Applying supervised learning to find out water user's behavior through ability to pay for irrigation water fee

Hoa-Thi-Thu Bui[1]

Abstract: Supervised machine learning is considered as one of the methods to find out variable relationships more informative compared to traditional statistical methods. In this article, both traditional statistical analysis and supervised machine learning approaches are used to study consumer behavior through their willingness to pay for irrigation services. 222 households in Nam Dinh, Thai Nguyen, and Phu Tho province were investigated. By the regression model, the results show that the variable area (DT) and yield of winter-spring crop (NS_DX) directly affect the household's willingness to pay. The result shows that the majority willing to pay of households for irrigation water are higher than the current level. However, by the supervised machine learning approach, the errors of the model, predictions of the household's payment level, and solutions to avoid overfitting are also shown, which could not be implemented if using traditional statistical analysis.

Keywords: Supervised learning, water user behavior, wilingness to pay.

1. Introduction

In the digital economy, all transaction activities are transformed into digital form, especially with the explosion of the Internet, cloud computing technology, which has made the ability to access data easily with high processing speed. Databases now are mostly big data. Big data has the characteristics of variety, volume, velocity and veracity. It is time to require the data analysts to change their point of view as well as methods to meet the demand of data analytics. This is a challenge as well as an opportunity for data analysts. Data analytics and exploratory data analysis is considered as one of the urgent requirements for many organizations, especially in business and economics. Data analytic is one of the most popular methods and getting useful information channels for discovering and interpreting in economics relationships. Normally, relationships are often analyzed, and predicted by regression models with traditional statistical approaches. Nowadays, the machine learning approach is considered as one of the important applications in data analytics as well.

Machine learning is intuitively understood as a way of learning knowledge from data. Machine learning is done through inductive inference, learning from data sets (John et.al,2020). The principle of machine learning is based on the data, it will learn how to build, design a model by the suitable algorithm in order to set up the appropriate parameters, and predict new situations based on previous experiences. Machine learning is done by algorithms such as automating the decision-making process by generalizing from known cases such as predicting the future, or algorithms that can generate results from input that the machine has not discovered yet, without any human help. The application of machine learning in various activities such as sorting spam email or customer clustering, customer behavior etc. What makes machine learning superior and unique over traditional

[1]Faculty of Economics and Management, Thuy loi University, Vietnam
Received 19th Jul. 2022
Accepted 31st Aug. 2022
Available online 31st Dec. 2022

statistical analysis is its ability to change models over time to fit new data. From data, through trial and error, repeated many times, machine learning will learn from failures and successes and know the probability, thereby helping to improve results or forecasts. Machine learning is a form of inferential statistics, by using data to understand patterns and relationships between variables. It is possible to make predictions for new variables on the basis of existing information. One of the most typical applications of machine learning is regression modeling which is known as supervised learning. Unlike the traditional statistical approach, machine learning will limit and overcome the assumptions or subjective assumptions of the analyst. Machine learning is considered to find out many potential relationships which are limited by subjective assumptions from analysts.

In water management, it is very important to understand the behaviors of water users through their ability to pay water fees. Their willingness to pay for the water fee depends on many factors. The application of different analytical approaches, especially the machine learning approach, aims to find out the hidden information behind the relationships, which is not implemented in traditional statistical analysis. In this article, the overview of the machine learning approach is introduced, as well as an application of supervised learning to find out the relationships between willingness to pay for water and other factors. The sample of 222 households are selected in the typical irrigation areas in Thai Nguyen, Phu Tho, and Nam Dinh district to study their willingness to pay farmers for irrigation water. This study is studied by using both the traditional statistics and machine learning in order to compare the differences and improvement of supervised learning approach.

2. Literature review and methodology

Machine learning is applied commonly in economic analysis and gradually replacing the traditional statistical analysis methods(März et al ., 2016 ; Crane-Droesch, 2017). The terms machine learning or Artificial Intelligence (AI) and Deep learning (DP) are often used interchangeably. Machine learning is a part of artificial intelligence which aims to learn from data and using statistical methods (Goodfellow et al., 2016).

Machine learning is defined as the automatic process of extracting patterns from data. However, unlike traditional computer programming, the output or decision is predetermined by the programmer, machine learning uses data as input to build a decision model. Decisions are made by deciphering relationships and patterns in data using probabilistic reasoning, trial and error, and other computationally intensive techniques. This means the output of the decision model is determined by the content of the input data but not by any preset rules from the analyst. The analyst is solely responsible for feeding the model data, choosing the appropriate algorithm, adjusting the settings to reduce calculation errors, and the machine will analyze and detect relationships from the data without interference, assumption or obligation.

Using machine learning, data is used both to develop and evaluate model performance, which is an objective approach compared to the traditional analytical approach, where the analyst makes certain assumptions to choose models. The data is divided into two or three parts: the training dataset and the validation dataset and the testing dataset (Hastie, Tibshirani and Friedman (2009) The training dataset usually accounts for the large percentage of existing data (about 60-80%). This data is used to develop models. In other words, training data is used to learn, detect from the experience

and aim to find out the important points to remember. The validation dataset is part of the training data but is used for checking the accuracy of the model during the training process. After the model has been developed based on the training data with a reasonable degree of accuracy of the predictions, one of the next requirements of machine learning is to test the model based on the remaining data, known as testing data. The testing data is known as the evaluation set, which is used to evaluate the performance of the model, after the machine has completed the learning process (John, et.al, 2020).

One of the common problems in machine learning is overfitting or underfitting. Overfitting and underfitting are understood as the phenomenon that the model is too fit or not enough to fit the training data, which leads to wrong predictions and poor model quality when using testing data to retest the model. In order to solve these problems, the validation dataset is used to check the accuracy of the machine learning model during training before using the testing data to evaluate the performance model. paradigm. The validation dataset is extracted from the training dataset to be used to evaluate the training error. The lower error from training and validation sets, the lower error can be predicted in testing data. The model with the smallest validation error will be the best model. Therefore, it can be intuitively seen that the higher the accuracy of the validation set, the higher the accuracy of the training set.Machine learning is one of the data analysis approaches used more popular today. However, to implement machine learning requires many algorithms, and choosing which algorithm is a big challenge. Therefore, to select and test specific algorithms, the classification and processing of input and output variables is necessary to understand. Machine learning is classified into four categories: Supervised Learning; Unsupervised Learning; Semi-supervised learning; Reinforcement Learning.

Supervised learning is the most common form of machine learning. Supervised learning is an algorithm to predict the output of new data based on previously known inputs and outputs. Based on the training data, the model will make predictions and correct when getting a level of accuracy. The popular algorithms used in supervised learning include regression analysis (linear regression, logistic regression, nonlinear regression), decision trees, classification, Random forest, K Nearest Neighbors, Neuron Network, etc.

Machine learning approach will be towards predictive models and to evaluate model selection. The performance indicators of the model will be evaluated based on the validation data set. The measures based on the validation dataset play a more objective role to evaluate the accuracy of the prediction than the training dataset. Those indicators are mean error (ME), root mean squared error/root mean square error (RMSE), mean absolute error (MAE), mean percentage error (MPE), mean absolute percentage error (MAPE) (Alexander, 2022). For the classical statistical approach, R^2 and standard error (SE) are used mainly to the performance evaluation to find a fit model. However, these indicators could not tell us much about the model's ability to predict new records.

With supervised learning, one of the questions is how the predictive model will adapt to apply new data. Therefore, the selection and comparison of model performance will help us to choose a good model for practical application. However, one of the common problems in machine learning to deal with is overfitting or underfitting as mentioned above. Overfitting is encountered when we put too much data or variables into the model. To overcome overfitting, the idea is to restrict the

magnitude of the coefficients. The higher the correlated predictors are, the higher the standard error of coefficients are. This is explained by the small changes in the training data that can change the correlation predictors, which represents a high standard error leading to poor predictability. By limiting the magnitude of the coefficients, the variance is reduced. To solve this problem, there are two reduction methods: Ridge regression and lasso regression (ridge regression and Lasso regression) . By Ridge regression , the penalty is based on the sum of squared coefficients (called the L_2 penalty). Lasso regression using the sum of absolute values (called the L_1 penalty), for the predicted p (excluding intercept). It shows that the lasso penalty effectively shrinks the coefficient to zero, thus leading to a subset of factors. Lasso regression is based on a linear regression model but L_1 penalty can be used to overcome overfitting. Therefore, we can fit a model that contains all possible predictors and use lasso to perform variable selection using the coefficient estimation adjustment technique (narrowing the estimate). In particular, the minimization goal includes not only the residual sum of squares (RSS) - like OLS regression - but also the sum of the absolute values of the coefficients. In practice, Python language can be applied to find out the relationship between variables by using machine learning algorithms. By Python programming language, regular linear regression can be replaced by the *Lasso* and *Ridge methods* in sklearn.linear_model. The penalty parameter α defines the threshold. *LassoCV*, *RidgeCV* and *BayesianRidge* methods will help to automatically select the penalty parameter. *LassoCV* and *RidgeCV* use cross validation to determine optimal values for the penalty parameter.

Machine -learning and artificial intelligence techniques are applied in studying irrigation water user's behavior. Li & Xuewei Chao (2020) reviewed the application of machine learning approaches such as artificial neural network for classification and yield forecasting, analysis of agricultural research surveys related to the farmer behaviors could be included. Kamilaris et al. (2018) applied deep learning- one of the methods of machine learning approach to study various agricultural and food production challenges. The willingness to pay for urban water supply was estimated under machine learning approach and traditional econometrics (multiple regression) (Malik et al. 1999). The results show that the forecasting error of the machine learning model was less than the traditional regression and found out the accuracy of the model as well. In Vietnam, most studies on water user's behaviors have been carried out mainly by applying the traditional statistical analysis approach. The OLS linear regression model is mainly used to estimate factors affecting their willingness to pay (Bui.2016, Bui & Doi.2020).

In this article, the factors affecting the willingness to pay for irrigation water services are found out by using both traditional statistical analysis and supervised machine learning approach. Willingness to pay for irrigation water was surveyed from 222 households in four provinces of Thai Nguyen, Phu Tho, Nam Dinh. The study conducted surveys and collected data from 2016 to 2017. Two typical irrigation systems which are gravity systems located in Thai Nguyen and Phu Tho province, and the other is a pumping irrigation system in Nam Dinh province are selected.The selection of typical survey households is based on the opinions of experts and experienced people in local authorities. A survey of 222 households was conducted to represent the populations in selected areas. The methodology to estimate the sample was followed by random sample selection of Smith (1986). The sampling method is a stratified method, from the provincial level,

classified according to the district and commune levels. The purpose of this survey aims to find out the water user's behaviors through their ability to pay for irrigation water fees as well as their perception for irrigation water management.

3. Application

To find out the economic relationships, supervised machine learning is implemented in order to present regression model analysis and compare it to the traditional statistical approach. So far, *the linear regression model* is quite commonly used according to data analysis. However, with the classical statistical approach, the models are mainly for inference purposes. Under machine learning approach, though the inference ability, the linear regression model is used to predict. The most commonly used regression model is the multiple regression model. The regression model describes the relationship between the dependent variable or the target variable Y with the set of input/independent variables X_1, $X_2 \ldots X_n$, which is expressed:

$$Y = \beta_0 + X_1 \beta_1 + X_2 \beta_2 + \ldots + X_n \beta_n + \varepsilon$$

where β_0, …, β_n are the *coefficients* and ε is the *noise* or *the unexplained part*. In predictive modeling, data is also used to evaluate the performance of the model.

Under the classical statistical analysis approach, regression models are often used to explain or quantify the average influence of the input factors on the outcome or output. The focus of the research is on the coefficients (β). To study the population, the data is selected randomly. Regression models are used to estimate from the sample to find out the population mean relationship. Furthermore, in explanatory modeling, all the data is used to estimate the model in order to find the population relationship hypothesis.

With machine learning approaches, by data mining the linear regression model is not only used to explain the relationship but also be applied to predict new records. The coefficients of the model are obviously not too concerned, but are aimed at predicting how the model can generate new assumptions. In order to find the model with the highest predictability through model evaluation which is based on a set of predictive data, the data used for the purpose of predicting the outcome of new records is usually divided into training data and validation data. The training data set is used to estimate the model. The validation set is used to evaluate the predictive performance of the model and is evaluated based on the accuracy of the model. The main purpose is to focus on predictions.

In this article, the author explores the factors affecting the willingness to pay for irrigation water services using traditional statistical analysis and supervised machine learning approach. Willingness to pay for irrigation water was surveyed from 222 households in four provinces of Thai Nguyen, Phu Tho, Nam Dinh. 82 households out of 222 households are selected from Thai Nguyen province, 91 householders from Phu Tho province, 49 households from Nam Dinh province. The sampling method is a stratified method, from the provincial level, classified according to the district and commune levels. Most of the cultivation areas are used for paddy with two main crops: Winter – Spring and Summer-Autumn. However, there is higher irrigated water in Winter – Spring compared to the Summer-Autumn, due to weather conditions. Therefore, the yield in the winter-spring crop is selected to reflect water- related behaviors of farmers clearer than that in the Summer-Autumn crop.

In order to find out which factors affect to their willingness to pay for irrigation water, there are several factors such as age, number member in the family, sex, area for paddy cultivation, yield, etc. However, there are only

area (DT) and yield in spring - winter crop (NS_DX) variable are the main factors affect to their willingness to pay with the estimated coefficient of the model is statistically significant. By method of least squares (OLS) and Python programming language, the regression model results show the relationship between willingness to pay and independent variables as below:

WTP = - 7872.42 + 4607.82*DT + 155.80NS_DX

The cultivated area (DT) and typical yield of winter-spring crop (NS_DX) variables are statistically significant, with p_values both less than 0.05. The equation shows that when the area is increased by 1 acre, the willingness to pay for irrigation water will increase by VND 3097.7, and increase the yield of winter-spring crop by 1 unit, the willingness to pay will also increase by VND 163.6/crop.

One of the basic requirements when performing the traditional linear regression analysis is checking required assumptions. In order to check these assumptions, we found that the yield and area variables are not random variables, and there is no error in the calculation. The remaining assumptions can be easily tested for the relationship between the variables and the residuals using a scatter plot. From Figure 1, the graph plots of the residual *ei* and the predictive value of willingness to pay shows that the residuals are clustered around the y = 0, so the assumption that εi has a mean of 0 is acceptable. The second graph shows the residual and expected values based on a normal distribution. The residuals are concentrated very close to the values on the standard curve. Therefore the assumption that εi is distributed according to the normal distribution, can also be met. The other graph shows the standardized residual and the value of . This graph shows that there is no difference between the standard residuals of , and the assumption with fixed variance σ^2 of ε_i for all x_i is also satisfied (Fig1). Based on the machine learning approach, the data is separated into training data (70%) and test data (30%). The relationship between the willingness to pay (WTP) and the area (DT) and yield of the winter-spring crop (NS_DX) is estimated by machine learning through a training dataset. The results show the regression model with error parameters as below:

WTP = 2098.8711 + 4031.8582 *DT + 122.7419 NS_DX

After the model has been established, the regression coefficients are used to predict the user's willingness to pay for irrigation water. By supervised learning approach, the accuracy of the prediction is interesting in applying new records. The following results describe a sample of predicted values of willingness to pay about 15 households in the validation set, with the estimated model. Furthermore, the predictions and errors are also compared to the actual level of those households (Figure 2).

The errors based on the training dataset will show how the model fits, the errors based on the validation dataset (known as the "prediction error") measure predictability of new data in the model (known as the predictive performance). The training errors normally will be smaller than the validation error because the model is generated from the training dataset. The more complex the model is, the more likely overfitting data is because of the large difference between training and validation errors. In the extreme case as above can be seen as overfitting, the training errors will be zero (perfect fit of the model to the training data) and the validation errors are non-zero and insignificant. Then it is necessary to compare the RMSE, MAE, etc indexes of the training and validation set. Table 1 shows the training set performance measures that appear to be slightly lower than those for the validation set.

By the supervised machine learning approach, the accuracy of the prediction is also specified. The mean error (ME) is – 1249 VND, representing lower average predictions than the actual outcome variable for willingness to pay for irrigation water. Most of the errors are within ± 20000 VND. The absolute average error is about 17607.23 VND.

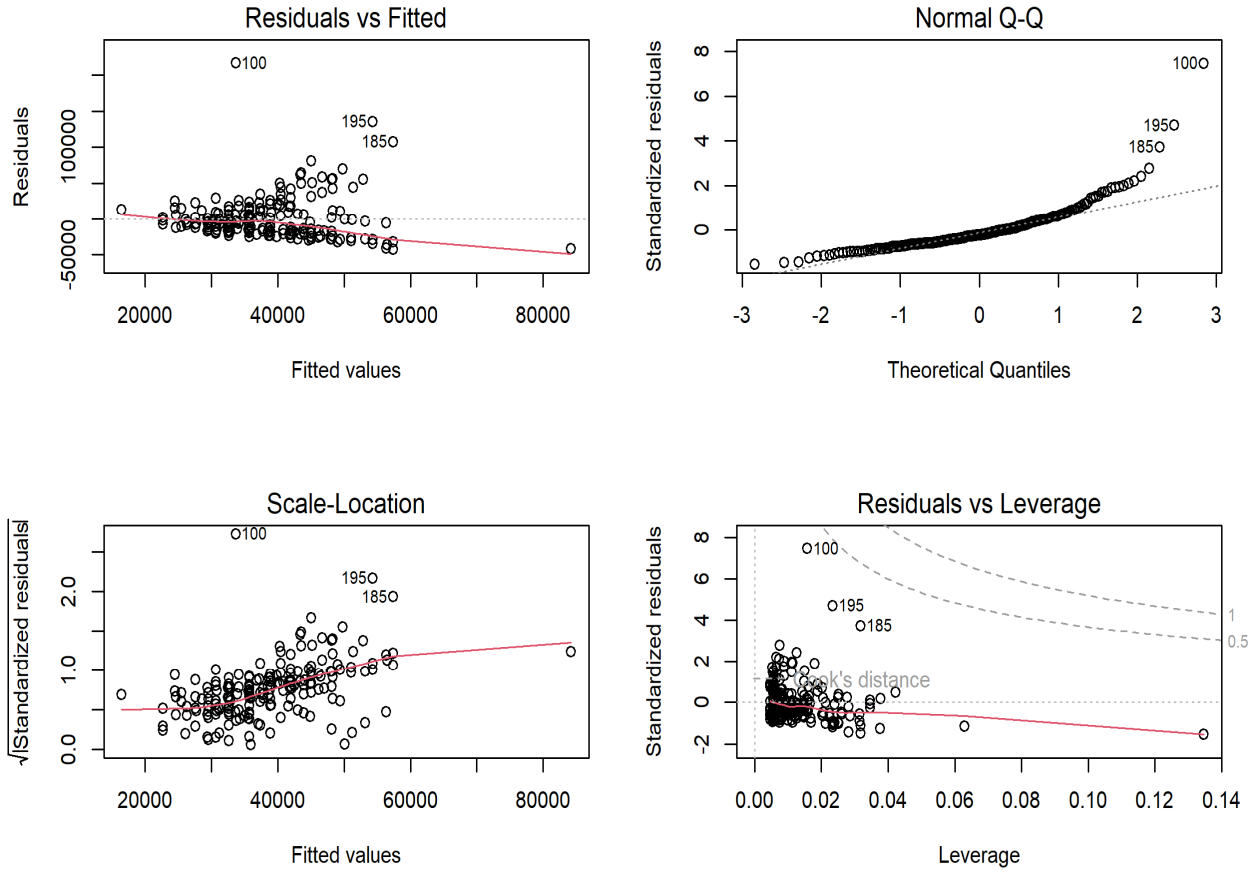

Figure 1. The hypothesis testings for OLS regression of the model

To avoid overfitting, ridge and lasso regression are used as techniques for regularization. The idea of these techniques is to penalize the magnitude of the coefficients in order to reduce the error between the predictions and the actual observations. Other regression algorithms that machine learning applies such as Lasso regression and Bayesian regression. The prediction results based on these algorithms are as follows (Table 1). Regression coefficients of area (DT) and yield of winter-spring crop (NS_DX), corresponding to the regression methods are shown in the Table 2 below.

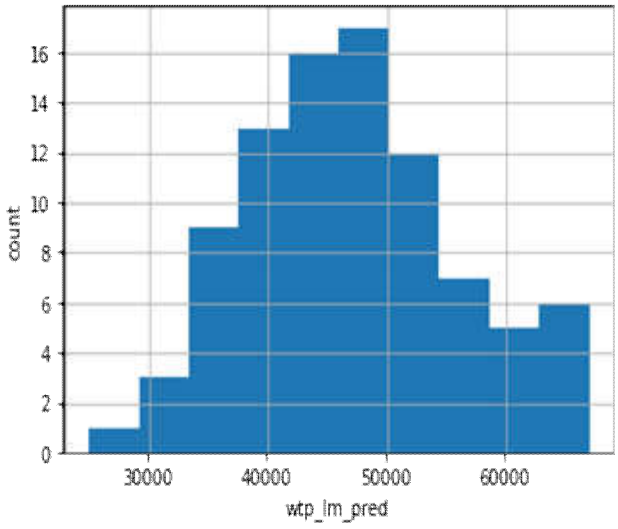

Figure 2. Predicted values of willingness to pay

Table 1. Comparing the performance indicators of various regression algorithms

Regression type	ME	RMSE	MAE	MPE	MAPE
Lasso	-1249.4697	22419.2123	17609.7499	-38.6853	59.3785
Ridge	-934.1537	23853.5459	18262.1297	-41.7522	62.6573
Bayes	-637.5771	26046.2245	19607.8398	-44.9666	67.4506

Table 2. Comparing the regression coefficients of various regression algorithms

Regression coefficient	OLS	Lasso	Ridge
DT	4031,858	3983,299	2029.33
NS_DX	122.742	118,734	64.509

The results show that Lasso regression has overcome the overfitting problem and has the lowest error indexes compared with other regression methods (Table 1). These results are not found and implemented in classical statistics because of depending on given assumptions. It can be seen that the supervised learning approach helps us to get a more accurate prediction, because this is based on the collection of data and evaluated continuously on the accuracy. Compared with the traditional statistical approach using the method of least squares (OLS), it is aimed at finding the regression line to minimize the sum of squares of errors. The machine learning approach has many other algorithms to check the prediction results. Through the performance evaluation indicators of the model, it is possible to choose the most appropriate prediction method. According to the supervised learning algorithm, the value of willingness to pay for each household is also predicted that could not be implemented if using traditional statistical analysis.

4. Conclusion

Based on the data, the most suitable model can be estimated by machine learning. In this study, supervised learning in the form of regression analysis is applied to understand the relationship between cultivation area (DT) and yield of winter-spring crop (NS_DX) variables and the willingness to pay water users. Besides the regression coefficients representing the relationships, the willingness to pay for irrigation water is predicted objectively with performance indicators of the model as well, compared to the traditional statistical method. The predictions of willingness to pay water users has normal distribution, with average value is around 47000 VND/*sao*, which is higher than the curent fee level. In this study, the mean error is lower than the actual outcome variable for willingness to pay. Most of the errors are within ± 20000 VND. In machine learning, the limitations of model building, specifically the case of overfitting, are also solved by various regression algorithms. The results show that Lasso regression has overcome the overfitting problem and has the lowest error indexes. This implies the machine learning approach is more objective and reliable than the traditional statistical approach. Applying machine learning in data analytics, especially in many fields such as economics and business, will provide a lot of useful information for decision makers.

References

Alexander Jung (2022). *Machine Learning*: The Basics. Springer.

Bui T.T.Hoa, Doi. T.Loan (2020). *Study the willingness to pay for domestic water*, *applied in Gia Lam district, Ha noi*, Journal of Water Resources & Environment Engineering. No 72, ISSN 1859-3941

Bùi Thị Thu Hòa (2016). *Nghiên cứu ý muốn thanh toán nước tưới của người nông dân phục vụ cho bài toán định giá nước tưới.* Tạp chí Kinh tế và Dự báo, số 23, ISSN 0866-7120

Crane-Droesch, A. (2017). *Technology diffusion, outcome variability, and social learning: evidence from a field experiment in Kenya a*. American Journal of Agricultural Economics 100: 955–974.

Goodfellow, I., Bengio, Y., Courville, *A. and Bengio, Y.* (2016). Deep Learning . Cambridge: MIT press

John D. Kelleher, *Brian Mac Namee, and Aoife D'Arcy (2020). Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies. 2ed.* The MIT Press Cambridge, Massachusetts London, England

Kamilar Andres & Francesc X. Prenafeta-Boldu (2018). *Deep learning in argriculture: A survey.* Computers and Electronics in Agriculture. Volume 147. Pages 70-90.

Malik Ranasinghe, Goh Bee-Hua & T. Barathithasan (1999). *Estimating willingness to pay for urban water supply: a comparison of artificial neural networks and multiple regression analysis.* Impact Assessment and Project Appraisal.SSN: 1461-5517 (Print) 1471-5465.

März, A., Klein, N., Kneib, T. and Musshoff, O. (2016). *Analyze farmland rental rates using Bayesian geoadditive quantile regression*. European Review of Agricultural Economics 43: 663–698.

Smith, M. J. (1986). *Contemporary communication research methods.* Belmout, CA: Wadsworth Publishing, pg 223- 225.

Yang Li & Xuewei Chao (2020). *ANN-Based Continual Classification in Agriculture.* Journal Agriculture 2020, 10, 178; doi:10.3390/agriculture10050178.