is essential for improving agricultural processes and ensuring the sustainable cultivation of crops like black pepper. In Malaysia, these challenges have resulted in losses of over 25% of the nation's pepper crop[18]. Therefore, this highlights the need for developing a hybrid solar dryer with optimized parameters to preserve the crop's nutritional value.

From an agricultural perspective, identifying the key parameters that influence moisture content removal provides valuable insights for improving drying performance and preserving the nutritional quality in black pepper. These parameters determine the drying rate and uniformity, which in turn affect the biochemical stability of bioactive compounds such as piperine and volatile oils. By identifying the significant factors that most strongly influence moisture diffusion and evaporation behavior, variable selection helps optimize drying conditions that maintain desirable product quality. However, this process is challenging since black pepper drying involves many interdependent parameters, such as temperature, humidity and air velocity, which can vary widely and introduce heterogeneity into the data.

One major issue in agricultural big data is heterogeneity, which denotes the degree of variability within the parameters due to factors like differences in parameters as well as varying units for factors such as solar radiation, relative humidity and temperature[17–19]. This variability introduces noise into the data, making it difficult to obtain reliable measurements from different sensors. When this inconsistent data is used in predictive models, especially those depending on factors like temperature and humidity, it can significantly lower their accuracy. Addressing this heterogeneity is crucial for optimizing drying efficiency, improving product quality, and ensuring consistent outcomes. Otherwise, it can result in incorrect predictions, poor decisions, financial losses, and lower product quality[20]. Heterogeneity could restrict result applicability due to lack of agreement at the study level[21]. Studies like[22] highlighted that the assumption of homogeneity poses a significant challenge, as it contributes to heterogeneity issues, which can lead to biased and inconsistent standard errors. Separately,[16] demonstrated that examining heterogeneity in seaweed big data enhanced the understanding of drying parameters dynamics and enabled more effective predictive modelling.

Another challenge is multicollinearity, which arises when two or more independent variables in a multiple linear regression model exhibit a strong linear association, potentially impacting the model's stability[23]. This issue may lead to inaccuracies in parameter evaluation within regression models by inflating the standard errors, causing some previously significant variables to be statistically insignificant[24]. As a result, the estimated parameters in regression models become inconsistent and lack reliability, leading to a decrease in their precision[25]. Consequently, the model becomes less effective at making accurate forecasts, and there is a higher risk of overfitting the data. To address multicollinearity, variable selection can be combined with machine learning techniques to improve parameter estimates[26].

Outliers tend to occur in agricultural data due to uncontrollable factors and natural variation[27]. These outliers, which deviate from the typical pattern or structure of the distribution, can arise from factors like human error, instrument errors, setup errors, mechanical problems and environmental changes[28]. The existence of outliers in data causes incorrect estimates of parameters, thus decreasing model precision and leading to unreliable results. Furthermore, they significantly affect sample mean and standard deviation in statistical analysis, leading to either overestimation or underestimation of values. This serves as a simple illustration of how undesired outliers might impact data analysis outcomes[29]. Given these challenges, particularly the simultaneous presence of multicollinearity and outliers, recent research has explored advanced techniques such as quantile-based robust ridge regression estimators, modified robust ridge M- estimators and penalized M- estimators which are designed to provide more reliable parameter estimates and enhance model precision[30–33]. On top of that, due to numerous factors affecting the moisture content removal of black pepper resulting in big data complexity, variable selection via machine learning algorithms is performed to identify significant factors. Insignificant variables are then removed to reduce overfitting and improve prediction accuracy.

Existing literature on black pepper, especially using solar dryers, is limited to a few studies by[15,34,35]. While some research has addressed multicollinearity for black pepper (for example, by[35]), heterogeneity and outliers in the context of moisture content removal for black pepper remain unexplored. Understanding of heterogeneity in the application of big data for black pepper in agriculture remains limited despite its obvious presence in actual

agricultural data. For instance,[16] investigated heterogeneity in seaweed drying, employing hybrid models using seven machine learning algorithms (Elastic Net, Ridge, Lasso, Bagging, Support Vector Machine, Random Forest, and Boosting) with robust regression methods (M Bi-Square, M Hampel, M Huber, MM and S) to identify the significant drying parameters influencing moisture content removal for 45 highest important variables before and after addressing heterogeneity. Similarly, a study by[36] focused on multicollinearity and heterogeneity in seaweed drying, applying Ridge, LASSO and Elastic Net machine learning algorithms, but uses only box plots to detect outliers. These studies, however, are specific to seaweed data, highlighting a notable gap in agricultural research on black pepper.

Additionally, a literature gap exists regarding the consideration of interaction terms in agriculture, particularly for black pepper. Studying the effects of interaction variables is important, as the relationship between two or more variables can be studied thus providing a meaningful result and preventing bias[27]. Furthermore, no prior work has explored the use of hybrid sparse and robust regression models to identify the most influential parameters affecting moisture content removal in black pepper. Understanding these parameters is an essential step toward optimizing drying efficiency and ensuring consistent product quality. This highlights the broader potential of statistical modelling approaches to enhance process optimization and data-driven agricultural applications.

To address gaps in the existing literature, this study aims to determine the significant drying parameters of black pepper that directly impact heterogeneity and assess their effects on moisture content removal, both before and after eliminating heterogeneity parameters. Additionally, hybrid models that combine sparse and robust regression techniques are proposed to address heterogeneity in black pepper big data and enhance the accuracy of predicting the moisture content removal during the black pepper drying process. The selection of significant parameters is conducted using the 25,35,45,55 and 100 highest ranking variables, which include interaction factors, by applying three sparse regression models, namely elastic net, ridge and LASSO. Furthermore, hybrid models are developed by integrating these sparse models with robust regression estimators (M Bi-Square, M Huber, M Hampel, S and MM estimators) to address multicollinearity and outliers effectively. The performance and accuracy of the models are first assessed using evaluation metrics, while the best model selection is determined using the eight selection criteria (8SC).

Methodology

Flowchart of study

The study is divided into 3 phases outlined in Fig. 1, mapping the study's objectives.

Phase I: This project begins by collecting data from the black pepper drying process, applying the Modified Hybrid Solar Dryer (MHSD) in Setiu, Terengganu, Malaysia. The computations cover all possible models, considering interactions up to the second order. R software is used to verify the assumptions related to linearity,

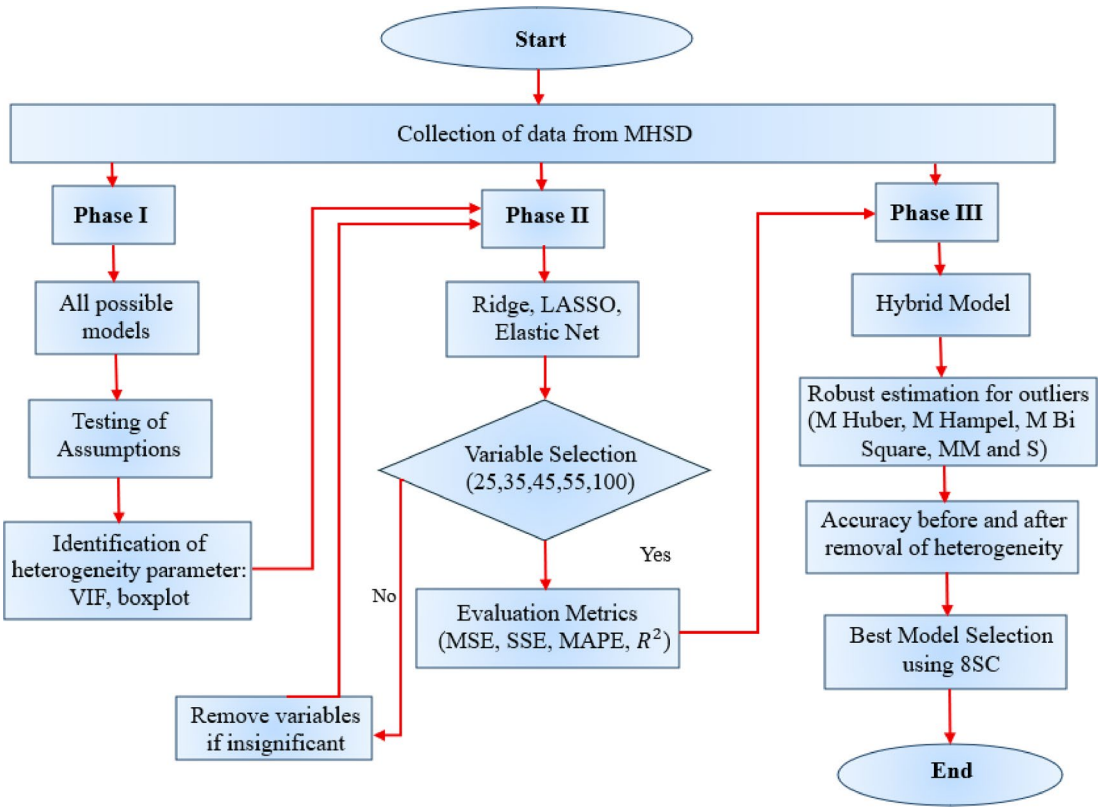

Fig. 1. Flowchart of the study.

errors, independent variables, and multicollinearity. Next, this study will use Variance Inflation Factor (VIF) and boxplot analysis to identify the significant parameters that show heterogeneity, as employed by[16] and[17]. Therefore, these techniques will be included in this study. The VIF is computed using the *vif*() function from the '*car*' package in R, involving the original dataset while considering only the main effects of the independent variables. R-squared values for main drying parameters can be determined through Eq. (1) with the VIF values obtained.

$$R^2 = 1 - \frac{1}{VIF} \tag{1}$$

Once the minimum and maximum R-squared values are identified, the average R-squared value for the main drying parameters is computed. This value then serves as a benchmark for detecting heterogeneity. When the R-squared value for the primary drying parameters falls below this benchmark, it suggests potential heterogeneity.

Phase II: Ridge, LASSO and elastic net are the three proposed sparse regression techniques that will be implemented in R software to select variables and identify key parameters affecting moisture ratio removal of black pepper. Consequently, these mentioned sparse regression techniques will independently select the highest-ranking 25, 35, 45, 55 and 100 variables. The machine learning algorithms proposed can provide information about the ranking of the important variables but do not specify the exact number of significant variables to include in a model[37]. The model's performance and accuracy will be assessed using the evaluation metrics Sum of Squared Error (SSE), Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), and R-squared.

Phase III: Hybrid models are then developed using the mentioned sparse regression techniques with robust regression estimators such as M Bi Square, M Huber, M Hampel, S and MM estimators. Outliers are then identified using robust estimation and the two—and three—sigma limits. The best models, before and after removing heterogeneity parameters, are selected using 8SC which include the Akaike Information Criterion (AIC), Final Prediction Error (FPE), Generalized Cross-Validation (GCV), Hannan-Quinn information criterion (HQ), Risk Inflation Criterion Estimate (RICE), SCHWARZ, SGMASQ, and SHIBATA.

Data description

The solar dryer used in this study is the Modified Hybrid Solar Dryer (MHSD), installed in Setiu, Terengganu, Malaysia, due to the predominant economic activities being the cultivation and drying of black pepper. Categorized as a forced convection indirect type, MHSD was adopted as the smart farming technology to dry the black pepper, as illustrated in Fig. 2.

This study examines the moisture content in black pepper will be observed as the dependent variable, while the independent variables include solar radiation, ambient temperature, ambient relative humidity, collector temperature, chamber temperature and chamber relative humidity. The data consists of 1924 observations with 29 independent variables and one dependent variable. Interaction variables up to second order will also be considered. For instance, T1T5 indicates the interaction between T1 and T5. As a result, the data includes 29

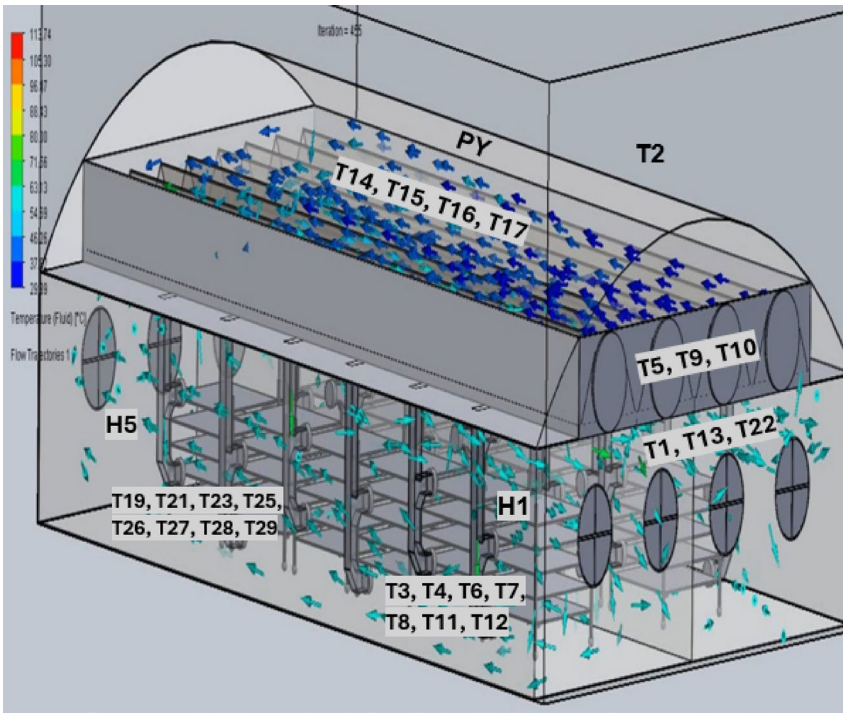

Fig. 2. MHSD simulation diagram.

main variables and 406 interaction variables, making an overall count of 435 independent variables that impact the moisture content removal of black pepper.

Multiple linear regression (MLR)

MLR model serves as a statistical method for analysing how a dependent variable y_i relates to multiple explanatory variables $x_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$ where p represents the explanatory variables. Consider an MLR with n observations:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2}$$

where $\boldsymbol{y}$ is a $n \times 1$ vector representing the dependent variables, while $\boldsymbol{X}$ is the $n \times p$ design matrix. The unknown parameters are represented by $\boldsymbol{\beta}$, a $p \times 1$ vector and $\boldsymbol{\varepsilon}$ is the $n \times 1$ error term, which is normally distributed with a mean of zero, consisting of uncorrelated errors and homoscedastic[38].

Ordinary Least Squares (OLS) is a regression analysis technique for estimating $\boldsymbol{\beta}$ by minimizing the sum of squared differences between the observed and the predicted values of the dependent variable $\boldsymbol{y}$[17]. According to[39,40], the OLS estimator of $\boldsymbol{\beta}$ is obtained by minimizing $\varepsilon\varepsilon'$ as follows:

$$\begin{aligned}
\varepsilon\varepsilon' &= (\boldsymbol{y} - \boldsymbol{X}\beta)\prime(\boldsymbol{y} - \boldsymbol{X}\beta) = \boldsymbol{y'y} - 2\beta'\boldsymbol{X}\prime\boldsymbol{y} + \beta'\boldsymbol{X}\prime\boldsymbol{X}\beta \\
\partial\left(\varepsilon\varepsilon'\right)/(\partial\beta) &= -2\boldsymbol{X'y} + 2\boldsymbol{X}\prime\boldsymbol{X}\beta = 0 \\
\boldsymbol{X}\prime\boldsymbol{X}\beta &= \boldsymbol{X'y} \\
\hat{\beta} &= (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}
\end{aligned} \tag{3}$$

From Eq. 2, if y_i represents the outcome, then $x_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^T$ denotes the predictor vector for the i^{th} case. In MLR model, parameter estimation is performed using the OLS method, where the coefficients $\beta = (\beta_0, \beta_1, \beta_2, \cdots, \beta_p)^T$ are determined to achieve the best fit by minimizing the sum of squared residuals (SSR)[41–43]. The SSR is expressed as:

$$SSR = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \tag{4}$$

Heterogeneity identification

As mentioned by[16,17,22], the following multiple linear regression model is considered:

$$Y_i = \beta_0 + \beta_1 T_{i,1} + \beta_2 T_{i,2} + \ldots + \alpha_j + \varepsilon_i \tag{5}$$

Here, Y_i for $i = 1,2, \ldots, n$ corresponds to the observed moisture content for the i^{th} case, estimates β′*s* denote the regression coefficients for the predictor variables (which refer to the drying parameters, *T*′*s*), α_j represents the parameters exhibiting heterogeneity for $j = 1,2, \ldots, f$ and ε is random error. Omitting an important variable from the regression equation can result in inaccurate and biased estimates of β. Additionally, there is a risk that variables may correlate with the error term, leading to a violation of regression assumptions. VIF is the most commonly used and simplest method to detect the presence of multicollinearity[26]. Stronger linear relationships between variables result in higher R^2 values and subsequently lead to increased VIF_i[44]. A higher VIF indicates more serious multicollinearity among variables, with values exceeding 10 indicating its presence. VIF is defined as:

$$VIF = \frac{1}{1 - R^2} \tag{6}$$

hence, $R^2 = 1 - \frac{1}{VIF}$.

Hybrid sparse and robust regression techniques

This study uses sparse regression models as the variable selection approach to select significant factors affecting the moisture ratio removal of black pepper. Robust regression models are employed to detect outliers effectively. Robust regression offers a more effective approach than traditional regression techniques, particularly for datasets containing outliers and heteroscedasticity, enabling more precise and efficient parameter estimation[20]. However, for sparse regression, standardizing both the independent and dependent variables before the estimation process is essential, ensuring they have a zero mean and unit variance. This way, the results do not depend on the measurement scale, ensuring that independent variables have equal consideration and are on a comparable scale[45]. The focus on standardization is on sparse regression because of its effectiveness in addressing multicollinearity due to the interaction of variables. Therefore, combining sparse and robust regression techniques can further improve forecasting and enhance black pepper moisture removal prediction by addressing multicollinearity and outliers.

Sparse regression models

Ridge regression

Ridge, referred to as L_2 regularization, is commonly used in statistics and machine learning, as it works by regularizing the estimated coefficients and is particularly effective in handling multicollinearity issues in the data[46]. By shrinking the coefficients towards zero, Ridge regression helps reduce overfitting, though the coefficients will never reach exactly zero. This improves prediction accuracy, but at the cost of a slight increase in bias[47].

According to[20,22], the ridge regression coefficients estimate $\hat{\beta}^{RR}$ minimize:

$$L^{RR}(\beta) = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = SSR + \lambda\sum_{j=1}^{p}\beta_j^2 \tag{7}$$

where $\lambda \geq 0$ serves as the regularization parameter that controls the shrinkage effect and L_2 penalty is the second term. Ridge regression determines the coefficient estimates that minimize the SSR while ensuring a good fit to the data. The term $\lambda\sum_{j=1}^{p}\beta_j^2$ serves as the shrinkage penalty. One advantage of ridge regression is its ability to minimize bias in large datasets. By constraining the coefficient estimates, ridge regression reduces the estimator's variance, introducing some bias as a trade-off. However, ridge regression applies continuous shrinkage to coefficients without setting any to zero, leading to a less interpretable model while retaining all predictors.

LASSO regression

LASSO regression or Least Absolute Shrinkage and Selection Operator regression, also known as L_1 regularization, is a regression analysis method that combines parameter shrinkage with variable selection[48,49]. Unlike Ridge regression, LASSO regression has the ability to shrink some coefficients to zero when the regularization parameter, λ is large[50]. In other words, it shrinks certain regression coefficients while completely eliminating others if they are insignificant, effectively performing variable selection[51]. As a result, LASSO produces a simpler, more efficient model by removing irrelevant data and reducing the number of parameters. This makes it particularly useful for handling multicollinearity and preventing overfitting[47].

The LASSO coefficient regression estimate $\hat{\beta}^{LASSO}$, minimize[52]:

$$L^{LR}(\beta) = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = SSR + \lambda\sum_{j=1}^{p}|\beta_j| \tag{8}$$

Here, λ represents a positive regularization parameter and the second term corresponds to the L_1 penalty[53]. LASSO regression stands out as a versatile and effective method for handling complex datasets and improving predictive modelling outcomes[51]. Nevertheless, the L_1 penalty term in LASSO applies the same penalty to all coefficients, which can introduce bias, especially for large coefficients. This approach can sometimes exclude important variables if they have relatively smaller coefficients.

In this study, the optimal tuning parameters (λ_1 for Lasso and λ_2 for Ridge regression) were selected using five-fold cross-validation implemented through the cv.glmnet() function in R. According to[54], the λ_2 penalty is first evaluated over a predefined grid of values, and for each λ_2 value, the Elastic Net solution path is obtained. The selected λ_2 is the value that produces the lowest cross-validation error. The second tuning parameter, λ_1, is then selected through five-fold cross-validation. Cross-validation is one of the most effective and widely used approaches for selecting tuning parameters, as it directly estimates prediction error. Five-fold cross-validation is adopeted due to computational efficiency, while maintaining relatively low bias and varianc. This process ensures a balanced trade-off between bias and variance, thereby improving predictive accuracy and model generalization[55].

Elastic net

Elastic Net regression was introduced to address the instability of LASSO when predictors are highly correlated, making it a robust solution for analysing high-dimensional data[56]. Elastic Net regression integrates the properties of Ridge (L_2) and Lasso (L_1) norms as a regularization technique. By integrating both penalties, it balances feature selection with coefficient stability, making it effective for handling datasets with highly correlated features. Moreover, it effectively addresses multicollinearity among predictor variables[57].

According to[54], the coefficient of the Elastic Net regression estimate $\hat{\beta}^{ENR}$ minimize:

$$\begin{aligned} L^{ENR}(\beta) &= \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda_1\sum_{j=1}^{p}|\beta_j| + \lambda_2\sum_{j=1}^{p}\beta_j^2 \\ &= SSR + \lambda_1\sum_{j=1}^{p}|\beta_j| + \lambda_2\sum_{j=1}^{p}\beta_j^2 \end{aligned} \tag{9}$$

where the parameter λ_1 represents L_1 regularization parameter, which determines the strength of the Lasso penalty, while λ_2 corresponds to the L_2 regularization parameter, managing the effect of the Ridge penalty.

The Elastic Net offers the advantage of enforcing sparsity while allowing flexibility in selecting variables[58]. Additionally, it also encourages grouping among highly correlated variables. However, one concern with this approach is the risk of double shrinkage in naive elastic nets, requiring careful consideration when applying it[17].

Robust regression estimations

M estimation

According to[24,33,59], M-estimation builds on the maximum likelihood estimation method while also providing a more robust approach. The M-estimator minimizes the function $\rho(\cdot)$, which operates on the residuals. It is given by:

$$\hat{\beta}_{ME} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \rho\left(e_i\left(\beta\right)\right) \tag{10}$$

The function ρ represents a $\rho-$ type M-estimator. Assuming σ is known, the residuals for β are estimated as $e_i = y_i - \beta^T x_i$. In M-estimation, β will minimize the objective function:

$$\sum_{i=1}^{n} \rho\left\{\frac{e_i\left(\beta\right)}{\sigma}\right\} \tag{11}$$

The σ is estimated in a robust way, and the scale of $\tilde{\sigma}_{ME}$ in M-estimator has a defined solution:

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{e_i}{\sigma}\right) = \frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{y_i - \beta^T x_i}{\sigma}\right) = k \tag{12}$$

where β is the $p \times 1$ parameter vector, and the function ψ yields:

$$\sum_{i} \psi\left(e_i\right) \frac{\partial e_i}{\partial \beta_i} for\ i = 1, 2, \ldots, p \tag{13}$$

The derivative $\psi\left(e\right) = \frac{\partial \rho(e)}{\partial(e)}$ impacts the function. The weight function is expressed as :

$$w\left(e\right) = \frac{\psi\left(e\right)}{e} \tag{14}$$

where the function $\psi\left(e\right)$ is defined as:
$\sum w\left(e_i\right) e_i \frac{\partial e_i}{\partial \beta_i} = 0$, for $i = 1, 2, \ldots, p$.
The objective is to solve the iterated re-weighted least square equation as follow:

$$min \sum_{i} w\left(e_i^{(k-1)}\right) e_i^2 \tag{15}$$

where k represents the iteration number.

The M robust regression is divided into M-Bi- Square Tukey, M-Huber and M-Hampel. Table 1 provides a summary of the three types of M-robust regression methods.

S-estimation

Based on the discussion by[59,60], S-estimators which is proposed by Rousseeuw and Yohai derives from the residual scale used in M- estimation. The main limitation of M-estimation is that it does not account for the overall data distribution since it relies only on the median as the weighted value, making it less representative of

Methods	Objective function	Weight function
Bi-Square	$\rho_{BS} = \begin{cases} \frac{k^2}{6}\left\{1 - \left[1 - \left(\frac{e}{k}\right)^2\right]^3\right\} for\ \lvert e\rvert \le k \\ \frac{k^2}{6} for\ \lvert e\rvert > k \end{cases}$	$w_{BS} = \begin{cases} \left[1 - \left(\frac{e}{k}\right)^2\right]^2 for\ \lvert e\rvert \le k \\ 0 for\ \lvert e\rvert > k \end{cases}$
Huber	$\rho_{Huber} = \begin{cases} \frac{1}{2}e^2 for\ \lvert e\rvert \le k \\ k\lvert e\rvert - \frac{1}{2}k^2 for\ \lvert e\rvert > k \end{cases}$	$w_{Hub} = \begin{cases} 1 for\ \lvert e\rvert \le k \\ \frac{k}{\lvert e\rvert} for\ \lvert e\rvert < k \end{cases}$
Hampel	$\rho_{Ham} = \begin{cases} \frac{e^2}{2}, 0 < \lvert e\rvert < a \\ a\lvert e\rvert - \frac{e^2}{2}, b < \lvert e\rvert \le c \\ \frac{-a}{2(c-b)}\left(c - e\right)^2 + \frac{a}{2}\left(b + c - a\right), b < \lvert e\rvert \le c \end{cases}$	$w_{Ham} = \begin{cases} 1 for 0 < \lvert e\rvert < a \\ \frac{a}{\lvert e\rvert} for b < \lvert e\rvert \le c \\ a\frac{\frac{c}{\lvert e\rvert} - 1}{c-b} for b < \lvert e\rvert \le c \end{cases}$

Table 1. Description of M-estimation robust regression.

the entire dataset. To overcome this, the method incorporates the residual standard deviation. The S-estimator is defined by:

$$\hat{\beta}_S = min_\beta \hat{\sigma}_{sd}\left(e_1, e_2, e_3, \ldots, e_n\right)$$

with identifying the minimum robust scale estimator $\hat{\sigma}_S$ and ensuring

$$min \sum_{i=1}^{n} \rho \left(\frac{y_i - \sum_{j=0}^{k} x_{ij}\beta}{\hat{\sigma}_{sd}} \right) \tag{16}$$

where

$$\hat{\sigma}_{sd} = \sqrt{\frac{1}{nK} \sum_{i=1}^{n} w_i e_i^2} \tag{17}$$

$K = 0.199, w_i = w_{\sigma(u_i)=} \frac{\rho(u_i)}{u_i^2}$ and the initial estimate is:

$$\hat{\sigma}_s = \frac{median\left|e_i - median\left(e_i\right)\right|}{0.6745} \tag{18}$$

The solution is determined by taking the derivative with respect to β:

$$\sum_{i=1}^{n} x_{ij} \psi \left(\frac{y_i - \sum_{j=0}^{k} x_{ij}\beta}{\hat{\sigma}_{sd}} \right) = 0,\ i = 0, 1, 2, \ldots, k \tag{19}$$

ψ is the function as derivatives of ρ:

$$\psi\left(u_i\right) = \rho'\left(u_i\right) \begin{cases} u_i \left[1 - \left(\frac{u_i}{c}\right)^2\right]^2, & |u_i| \le c \\ 0, & |u_i| > c \end{cases} \tag{20}$$

S-estimators are known to be more robust than M-estimators because they have lower asymptotic bias and variance, especially when dealing with contaminated data.

MM-estimation

MM-estimation integrates S-estimation, which has a high breakdown point, with M-estimation[61]. A study by[60] compared the S, M and MM methods and found that MM was the most effective, as it had the smallest bias and mean square error (MSE). According to[59], the MM estimation process consists of two steps. First, S-estimation estimates the regression coefficients by minimizing residual scale from M-estimation, followed by the application of M-estimation. MM estimator can be described as:

$$\sum_{i=1}^{n} \rho_1'\left(u_i\right) X_{ij} = 0 \text{ or } \sum_{\mathrm{i=1}}^{\mathrm{n}} \rho_1' \left(\frac{Y_i - \sum_{j=0}^{k} X_{ij}\hat{\beta}_j}{SD_{MM}} \right) X_{ij} = 0 \tag{21}$$

Here, SD_{MM} refers to the standard deviation determined based S estimation residuals and ρ represents Tukey's biweight function, given by:

$$\rho\left(u_i\right) = \begin{cases} \frac{u_i^2}{2} - \frac{u_i^4}{2c^2} + \frac{u_i^6}{6c^2}, & -c \le u_i \le c \\ \frac{c^2}{6}, & u_i \left\langle -cor u_i \right\rangle c. \end{cases} \tag{22}$$

Model evaluation

Assessing a model's precision is essential in regression analysis. This study evaluates model performance using SSE, MSE, MAPE and R-squared as evaluation metrics. These indicators facilitate the comparison of how well each regression model predicts moisture content removal. In general, lower MSE, SSE, and MAPE values indicate higher prediction accuracy, while a greater R-squared signifies stronger model fit to the data.

Each of these metrics serves a distinct purpose in evaluating model performance. MSE serves as a widely used performance indicator that evaluates the average squared deviation between actual and predicted values in the dataset. It is particularly useful for models that predict a continuous variable due to its connection to the principle of cross-entropy from information theory[62]. SSE quantifies the difference between the observed data and a predictive model, with lower SSE values indicating that the model can more accurately explain the data[63]. MAPE is frequently used as a performance metric for regression models due to its straightforward interpretation in relation to relative error. It represents the average absolute error expressed as a percentage over a sample[64]. A MAPE value below 10 indicates a highly accurate forecast, while values exceeding 50 suggest an inaccurate

Range of R^2 Values	Description
$85\% \leq R^2 \leq 100\%$	Very good
$70\% \leq R^2 < 85\%$	Good
$50\% \leq R^2 < 70\%$	Reasonably good
$30\% \leq R^2 < 50\%$	Reasonably bad
$15\% \leq R^2 < 30\%$	Bad
$0\% \leq R^2 < 15\%$	Very bad

Table 2. Range of R^2 values.

Range of R^2 Values	Description	References
Mean squared error (MSE)	$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.	33
Sum of squared error (SSE)	$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.	33
Mean absolute percentage error (MAPE)	$MAPE = \left[\frac{100}{n}\right]\sum_{i=1}^{n}\left\|\frac{(y_i-\hat{y}_i)}{y_i}\right\|$.	31
R-squared	$R^2 = \frac{SSR}{SST} = \frac{SST-SSE}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2}$.	33

Table 3. Evaluation metrics formulas.

forecast[65]. Meanwhile, R-squared shows the proportion of variance in the dependent variable that is explained by the variation in the independent variables[66].

According to[17,67], this study will apply the R-squared value ranges outlined in Table 2, to assess regression model quality:

Table 3 presents the formulas for the evaluation metrics, where y_i denote the actual observations, $\widehat{y}_i$ denote the predicted values, $\overline{y}$ signifies the average of all observations, and n refers to the total count of observations.

Eight selection criteria

In the next step, the best model from each group is identified using the Eight Selection Criteria (8SC). Based on[27,68] and[69], the 8SC consists of Akaike Information Criterion (AIC), Final Prediction Error (FPE), Generalized Cross-Validation (GCV), Hannan-Quinn information criterion (HQ), Risk Inflation Criterion Estimate (RICE), SCHWARZ, SGMASQ, and SHIBATA. The optimal model is determined by selecting the one that yields the highest number of minimum values across these criteria.

The formulas for 8SC are presented in Table 4. Here, SSE represents the sum of squared errors, $k+1$ corresponds to the number of estimated parameters, and n denotes the sample size. As noted by[70], these criteria can be applied only when the condition $2(k+1) < n$ is satisfied.

Results and discussion

Table 5 highlights that parameters T7 and T11 exhibit heterogeneity, as their R-squared values fall below the benchmark, which is the average R-squared of 0.8372. Moreover, the VIF values are notably high, with the highest reaching 76,050.9483, indicating a high level of multicollinearity.

Additionally, the box plot serves as supporting evidence for detecting heterogeneity within the drying parameters. It is particularly useful for analyzing symmetry, variability, and identifying potential outliers. The boxplot also helps visualize data distribution by showing the median and quartiles for location and using the interquartile range to capture variability[79]. Figure 3 illustrates the variability in the 29 primary drying parameters of black pepper. Each box plot corresponds to a specific drying parameter of black pepper, providing insight into the variability among the key parameters[19]. The box plot reveals that the variables T7, T11, H1, H5 and PY show variation. Notably, the patterns observed for T7 and T11 align with the earlier R-squared results, confirming their heterogeneity. However, while H1, H5 and PY appear heterogeneous in the box plot, their R-squared values exceed the benchmark of 0.4843, suggesting that they do not show heterogeneity based on the previous findings.

While data visualization enhances decision-making by improving speed and quality[80], it simplifies complex data into visual representation, which may lead to some loss of detail. On the other hand,[81] highlighted that quantitative results from numerical operations and statistical analysis tend to ensure greater accuracy and

Model selection criteria	Description	References
AIC	$\left(\frac{SSE}{n}\right)(e)^{2(k+1)/n}$	71
FPE	$\left(\frac{SSE}{n}\right)\frac{n+(k+1)}{n-(k+1)}$	72
GCV	$\left(\frac{SSE}{n}\right)\left[1-\left(\frac{k+1}{n}\right)\right]^{-2}$.	73
HQ	$\left(\frac{SSE}{n}\right)(\ln n)^{2(k+1)/n}$	74
RICE	$\left(\frac{SSE}{n}\right)\left[1-\left(\frac{2(k+1)}{n}\right)\right]^{-1}$	75
SCHWARZ	$\left(\frac{SSE}{n}\right)n^{(k+1)/n}$	76
SGMASQ	$\left(\frac{SSE}{n}\right)\left[1-\left(\frac{k+1}{n}\right)\right]^{-1}$	77
SHIBATA	$\left(\frac{SSE}{n}\right)\frac{n+2(k+1)}{n}$	78

Table 4. Eight selection criteria formulas.

Lowest VIF	Highest VIF	Lowest R^2	Highest R^2	Average R^2	Parameters exhibiting heterogeneity
3.0711	76,050.9483	0.6744	0.9999	0.8372	T7, T11

Table 5. Heterogeneity Identification.

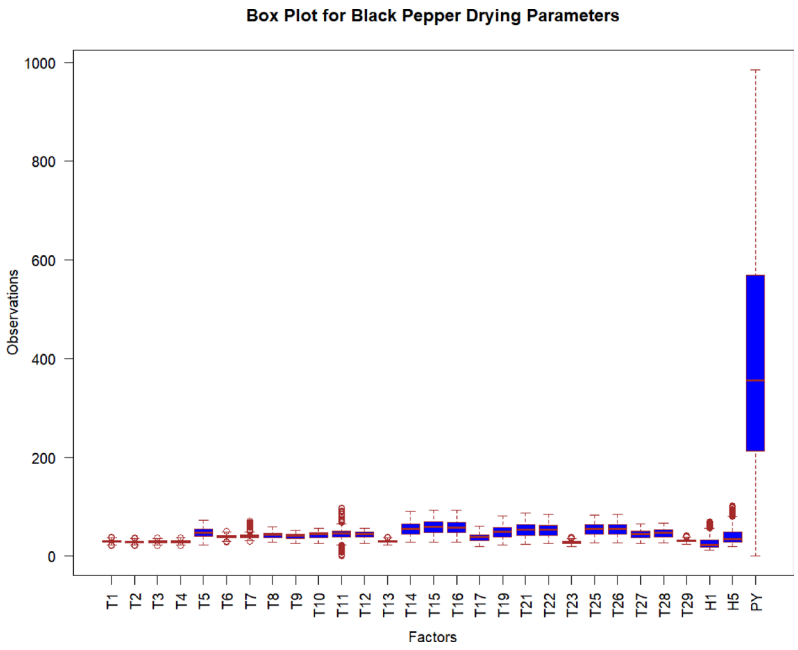

Fig. 3. Box plot for black pepper drying parameters.

reliability in predictions. Hence, the average R-squared value will be employed in this context to identify heterogeneity among the drying parameters, indicating that T7 and T11 exhibit heterogeneity. Once these two parameters contributing to heterogeneity and their second-order interaction are removed from the model, 378 parameters remain for determining the moisture content removal of black pepper. However, note that the ridge model can only select significant parameters up to the 89 highest-ranking variables after removing heterogeneity parameters, as the remaining ones are insignificant.

Table 6 provides a summary of the analysis comparison based on evaluation metrics, showing the results before and after removing heterogeneity parameters. Tables 7 and 8 present the outcomes of the 8SC for the models before and after the removal of heterogeneity parameters respectively. Before accounting for heterogeneity, the Elastic Net model shows a downward trend in SSE, MSE and MAPE values, while R-squared values increases as more high-ranking variables are included. For instance, with the 25 highest-ranking variables, the MSE is

Machine Learning Models	Highest Important Variables	Evaluation Metrics before heterogeneity				Evaluation Metrics after heterogeneity			
		SSE	MSE	MAPE	R^2	SSE	MSE	MAPE	R^2
Elastic Net	25	120,669.2	62.71790	14.61765	0.8201228	131,564.8	68.38090	14.56960	0.8038811
	35	108,149.6	56.21083	13.67150	0.8387853	116,171.3	60.38012	13.73545	0.8268276
	45	100,572.8	52.27275	13.07322	0.8400798	114,426.0	59.47298	13.75388	0.8294293
	55	97,914.47	50.89109	13.06336	0.8540425	114,021.0	59.26248	13.71846	0.8300330
	100	93,810.08	48.75784	12.90658	0.8601607	106,783.8	55.50094	13.32597	0.8408213
Ridge	25	187,316.2	97.35771	19.69328	0.7207745	145,198.8	75.46713	16.57897	0.7835575
	35	116,333.3	60.46430	14.38491	0.8265862	127,026.6	66.02213	14.71922	0.8106461
	45	111,069.1	57.72822	13.90969	0.8344334	118,844.8	61.76964	14.10132	0.8228424
	55	107,532.8	55.89021	13.63337	0.8397048	118,427.9	61.55299	14.08642	0.8234638
	100	98,948.36	51.42846	13.17582	0.8525013	110,705.2	57.53911	13.52590	0.8349757
LASSO	25	121,560.7	63.18123	14.68598	0.8187939	121,990.1	63.40441	14.37650	0.8181538
	35	113,221.8	58.84711	13.94504	0.8312243	115,746.6	60.15934	13.90809	0.8274608
	45	101,489.5	52.74919	13.28239	0.8487134	114,349.3	59.43310	13.69018	0.8295437
	55	99,817.19	51.88004	13.15495	0.8512062	112,408.1	58.42416	13.66430	0.8324374
	100/89 for after	94,122.63	48.92029	12.95327	0.8596948	107,155.0	55.69389	13.34491	0.8402679

Table 6. Comparison before and after removing heterogeneity parameters.

Machine learning models	Highest Important Variables	AIC	FPE	GCV	HQ	RICE	SCHWARZ	SGMASQ	SHIBATA
Elastic Net	25	64.43607	64.43618	64.44795	66.24278	64.46004	69.46604	63.57703	64.41296
	35	58.35419	58.35444	58.37488	60.63182	58.39611	64.75461	57.28263	58.31433
	45	54.83302	54.83352	54.86488	60.29664	54.89782	62.63196	53.55314	54.77229
	55	54.15528	54.16052	54.31018	57.58240	54.47740	62.51839	52.45918	53.87690
	100	53.94150	53.94238	53.98812	57.25186	54.03668	63.42084	51.41674	53.85357
Ridge	25	100.0249	100.0250	100.0433	102.8294	100.0621	107.83293	98.69136	99.98898
	35	62.76986	62.77013	62.79212	65.21984	62.81496	69.6546	61.61721	62.72699
	45	60.55568	60.55623	60.59086	63.59199	60.62724	69.16855	59.14223	60.48862
	55	59.24027	59.24125	59.29148	62.87582	59.34481	69.65079	57.56574	59.14371
	100	57.12154	57.12706	57.28492	63.59928	57.46130	76.49045	54.27776	56.82792
LASSO	25	64.91212	64.91223	64.92409	66.73218	64.93627	69.97925	64.04673	64.88884
	35	61.09099	61.09126	61.11265	63.47544	61.13488	67.79159	59.96917	61.04927
	45	55.33281	55.33331	55.36496	58.10725	55.39820	63.20283	54.04127	55.27153
	55	54.98971	54.99062	55.03724	58.36440	55.08675	64.65326	53.43533	54.90008
	100	54.33572	54.34096	54.49113	60.49753	54.65890	72.76000	51.63063	54.05641

Table 7. The 8SC for models before removing heterogeneity parameters.

62.7179 and the MAPE is 14.61765. When the number of high-ranking variables increases to 100, the MSE declines to 48.75784 and the MAPE drops to 12.90658. Meanwhile, the R-squared value increases from 0.8201 for the 25 highest-ranking variables to 0.8602 for the 100 highest-ranking variables. A similar trend is observed in the LASSO and Ridge models, where SSE, MSE and MAPE consistently decrease, while R-squared values increase with the inclusion of more high-ranking variables before heterogeneity. In addition, the 8SC values also decrease across all criteria as the number of high-ranking variables increases. The 8SC further strengthen the evaluation by confirming which model captures the factors that influence the moisture removal in black pepper. Since drying efficiency depends on interactions among temperature, airflow and humidity, the model with the lowest 8SC values is more likely to represent these relationships accurately. This ensures that the selected model not only minimizes error metrics but also provides a reliable representation of the moisture reduction process. Overall, adding more variables consistently improves model performance, as indicated by lower error metrics, higher R-squared values and decreased 8SC values. After addressing heterogeneity, this pattern remains consistent across all three models. Generally, selecting a larger number of high-ranking variables improves predictive accuracy, which confirms the findings of[19,33].

The Elastic Net model consistently performs better than the Ridge and LASSO models for the 25, 35, 45, 55 and achieves its best results for 100 highest ranking variables before the removal of heterogeneity parameters. This advantage is evident from the Elastic Net model's lower SSE, MSE and MAPE, along with its higher R-squared values. The higher R-squared values indicate that the model explains a greater percentage of variation in moisture content removal for black pepper compared to Ridge and LASSO. For example, achieving an R-squared value of 0.8602 for the 100 highest important variables, the Elastic Net model accounts for 86.02%

Machine learning models	Highest Important Variables	AIC	FPE	GCV	HQ	RICE	SCHWARZ	SGMASQ	SHIBATA
Elastic Net	25	70.25421	70.25432	70.26715	72.22405	70.28034	75.73834	69.31760	70.22900
	35	62.68245	62.68272	62.70468	65.12901	62.72748	69.55760	61.53141	62.63964
	45	62.38588	62.38645	62.42213	65.51397	62.45961	71.25907	60.92971	62.31679
	55	62.81465	62.81568	62.86894	66.66956	62.92550	73.85331	61.03908	62.71226
	100	61.64484	61.65079	61.82115	68.63553	62.01150	82.54751	58.57586	61.32796
Ridge	25	77.53462	77.53474	77.54891	79.70859	77.56346	83.58707	76.50095	77.50680
	35	68.53964	68.53993	68.56394	71.21481	68.58888	76.05722	67.28104	68.49282
	45	64.79505	64.79564	64.83270	68.04393	64.87162	74.01089	63.28264	64.72329
	55	65.24243	65.24350	65.29882	69.24633	65.35756	76.70773	63.39823	65.13609
	100	63.90861	63.91478	64.0914	71.15602	64.28873	85.57889	60.72693	63.58009
LASSO	25	65.14142	65.14153	65.15342	66.96790	65.16565	70.22644	64.27297	65.11805
	35	62.45330	62.45357	62.47544	64.89091	62.49816	69.30331	61.30646	62.41064
	45	62.34407	62.34463	62.38029	65.47005	62.41774	71.21130	60.88887	62.27502
	55	61.92610	61.92712	61.97962	65.72647	62.03538	72.80860	60.17564	61.82516
	89	61.85913	61.86510	62.03605	68.87412	62.22706	82.83446	58.77948	61.54114

Table 8. The 8SC for models after removing heterogeneity parameters.

of the variation, demonstrating strong predictive ability. Since its R-squared values consistently exceed 80%, the model quality ranges from good to very good. Although all three models have MAPE values between 10 and 20, which suggests a good level of forecasting accuracy, the Elastic Net model maintains lower MAPE values across all sets of high-ranking variables, significantly outperforming Ridge and LASSO. In line with these findings, all 8SC also identify Elastic Net as the best model, as it consistently records the lowest values across all criteria and variable sets. This further confirms that Elastic Net provides the most reliable model fit before addressing heterogeneity. Generally, before the removal of heterogeneity parameters, the Elastic Net model which is a combination of L_1 and L_2 regularization proves to be the superior choice, as it maintains a balance between variable selection and model stability. This combination enables it to manage highly correlated predictors and handle multicollinearity effectively[56].

After removing the heterogeneity parameters, the LASSO model shows better predictive performance with better accuracy and reduced error for the 25, 35, 45 and 55 highest-ranking variables compared to the other two models. Although the Elastic Net model, with 100 variables, records a slightly higher R-squared (0.8408) than LASSO with 89 variables ($R^2 = 0.8403$), the difference is minimal. Consistent with these results, the 8SC show that LASSO achieves the lowest values for the 25, 35, 45, and 55 variable sets, confirming it as the best model in these cases. The 8SC also indicate that Elastic Net has the lowest values for the 100-variable set, although the improvement over LASSO with 89 variables is minimal. This suggests that while Elastic Net offers a slightly better fit at higher variable counts, LASSO remains the more efficient and stable option after heterogeneity is removed. Since LASSO reaches almost the same level of accuracy with fewer significant predictors, it proves to be more efficient and stable after the removal of heterogeneity parameters. Similarly to the analysis before excluding parameters exhibiting heterogeneity, all three models have MAPE values ranging between 10 and 20, suggesting a good level of forecasting accuracy. However, LASSO achieves the lowest MAPE values for the 25, 35, 45 and 55 highest ranking variables, confirming it strong predictive ability. This also supports the view that LASSO works well with high-dimensional data by reducing redundancy, preventing overfitting, and improving interpretability[47]. Elastic Net, on the other hand, continues to perform well as the number of high-ranking variables increases due to its strength in managing multicollinearity.

Interestingly, the results indicate that the accuracy of the regression model is unexpectedly reduced by the removal of heterogeneity parameters. When parameters are removed from the model, it can lead to a loss of meaningful variability, causing the model to overlook important relationships among predictors. Such elimination can introduce specification bias in the model, as it may not adequately consider crucial factors that influence moisture content removal, ultimately affecting its predictive performance[17]. Overall, the superior performance of the Elastic Net before eliminating heterogeneity parameters implies that retaining this natural variability helps capture meaningful interactions among drying factors such as temperature, humidity, and airflow. Its minimal sensitivity to heterogeneity further explains why Elastic Net remains stable and reliable under varying conditions. Preserving these interdependencies allows the model to more accurately represent real drying conditions, leading to improved variable selection and more reliable parameter estimation. In practical terms, these findings highlight the importance of considering parameter relationships when optimizing drying systems to enhance efficiency and ensure consistent product quality.However, once heterogeneity parameters are removed, the data becomes cleaner and less correlated. Consequently, the LASSO model emerges as the better predictive model due to its strong regularization effect, which enables it to focus on the most important predictors while shrinking less relevant ones to zero. This may explain why LASSO slightly outperformes Elastic Net after accounting for heterogeneity.

Figures 4 and 5 present the standardized residual plots of the optimal Elastic Net, Ridge and LASSO models with the robust estimators, both before and after removing the heterogeneity parameters, respectively. Table 9 compares the outlier counts and their corresponding percentages exceeding the 2-sigma and 3-sigma limits for

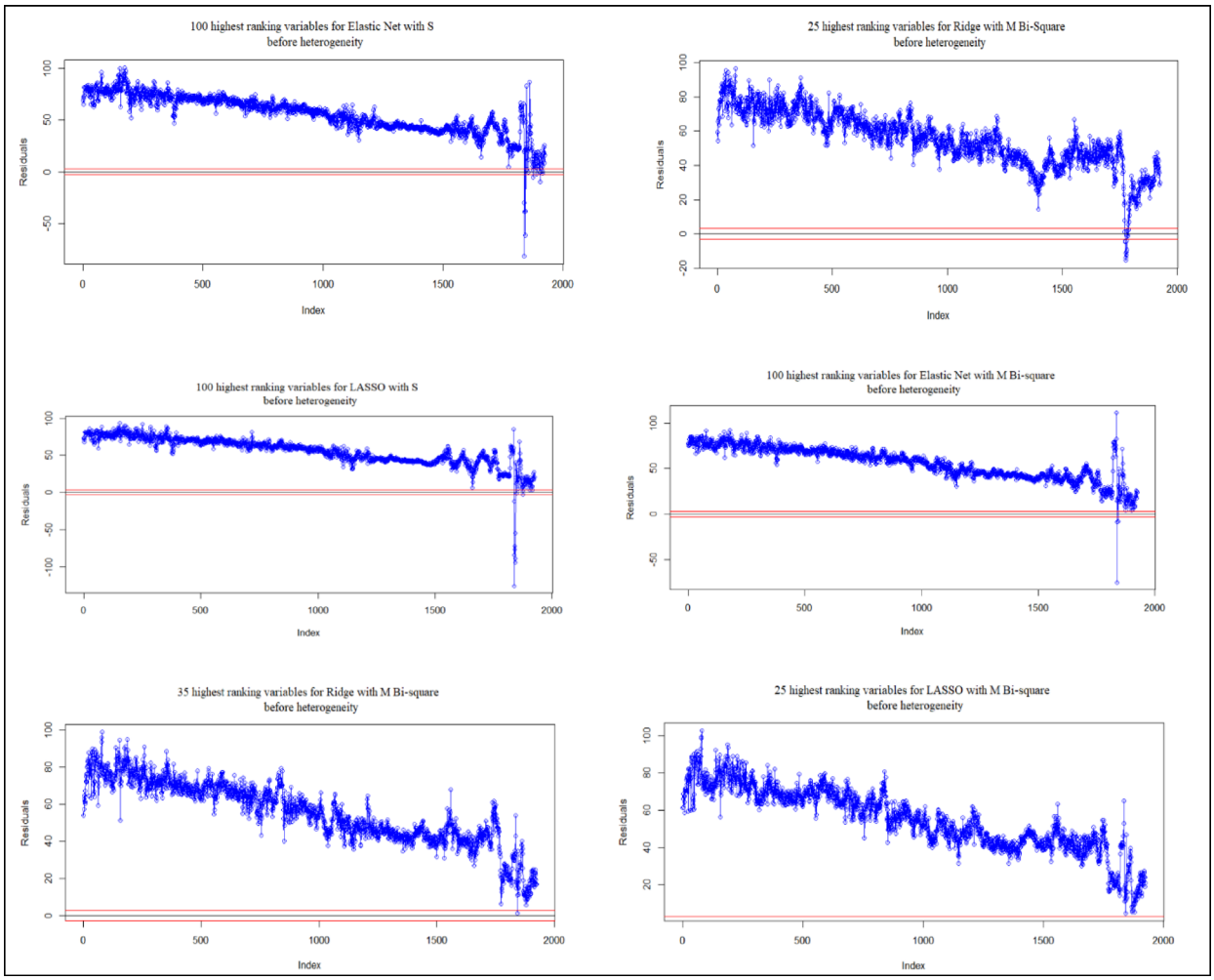

Fig. 4. Standardized residual plots for the optimal model with robust method before heterogeneity.

Elastic Net, Ridge and LASSO, applying robust methods before and after addressing heterogeneity, for 25, 35, 45, 55 and 100 highest-ranking variables. Figures 6, 7 and 8 depict the plots for showing the performance of elastic net before and after heterogeneity, ridge before and after heterogeneity and LASSO before and after heterogeneity, respectively. Before heterogeneity, the hybrid model shows significant improvements in reducing outliers across all three sparse regression models (Elastic Net, Ridge and LASSO). For Elastic Net, the S estimator performs best for the 100 highest ranking variables at the 2-sigma limits, reducing outliers from 113 in the original model to 19, with an 83% reduction. At the 3-sigma limits, the M Bi-Square estimator proves most effective, lowering outliers from 4 to 3 (a 25% reduction). In the Ridge model, the M Bi-Square estimator achieves the highest reduction of 90% for the 25 highest-ranking variables at the 2-sigma limits, identifying only 14 outliers compared to 144 in the original model. Although the MM estimator with 100 variables also performs well with an 84% reduction, the improvement is slightly lower, suggesting that the Ridge M Bi-Square hybrid model works most efficiently with 25 variables at 2 sigma-limits. Meanwhile, at the 3-sigma limits, the hybrid model with the M Bi-Square estimator completely eliminates outliers, lowering the count from 5 to 0 for the 35 highest-ranking variables. For LASSO, the S estimator performs best for the 89 highest ranking variables at the 2-sigma limits, detecting 15 outliers compared to 115 in the original (87% reduction). At the 3-sigma limit, M Bi-Square reduces outliers from 7 to 4 for the 25 highest-ranking variables, indicating a 42% decrease.

Following the removal of heterogeneity parameters, the hybrid models continue to show strong reductions in outliers. For Elastic Net, the MM and S estimators perform best at the 2-sigma limits, identifying 23 outliers among the 55 highest-ranking variables, approximately 85% reduction from the original 158 outliers. At the 3-sigma limits, the S estimator eliminates all outliers for the 45 and 55 highest-ranking variables, while the original model reports 12 and 11 outliers, respectively. This suggests that heterogeneity is a major source of outliers, and its removal makes the data appear more normally distributed, reducing the number of detected outliers[82]. In the Ridge model, the S estimator proves to be the most effective for the 45 highest-ranking variables at the 2-sigma limits, reducing outliers from 164 to 21 (87% reduction). Similarly, at the 3-sigma limits, it remains the best estimator for the 55 high-ranking variables, decreasing outliers from 12 to 1, representing a 91% reduction. For LASSO, the S estimator again performs best at both 2-sigma and 3-sigma limits for the 45 highest-ranking variables, reducing outliers from 163 to 16 (90% reduction) and 14 to 1, with roughly 92% decrease.

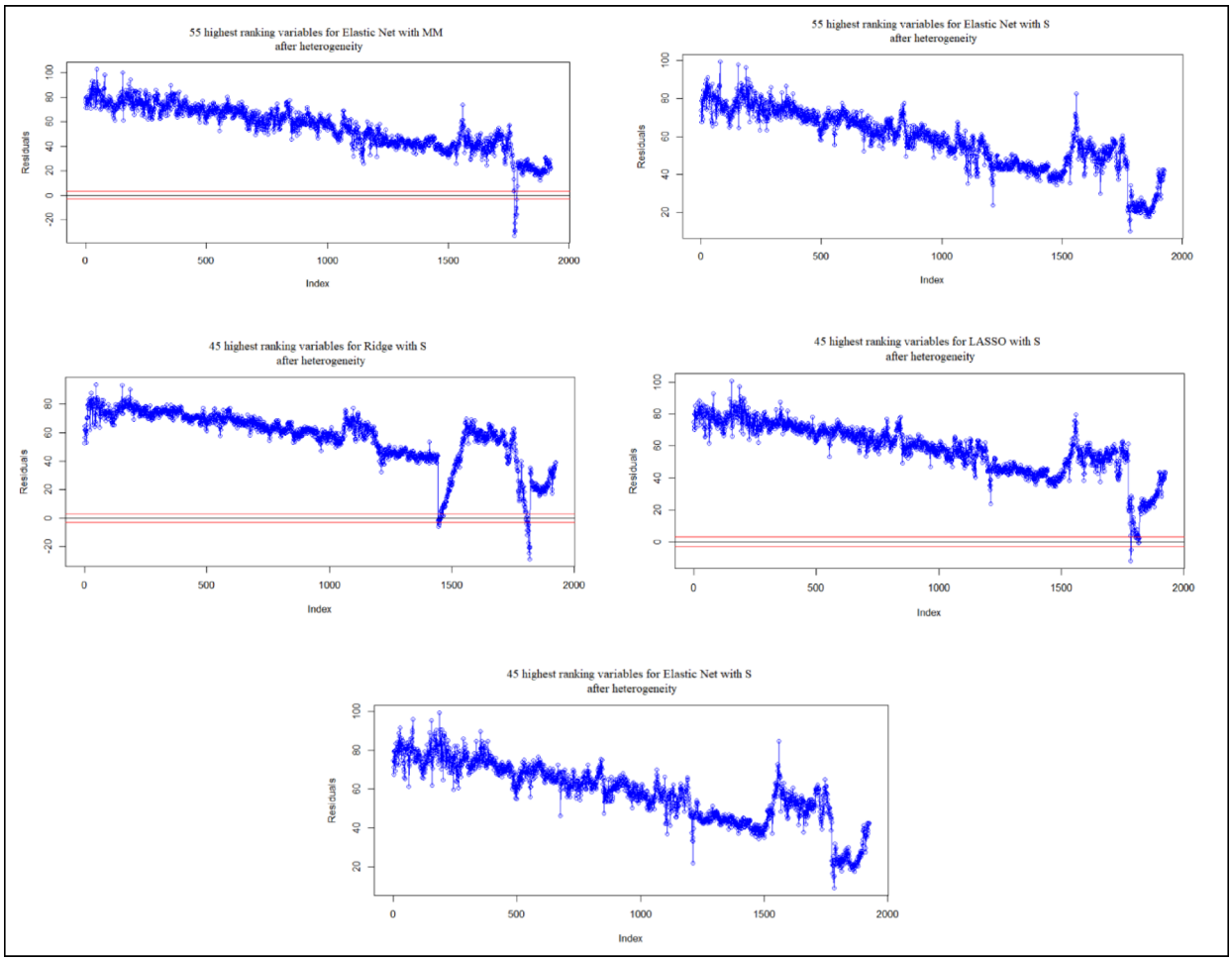

Fig. 5. Standardized residual plots for the optimal model with robust method after heterogeneity.

Overall, although outlier counts increase after removing heterogeneity parameters especially at 2-sigma limits, the number of outliers significantly declines from the 2-sigma to the 3-sigma limits across all models, with hybrid models demonstrating stronger performance compared to the original models. This is likely because the 3-sigma rule sets a high threshold for detecting outliers, making it less sensitive after applying the robust method[80]. Before removing heterogeneity parameters, the hybrid Ridge model with M Bi-Square estimator performs best at both 2-sigma and 3-sigma limits, completely eliminating outliers at the 3-sigma limit. After heterogeneity removal, the LASSO model combined with the S estimator is the best, achieving the highest reduction at both limits, particularly at the 3-sigma level. These findings confirm that hybrid models combining robust estimators with sparse regression can handle heterogeneity effectively both before and after its removal, maintaining strong performance across both conditions.

Conclusion

This study explores heterogeneity in the drying parameters of black pepper and proposes hybrid models that combine sparse and robust regression estimators to improve the accuracy of estimating moisture content removal. Elastic Net, Ridge and LASSO are applied for variable selection, followed by the development of hybrid models integrating these sparse regression techniques with robust regression estimators to detect and minimize the influence of outliers. Before heterogeneity removal, hybrid Ridge model with M Bi-Square demonstrates the best performance, whereas after removing heterogeneity, the hybrid LASSO model with S achieves the highest accuracy and stability, making it the most effective model. The consistent improvements observed at both 2-sigma and 3-sigma limits highlight the robustness of these methods, with the 3-sigma limit proving especially effective in reducing outliers and improving prediction reliability. The findings confirm that hybrid sparse-robust models are crucial for maintaining stable performance in the presence of heterogeneity. By improving predictive accuracy and reliability, these models offer valuable insights for optimizing sensor placement, drying conditions, and energy use in IoT-based solar drying systems. Ultimately, the proposed hybrid models enhance black pepper drying by incorporating significant parameters and their interactions, enabling farmers to produce high-quality dried pepper with uniform moisture content, improved yield, and shorter drying periods. This will boost efficiency and the income across the black pepper industry.

Machine learning models	Robust methods	Highest important variables	Before heterogeneity		After heterogeneity	
			$\mu \pm 2\sigma$ (%)	$\mu \pm 3\sigma$ (%)	$\mu \pm 2\sigma$ (%)	$\mu \pm 3\sigma$ (%)
Elastic Net	Original	25	142(7.42)	6(0.31)	161(8.41)	12(0.63)
		35	125(6.53)	10(0.52)	167(8.73)	12(0.63)
		45	120(6.27)	6(0.31)	163(8.52)	12(0.63)
		55	117(6.11)	9(0.47)	158(8.25)	11(0.57)
		100	**113(5.90)**	**4(0.21)**	**163(8.52)**	**8(0.42)**
	M Bi-Square	25	44(2.30)	11(0.57)	57(2.97)	17(0.89)
		35	51(2.66)	12(0.63)	49(2.56)	18(0.94)
		45	44(2.30)	14(0.73)	52(2.72)	13(0.68)
		55	26(1.36)	11(0.57)	58(3.03)	17(0.89)
		100	26(1.36)	**3(0.16)**	49(2.56)	15(0.78)
	M Hampel	25	44(2.30)	14(0.73)	60(3.13)	17(0.89)
		35	58(3.03)	10(0.52)	53(2.77)	16(0.84)
		45	54(2.82)	17(0.89)	50(2.61)	14(0.73)
		55	34(1.78)	12(0.63)	58(3.03)	16(0.84)
		100	33(1.72)	8(0.42)	49(2.56)	15(0.78)
	M Huber	25	46(2.40)	13(0.68)	57(2.98)	17(0.89)
		35	57(2.98)	10(0.52)	48(2.51)	16(0.84)
		45	55(2.87)	16(0.84)	51(2.66)	16(0.84)
		55	43(2.25)	11(0.57)	62(3.24)	15(0.78)
		100	44(2.30)	10(0.52)	45(2.35)	15(0.78)
	MM	25	42(2.19)	9(0.47)	60(3.13)	17(0.89)
		35	33(1.72)	8(0.42)	47(2.46)	16(0.84)
		45	46(2.40)	14(0.73)	56(2.93)	14(0.73)
		55	41(2.14)	8(0.42)	23(1.20)	14(0.73)
		100	20(1.04)	10(0.52)	**39(2.04)**	15(0.78)
	S	25	27(1.41)	16(0.84)	40(2.09)	8(0.42)
		35	45(2.35)	11(0.57)	38(1.99)	7(0.37)
		45	37(1.93)	15(0.78)	36(1.88)	0(0)
		55	27(1.41)	24(1.25)	23(1.20)	0(0)
		100	**19(0.99)**	6(0.31)	50(2.61)	13(0.68)

Continued

Machine learning models	Robust methods	Highest important variables	Before heterogeneity		After heterogeneity	
			$\mu \pm 2\sigma$ (%)	$\mu \pm 3\sigma$ (%)	$\mu \pm 2\sigma$ (%)	$\mu \pm 3\sigma$ (%)
Ridge	Original	25	144(7.52)	9(0.47)	148(7.73)	5(0.26)
		35	133(6.95)	7(0.37)	152(7.94)	8(0.42)
		45	135(7.05)	5(0.26)	162(8.46)	10(0.52)
		55	135(7.05)	6(0.31)	164(8.57)	12(0.63)
		100	**121(6.32)**	**5(0.26)**	**168(8.78)**	**12(0.63)**
	M Bi-Square	25	14(0.73)	1(0.05)	40(2.09)	7(0.37)
		35	23(1.20)	0(0)	53(2.77)	12(0.63)
		45	24(1.25)	13(0.68)	51(2.66)	12(0.63)
		55	24(1.25)	12(0.63)	46(2.40)	14(0.73)
		100	25(1.31)	9(0.47)	34(1.78)	14(0.73)
	M Hampel	25	22(1.15)	2(0.10)	37(1.93)	7(0.37)
		35	42(2.19)	7(0.37)	52(2.72)	15(0.78)
		45	20(1.04)	12(0.63)	52(2.72)	12(0.63)
		55	34(1.78)	10(0.52)	48(2.51)	14(0.73)
		100	26(1.36)	11(0.57)	46(2.40)	17(0.89)
	M Huber	25	16(0.84)	1(0.05)	39(2.04)	8(0.42)
		35	37(1.93)	3(0.16)	48(2.51)	13(0.68)
		45	28(1.46)	10(0.52)	50(2.61)	12(0.63)
		55	43(2.25)	12(0.63)	45(2.35)	14(0.73)
		100	39(2.04)	11(0.57)	40(2.09)	15(0.78)
	MM	25	45(2.35)	10(0.52)	42(2.19)	8(0.42)
		35	21(1.10)	1(0.05)	53(2.77)	12(0.63)
		45	15(0.78)	1(0.05)	33(1.72)	12(0.63)
		55	28(1.46)	23(1.20)	42(2.18)	12(0.63)
		100	**19(0.99)**	6(0.31)	45(2.35)	14(0.73)
	S	25	42(2.19)	11(0.57)	34(1.78)	6(0.31)
		35	38(1.99)	26(1.36)	28(1.46)	2(0.10)
		45	18(0.94)	1(0.05)	17(0.89)	5(0.26)
		55	28(1.46)	17(0.89)	21(1.10)	1(0.05)
		100	20(1.04)	8(0.42)	**33(1.72)**	13(0.68)

Continued

Machine learning models	Robust methods	Highest important variables	Before heterogeneity		After heterogeneity	
			$\mu \pm 2\sigma$ (%)	$\mu \pm 3\sigma$ (%)	$\mu \pm 2\sigma$ (%)	$\mu \pm 3\sigma$ (%)
LASSO	Original	25	150(7.84)	7(0.37)	167(8.73)	9(0.47)
		35	143(7.47)	5(0.26)	162(8.46)	8(0.42)
		45	136(7.11)	7(0.37)	163(8.52)	14(0.73)
		55	131(6.84)	6(0.31)	158(8.25)	9(0.47)
		100/89 for after	**115(6.01)**	**6(0.31)**	**161(8.41)**	**8(0.42)**
	M Bi-Square	25	35(1.83)	**4(0.21)**	44(2.30)	10(0.52)
		35	56(2.93)	17(0.89)	54(2.82)	16(0.84)
		45	57(2.98)	20(1.04)	53(2.77)	14(0.73)
		55	59(3.08)	14(0.73)	56(2.93)	19(0.99)
		100/89 for after	24(1.25)	9(0.47)	52(2.72)	18(0.94)
	M Hampel	25	41(2.14)	7(0.37)	42(2.19)	9(0.47)
		35	54(2.82)	10(0.52)	57(2.98)	20(1.04)
		45	55(2.87)	16(0.84)	53(2.77)	17(0.89)
		55	56(2.93)	13(0.68)	52(2.72)	19(0.99)
		100/89 for after	30(1.57)	7(0.37)	53(2.77)	15(0.78)
	M Huber	25	39(2.04)	6(0.31)	46(2.40)	10(0.52)
		35	55(2.87)	10(0.52)	55(2.87)	17(0.89)
		45	54(2.82)	16(0.84)	52(2.72)	18(0.94)
		55	56(2.93)	12(0.63)	55(2.87)	16(0.84)
		100/89 for after	34(1.78)	8(0.42)	54(2.82)	14(0.73)
	MM	25	31(1.62)	8(0.42)	43(2.25)	10(0.52)
		35	45(2.35)	10(0.52)	56(2.93)	14(0.73)
		45	32(1.67)	15(0.78)	54(2.82)	15(0.78)
		55	28(1.46)	23(1.20)	28(1.46)	13(0.68)
		100/89 for after	30(1.57)	23(1.20)	52(2.72)	18(0.94)
	S	25	23(1.20)	12(0.63)	29(1.52)	9(0.47)
		35	28(1.46)	14(0.73)	38(1.99)	2(0.10)
		45	29(1.52)	23(1.20)	**16(0.84)**	**1(0.05)**
		55	29(1.52)	24(1.25)	20(1.04)	9(0.47)

Table 9. Comparison of outliers counts and percentages outside 2 and 3—sigma limits in original and hybrid models before and after addressing heterogeneity. Significant values are in bold.

This study has several limitations; for example, the sensors determine the variables to be captured, and some variables were not captured due to measurement errors. Additionally, the findings are based on specific environmental conditions, which may not represent all possible use cases for the dryer. Differences in sensor accuracy and placement may also affect the results. This study is also limited to the main effects of the drying parameters and second -order interaction due to the constraints related to time, feasibility, and complexity of the models. Hence, it can be inferred that the moisture content removal of black pepper is determined by a total of 435 independent variable models. In the context of big data, determining the precise number of significant variables to include in a model is challenging. While the proposed algorithms can indicate the relative importance of variables, they do not explicitly determine how many of these should be selected, as feature selection methods only provide a ranking rather than a definitive count of significant variables[83]. Therefore, the 25, 35, 45, 55 and 100 highest-ranking variables are selected to determine the moisture content removal of black pepper. Future studies could explore other sparse regression models, such as adaptive LASSO, adaptive group LASSO, Minimax Concave Penalty and Smoothly Clipped Absolute Deviation (SCAD) for variable selection and the number of drying parameters chosen could also be expanded. Alternative robust estimators, like least median of squares (LMS) and least absolute deviations (LAD) could be applied to build a hybrid model. Hybrid model can also be explored to handle imbalanced data or missing values.

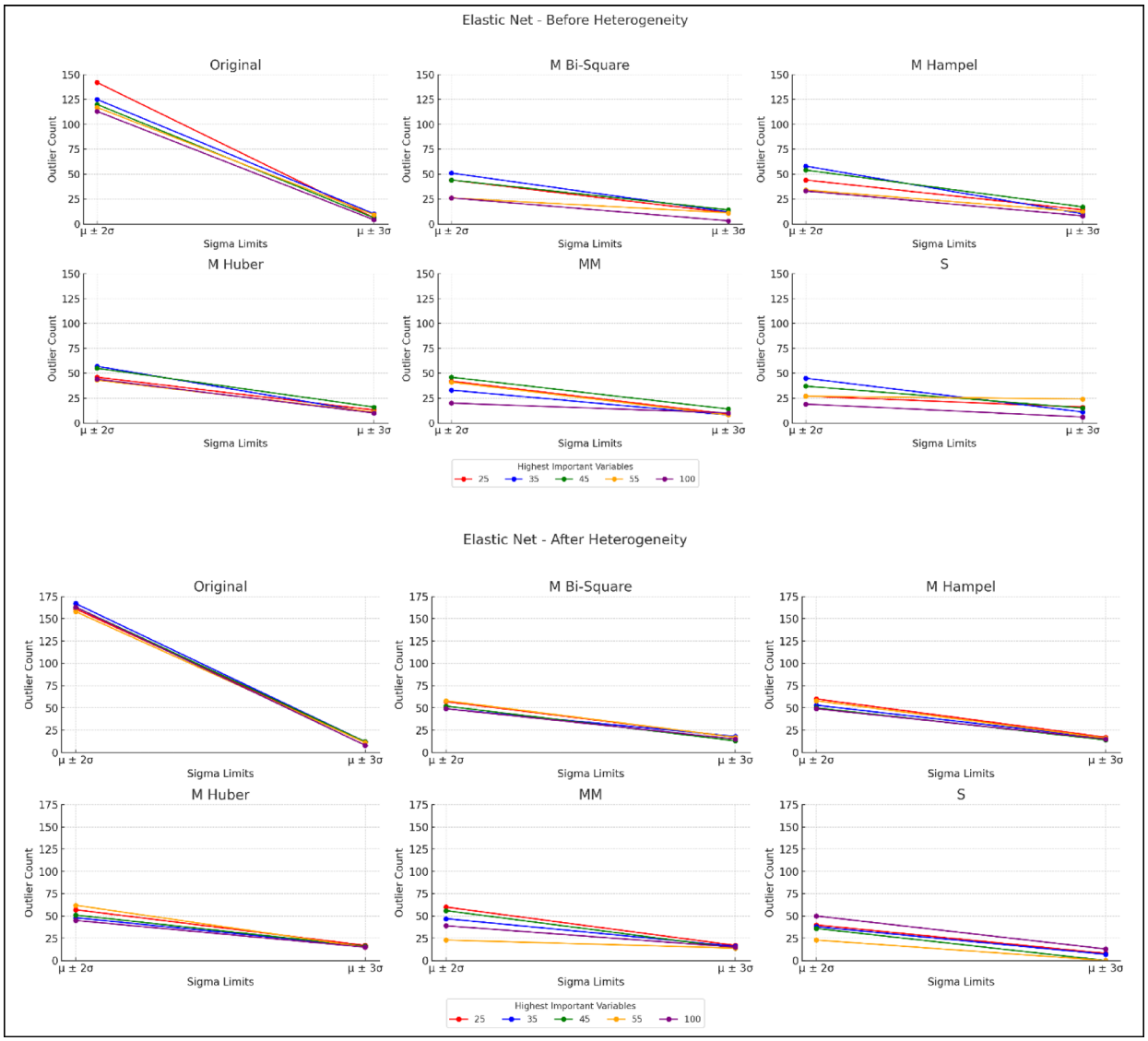

Fig. 6. Plots showing the performance of elastic net for before and after heterogeneity.

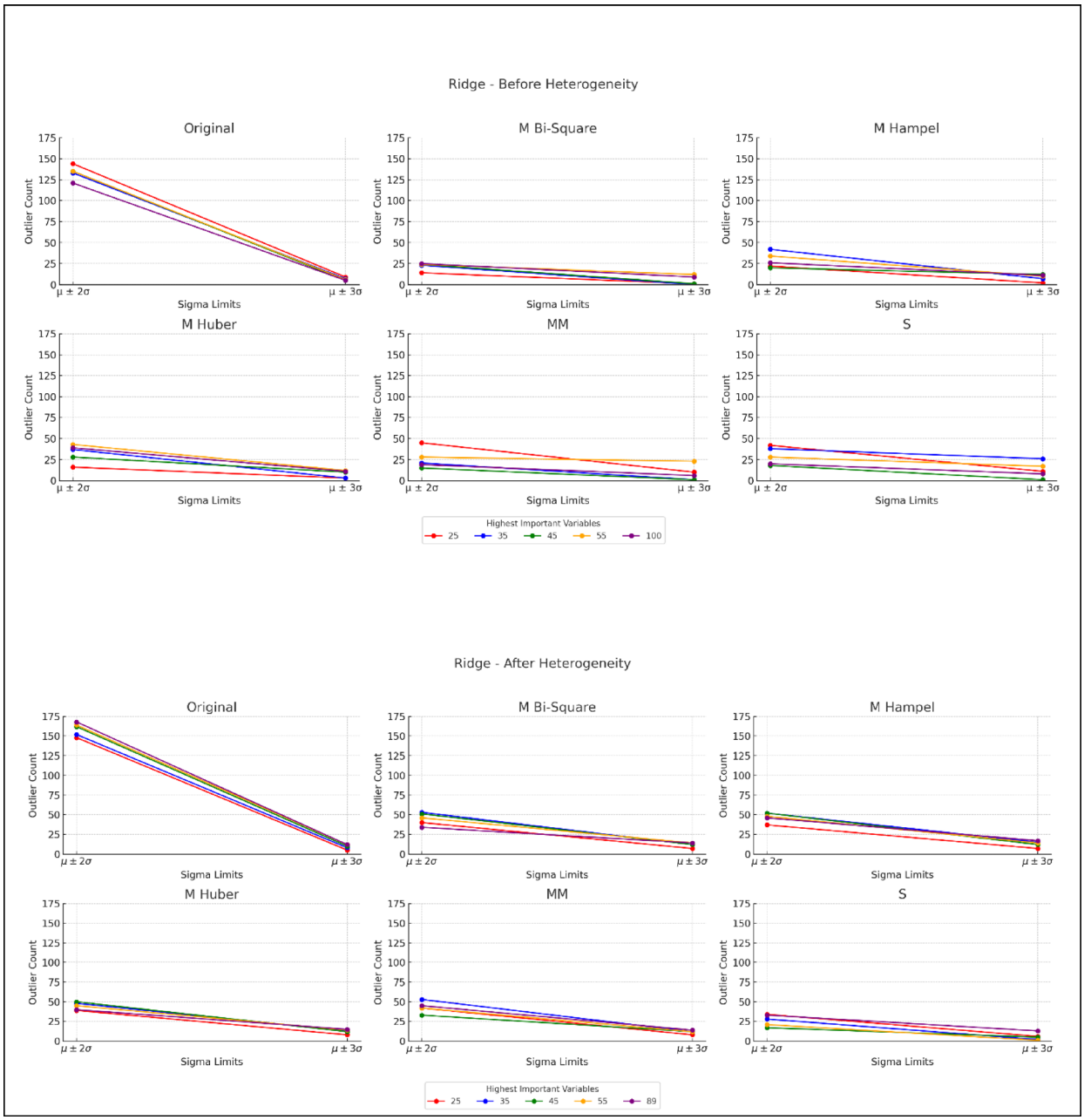

Fig. 7. Plots showing the performance of ridge for before and after heterogeneity.

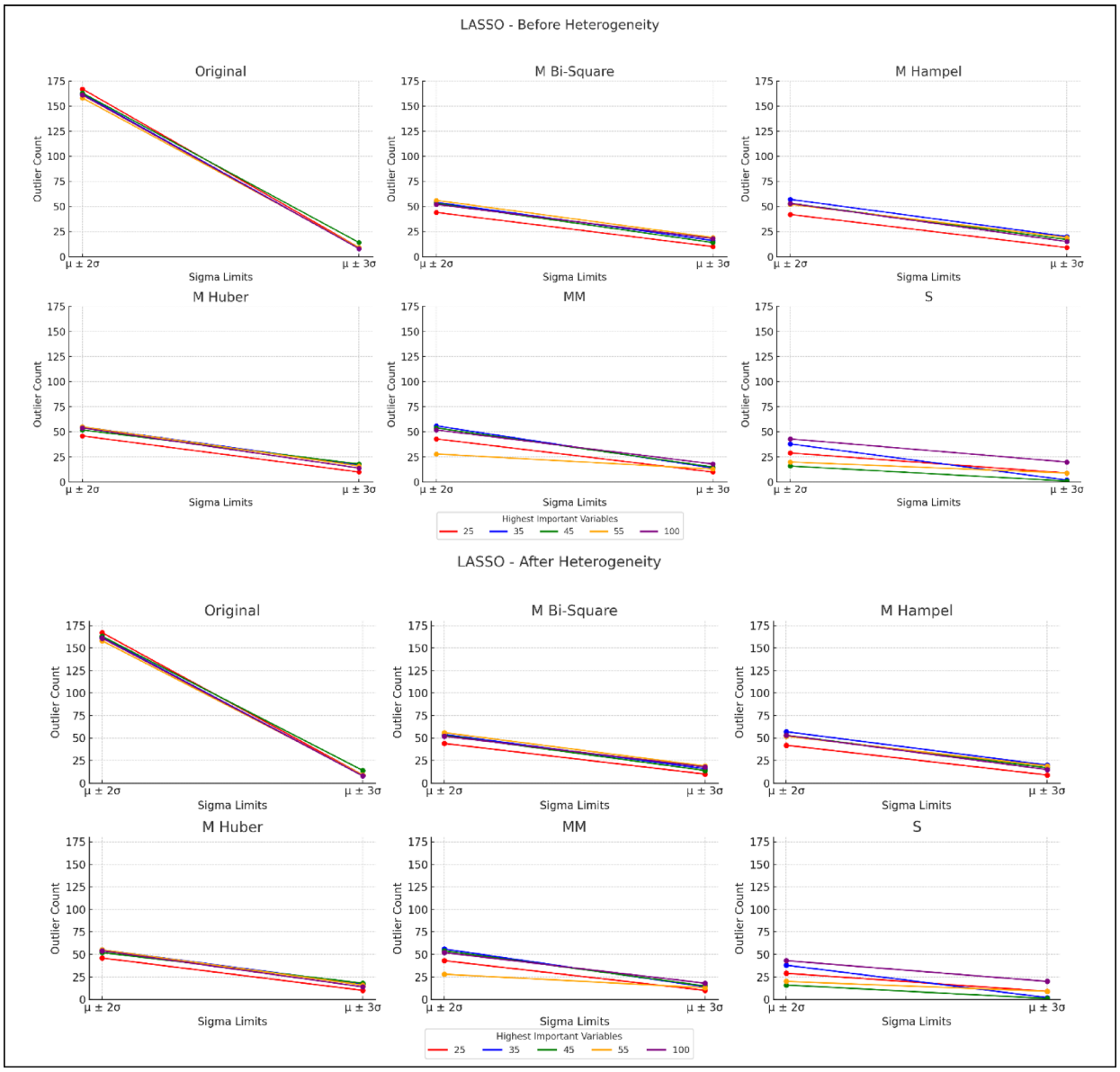

Fig. 8. Plots showing the performance of LASSO for before and after heterogeneity.

Data availability

All data are included in this article.

Received: 10 March 2025; Accepted: 4 February 2026
Published online: 03 April 2026

References

1. Kumari, A., Golyan, A., Shah, R., & Raval, N. Introduction to Data Analytics. In *Advances in ystems analysis, software engineering, and high performance computing book series* (pp. 1–14). https://doi.org/10.4018/979-8-3693-3609-0.ch001 (2024).
2. Gandhi, P. Towards data science in agriculture with big data management. *Res. Sq. (Res. Sq.)* https://doi.org/10.21203/rs.3.rs-4766405/v1 (2024).
3. Das, P., Banerjee, R., Bharti, R. A. & Varshney, N. *An Overview of Data Analytics in Smart Agriculture* (Bhumi Publishing, India, 2023).
4. Weraikat, D., Šorič, K., Žagar, M. & Sokač, M. Data analytics in agriculture: enhancing decision-making for crop yield optimization and sustainable practices. *Sustainability.* **16**(17), 7331. https://doi.org/10.3390/su16177331 (2024).
5. Pathakakula, S. *Chapter -16 Ripening Mechanism In Black Pepper*. 2nd edn. In Advances in horticulture and allied sciences (Vol. **2**). (Royal Book Publishing, 2023).
6. Priya & Garg, A.P. Biomedical Applications of Black Pepper, The King of Spices: A review. *Biomed. J. Sci. Tech. Res.* **53**(1). https://doi.org/10.26717/bjstr.2023.53.008353 (2023).
7. Madhumathy, S. Water activity and its impacts on food stability. *Int. J. Food Nutr. Sci.* **10**(12), 832–850 (2021).

8. Tapia, M. S., Alzamora, S. M. & Chirife, J. Effects of water activity (a_w) on microbial stability: As a hurdle in food preservation. *Water Act. Foods Fundam. Appl.* https://doi.org/10.1002/9781118765982.ch14 (2020).
9. Rodrigues, S. S. Q. et al. Use of black pepper essential oil to produce a healthier chicken pâté. *Appl. Sci.* **15**(4), 1733. https://doi.org/10.3390/app15041733 (2025).
10. Vieira, L. V. et al. The effects of drying methods and harvest season on piperine, essential oil composition, and multi-elemental composition of black pepper. *Food Chem.* **390**, 133148. https://doi.org/10.1016/j.foodchem.2022.133148 (2022).
11. Roslan, N. F. M. & Yudin, A. S. M. Drying process of black pepper in a swirling fluidized bed dryer using experimental method. *IOP Conf. Series Mater. Sci. Eng.* **863**(1), 012047. https://doi.org/10.1088/1757-899x/863/1/012047 (2020).
12. Ali, M. K. M., Sulaiman, J., Md Yasir, S. & Ruslan, M. Cubic spline as a powerful tools for processing experimental drying rate data of seaweed using solar drier. *Article Malays. J. Math. Sci.* **11**, 159–172 (2017).
13. Ali, M. K. M., Fudholi, A., Muthuvalu, M., Sulaiman, J., & Yasir, S. M. Implications of drying temperature and humidity on the drying kinetics of seaweed. *Proceedings of the 13th IMT-GT International Conference on Mathematics, Statistics and their Applications (ICMSA2017)*. **1905**(1), 050004–1- 050004–7. https://doi.org/10.1063/1.5012223 (2017b).
14. Ali, M. K. M., Fudholi, A., Sulaiman, J., Muthuvalu, M. S., Ruslan, M. H., Yasir, S. Md., & Hurtado, A. Q. Post-harvest handling of eucheumatoid seaweeds. In Tropical seaweed farming trends, problems and opportunities. *Springer International Publishing*. 131–145. https://doi.org/10.1007/978-3-319-63498-2_8 (2017c).
15. Shreelavaniya, R., Pangayarselvi, R. & Kamaraj, S. Mathematical modeling of drying characteristics of black pepper (Piper nigrum) in indirect type solar-biomass hybrid dryer. *Int. J. Curr. Microbiol. Appl. Sci.* **6**(11), 2634–2644. https://doi.org/10.20546/ijcmas.2017.611.309 (2017).
16. Ibidoja, O. J., Shan, F. P., Sulaiman, J. & Ali, M. K. M. Detecting heterogeneity parameters and hybrid models for precision farming. *J. Big Data.* **10**, 130. https://doi.org/10.1186/s40537-023-00810-8 (2024).
17. Kumar, P. R., Ali, M. K. M. & Ibidoja, O. J. Identifying heterogeneity for increasing the prediction accuracy of machine learning models. *J. Niger. Soc. Phys. Sci.* https://doi.org/10.46481/jnsps.2024.2058 (2024).
18. Department of Statistics Malaysia, Malaysian Open Data Portal-Information on Black Pepper Industry https://www.data.gov.my/ (2019).
19. Nunes, A., Trappenberg, T. & Alda, M. The definition and measurement of heterogeneity. *Transl. Psychiatry.* https://doi.org/10.31234/osf.io/3hykf (2020).
20. Afouna, N. H. & Ali, M. K. Optimizing precision farming: Enhancing machine learning efficiency with robust regression techniques in high-dimensional data. *J. Niger. Soc. Phys. Sci.* https://doi.org/10.46481/jnsps.2025.2314 (2024).
21. Mikolajewicz, N. & Komarova, S. V. Meta-analytic methodology for basic research: A practical guide. *Front. Physiol.* https://doi.org/10.3389/fphys.2019.00203 (2019).
22. Ibidoja, O. J., Shan, F. P. & Ali, M. K. M. Modified sparse regression to solve heterogeneity and hybrid models for increasing the prediction accuracy of seaweed big data with outliers. *Sci. Rep.* https://doi.org/10.1038/s41598-024-60612-7 (2024).
23. Sundus, K., Hammo, B., Al-Zoubi, M. I. & Al-Omari, A. Solving the multicollinearity problem to improve the stability of machine learning algorithms applied to a fully annotated breast cancer dataset. *Inf. Med. Unlocked.* **33**, 101088. https://doi.org/10.1016/j.imu.2022.101088 (2022).
24. Ali, M., Bin, M. K., Ismail, M. & Fudholi, A. Accurate and hybrid regularization—Robust regression model in handling multicollinearity and outlier using 8SC for big data. *Math. Model. Eng. Probl.* **8**(4), 547–556. https://doi.org/10.18280/mmep.080407 (2021).
25. Daoud, J. I. Multicollinearity and regression analysis. *J. Phys.* **949**, 012009. https://doi.org/10.1088/1742-596/949/1/012009 (2017).
26. Chan, J. Y. et al. Mitigating the multicollinearity problem and its machine learning approach: A review. *Mathematics.* **10**(8), 1283. https://doi.org/10.3390/math10081283 (2022).
27. Lim, H. Y., Fam, P. S., Javaid, A. & Ali, M. Ridge regression as efficient model selection and forecasting of fish drying using V-groove hybrid solar drier. *Pertanika J. Sci. Technol.* https://doi.org/10.47836/pjst.28.4.04 (2020).
28. Ayadi, A., Ghorbel, O., Obeid, A. M. & Abid, M. Outlier detection approaches for wireless networks: A survey. *Comput. Netw.* **129**, 319–333. https://doi.org/10.1016/j.comnet.2017.10.007 (2017).
29. Perez, H. & Tah, J. H. M. Improving the accuracy of convolutional neural networksby identifying and removing outlier images in datasets using t-SNE. *Mathematics.* https://doi.org/10.3390/MATH8050662 (2020).
30. Wasim, D. et al. Quantile-based robust Kibria-Lukman estimator for linear regression model to combat multicollinearity and outliers: Real life applications using T20 cricket sports and anthropometric data. *Kuwait J. Sci.* **52**, 100336. https://doi.org/10.1016/j.kjs.2024.100336 (2024).
31. Wasim, D., Khan, S. A. & Suhail, M. Modified robust ridge M-estimators for linear regression models: An application to tobacco data. *J. Stat. Comput. Simul.* **93**(15), 2703–2724. https://doi.org/10.1080/00949655.2023.2202913 (2023).
32. Wasim, D., Suhail, M., Albalawi, O. & Shabbir, M. Weighted penalized m-estimators in robust ridge regression: An application to gasoline consumption data. *J. Stat. Comput. Simul.* **94**(15), 3427–3456. https://doi.org/10.1080/00949655.2024.2386391 (2024).
33. Wasim, D., Khan, S. A., Suhail, M. & Shabbir, M. New penalized M-estimators in robust ridge regression: Real life applications using sports and tobacco data. *Commun. Stat. Simul. Comput.* https://doi.org/10.1080/03610918.2023.2293648 (2023).
34. Joy, C. M., Pittappillil, G. P. & Jose, K. P. Drying of black pepper (Piper nigrum L.) using solar tunnel dryer. *Pertanika J. Trop. Agric. Sci.* **25**(1), 39–45 (2002).
35. Padmanaban, K., Mishra, P., Dubey, A. & Tiwari, P. Study of factors of production on productivity of black pepper and its sustainability. *Acta Sci. Agric.* **2**(12), 138–143 (2018).
36. Afouna, N. A. & Ali, M. K. M. The impact of heterogeneity in high-ranking variables using precision farming. *Malays. J. Fundam. Appl. Sci.* **20**(6), 1344–1362. https://doi.org/10.11113/mjfas.v20n6.3564 (2024).
37. Drobnič, F., Kos, A. & Pustišek, M. On the interpretability of machine learning models and experimental feature selection in case of multicollinear data. *Electronics (Switzerland)* **9**(5), 761. https://doi.org/10.3390/electronics9050761 (2020).
38. Ibidoja, O. J., Shan, F. P., Mukhtar, N., Sulaiman, J. & Ali, M. Robust M estimators and machine learning algorithms for improving the predictive accuracy of seaweed contaminated big data. *J. Niger. Soc. Phys. Sci.* https://doi.org/10.46481/jnsps.2023.1137 (2023).
39. Gujarati, D. N. & Porter, D. N. *Basic Econometrics* 4th edn. (The McGraw-Hill Companies, 2004).
40. Obadina, O. G., Adedotun, A. F. & Odusanya, O. A. Ridge estimation's effectiveness for multiple linear regression with multicol linearity: An investigation using Monte-Carlo simulations. *J. Niger. Soc. Phys. Sci.* **3**(4), 278–281. https://doi.org/10.46481/jnsps.2021.304 (2021).
41. Yildirim, H. & Revan Özkale, M. The performance of ELM based ridge regression via the regularization parameters. *Expert Syst. Appl.* **134**, 225–233. https://doi.org/10.1016/j.eswa.2019.05.039 (2019).
42. Moreno-Salinas, D., Moreno, R., Pereira, A., Aranda, J. & de la Cruz, J. M. Modelling of a surface marine vehicle with kernel ridge regression confidence machine. *Appl. Soft Comput. J.* **76**, 237–250. https://doi.org/10.1016/j.asoc.2018.12.002 (2019).
43. Melkumova, L. E. & Shatskikh, S. Y. Comparing Ridge and LASSO estimators for data analysis. *In Procedia Eng.* https://doi.org/10.1016/j.proeng.2017.09.615 (2017).
44. Jiehong, C., Sun, J., Yao, K., Min, X. & Yan, C. A variable selection method based on mutual information and variance inflation factor. *Spectrochimica Acta Part A: Mol. Biomol. Spectrosc.* **268**, 120652. https://doi.org/10.1016/j.saa.2021.120652 (2022).
45. Frost, J. When Do You Need to Standardize the Variables in a Regression Model? *Statistics by Jim*. https://statisticsbyjim.com/regression/standardize-variables-regression/ (2017).

46. Duzan, H. & Shariff, N. S. B. M. Ridge regression for solving the multicollinearity problem: Review of methods and models. *J. Appl. Sci.* **15**(3), 392–404. https://doi.org/10.3923/jas.2015.392.404 (2015).
47. Safi, S. K. et al. Optimizing linear regression models with lasso and ridge regression: A study on UAE financial behavior during COVID-19. *Migr. Lett.* **20**(6), 139–153. https://doi.org/10.59670/ml.v20i6.3468 (2023).
48. Tibshirani, R. Regression shrinkage and selection via the lasso. In *Source: J. Royal Stat. Soc. Series B (Methodological)*. **58**(1). (1996).
49. Tibshirani, R. Regression shrinkage and selection via the lasso: a retrospective. In *J. R. Statist. Soc. B*. **73**. (2011).
50. Enwere, K., Nduka, E. & Ogoke, U. Comparative analysis of ridge, bridge and lasso regression models in the presence of multicollinearity. *IPS Intelligentsia Multidiscip. J.* **3**(2), 1–8. https://doi.org/10.54117/iimj.v3i1.5 (2023).
51. Usman, M., Doguwa, S. & Alhaji, B. Comparing the prediction accuracy of ridge, lasso and elastic net regression models with linear regression using breast cancer data. *Bayero J. Pure Appl. Sci.* **14**(2), 134–149. https://doi.org/10.4314/bajopas.v14i2.16 (2022).
52. García-Nieto, P. J., García-Gonzalo, E. & Paredes-Sánchez, J. P. Prediction of the critical temperature of a superconductor by using the WOA/MARS, Ridge, Lasso and Elastic-net machine learning techniques. *Neural Comput. Appl.* **33**(24), 17131–17145. https://doi.org/10.1007/s00521-021-06304-z (2021).
53. Zou, H. The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.* **101**(476), 1418–1429. https://doi.org/10.1198/016214506000000735 (2006).
54. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. Royal Stat. Soc. Series B (Stat. Methodol.)* **67**(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x (2005).
55. Le, C. V. How to choose tuning parameters in lasso and ridge regression?. *Asian J. Econ. Banking* **4**(1), 61–76 (2020).
56. Ogutu, J. O., Schulz-Streeck, T. & Piepho, H. P. Genomic selection using regularized linear regression models: Ridge regression, lasso, elastic net and their extensions. *BMC Proc.* https://doi.org/10.1186/1753-6561-6-S2-S10 (2012).
57. Al-Jawarneh, A. S., Ismail, M. T. & Awajan, A. M. Elastic net regression and empirical mode decomposition for enhancing the accuracy of the model selection. *Int. J. Math. Eng. Manag. Sci.* **6**(2), 564–583. https://doi.org/10.33889/ijmems.2021.6.2.034 (2021).
58. Schreiber-Gregory, D. N. Regulation Techniques for Multicollinearity: Lasso, Ridge, and Elastic Nets. In *Proceedings of the SAS Conference Proceedings: Western Users of SASSoftware*. 1–23. https://api.semanticscholar.org/CorpusID: 189925961. (2018).
59. Susanti, Y., Pratiwi, H. & H, S. S., & Liana, T.,. M estimation, S estimation, and MM estimation in robust regression. *Int. J. Pure Appl. Math.* https://doi.org/10.12732/ijpam.v91i3.7 (2014).
60. Almetwally E.M., & Almongy H.M. Comparison between M-estimation, S-estimation, and MM estimation methods of robust estimation with application and simulation. *Int. J. Math. Arch.* **9**(11). (2018).
61. Singgih, M. N. A. & Fauzan, A. Comparison of M estimation, S estimation, with MM estimation to get the best estimation of robust regression in criminal cases in Indonesia. *Jurnal Matematika Statistika Dan Komputasi.* **18**(2), 251–260. https://doi.org/10.20956/j.v18i2.18630 (2022).
62. Hodson, T. O., Over, T. M. & Foks, S. S. Mean squared error, deconstructed. *J. Adv. Model. Earth Syst.* https://doi.org/10.1029/2021ms002681 (2021).
63. Kim, S. & Kim, H. A new metric of absolute percentage error for intermittent demand forecasts. *Int. J. Forecast.* **32**(3), 669–679. https://doi.org/10.1016/j.ijforecast.2015.12.003 (2016).
64. De Myttenaere, A., Golden, B., Grand, B. L. & Rossi, F. Mean absolute percentage error for regression models. *Neurocomputing* **192**, 38–48. https://doi.org/10.1016/j.neucom.2015.12.114 (2016).
65. Moreno, J., Pol, A. L. P., García-Labiano, F. & Blasco, B. C. Using the R MAPE index as a resistant measure of forecast accuracy. *PubMed.* **25**(4), 500–506. https://doi.org/10.7334/psicothema2013.23 (2013).
66. Ijomah, M. A. On the misconception of R2 for (r)2 in a regression model. *Int. J. Res. Sci. Innovation.* **6**(12), 2321–2705 (2019).
67. Arsad, Z. *Multiple Linear Regression* 10–31 (Regression Analysis, 2023).
68. Abdullah, N., Kiu, A. L. L., & Lintangah, W. Log production prediction model: A comparison between Malaysia and Indonesia using multiple regression technique. *UMS Institutional Repository (Universiti Malaysia Sabah).* http://eprints.ums.edu.my/14146/7/LOG%20PRODUCTION%20PREDICTION%20MODEL%20A%20COMPARISON%20BETWEEN%20MALAYSIA%20AND.pdf (2016).
69. Abdullah, N., Lee, C. L. & Jubok, Z. H. Factors on palm oil fruit bunches production volume for biomass fuel and biofuel during cogeneration processes. *J. Jpn Inst. Energy* **94**(12), 1428–1439. https://doi.org/10.3775/jie.94.1428 (2015).
70. Hajijubok, Z. & Gopal, P. K. Procedure in getting best model using multiple regression. *J. Borneo Sci.* **23**, 47–63 (2008).
71. Akaike, H. Fitting autoregressive models for prediction. *Ann. Inst. Stat. Math.* **21**(1), 243–247. https://doi.org/10.1007/bf02532251 (1969).
72. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control* **19**(6), 716–723. https://doi.org/10.1109/tac.1974.1100705 (1974).
73. Golub, G. H., Heath, M. & Wahba, G. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21**(2), 215–223. https://doi.org/10.1080/00401706.1979.10489751 (1979).
74. Hannan, E. J. & Quinn, B. G. The determination of the order of an autoregression. *J. Royal Stat. Soc. Series B (Stat. Methodol.)* **41**(2), 190–195. https://doi.org/10.1111/j.2517-6161.1979.tb01072.x (1979).
75. Rice, J. Bandwidth choice for nonparametric regression. *Ann. Stat.* https://doi.org/10.1214/aos/1176346788 (1984).
76. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **6**(2), 15–18. https://doi.org/10.1214/aos/1176344136 (1978).
77. Ramanatam, R. *Introductory Econometrics with Application* 5th edn. (Harcourt College Publishers, 2002).
78. Shibata, R. An optimal selection of regression variables. *Biometrika* **68**(1), 45–54. https://doi.org/10.1093/biomet/68.1.45 (1981).
79. Morales, C. S., Giraldo, R. & Torres, M. E. Boxplot fences in proficiency testing. *Accred. Qual. Assur.* **26**, 193–200. https://doi.org/10.1007/s00769-021-01474-8 (2021).
80. Eberhard, K. The effects of visualization on judgment and decision-making: A systematic literature review. *Manag. Rev. Q.* **73**(1), 167–214. https://doi.org/10.1007/s11301-021-00235-8 (2021).
81. Almeida, F., Faria, D. & Queirós, A. Strengths and limitations of qualitative and quantitative research methods. *Eur. J. Educ. Stud.* **3**(9), 369–387. https://doi.org/10.5281/zenodo.887089 (2017).
82. Lin, L., Chu, H. & Hodges, J. S. Alternative measures of between-study heterogeneity in meta-analysis: Reducing the impact of outlying studies. *Biometrics* **73**(1), 156–166 (2017).
83. Drobnič, F., Kos, A. & Pustišek, M. On the interpretability of machine learning models and experimental feature selection in case of multicollinear data. *Electronics* **9**(5), 761. https://doi.org/10.3390/electronics9050761 (2020).

Author contributions

Paavithashnee Ravi Kumar: Conceptualization; Data curation; Formal analysis; Methodology; Project Writing—original draft; and Writing. Olayemi Joshua Ibidoja: Supervision; Writing—review & editing. Majid Khan Majahar Ali: Data curation, Writing—review & editing, Supervision. Wan Rosli Wan Ishak: Supervision, Writing—review & editing.

Funding

The authors are grateful for the financial assistance from the "Ministry of Higher Education Malaysia under the Fundamental Research Grant Scheme (FRGS), with Project Code: FRGS/1/2023/STG06/USM/02/6".

Declarations

Competing interests
The authors declare no competing interests.

Ethics approval
Ethics approval was not obtained as humans/ animals were not used in the study.

Additional information
Correspondence and requests for materials should be addressed to M.K.M.A. or W.R.W.I.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.