https://doi.org/10.17221/399/2018-AGRICECON

Evaluation of influencing factors on tea production based on random forest regression and mean impact value

Yihui Chen[1,2,3]*, *Minjie Li*[1]

[1]*School of Economics and Management, Fuzhou University, Fuzhou, China*
[2]*Cooperative Innovation Centre of Modern Agricultural Industrial Park, Quanzhou, China*
[3]*Anxi College of Tea Science, Fujian Agriculture and Forestry University, Fuzhou, China*
**Corresponding author: chenyihui@fafu.edu.cn*

Citation: Chen Y., Li M. (2019): Evaluation of influencing factors on tea production based on random forest regression and mean impact value. Agricultural Economics – Czech, 65: 340–347.

Abstract: Overproduction of tea in the major producing countries is an important factor which restricts the development of tea. Therefore, the factors from the economic, social and environmental system affecting tea production have become the focus of both academia and practice. Random forest regression (RFR) and mean impact value (MIV) were applied to evaluate the weights of variables. Firstly, RFR was preliminarily used to build a well-trained model, and then the weights of variables combining with MIV were calculated. Then, a well-trained model was constructed after variable selection to evaluate the importance of tea production from 2007 to 2016. The results revealed that the economic system and the social system are the main factors that affect tea production. The net production value and total population have little negative effects on tea production, while the area harvested has a little positive effect. Based on the research findings, governments and enterprises should develop and upgrade tea production technology, promote the exchange and cooperation in the international tea trade, then ultimately achieve sustainable development of the tea industry.

Keywords: agricultural production; machine learning; sustainable development; weights

Camellia sinensis (L.) is a species of evergreen shrub or small tree whose leaves and leaf buds are used to produce tea. Tea has the great potential to increase income and social benefits for resource-poor small-scale farmers. Many existing studies in the literature have focused on the impacts of various variables related to the economic system, social system and environmental system on tea production (Qiao et al. 2016; Gunathilaka et al. 2018; Xiao et al. 2018). Compared with other factors, the increase in the tea plantations area is considered to be the most important factor leading to the increase in tea production (Xiao et al. 2018). Limited opportunities for crop switching and lengthy pre-harvesting periods, make the plantation sector particularly vulnerable to climate change (Gunathilaka et al. 2018). Changes in temperature, rainfall, and the occurrence of extreme weather events, for instance, have adversely affected tea production (Gunathilaka et al. 2017). Irrigation is essential to overcome insufficient rainfall and to achieve a stabilised yield for tea production (Hong and Yabe 2017). Life cycle assessment and data envelopment analysis were used to evaluate the energy efficiency and aid in the reduction of environmental burdens for tea production (Kouchaki-Penchah et al. 2017). The stochastic frontier analysis was applied to measure the environmental efficiency of Vietnamese tea farms while chemical fertilisers and pesticides are widely used by tea farmers to increase yield (Hong et al. 2016).

The artificial neural network is not efficient in modelling high-dimensional data while a nonlinear support

Supported by the Social Science Planning Project of Fujian Province, China, Project No. FJ2018C045, and the Ministry of Agriculture and Rural Affairs, China, Project No. KMD18003A.

https://doi.org/10.17221/399/2018-AGRICECON

vector machine is not robust to the presence of noisy data (Ghasemi and Tavakoli 2013). Compared with other statistical approaches, such as linear regression (Grömping 2009) and support vector (Ishak 2016), the random forest has very high accuracy, ability to model complex interactions among predictor variables, and flexibility to perform several types of statistical data analysis (Culter et al. 2007). Random forest is a non-parametric, non-linear regression and classification algorithm (Breiman 2001) and widely used in chemometrics (Ghasemi and Tavakoli 2013), biomedical science (Wu et al. 2017), computer mathematics (Mendez and Lohr 2011) and other research fields (Hutengs and Vohland 2016). Initially proposed by Dombi et al. (1995), mean impact value is a feature selection method to evaluate the importance degree and has been considered as one of the best indicators for evaluating variable relevance (Fan et al. 2018). Mean impact value is usually combined with other statistical methods, such as support vector regression (Fan et al. 2018), back propagation (BP) model, factor analysis (Zhang and Jin 2018) and artificial neural network (Luo et al. 2016).

METHODOLOGY AND RESEARCH PROCESS

Methodology

BP neural network. The three-layer BP neural network with enough nodes in the hidden layer has the ability to simulate any complex nonlinear mappings. The particular three-layer BP neural network consists of three parts: the input layer, the hidden layer and the output layer. There is a certain number of nodes on each layer, and one node represents one neuron. For the input signals, they must first be propagated forward to the nodes in the hidden layer, then the output signals of hidden layer nodes are propagated to the output nodes after the ordinary transfer function, and finally, the output predictions are obtained. The initial weights are randomly set, and the fitting output values can be obtained by certain rules during the forward training process. Then, according to the differences between fitting data and actual data, the weights are modified in the backward process.

Radial basis function neural network. A typical radial basis function (RBF) neural network is a three-layer neural network, which includes the transparent input layer, the hidden layer with a sufficiently large number of nodes and the output layer. RBF neural network consists of large quantities of interconnected artificial neurons. More importantly, every neuron in the radial symmetric basis function network is wholly connected with each neuron of the next layer. RBF neural network requires three parameters: the centre of the basis function, width and the weight parameters from the hidden layer to the output layer. In the input layer, the size of the nodes is mainly determined by the input vector dimension. Meanwhile, the hidden layer is connected to the input nodes, where the output data dimension equals the size of the nodes. RBF neural network performs an arbitrary nonlinear mapping from the input space to the output space.

Support vector regression. While compared to an artificial neural network, the predictive ability of support vector regression (SVR) is demonstrated to maintain interpretability of the model and have better performance. SVR is an extensively used machine learning algorithm grounded in statistical learning theory and for which training is efficient due to the convexity of the training problem. Furthermore, the SVR formulation provides for an extension to nonlinear regression through kernel functions. In the SVR model, the original data is nonlinearly mapped to a higher dimensional feature space.

Random forest regression. Two parameters, the number of trees in the forest (*ntree*) and candidate features in each tree (*mtry*), are important in the random forest regression algorithm. The default number of *ntree* is 500 (Adam et al. 2014), while *mtry* is usually equal to $p/3$ (Grömping 2009); p represents the total number of predicted variables. The process of building random forest regression can be divided into three steps as follows: i) extract the n sample training set randomly from the original sample by adopting bootstrap method to form *ntree* regression trees, organise the samples not be collected each time as out-of-bag (OOB) data sets and treate them as test set for the validation; ii) extract *mtry* explanatory variables that are most effective in partitioning data from explanatory variables while *mtry* depends on the minimum aggregate error rate of the OOB data sets; iii) integrate the generated regression trees into the random forest, then select the average value of all decision trees as the final prediction value.

Mean impact value. The brief calculation procedures of mean impact value can be divided into four steps as follows: i) divide original samples into training set and testing set with a definite proportion, then train and simulate an effective model base on scientific method; ii) transform each independent variable

https://doi.org/10.17221/399/2018-AGRICECON

in training sample by a given symmetric variation, which means that one independent variable is changed at a time, and other variables keep unchanged, to obtain a pair of simulation prediction based on the well-trained model; iii) calculate the differentials between the former pair of simulation prediction and note as $iv(p)_i$; iv) average $iv(p)_i$ as the $miv(p)$ which reflects the importance of variable p. A relatively higher absolute values of $miv(p)$ are considered to have greater influence. A positive value of $miv(p)$ indicates a positive influence of a specific independent variable p on the response variable, whereas *vice versa*.

Combination of methodologies

The final set of variable importance measures of random forest regression may not include covariate of interest and lacks interpretability. Therefore, avoiding using the three indicators mentioned before, random forest regression and mean impact value are

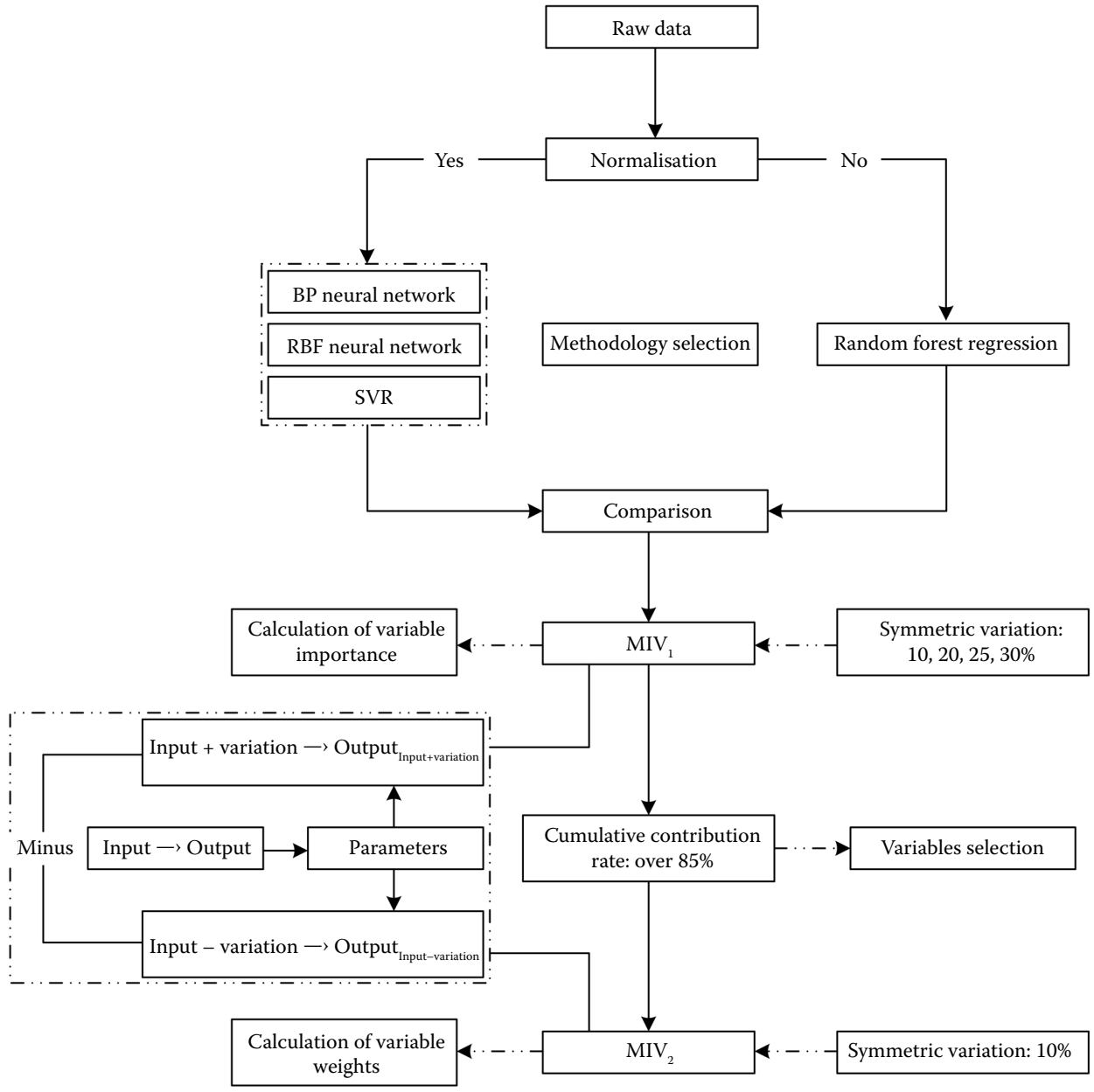

Figure 1. Research process

BP – back propagation; RBF – radical basis function; SVR – support vector regression; MIV – mean impact value

Source: drawn by authors

https://doi.org/10.17221/399/2018-AGRICECON

Table 1. Descriptive statistics of all variables

Index	Code	Unit	Mean	Std. deviation	Skewness	Kurtosis
Production	Y	10^4 t	14.125	35.982	3.884	16.027
Area harvested	X_1	10^4 ha	9.991	31.099	4.841	24.775
Gross production value	X_2	10^8 USD	3.634	14.151	5.881	36.613
Export quantity	X_3	10^4 t	4.946	9.826	2.203	3.648
Import quantity	X_4	10^4 t	1.464	3.382	3.818	15.126
Net production value	X_5	10^7 USD	6.362	18.317	5.062	28.170
GDP deflator	X_6	10^2 USD	1.021	0.161	0.212	–0.274
Value added deflator	X_7	10^2 USD	1.028	0.197	0.262	0.517
Exchange rate	X_8	10^3 per USD	1.619	4.896	3.642	13.409
Total population	X_9	10^7 persons	13.355	31.105	3.437	10.426
Rural population	X_{10}	10^7 persons	6.939	18.114	3.583	11.507
Temperature change	X_{11}	˚C	0.827	0.470	0.679	1.352

Source: collation by authors

combined to calculate and rank the weight of selection variables impacting on tea production.

Research process

The flow chart of this research is shown in Figure 1. This paper can be divided into four steps as follows: i) normalise the original data to facilitate machine learning algorithm; ii) based on the goodness of fit, evaluate the various methods and select the best methodology for the subsequent calculation; iii) combine the methodology selected above with mean impact value, and calculate the weights of all primary characteristics without time dimension; iv) combine the most appropriate methodology with mean impact value, then calculate the weights in every year of all characteristics selected above and reveal the effects of variables on tea production.

Methodology selection. The coefficient of determination (R^2) (Wang et al. 2015), the mean absolute error (MAE) (Peng et al. 2018), the root mean square error (RMSE) (Zhu et al. 2017; Peng et al. 2018), and the mean absolute per cent error (MAPE) (Guo et al. 2011; Ren et al. 2014) which can be evaluated respectively as follows are applied to estimate the effect on forecasting in this study. When the R^2 value closes to 1, meanwhile the MAE, the RMSE and the MAPE values verge on 0, the performance of the prediction model is outstanding.

Variables selection. In this process, the key is the application of mean impact value. When calculating the MIV_1 in Figure 1, the symmetric variations are specified as 10, 15, 20, 25 and 30%, respectively. By contrast, the symmetric variation is specified as 10% when calculating the MIV_2. Based on the calculation results of MIV_1, the variables with a cumulative contribution of more than 85% are selected. In the calculation of MIV, the initial parameters are used for the variable input values, and then the weight of the variables is measured by comparing the changes of the output values.

DATA AND VARIABLES

In order to maintain accuracy and continuity, the primary data of tea production in global major agricultural countries for the past 10 years with a span from 2007 to 2016 were selected. In the end, 320 original samples from 32 countries were obtained and integrated. According to existing literature (Gunathilaka et al. 2018; Xiao et al. 2018), area harvested, gross production value, export quantity, import quantity, net production value, gross domestic product (GDP) deflator, value-added deflator and exchange rate were chosen as economic system variables prepared for inputs (Table 1). Total population and rural population are regarded as social system variables, while temperature change reflects environmental system. All raw data without any other transformations are downloaded from the database of the Food and Agriculture Organization of the United Nations (UNFAO 2018).

RESULTS

By simulating four different machine learning models several times, the comparisons of four methodologies' goodness of fit can be seen in Table 2. Comparing goodness of fit of methodologies, random forest regression

https://doi.org/10.17221/399/2018-AGRICECON

Table 2. Comparisons of methodologies

Methodologies/goodness of fit	R^2	MAE	RMSE	MAPE
BP-neural network	0.999	135.900	311.276	4.324
RBF-neural network	0.997	3709.257	20195.990	1.322
Support vector regression	0.999	75.974	92.632	6.826
Random forest regression	0.999	0.768	1.060	0.072

R^2 – coefficient of determination; MAE – mean absolute error; RMSE – root mean square error; MAPE – mean absolute per cent error; BP – back propagation; RBF – radial basis function

Source: collation by authors

is chosen finally and applied to calculate the weights of variables by combining with mean impact value. The results are shown in Table 3.

In order to accurately calculate the weight of each factor influencing tea production, random forest regression model and mean impact value method were combined in this study. The variation explained by random forest regression based on raw data reached 99.06%, which indicated that the established model is extremely practical and effective.

By further optimising the calculation process, six independent variables with a cumulative contribution rate over 85% based on the weights ranked are selected, and the results are shown in Table 4. The magnitude of the variation of the independent variables is set to 10%, that is, the weights of the independent variables are calculated based on 10% increase and 10% decrease respectively.

The economic system and social system are the main factors that affect tea production, which is consistent with the observation from Xiao et al. (2018). Somewhat different from the results of Table 3, the results of Table 4 show that there is a fluctuation to a certain extent in weight. Notably, the direction of export quantity, net production value and total population affecting tea production have been changed over time. Meanwhile, area harvested, GDP deflator and value-added deflator remain positive weights in any year, indicating positive influences. The weight of the area harvested is small on the whole, indicating that with the development of science and technology, tea production is not entirely limited by planting and harvesting areas. With a cumulative contribution rate over 91%, export quantity, GDP deflator and value-added deflator play important roles and have positive influences on tea production in 2016. By contrast, the negative influences of net production value and the total population on tea production are the most obvious about ten years ago. The favourable economic development circumstances and smooth export trade have enabled

Table 3. Weights of all variables

Variables	Amplitude of symmetric variations (%)				
	10	15	20	25	30
Area harvested	0.077	0.082	0.084	0.100	0.144
Gross production value	0.056	0.059	0.055	0.062	0.067
Export quantity	0.101	0.091	0.100	0.083	0.092
Import quantity	0.009	0.004	0.006	0.006	0.005
Net production value	0.212	0.209	0.227	0.245	0.283
GDP deflator	0.089	0.103	0.090	0.101	0.099
Value added deflator	0.077	0.095	0.105	0.121	0.102
Exchange rate	–0.014	–0.023	–0.016	–0.004	–0.000
Total population	0.347	0.315	0.278	0.225	0.173
Rural population	–0.007	–0.008	–0.033	–0.048	–0.032
Temperature change	0.011	0.011	0.006	0.005	0.003

Source: collation by authors

https://doi.org/10.17221/399/2018-AGRICECON

Table 4. Weights of all selected variables

Variables	Year									
	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
X_1	0.020	0.015	0.003	0.005	0.018	0.015	0.009	0.010	0.008	0.006
X_3	–0.081	–0.149	0.217	0.233	0.232	–0.112	–0.172	–0.120	–0.106	0.262
X_5	–0.318	–0.328	0.170	0.224	–0.011	–0.382	–0.405	–0.420	–0.384	–0.054
X_6	0.089	0.031	0.225	0.261	0.341	0.034	0.032	0.006	0.017	0.323
X_7	0.088	0.033	0.229	0.263	0.353	0.033	0.030	0.005	0.018	0.331
X_9	–0.404	–0.444	0.156	0.014	–0.045	–0.424	–0.352	–0.439	–0.467	–0.024

X_1 – area harvested; X_3 – export quantity; X_5 – net production value; X_6 – GDP deflator; X_7 – value added deflator; X_9 – total population; for explanation of variables see Table 1

Source: collation by authors

tea to be sold all over the world in recent years, especially to large tea-consuming countries that do not produce tea. Another interesting finding is that the positive and negative signs of the independent variables in 2007 are the same as in 2008, 2012, 2013, 2014 and 2015. The situation in 2016 and 2011 is extraordinarily similar, which means that the weights of factors affecting tea production will change somewhat in the next few years. There is thereby an urgent need, but it is still a significant challenge to clarify the influencing factors of tea production based on scientific methods.

DISCUSSION

From the above analysis, the results provide policy reference for improving the effectiveness of tea production and balancing the production and marketing of tea industry in the world. These policies can increase the income of local tea farmers, promote the improvement of living standards, and ultimately reduce or even eliminate poverty. These effective recommended policies are as follows:

i) Some governments from the main tea producing areas should vigorously develop the tea industry and promote the development of agricultural economy according to their own advantages of the natural environment and climate conditions. In addition, the governments should encourage the promotion of tea production technology, and even as a major investor, they can subsidise the research and development of tea production technology which is regarded as a basic public technology, so that tea practitioners can benefit from it.

ii) Enterprises can improve the tea production process and actively develop new products based on tea raw materials to enhance the recognition of consumers, especially the newest generation of youths. In addition, enterprises should fully develop technology to improve tea production efficiency, promote the exchange and cooperation in the international tea trade, ultimately promote the sustainable development of the tea industry.

iii) As grass-roots farmers in tea production, they can control the planting area and do not blindly open up wasteland for farming without planning. Tea farmers can reduce the excessive use of chemical fertilisers and pesticides by introducing scientific management system, increase the unit yield of existing planting area, and make the quality of tea meet the health requirements.

CONCLUSION

The results reflect the tea production and trade situation of main tea producing countries and regions more intuitively. In addition, the methods have a certain reference value in other regions with the same economic crop. The main conclusions are drawn as follows:

i) The results show that tea production has been increasing continuously in the past ten years, but the differences between countries and regions are obvious, that is, tea production is concentrated in a few countries and regions. Tea production and its influencing factors all show the distribution characteristics contrary to the normal distribution, which indicates that the subject of this study is not suitable to adopt the traditional methods with the hypothesis of normal distribution.

ii) The results show that exchange rate and rural population have negative influences on tea production

https://doi.org/10.17221/399/2018-AGRICECON

while others have positive influences within the selected time frame. It is noteworthy that net production value and total population are the main factors influencing tea production within the selected countries and regions. The results state clearly that population composition and external economy are the important breakthroughs for the adjustment of tea production and consumption in the future, which is consistent with the actual situation.

iii) The results reflect a more complicated situation, that is, there is a fluctuation to a certain extent in weights of variables over time. Area harvested, GDP deflator and value-added deflator remain positive weights and influences while other factors are negative sometimes. How all the independent variables affect tea production and their positive and negative correlation over time need to be further analysed in the future.

REFERENCES

Adam E., Mutanga O., Abdel-Rahman E.M., Ismail R. (2014). Estimating standing biomass in papyrus (*Cyperus papyrus* L.) swamp: exploratory of *in situ* hyperspectral indices and random forest regression. International Journal of Remote Sensing, 35: 693–714.

Breiman L. (2001): Random forests. Machine Learning, 45: 5–32.

Culter D.R., Edwards T.C., Beard K.H., Cutler A., Hess K.T., Gibson J., Lawler J.J. (2007): Random forests for classification in ecology. Ecology, 88: 2783–2792.

Dombi G.W., Nandi P., Saxe J.M., Ledgerwood A.M., Lucas C.E. (1995): Prediction of rib fracture injury outcome by an artificial neural network. The Journal of Trauma: Injury, Infection, and Critical Care, 35: 915–921.

Fan H., Xuan J., Zhang K., Jiang J. (2018): Anticancer component identification from the extract of *Dysosma versipellis* and *Glycyrrhiza uralensis* based on support vector regression and mean impact value. Analytical Methods, 10: 371–380.

Ghasemi J.B., Tavakoli H. (2013): Application of random forest regression to spectral multivariate calibration. Analytical Methods, 5: 1863–1871.

Grömping U. (2009): Variable importance assessment in regression: linear regression versus random forest. American Statistician, 63: 308–319.

Gunathilaka R.P.D., Smart J.C.R., Fleming C.M. (2017): The impact of changing climate on perennial crops: the case of tea production in Sri Lanka. Climatic Change, 140: 577–592.

Gunathilaka R.P.D., Smart J.C.R., Fleming C.M., Hasan S. (2018): The impact of climate change on labour demand in the plantation sector: the case of tea production in Sri Lanka. Australian Journal of Agricultural and Resource Economics, 62: 480–500.

Guo Z.H., Wu J., Lu H.Y., Wang J. (2011): A case study on a hybrid wind speed forecasting method using BP neural network. Knowledge-Based Systems, 24: 1048–1056.

Hong N.B., Takahashi Y., Yabe M. (2016): Environmental efficiency and economic losses of Vietnamese tea production: implications for cost savings and environmental protection. Journal of the Faculty of Agriculture Kyushu University, 61: 383–390.

Hong N.B., Yabe M. (2017): Improvement in irrigation water use efficiency: a strategy for climate change adaptation and sustainable development of Vietnamese tea production. Environment Development and Sustainability, 19: 1247–1263.

Hutengs C., Vohland M. (2016): Downscaling land surface temperatures at regional scales with random forest regression. Remote Sensing of Environment, 178: 127–141.

Ishak A.B. (2016): Variable selection using support vector regression and random forests: a comparative study. Intelligent Data Analysis, 20: 83–104.

Kouchaki-Penchah H., Nabavi-Pelesaraei A., O'Dwyer J., Sharifi M. (2017): Environmental management of tea production using joint of life cycle assessment and data envelopment analysis approaches. Environmental Progress & Sustainable Energy, 36: 1116–1122.

Luo S., Cheng J., Wei K. (2016): A fault diagnosis model based on LCD-SVD-ANN-MIV and VPMCD for rotating machinery. Shock and Vibration, 2016: 1–10.

Mendez G., Lohr S. (2011): Estimating residual variance in random forest regression. Computational Statistics & Data Analysis, 55: 2937–2950.

Peng Y., Albuquerque P.H.M., de Sá J.M.C., Padula A.J., Montenegro M.S. (2018): The best of two worlds: Forecasting high frequency volatility for cryptocurrencies and traditional currencies with Support Vector Regression. Expert Systems with Applications, 97: 177–192.

Qiao Y.H., Halberg N., Vaheesan S., Scott S. (2016): Assessing the social and economic benefits of organic and fair trade tea production for small-scale farmers in Asia: a comparative case study of China and Sri Lanka. Renewable Agriculture and Food Systems, 31: 246–257.

Ren C., An N., Wang J., Li L., Hu B., Shang D. (2014): Optimal parameters selection for BP neural network based on particle swarm optimization: A case study of wind speed forecasting. Knowledge-Based Systems, 56: 226–239.

UNFAO (2018): Database of Food and Agriculture Organization of the United Nations. UNFAO.

Wang H., Kong C., Li D., Qin N., Fan H., Hong H., Luo Y. (2015): Modeling quality changes in brined bream (*Mega-*

https://doi.org/10.17221/399/2018-AGRICECON

lobrama amblycephala) fillets during storage: comparison of the Arrhenius model, BP, and RBF neural network. Food and Bioprocess Technology, 8: 2429–2443.

Wu H.Y., Cai Y.P., Wu Y.S., Zhong R., Li Q., Zheng J., Lin D., Li Y. (2017): Time series analysis of weekly influenza-like illness rate using a one-year period of factors in random forest regression. BioScience Trends, 11: 292–296.

Xiao Z., Huang X., Zang Z., Yang H. (2018): Spatio-temporal variation and the driving forces of tea production in China over the last 30 years. Journal of Geographical Sciences, 28: 275–290.

Zhang Z., Jin X. (2018): Prediction of peak velocity of blasting vibration based on artificial neural network optimized by dimensionality reduction of FA-MIV. Mathematical Problems in Engineering, 2018: 1–12.

Zhu B., Han D., Wang P., Wu Z., Zhang T., Wei Y.M. (2017): Forecasting carbon price using empirical mode decomposition and evolutionary least squares support vector regression. Applied Energy, 191: 521–530.

Received December 25, 2018
Accepted February 3, 2019