MODIFICATION OF REGRESSION MODELS TO SOLVE HETEROGENEITY PROBLEM USING SEAWEED DRYING DATA

IBIDOJA OLAYEMI JOSHUA

UNIVERSITI SAINS MALAYSIA

2023

MODIFICATION OF REGRESSION MODELS TO SOLVE HETEROGENEITY PROBLEM USING SEAWEED DRYING DATA

by

IBIDOJA OLAYEMI JOSHUA

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy

September 2023

ACKNOWLEDGEMENT

I sincerely appreciate the supreme God for His grace, sustenance, strength, and, above all, His faithfulness, provision, and love from the beginning of my academic life up to this doctoral stage. I would also like to express my sincere gratitude to my supervisor, Dr Fam Pei Shan and my co-supervisor, Dr Majid Khan bin Majahar Ali, for their invaluable advice, continuous support, mentoring and patience during my PhD study. Their huge knowledge and abundant experience have inspired me during my academic research. I sincerely appreciate and acknowledge the financial support from the Tertiary Education Trust Fund (TETFund) with the award number TETF/ES/UNIV/ZAMFARA/TSAS/2019 to study at the Universiti Sains Malaysia. I am also indebted to my employer Federal University Gusau, Nigeria for their support and study leave. I would like to express my gratitude to my beautiful and caring wife Victoria, children Tope and Seun, and parents. Without their tremendous understanding, sacrifice, support, prayers, and care in the past few years, it would have been difficult for me to complete my research. Finally, I appreciate my colleagues and friends for their support.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

GPS	Global Positioning System
IoT	Internet of Thing
LASSO	Least Absolute Shrinkage and Selection Operator
MAPE	Mean absolute percentage error
MSE	Mean square error
PF	Precision Farming
R^2	Coefficient of Determination
SFTs	Smart Farming Technologies
SSE	Sum of square error
v-GHSD	v- Groove Hybrid Solar Drier
VIF	Variance Inflation Factor

LIST OF APPENDICES

PENGUBAHSUAIAN MODEL REGRESI UNTUK MENYELESAIKAN MASALAH HETEROGEN MENGGUNAKAN DATA PENGERINGAN RUMPAI LAUT

ABSTRAK

Semasa proses pengeringan rumpai laut, banyak parameter pengeringan terlibat. Salah satu masalah dalam analisis regresi ialah kesan parameter heterogen. Data rumpai laut dikumpul dengan menggunakan teknologi pertanian pintar sensor yang dipasang pada Pengering Solar Hibrid v-Groove. Kaedah yang dicadangkan menggunakan faktor inflasi varians untuk mengenal pasti parameter heterogen. Untuk menentukan 15, 25, 35, dan 45 parameter penting berpangkat tinggi untuk rumpai laut, model seperti rabung, hutan rawak, mesin vektor sokongan, pembungkusan, penggalak, LASSO, dan jaring elastik digunakan sebelum heterogen, selepas heterogen, dan untuk model yang diubah suai. Untuk mengurangkan pencilan, regresi teguh seperti M Huber, M Hampel, Kuasa Dua M Bi, MM dan penganggar S digunakan. Sebelum parameter heterogen dikecualikan daripada model, model hibrid rabung dengan penganggar M Hampel menunjukkan bahawa keputusan signifikan yang lebih baik diperolehi dengan pencilan 2.14%. Selepas parameter heterogen dikecualikan daripada model, mesin vektor sokongan dengan penganggar MM menunjukkan bahawa keputusan signifikan yang lebih baik diperolehi dengan pencilan 2.09%. Bagi model yang diubah suai, LASSO dengan penganggar kuasa dua M Bi menunjukkan keputusan signifikan yang lebih baik diperolehi dengan pencilan 1.31%. Untuk kajian masa depan, kesan heterogen menggunakan model hibrid dengan data tidak seimbang atau nilai yang hilang boleh diselidiki. Algoritma pembelajaran mesin ensemble seperti stacking, XGBoost dan AdaBoost boleh digunakan.

MODIFICATION OF REGRESSION MODELS TO SOLVE HETEROGENEITY PROBLEM USING SEAWEED DRYING DATA

ABSTRACT

During the seaweed's drying process, a lot of drying parameters are involved. One of the problems in regression analysis is the impact of heterogeneity parameters. The seaweed data was collected using sensor-smart farming technology attached to the v-Groove Hybrid Solar Drier. The proposed method used the variance inflation factor to identify the heterogeneity parameters. To determine the 15, 25, 35, and 45 high-ranking important parameters for the seaweed, models such as ridge, random forest, support vector machine, bagging, boosting, LASSO, and elastic net are used before heterogeneity, after heterogeneity, and for the modified model. To reduce the outliers, robust regressions such as M Huber, M Hampel, M Bi Square, MM, and S estimators are used. Before the heterogeneity parameters were excluded from the model, the hybrid model of the ridge with the M Hampel estimator showed that better significant results were obtained with 2.14% outliers. After the heterogeneity parameters were excluded from the model, the support vector machine with the MM estimator showed that better significant results were obtained with 2.09% outliers. For the modified model, LASSO with M Bi square estimator showed that better significant results were obtained with 1.31% outliers. For future studies, the impact of heterogeneity using a hybrid model with imbalanced data or missing values can be investigated. Ensemble machine learning algorithms such as stacking, XGBoost, and AdaBoost can be used.

CHAPTER 1

INTRODUCTION

1.1 Background of the Study

The linear regression, which is the most usual regression analysis assumes that there is a linear relationship between an independent and a dependent variable (Ambrosius, 2007; Boldina & Beninger, 2016; Lee, 2022). Regression is used in many fields, including computer science, economics, and social sciences (Ali et al., 2020; Ambrosius, 2007; Boldina & Beninger, 2016; Flores-Sosa et al., 2022; Laanaya et al., 2017; Mata, 2011; Oukawa et al., 2022; Uyanık & Güler, 2013; Wang et al., 2020; Xie et al., 2021). Regression is also used in precision farming (Akhter & Sofi, 2022; Groher et al., 2020; Mancipe-Castro & Gutiérrez-Carvajal, 2022; Segarra et al., 2022; Sugirbay et al., 2020).

Precision farming (PF) is part of smart farming technologies (SFTs) that deal with information systems, farm management, internet of things (IoT), cloud computing, precision agriculture systems, artificial intelligence, robotics, and automation of agriculture and wireless sensor networks (Balafoutis et al., 2020; Klerkx et al., 2019; Montalcini et al., 2022; Moysiadis et al., 2021; Neethirajan, 2020; Rose & Chilvers, 2018; Vecchio et al., 2022). The merit of the approach is that it increases farm profits and reduces the cost of production (Sharma et al., 2021). The conventional methods used by agriculturalists are not precise, which leads to physical labour and consumes a lot of time (Durai & Shamili, 2022).

The v- Groove Hybrid Solar Drier (v-GHSD) simulation diagram in Figure. 1.1 was used as the smart farming technology to dry the seaweed. The sensors are

positioned to capture the data for the parameters. There are many parameters involved in the seaweed drying and the data are sent to the cloud at a faster rate.

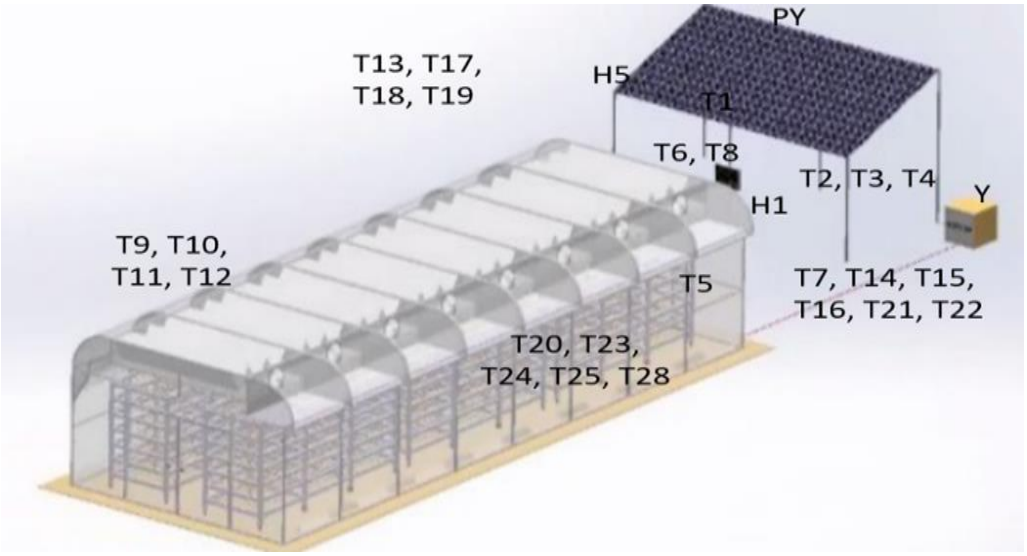

Figure 1.1 v-GHSD simulation diagram (Ali, et al., 2017)

Big data refers to the information asset characterized by velocity, volume, variability, veracity, value and variety that requires specialized technology and analytical methods for transformation into value (De Mauro et al., 2016; Ur Rehman et al., 2016). Social networks, digital behaviour and health data analytics make use of big data (Yousef, 2021). Furthermore, big data is used in many fields, like agriculture, biology, medicine, economic modelling, meta-analysis, stem review, ultra-dimension, etc (Bautista Villalpando et al., 2014a; Belias et al., 2021; Hassani et al., 2020; Hung et al., 2020; Li et al., 2021; Sagiroglu & Sinanc, 2013; Shi & Abdel-Aty, 2015a; Srivastava & Gopalkrishnan, 2015; Vogel et al., 2019; Yang et al., 2020). Hence, big data analysis has gained popularity in a variety of fields, including geoscience, disaster management, health science and business (Arinta & Andi, 2019; Bautista Villalpando et al., 2014b; Lakerveld et al., 2020; Martínez–Álvarez & Morales–Esteban, 2019; Rahman et al., n.d.; Shi & Abdel-Aty, 2015b; Tamiminia et al., 2020). In the field of agriculture, big data has the potential to broaden and deepen tools for evaluating farm-

level decisions, as well as the ability to assess the impact of policy interventions on the agricultural sector (Coble et al., 2018; Lioutas & Charatsari, 2020; Neethirajan, 2020). Although data-centric technology has not been widely adopted in production agriculture, it could solve the problems of insufficient food and food insecurity and improve farm profitability (White et al., 2021). However, to solve the problems of climate change and food shortage, big data are now used in agriculture which brings in many opportunities and reduce cost (Sadiku et al., 2020).

Hybrid models will be developed to combine the machine learning algorithms and robust statistical methods. Hybrid model was used to combine user-product relations and user features for spammer detection (Wu et al., 2016)**.** Hybrid model was used to achieve a recommendation based on review texts, ratings, and social networks (Ji et al., 2019). Hybrid mathematical programming approach and machine learning was used to predict the accuracy of the emission of greenhouse gases (Javanmard & Ghaderi, 2022). Hybrid Leader-based Optimization based on DL-driven Weed Detection using IoT-enabled Smart Agriculture (HLBODL-WDSA) model was used to recognize the presence of the weeds (Alrowais et al., 2022). A hybrid model using deep learning techniques for the recognition and classification of sunflower diseases was used (Malik et al., 2022).

1.2 Problem Statement

Drying seaweed using a solar drier with smart monitoring systems (IoT) will involve a lot of parameters and their interactions which leads to big data complexity during data recording in a cloud database. The drying parameters in the v-Groove Hybrid Solar Drier (v-GHSD) are shown in Figure 1.1 was built, which comprises sections of the solar collector, drying chamber, forced fans, and outlet flushing system

components. A total of 29 sensors were positioned in the drier to collect the data for the drying parameters. The data is characterised by volume, velocity, variety, varaiability and value , whichs makes it a big data (Yaseen & Obaid, 2020). Because of the large number of sensors, the selection of the important drying parameters is important. It is very rare to find a study where $p > n$ (number of parameters is greater than the number of observations).

In aquaculture, post-harvest monitoring systems are crucial for production sustainability (Stedt et al., 2022). However, there are few issues in post-harvest management with smart monitoring systems. One of the main issues is heterogeneity. It remains a crucial problem for machine analytics because it is one of the characteristics of big data and heterogeneity is a problem in big data analytics and data integration (Ahsaan et al., 2022). The different parameters, difference in units value for the temperature, relative humidity, wind, solar radiation, and variability in variances cause heterogeneity (Chaney et al., 2018). Heterogeneity is the degree of variability within the data (Fitch et al., 2015). It is the degree to which a system differs from conformity or from an ideal state. Heterogeneity makes it difficult for classical learning algorithms to handle big data (Somwya & Suneetha, 2017). Numerous sophisticated, effective, and intelligent learning algorithms are needed to handle the enormous heterogeneous data (Al Nuaimi et al., 2015). The observed and unobserved heterogeneities can cause a bias in the efficiency of the results, impact of the errors before and after the model (Gormley & Matsa, 2014). Heterogeneity will help to comprehend the dynamics of the large number of the drying parameters of the variability and offers a way to leverage the data for efficient predictive modelling (Caiado et al., 2016; Drnevich & Kriauciunas, 2011; Nair-Reichert & Weinhold, 2001).

To determine the significant drying parameters, 15, 25, 35 and 45 high ranking parameters will be selected from the drying parameters. In features selection, only the ranks of important variables are provided and not only the number of the significant factors (Drobnič et al., 2020). Similarly, there is no rule to choose the number of parameters to be incorporated in a prediction model (Chowdhury & Turin, 2020). Additionally, the algorithms can only tell the ranks and not the number of significant parameters (Kaneko, 2021). All the possible models are computed up to the second order interaction. The interaction variables effects need to be considered, because interaction helps to understand the relationships among the variables available in the model, and more hypotheses can be tested (Whisman & McClelland, 2005). Although it is challenging to study the asymptotic and statistical inference of second order because of their complex covariance structure (Hao & Zhang, 2017). Figure 1.1 shows more information about the drying parameters. It has 29 independent variables and one dependent variable. The data has main effects of 29 variables with interaction effects of 406 variables, one independent variable Y. Which means there is a total of 435 independent variable that determine the moisture content Y.

Apart from the problem of heterogeneity in agriculture, other problem include multicollinearity. Multicollinearity occurs because of the high number of the main and interactions parameters. This may result in overfitting, where noise or pointless variables may have an improper influence on the model's predictions (Peralta et al., 2015). In addition, Omara et al. (2018) discovered that a major issue in many empirical analyses is the inclusion of a group of variables that are not significantly contributing to describe the phenomenon under study. The existing machine learning models will be used to select the high ranking variables and solve for multicollinearity and measures of accuracies will be used.

Another problem is outliers. Observations that vary from the distribution's common pattern or shape are called outliers (Ayadi et al., 2017; Yusuf et al., 2021). Data have outliers since the factors that cannot be controlled or regulated, and the outliers will add to the standard errors (Lim et al., 2020; Rajarathinam & Vinoth, 2014). In statistical analysis, the occurrence of outliers in the data can greatly affect the estimation of the sample's mean and standard deviation, which can result in either over- or under-estimated values. This is a straightforward example of how unwanted outliers can affect the results of data analysis (Perez & Tah, 2020).

1.3 Objectives of the Study

The objectives of the study are:

i. To identify significant parameters that directly impact heterogeneity.

ii. To compare the (SSE, MSE and MAPE) for significant parameters selection using 15, 25, 35, and 45 highest important variables using 7 existing machine learning models (ridge, random forest, bagging, boosting, elastic net, LASSO, and support vector machine).

iii. To ascertain the impacts of heterogeneity before, after and modified heterogeneity through metric validation from machine learning methods such as (ridge, random forest, bagging, boosting, elastic net, LASSO, and support vector machine) to reduce the multicollinearity.

iv. To develop a hybrid of machine learning and robust techniques to reduce outliers.

1.4 Scope and Limitation

This thesis concentrates on heterogeneity using big data in the field of agriculture, specifically seaweed drying big data. The data are collected by the sensors for the different drying parameters. During the process, the data is sent to the cloud platform. Hence, primary data is used in this study.

For accurate prediction and a reliable estimate, heterogeneity is very important in big data. To achieve the purpose, the heterogeneity parameters are identified among the 29 independent variables that determine the moisture content removal of the seaweed. For efficiency, optimization, model improvement, resources conservation and accuracy improvement, all possible models of the 15, 25, 35 and 45 highest important variables are chosen for the modelling. Furthermore, the drying parameters that exhibit heterogeneity will be excluded from the model to determine the impacts of the heterogeneity in the model, because it can lead to wrong results and scientific conclusions.

For the identification of the heterogeneity parameters, this study proposes a method using the variance inflation factor to identify the heterogeneity parameters with some constraints. For the selection of the 15, 25, 35 and 45 high ranking variables that determine the moisture content removal, the methods proposed are ridge, random forest, support vector machine, bagging, boosting, elastic net and LASSO. Data in agriculture has outliers because of factors that cannot be controlled, and these outliers will raise the standard errors (Lim et al., 2020; Rajarathinam & Vinoth, 2014). The presence of outliers is a problem in data analysis, because it affects the performance of OLS, and a robust regression is proposed (Jegede et al., 2022). In this study, the impact of heterogeneity is compared before and after using metric validations. The

modified heterogeneity is also compared with them. Robust estimators such as M Bi-Square, M Hampel, M Huber, MM and S are proposed to develop the hybrid models to reduce the outliers, by incorporating sigma limits into the model to improve the prediction accuracy.

This study has some limitations, for example, the sensors determine the variables to be captured, and some variables were not captured due to measurement errors. This study is also limited to the main effects of the drying parameters and second order interaction, with this, there are 435 independent variables. Due to the time, feasibility and complexity of the models, it is restricted to the main effects and second order interactions. The machine learning algorithms proposed cannot tell the number of significant variables to use, because features selection can only provide the rank of important variables (Drobnič et al., 2020). Therefore, the 15, 25, 35 and 45 selection of high ranking variables are proposed to determine the moisture content removal. The data will be processed and analysed with Microsoft Excel 365 and RStudio version 2022.7.1.554.

1.5 Significance of the study

The identification of significant parameters that impact heterogeneity will improve the qaunatity and quality of seaweed. Globally, the demand for food is increasing every day. The United Nations world population index predicts that by 2050, there will be 9.7 billion people on the planet (Namana et al., 2022). Rahimi et al. (2022) stated that the effect of COVID-19 affected treatment for animals, the lockdown affected the production of food and the supply of labour. In Malaysia, the effects of Covid-19 and climatic anomalies have caused interruptions in the food supply and prices have skyrocketed.

By comparing the significant parameters selection using ridge, random forest, bagging, boosting, elastic net, LASSO, and support vector machine. The best predictive model to determine the moisture content of the seaweed will be determined. One of the problems with seaweed extraction during drying is the high moisture content (Liu et al., 2022).

To ascertain the impacts of heterogeneity before, after and modified heterogeneity through metric validation from machine learning methods such to reduce the multicollinearity. The effects of the irrelevant variables and impact of the relevant variables will be determined for the seaweed drying.

To develop a hybrid of machine learning and robust techniques to reduce outliers. The hybrid model will be used to develop a better predictive model for the seaweed drying. The scientists, farmers, agronomists, researchers, companies, and people who consume seaweed will benefit from this research. Finally, the solution to heterogeneity will improve the food production to support the Sustainable Development Goal 2 (SDG 2).

1.6 Thesis Framework

The thesis is organised as follows: Chapter 1 covers the background to the problem, the problem statement, the objectives of the study, the scope and limitations, and the significance of the study. Chapter 2 discusses seaweed and seaweed drying, the literature review on drying parameters for food, heterogeneity, multicollinearity, outliers, sparse regression, robust regression, and machine learning. Chapter 3 presents the flowchart of the research, data collection, and models used for the study. Chapter 4 is the results and discussion; it provides the results of the objectives accordingly.

Chapter 5 contains the conclusions, contribution of the study, limitations of the study, and future work.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

In the post-harvest monitoring systems, the drying parameters are crucial for production sustainability. However, there are few issues in post-harvest management with smart monitoring systems. One of the main issues is heterogeneity. It remains a crucial problem for machine analytics. The issue of food drying has been stated in Section 2.2. The literature review on heterogeneity has been stated in Section 2.3. Section 2.4 illustrates the literature review on multicollinearity. Section 2.5 reviews the literature on outliers. Section 2.6 illustrates the literature review on robust regression. Section 2.7 illustrates the literature review on sparse regression. Section 2.8 presents a literature review on machine learning for dealing with big data. Finally, in Section 2.9, the summary of the literature review is presented.

2.2 Introduction to Seaweed and Seaweed Drying

Seaweeds are macroalgae that are an essential part of the marine and coastal ecosystem, enhancing both their rich biodiversity and the biosphere as a whole (Ali et al., 2021). Seaweed can be used as food, fertilizer, cosmetics, biofuel, and medicine (Ali, et al., 2017; Ali et al., 2020; Echave et al., 2022; Jesumani et al., 2019; Lomartire et al., 2021; Pati et al., 2016; Pradhan et al., 2022; Rammou et al., 2021). The global seaweed industry provides a diversity of products directly or indirectly for human consumption, with a total value of approximately US$ 10 billion a year (Bixler & Porse, 2011). However, main problems with seaweed extraction during drying is the high moisture content (Liu et al., 2022).

One of the post-harvest problems with seaweed is its high moisture content. According to Pradana et al. (2019), seaweed is easily damaged when it is very fresh. Therefore, this demands that seaweed be dried after harvesting. Drying is used to conserve food, and the aim is to obtain food of high quality (Shishir & Chen, 2017). The drying will prevent the seaweed from being rotten and moldy (Huang et al., 2021). The drying of seaweed is used to reduce the moisture content (M. K. M. Ali, Sulaiman, et al., 2017). The biomass weight of seaweed during transportation will be decreased, which makes it available for additional processing (Oirschot et al., 2017). Drying is vital to prevent food global shortages. The reproduction of microbes' multiplication in the food can be reduced by removing the moisture level. Drying also reduces storage, transportation, and processes to prevent losses and increase value (Xiao & Mujumdar, 2020). The types of drying are freeze-drying (direct drying method), conventional drying and microwave - assisted drying (solar). A solar drier is the most efficient drying method for seaweed and can dry the water content faster (Suherman et al., 2018). These authors (Ali et al., 2017; Hao et al., 2020; Lakshmi et al., 2018; Nabnean & Nimnuan, 2020; Pankaew et al., 2020; Vijayan et al., 2020) have employed solar driers in their studies. The drying parameters using the v-GHSD were monitored effectively. Furthermore, the IoT based solar drying system using the v-GHSD was more effective in monitoring the drying behaviour (Ali et al., 2017). All the parameters involved in solar drying should be studied to reduce the moisture content of seaweed, improve food quality and quantity.

Table 2.1 shows the assessment of drying parameters measured for the existing dryers by other scholars. The former studies in Table 2.1 measured the period of the drying time, temperature, solar radiation, and relative humidity.

Table 2.1 Study about drying parameters for food

Author(s) Year(s)	Drying Factors							Food	Types of drying methods
	Solar Radiation	Temperature	Relative Humidity	Air Velocity	Mass	Time	Others		
Jafari et al., (2021)	-	-	-	√	-	-	Powder nanoparticles	Vitamins and Minerals	Conventional spray
Zhang et al. (2021)	-	√	√	√	-	√	Imaging techniques	Plant	Terahertz Spectroscopy
Tümay and Ünver (2021)	√	√	√	√	√	√	Drying kinetics	Strawberries, carrot, and Pear	Smart and automatic oven
Javaid et al. (2020b)	√	√	-	-	-	√	Collector Efficiency	Seaweed	Indirect type forced convection
Javaid et al. (2019)	√	√				√	Collector Efficiency	Seaweed	Indirect type forced convection
Essalhi et al. (2018)	√	√	√	√	-	√	-	Grapes	Indirect type forced convection

Table 2.1 (Continued)

Author(s) Year(s)	Drying Factors							Food	Types of drying methods
	Solar Radiation	Temperature	Relative Humidity	Air Velocity	Mass	Time	Others		
Wang et al. (2018)	√	√	√	√	√	√	-	Mango	Indirect type forced convection
Castillo-Téllez et al. (2017)	√	√	√	√	√	√	Wind direction	Red Chili	Indirect type forced convection
Fudholi et al. (2016)	√	√	√	√	-	√	-	Silver jewfish	Indirect type forced convection
Khama et al. (2016)	√	√	√	√	√	√	-	Tomatoes	Indirect type forced convection
Fudholi et al. (2015)	√	√	√	√	√	√	-	Palm oil fronds	Indirect type forced convection
El-Sebaii and Shalaby (2013)	√	√	√	√	√	√	Pressure	Thymus and mint	Indirect type forced convection

Table 2.1 (Continued)

Author(s) Year(s)	Drying Factors							Food	Types of drying methods
	Solar Radiation	Temperature	Relative Humidity	Air Velocity	Mass	Time	Others		
Gupta et al. (2012)	√	√	√	√	√	√	-	Tomatoes	Indirect type natural and forced convection
Chandrasekar and Mohanraj, (2009)	√	√	√	√	√	√	Energy Consumption	Chili	Indirect type forced convection
Akpinar and Bicer (2008)	√	√	√	√	√	√	-	Long green pepper	Indirect type forced convection

2.3 Heterogeneity

Heterogeneity is a problem among the drying parameters. Heterogeneity is the degree of variability within the data. It is the degree to which a system differs from conformity or from an ideal state. Heterogeneity causes the standard errors to be biased and inconsistent.

This study requires a deep knowledge of the work that has been done on heterogeneity. Feczko and Fair (2020) stated that the problem of homogeneity assumption is a big challenge and leads to the heterogeneity problem. In medicine, the homogeneity assumption is difficult because disorders are not identical discrete entities. Researchers continue to treat cohorts as homogeneous discrete entities.

Feczko et al. (2019) stated that the heterogeneity problems were investigated in human cognition and mental health and emphasize that the question or outcome should be linked to that subtype identification. In the treatment of mental health illnesses, the imprecise nature of psychiatric health nosology restricts how to distinguish and treat the mental disorders. The supervised and unsupervised statistical approaches were emphasized to study heterogeneity problem for researchers trying to understand biological and physiological correlates of mental health and cognition.

Assunção et al. (2015) stated that the issue of heterogeneity is usually considered in regression analysis, the issue of missing data affects the ability to control the significant variables in the estimation of the regression coefficients, which affects the stability of the inference. A semiparametric framework using a group of proxy variables to control for unobserved effects and heterogeneity was proposed. The proposed method uses a series (sieve) expansion of possibly nonlinear terms and is

unknown. In the sieve expansion method, multiple explanatory variables can be handled and has better approximation capabilities when compared to kernel methods.

Gormley and Matsa (2014) claimed that a vital problem in financial empirical research is the issue of how and whether to control unobserved heterogeneity. A lot of variables cannot be observed by the econometrician, and corporate policies are affected by the unobserved factors. In a situation where the factors are associated with the variables of interest, if heterogeneity is not considered and solved, some variables will be omitted from the model and will affect the unknown parameters and prevent causal inference. According to the literature, there are many estimation methods used in empirical finance to control unobserved group heterogeneity. The adjusted Y is an estimation method that is used to demean the response variable, to the other group before estimating the model with ordinary least squares (OLS).

Lewbel (2005) considered heterogeneity from a statistical and structural perspective. The heterogeneity here means unobserved heterogeneity which is the behavioural variation in the dependent variable that is not explained by the variation in the independent variables. The statistical heterogeneity means the unobserved heterogeneity is the conditional distribution of the dependent variable given the independent variables. It was stated that it can be estimated nonparametrically. For the structural model, another parameter that has unobservable heterogeneity. The structural model needs to be estimated since it relies on the economic model and affects the policy of the economy.

This study is different from other studies because it proposes a method to identify the heterogeneity parameters, compare the impacts of the heterogeneity before

removing the heterogeneity parameters, after removing the heterogeneity and modified model.

Table 2.2 shows a summary of the literature found in different fields related to the problem of heterogeneity.

Table 2.2 Study on Heterogeneity

Author (Year)	Type of Paper		Application						Remarks
	Review	Research	Agriculture	Big Data	Economic Model	Machine Learning	Biomedical	Statistics	
Khoda and Ahmed(2021)		√						√	$Y_{ij} = \beta_{00} + \beta X + u_{oj} + e_{ij}$
Renner et al., (2021)		√	√						$lny_{it} = lnf(x_{it})\|_j + v_{ut}\|_j - u_{it}\|_j$ The vertical bar is to show the different class models each latent class.
Marenya et al., (2020)		√	√						$Y_{ij} = Z_{ij}\alpha_j + Z_{ij}\theta_j + v_{ij}$
Mikolajewicz and Komarova (2019)		√					√		$y_i = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_n x_n + a_i + +\varepsilon_i$
Wossen et al., (2019)		√			√				V_{pi}=f($F_{pi}, Q_{pi}, Z_{pi}, w_{pi}, \gamma$) $w_{pi} = e_{pi} + h_i$
Ryan, (2016)	√						√		Systematic reviews -Data synthesis -I square, Chi-square -Forest Plot

Table 2.2 (Continued)

Author (Year)	Type of Paper		Application						Remarks
	Review	Research	Agriculture	Big Data	Economic Model	Machine Learning	Biomedical	Statistics	
Assunção et al. (2015)		√			√	√		√	y_i=$\eta_0(z_i)$+$u_i + \beta_0^{\cdot} x_i + e_i$ -Extremum estimates Model selection
Krueger et al. (2015)		√			√				Household heterogeneity $Y = ZK^{\alpha}N^{1-\alpha}$
Gormley and Matsa (2014)		√			√				Adjusted Y and Average Effect $Q_{i,j,t} = \alpha + \beta' x_{i,j,t} + e_{i,j,t}$ $y_{i,j} = \beta x_{i,j}$+f_i+$e_{i,j}$
Juutinen et al., (2012)		√			√				$U_{ij} = j + \beta_i' x_{ij}$ Where $j = \beta$+Φv_i
Kopsakangas -Savolainen and Svento (2011)		√			√				$c_{it} = (\alpha + w_i) + \beta' x_{it}$+$v_{it}$+$u_{it}$ -Stochastic frontier analysis -Randomizing frontier parameters -Cobb-Douglass specifications

Table 2.2 (Continued)

Author (Year)	Type of Paper		Application						Remarks
	Review	Research	Agriculture	Big Data	Economic Model	Machine Learning	Biomedical	Statistics	
Abdulai and Tietje (2007)		√	√		√				$y_{it} = x'_{it}\beta + \alpha' x_i + \overline{z}_i + e_{it}$ -Panel data
Fletcher (2007)	√						√		-Forest plot , Methodological, Statistical, -Meta analysis
Lewbel (2005)		√			√				-Structural heterogeneity $y = g(X, \Theta, U)$ -Statistical heterogeneity $F(Y/X)$ -Theoretical Point

2.4 Multicollinearity

Heterogeneity is the degree of variability within the data. The concept of multicollinearity can be used to identify the varaiables that exhibit heterogeneity. Multicollinearity means the existence of a linear relationship among two or more variables. It is a problem among the drying parameters and can cause serious problems with the estimate of the parameters. It can also cause redundancy in information about the dependent variable.

The goal of this section is to explain the phenomenon of multicollinearity. Multicollinearity is a challenge in regression. Regression is an important component of machine and statistical learning, and has been used in many fields, including the field of agriculture.

One of the methods employed in supervised machine learning is regression. It is used to develop a model and evaluate its efficiency centered on the connection between a number of variables (Xiao et al., 2015). Regression is used to achieve a prediction function that forecasts an independent variable, Y, using the dependent variables X_1,…, X_n. The relationship between the observed and the response variable can be estimated by regression (Lim et al., 2020).

One of the assumptions regarding the regression model is the issue of multicollinearity (Chatterjee & Hadi, 2012; Ernst & Albers, 2017). The model has to be tested for the assmption to avoid violation. The violation of the assumptions makes the model unreliable to estimate the population parameters (Javaid et al., 2020a; Tan et al., 2021). The multicollinearity crisis is commonly recognized in the regression model. Multicollinearity can reduce the precision of the estimates of the parameters in the regression models. Multicollinearity is a problem that happens in the sample,

therefore, there is no single method to identify multicollinearity. Multicollinearity should be checked when assessing data.

Big data means datasets with a large number of observations, complexity, and large numbers of variables. The extraneous variables in big data introduces bias to a research and can lead to erroneous conclusions about the relationship between dependent and independent variables. Moreover, the statistical analysis is affected significantly by the model, and the effect on the model can result in severe multicollinearity (Ullah et al., 2019). Extraneous variables introduce research bias to a study.

According to Daoud (2018), the regression model relates the dependent variable with the independent variables. But the correlation between the independent variables is not appropriate. The correlation between the independent variables is called multicollinearity. This will affect the estimate of the parameters and makes it difficult to fit a reliable model. Principal Component Analysis (PCA) is one of the methods used to tackle the problem of multicollinearity. The factors that exhibit high correlation with other factors can also be excluded from the model.

Shrestha (2020) stated that correlation coefficients, variance inflation factors (VIF), and eigenvalue techniques can be used to detect the presence of multicollinearity. In addition, principal component regression, ridge regression and weighted regression can be used to address the issue of multicollinearity. These three methods can be used to eliminate multicollinearity among the independent variables.

The multicollinearity problem was solved by using the ridge regression method (Bager et al., 2017; Sami et al., 2022). Sami et al. (2022) used ridge regression estimators in the Conway-Maxwell Poisson (COMP) regression model. The

equidispersion, overdispersipn and underdispersion were introduced. To estimate the parameters in the COMP regression model, iterative reweighted least square was used. Bager et al. (2017) also applied ridge regression to solve the multicollinearity problem. A comparison of the results was done by comparing the OLS methods used to build the feasible model in the study. The Variance Inflation Factors (VIF) and Condition Number (CN) 11 were used to identify the multicollinearity. Hence, the use of ridge regression methods in more studies is advised because it gives a more accurate estimate than the OLS method when the independent variables exhibit multicollinearity.

Table 2.3 gives a summary of the literature review related to the problem of multicollinearity.