[Volume 4 issue 06 June 2016]
PageNo.1502-1513 ISSN : 2320-7167

INTERNATIONAL JOURNAL OF MATHEMATICS AND COMPUTER RESEARCH

Modelling Procedures in Determining Heavy Metals Concentration: A Case Study Using Barks of the Cinnamon Tree

Noraini A.[1], Liew L. K.[2], Tan W. H.[3], Zainodin H.J.[4] and N. Surugau[5]

[1,2,4] Mathematics with Economics Programme
[3, 5] Industrial Chemistry Programme
Faculty of Science and Natural Resources
Universiti Malaysia Sabah
88400 Kota Kinabalu, Sabah
Malaysia
Corresponding email: norainiabdullah.ums@gmail.com

ABSTRACT

High concentrations of heavy metals may present as toxins to living organisms. Hence, heavy metal absorption by the cinnamon tree (*Cinnamomum iners*) grown in Universiti Malaysia Sabah (UMS), Malaysia was investigated in order to assess its composition, concentration and dynamics. The relationship of heavy metals concentration in barks of *C.iners* was determined using the multiple regression (MR) technique. The model building procedures were illustrated and discussed. Five independent variables were considered during field and experimental data collection which were namely, diameter of breast height, stem height, average ppm in bark, average ppm in soil and concentration of heavy metal in soil. The concentrations of heavy metals in the bark form the six dependent variables, and they were Cadmium (Cd), Copper (Cu), Iron (Fe), Lead (Pb), Nickel (Ni) and Zinc (Zn). The non-parametric bootstrapping method was used to generate the small sample size (n=28) into 500 observations. The 80 multiple regression (MR) models were developed up to the fourth-order interactions. Results obtained were subjected to statistical modelling, enhanced by the four phase model-building procedures and the process of getting the best model based on the eight selection criteria (8SC). The progressive elimination of variables using the variance inflation factor (VIF) was used to remove collinearity variables from these models. The forecasting criteria of mean absolute error (MAE), root mean square error (RMSE) and mean absolute percentage error (MAPE) were compared and discussed. Comparisons were made to have the best model equation from the six respective heavy metals concentrations. The model M73.20.0 with the toxic heavy metal concentration of Copper (Cu) was obtained as the best model. This thus indicated that Copper was the most toxic heavy metal concentration absorbed in this bark of *C. iners.*

Keywords: heavy metals concentrations, toxic, multiple regression models (MR), bootstrap, VIF, collinearity.

INTRODUCTION

The term ‘heavy metal’ refers to any metallic chemical element that exhibit metallic properties with relatively high atomic weight and has density that greater than 5g/ml (Agarwal, 2009). Heavy metals are natural components of the Earth’s crust which cannot be degraded or destroyed. The most common heavy metals are cadmium (Cd), lead (Pb), nickel (Ni), zinc (Zn), copper (Cu), mercury (Hg) and many others (Hogan, 2010). Some of these metals are essential for plant growth as micronutrients such as zinc, copper, manganese, nickel and cobalt whereas cadmium, lead and mercury have unknown biological functions. If the concentrations of heavy metals exceed a certain dose, it may demonstrate some toxic effects on the living organism via metabolic interference and mutagenesis. In other words, the ecosystem and human health may be affected by the high concentrations of heavy metals present in the animals, plants parts and many other organisms (Mitsios *et al.*, 2010; Zhao *et al.*, 2010).

Plants are the first compartment of the terrestrial food chain in the agro-system; taking heavy metals from soils through absorption, ionic exchange, redox reactions and precipitation. Hence, the importance to comprehend the capacity of toxic metals accumulated by plants parts which will directly influenced to the living beings which consume them. Generally, plants will show certain reaction to the increasing toxic elements concentrations in the soils by depending upon the sensitivity of the plants exposure intensity and chemical species. Some researches had showed that herbs absorb less metal than faster growing plants such as lettuce, carrot and spinach. The concentrations of heavy metals will be different in each plant part, such as roots and leaves which are found to contain higher level of heavy metals than the flower buds and fruits. Moreover, different type of plants species has specific threshold value for the heavy metals where it exerts toxicity (Smical *et al*., 2008).

MATERIALS AND METHOD

Data Samples from Site

In this research, a commonly grown tropical tree, *Cinnamomum iners* which is grown in the main campus of the Universiti Malaysia Sabah (UMS) was chosen. Data mensuration variables were measured nondestructively; namely, diameter at breast height and stem height of tree from twenty-eight *C.iners* trees, planted along the main road of the UMS gateway. Figure 1 below showed the schematic diagram of the sampling area along the UMS gateway. The stem height of each tree was measured from the land to where twigs started to grow using a clinometer, while the diameter at breast height of each tree with a girth tape. Diameter at breast height (Dbh) is often quoted by foresters as technically 1.3 meters from the ground level.

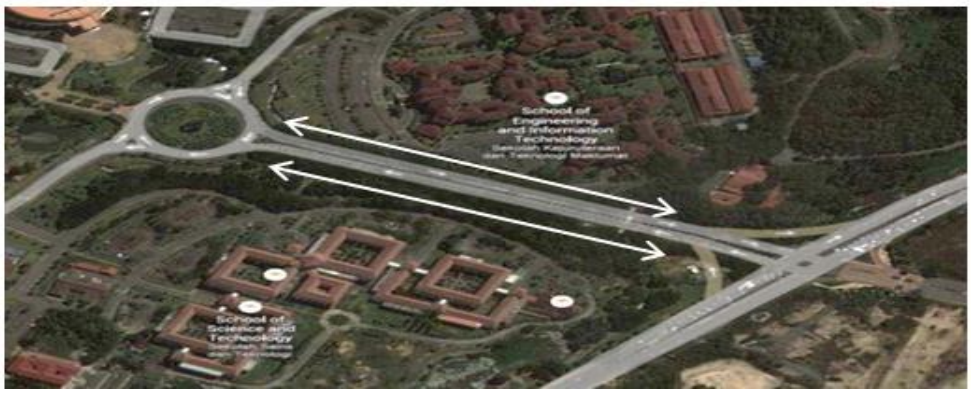

Figure 1. Schematic diagram of sampling area along UMS gateway
(Source: Map Data @ 2013 Google MapIT).

The bark samples of 5-10g were collected at an average height of 1.5 meters above ground level, while the soil samples (5-10g also) were collected within 0-20cm depth from the soil surface in three locations at approximately 1.5 meters radius around the same tree where the barks were taken (Tiina *et al.,* 2013). For the bark sample collection, the outer bark of the *C.Iners* tree was removed first. Then, the inner bark was carefully peeled off using a stainless steel knife at about 5-10g, and then kept in plastic bags with labels corresponding to each tree.

The soil sample was collected from four locations at about 10 cm radially around each tree and within 0-10cm in depth from the surface. The four sites of each soil sample taken was plotted as shown in the Figure 2. The soils were then mixed and combined as one sample from each tree, weighing about 5-10g as the bark samples. There were 28 soil samples from 28 trees collected, and these samples were also kept in plastics and labelled.

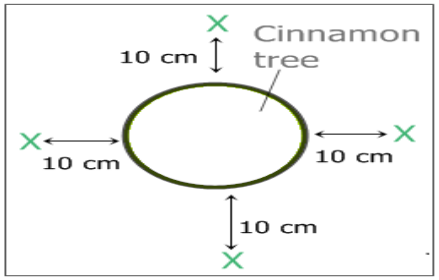

Figure 2. Diagram show the four sites which the soil samples were taken.

Experimental Procedures

Throughout this research, the instrument used was PerkinElmer 5000 Inductively Coupled Plasma Optical Emission Spectrometer (ICP-OES). The chemicals that were used during the experiments was Scanlab brand and 65% nitric acid (HNO_3). The labelled samples were analyzed for the presence of heavy metals using ICP-OES.

Experimental data were collected using the barks and soil samples. Both bark and soil samples were first dried in an oven with a temperature of 70^{o}C for 24 hours and 105^{o}C overnight respectively. After the process of drying, each of the bark and soil sample was grinded into powder and weighed.The soil sample weighed 1.0g, meanwhile the bark sample weighed 0.5g. The powdered samples were then further analysed in a liquid state.

The sample was thus transferred into a 100 ml beaker. The powdered samples were then digested with 20ml of 65% nitric acid (HNO_3) and heated using a hot plate at 170^{o}C in the fume hood. It was heated until a small quantity of solution left in the beaker. Another 10 ml of 65% nitric acid (HNO_3) was added into the remaining solution and it was left to cool. 10 ml of solution was then diluted with distilled water into a 100 ml volumetric flask. Then, the digested sample was filtered using a 0.45μm membrane and heavy metals concentrations were analyzed using Perkin Elmer Optima 5300DV Inductively Coupled Plasma–Optical Emission Spectrometer (ICP-OES) (Huseyin & Mustafa, 2011). Quality Control Standard 21 (Perkin Elmer Pure) was used as Standard Reference Material for instrument recovery. The concentration of heavy metals were calculated by using the following formula (Skoog *et al.*, 2004).

Concentration of heavy metal $= \frac{CVD}{W}$ where,

V= Final volume of solutions
D= Dilution factor of 10
C= Concentration obtained by ICP-OES, mg/L
W= Weight of sample, kg

The process of heavy metal determination were carried out after all the processes of drying and wet digestion had been done. There were six heavy metals needed to be determined from the samples collected. The heavy metals were cadmium (Cd), copper (Cu), iron (Fe), lead (Pb), nickel (Ni) and zinc (Zn).

Mathematical Modelling

Mathematical modeling is an activity or process that allows a mathematician to be a chemist, an ecologist, an economist, a physiologist and so on, by instead of undertaking experiments in the real world, a modeller undertakes experiments on mathematical representations of the real world, herewith, the effects of the different components can be studied and hence, to make predictions about the behaviour (Vries, 2001). In this study, the heavy metal relationships between the tree and soil are exemplified with all the basic theory and mathematical formulations.

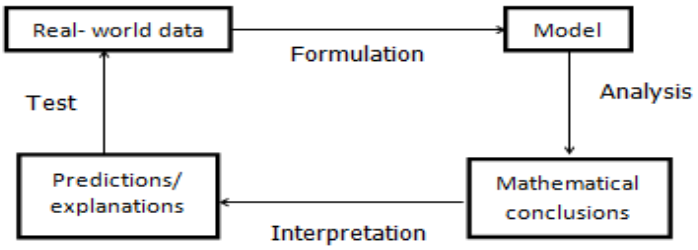

Figure 3. Basic processes in mathematical modeling (Source: Vries, 2001)

In this study, the relationship of heavy metals concentration in barks of the cinnamon tree was determined using the multiple regression (MR) technique. Firstly, the variable selection from site and experiments were made, followed by variables descriptive statistics, and data generation based on the concept of non-parametric bootstrap which will be explained later in this section. Afterwards, model building was also illustrated via the four phase procedures. Lastly, the prediction efficiency was calculated to determine the model's predictive capability.

The calculations on the heavy metals concentrations were used as dependent variable in the multiple regression technique. The dependent variable, Y_i with i=1,2,…,6 , based on the heavy metals were as follows:

a) Concentration of Cadmium in bark (Y_1)
b) Concentration of Copper in bark (Y_2)
c) Concentration of Iron in bark (Y_3)
d) Concentration of Lead in bark (Y_4)
e) Concentration of Nickel in bark (Y_5)
f) Concentration of Zinc in bark (Y_6)

The independent variables based on the site and experimental variables were as follows:

a) Diameter of breast height (w_1)
b) Stem height(w_2)
c) Average ppm in bark (w_3)
d) Average ppm in soil (w_4)
e) Concentration of heavy metal in soil (w_5)

Descriptive statistics were used to describe the basic features of the data by providing simple summaries about the sample and measures with the simple graphic analysis. Quantitative descriptions were also presented in a manageable form and this helped to simplify large amounts of data in a sensible way (Trochim, 2006).

Normality test was carried out for the variables distributions using the Kolmogorov-Smirnov test (KS) and Shapiro-Wilk test (SW). As a guideline, the KS test will be used when the number of observations is large (n>50), while SW test is used when the number of observations is small (n<50).

Data generation refers to the theory and methods used by researchers to create data from a sampled data source in a qualitative study. To generate data from a sampled data source, researcher interacts with the data source using qualitative research methods within an overall strategy of inquiry (Garnham, 2008). Bootstrapping is a numerical sampling technique where data sampled are resampled with replacement and it is mostly used for estimating variance when sampling from an empirical distribution of the observed data (John, 2011). The nonparametric bootstrap is the usual method where it resamples the observations from the original samples while parametric bootstrap method generates the bootstrap observations by a parametric distribution (Saeid *et al.*, 2008). However, non-parametric bootstrap does not require distributional assumptions such as normality distribution. Nonetheless, past research showed that an analysis of the resampling the small sample size to bigger sample size will decrease the proportion of sample to be normality distributed. For example, sampling by 20 observations will generate 74% normality distributed sets, while sampling by 50 observations will only generate 24% normality distributed (Igor *et al.*, 2010). According to Saeid *et al.* (2008), the nonparametric bootstrap is better than parametric bootstrap if the sample kurtosis is less than the kurtosis of distribution.

Multiple Regressions is the extension of simple regression to the case of two or more independent variables relating dependent variable Y_i to k predictor variables $w_1, w_2, \ldots, w_k$. The basic general equation of multiple regressions will be given as:

$$Y_i = \Omega_0 + \Omega_1 W_1 + \Omega_2 W_2 + \ldots + \Omega_k W_k + u_k \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

with the assumption that the random deviation u_k is normally distributed, with zero mean and variance σ^2 for any values of $w_1, w_2, \ldots, w_k$. The outcome or dependent variable is denoted by Y_i where i=1,2,…, 6,

while $w_1, w_2, \ldots, w_k$ represents the set of predictor variables which can be in the form of single independent variables, interaction variables, generated dummy variables and transformed variables where j=1, 2,…,k. The parameter, Ω_0 represents the intercept of the regression equation, and $\Omega_1, \Omega_2, \ldots, \Omega_k$ are the regression coefficients for each predictor variables, $w_1, w_2, \ldots, w_k$ respectively (Zainodin *et al.*, 2011).

In this research, the dependent variable, Y_i stands for the concentration of heavy metals in the *C.Iner*'s bark where i=1,2,…,6 with 1 represents Cadmium, 2 represents Copper, 3 represents Iron, 4 represents Lead, 5 represents Nickel and 6 represents Zinc, while W_1 represents the diameter of breast height, W_2 represents the stem height, W_3 represents the average of ppm in bark, W_4 represents the average of ppm in soil and W_5 represents the concentration of heavy metal in soil, where $W_1, W_2, \ldots, W_5$ are the independent variables before normality test are carried out.

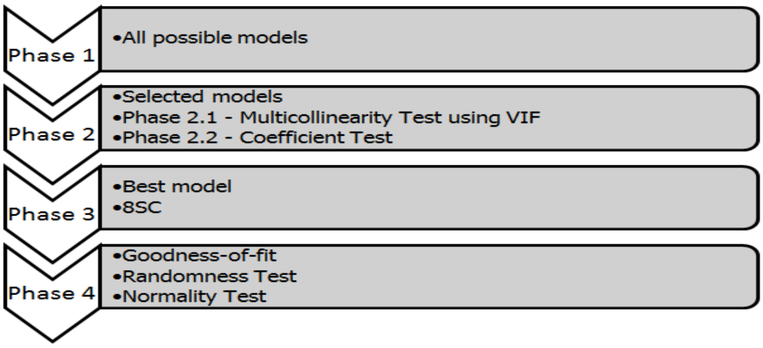

Figure 4. The Four Phases in Model Building Procedure (Source: Zainodin *et al.*, 2011).

In Phase 1, all possible models have to be listed out before analysis is carried out. The number of all possible models can be calculated as follows:

$$N = \sum_{j=1}^{q} j\ ({}^{q}C_{j}) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2)$$

where, N is the number of possible models and q is the number of single quantitative independent variables for j = 1, 2, 3,…, q, which exclude the dummy variables (Zainodin & Khuneswari, 2009).

In Phase 2, multicollinearity is said to exist when one or more of the independent variables is highly correlated with one or more of the other independent variables. This multicollinearity problem can be identified in two ways which are namely, by testing the correlation value (R^2) or through the Variance Inflation Factor (VIF). This study will focus on the VIF rather than the correlation matrix test. Variance Inflation Factor (VIF) is used to measure the impact of collinearity among the independent variables in a multiple regression model. It shows how much the variance of the coefficient estimate is being inflated by the multicollinearity. Kutner *et al.* (2008) stated that "the largest VIF value among all the independent variable is often used as an indicator of the severity of multicollinearity. A maximum VIF value in excess of 10 is frequently taken as an indication of multicollinearity that may be unduly influencing the least squares estimate. "As a rule of thumb, multicollinearity may not be a serious issue if VIF does not exceed 10, although some authors use a more conservative rule that VIF does not exceed 5" (Mohammad & Hong, 2010). The collinearity variable will then be removed from the model.

The VIF value used in this research was set at 5. When an independent variable had the greatest VIF value which is greater than 5 (VIF > 5), it will be removed from the regression model (Zainodin *et al.*, 2015). This process will be carried out using the SPSS and Excel program. For regression models, to determine the VIF value for each independent variable, there are three possibilities which can occur. They are as follows:

Case 1: None of the independent variable has VIF greater than 5. In this case, it will directly proceed to the next step which is the elimination of insignificant variable using the coefficient test.

Case 2: One of the independent variable has VIF greater than 5. In this case, remove the independent variable with VIF greater than 5, and rerun the model.

Case 3: More than one independent variable has VIF greater than 5. In this case, choose the independent variable with the highest VIF value and remove it. Next, rerun the reduce model again. If a tie occurs, remove the variable with higher standard error. The overall VIF test procedures of removing the variables with multicollinearity are as shown in Figure 5.

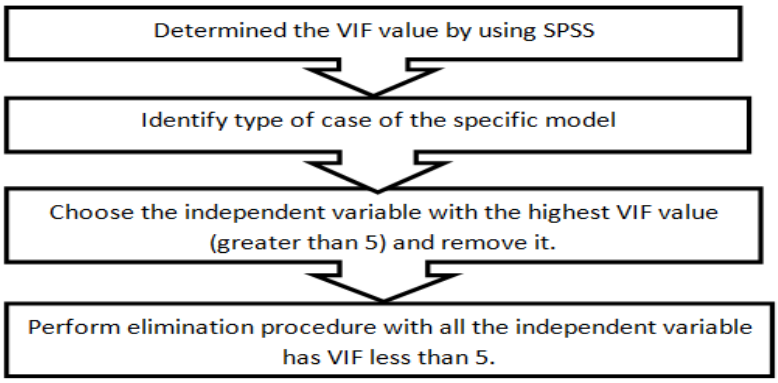

Figure 5. VIF Test Procedures

Coefficient test in Phase 2 is used to test the coefficient of the corresponding variables. Determine the variables which are insignificant variables and eliminate the variable with highest p-value with a condition higher than $\alpha = 0.05$. variable with the smallest $|t_{cal}|$ and nearest to zero will be eliminated from the model. The elimination process is repeated until there is no more insignificant variable in the models (Noraini *et al.*, 2011).

In Phase 3, the Eight Selection Criteria (8SC) were used so as to get the best model. The Finite Prediction Error (FPE) and Akaike Information Criterion (AIC) were developed by Akaike in year 1970 and year 1974 respectively. The Generalised Cross Validation (GCV) was developed by Golub *et.al* in 1979, while HQ criterion was suggested by Hannan and Quinn in the same year. The SHIBATA criterion was suggested by Shibata in 1981. In addition, the RICE criterion was discussed by Rice in 1984 and the SCHWARZ criterion was discussed by Schwarz in 1987. Furthermore, SGMASQ was developed by Ramanathan in 2002. The Eight Selection Criterion (8SC) were shown in the Table 1 below (Zainodin & Khuneswari, 2009), where *n* is number of observations, *(k+1)* is number of estimated parameters, and *SSE* is sum of square error. The model with the least value in majority of the criteria will be chosen as the best model.

Table 1. Eight Selection Criterion (8SC)

AIC : $\left(\frac{SSE}{n}\right) e^{\frac{2(k+1)}{n}}$	RICE : $\left(\frac{SSE}{n}\right)\left(1-\frac{2(k+1)}{n}\right)^{-1}$
FPE : $\left(\frac{SSE}{n}\right)\frac{n+k+1}{n-(k+1)}$	SCHWARZ : $\left(\frac{SSE}{n}\right)(n)^{\frac{(k+1)}{n}}$
GVC : $\left(\frac{SSE}{n}\right)\left(1-\frac{k+1}{n}\right)^{-2}$	SGMASQ: $\left(\frac{SSE}{n}\right)\left(1-\frac{k+1}{n}\right)^{-1}$
HQ : $\left(\frac{SSE}{n}\right)(\ln n)^{\frac{2(k+1)}{n}}$	SHIBATA: $\left(\frac{SSE}{n}\right)\frac{n+2(k+1)}{n}$

Finally, the normality of the regression model in Phase 4 can be obtained by using the Kolmogorov-Smirnov statistics (KS) and the histogram of the standardized residuals. The Kolmogorov-Smirnov statistics was used to test the normality of the residuals since the number of observations was large (n=450) after bootstrapping. If the p-value was greater than α=0.05, the null hypothesis would be accepted where it showed that the residual were assumed to be normally distributed. The histogram of the standardized residuals would looked like a bell-shape and can be used as a supporting evidence for the normality test (Noraini *et al.*, 2011). Similar modelling procedures were carried out on all the heavy metals concentrations in barks of the *C.iners*.

Lastly, the prediction efficiency is used to forecasting the value of some reserved variable, and some others actual value for the same variable. Let F_t denote the forecast value and let A_t denote the actual value of the variable, and n as the number of reserved observations for prediction. A forecasting criteria, namely the mean absolute percentage error (MAPE) is calculated in order to indicate that the best model can give

the most accurate result on the relationship between the heavy metal concentration in the *C.iners* bark and soil. MAPE is the more objective statistics indicator because the measure is in relative percentage and will not be affected by the unit of the forecasting series. The closer the MAPE approaches zero, the better would be the forecasting results. MAPE is commonly used in quantitative forecasting methods because it produces a measure of the relative overall fit. The MAPE value is calculated using the following formula:

$$MAPE = \left(\frac{100\ \%}{n}\right)\sum_{t=1}^{n}\left|\frac{(A_t - F_t)}{A_t}\right| \quad \text{.....(3)}$$

Where A_t is the actual value of the reserved data, F_t is the forecasted value of the reserved data, n is the number of reserved data with t=1,….,n (Noraini *et al*., 2015).

RESULTS AND DISCUSSIONS

The field and experimental data variables were initially tested for their normality distributions. In this study, the sample size was small (n=28) had been collected Since the sample size of the data collected were small, hence not conforming to normality. They thus had to undergo the method of non-parametric bootstrap resampling which was generated up to 500 observations where 450 observations were used for modelling, while the other 50 observations were randomly selected and used for prediction efficiency. From Table 2, the entire sample kurtosis statistics were seen to be less than the kurtosis of the distribution. It therefore fulfilled the condition that nonparametric bootstrap was better than the parametric bootstrap (Saied *et al.*, 2008).

Table 2. Kurtosis of Dependent Variable (Copper) and Independent Variables

Variables	Kurtosis	Sample kurtosis	Non-parametric
$\mathbf{Y_2}$	0.6127	0.3504	✓
$\mathbf{W_1}$	-1.0137	-1.0522	✓
$\mathbf{W_2}$	0.0159	-0.0923	✓
$\mathbf{W_3}$	0.5900	-0.2756	✓
$\mathbf{W_4}$	-0.5824	-0.6951	✓
$\mathbf{W_5}$	-0.5824	-0.7232	✓

After the bootstrapp resampling, all the MR models had to undergo the four phase model building procedures. From equation (2), the number of all possible models based on 5 single independent variables were 80 models. Phase 2 involves the removals of multicollinearity source variables. In this phase, VIF value greater than 5 were then removed from the parent model. Next, the coefficient test were carried out by eliminating the source variables which had p-values greater than 0.05. For illustration purposes, the dependent variable Y_2 (Copper) and parent model M73 was thus chosen as:

M73-Y_2 = f (W_1,W_2, W_3, W_4, W_5, W_{12}, W_{13}, W_{14}, W_{15}, W_{23}, W_{24}, W_{25}, W_{34}, W_{35}, W_{45}, W_{123}, W_{124}, W_{125}, W_{134}, W_{135}, W_{145}, W_{234}, W_{235}, W_{245}, W_{345}) …… (4)

Table 3 show the regression output for the model M73 for the dependent variable, Y_2. By looking at the VIF column, variable $\boldsymbol{W_3}$ has the highest VIF value among the other variables. Hence, variable $\boldsymbol{W_3}$ was removed from the model and rerun the model again. Now the model M73 will thus change to M73.1, which indicates one variable has been removed from the model.

Table 3. The regression output for model M73.

Variables	Coefficient	Std. Error	P-value	VIF	Action
Constant	10.2114	15.2458	0.5034		
$\boldsymbol{W_1}$	-0.1845	0.5066	0.7159	6790.3439	
$\boldsymbol{W_2}$	-2.2100	6.7212	0.7425	1403.6223	
$\boldsymbol{W_3}$	722.1238	889.8237	0.4175	9637.0897	**Removed**

W_4	-204.7467	177.0946	0.2483	7880.7214
W_5	-1.9262	1.8119	0.2883	7606.6995
W_{12}	0.0100	0.2267	0.9650	7080.6368
W_{13}	-26.8581	23.4314	0.2523	5066.1883
W_{14}	6.9478	5.2444	0.1859	5558.2762
W_{15}	0.0330	0.0596	0.5801	7242.8308
W_{23}	-412.4122	379.3438	0.2776	8787.8700
W_{24}	82.2754	76.7771	0.2845	7285.5454
W_{25}	0.6511	0.8030	0.4179	7417.9125
W_{34}	-5,912.3325	8,282.2048	0.4757	5206.1112
W_{35}	80.0803	80.0362	0.3176	4562.8398
W_{45}	16.9298	17.5925	0.3364	4709.3777
W_{123}	17.4627	10.0847	0.0841	4512.7181
W_{124}	-2.7208	2.1497	0.2063	4551.5194
W_{125}	-0.0034	0.0254	0.8936	6433.5901
W_{134}	-46.0837	116.7126	0.6932	623.2095
W_{135}	-1.6202	1.1628	0.1643	609.7922
W_{145}	-0.0719	0.2686	0.7892	724.2550
W_{234}	2,813.5275	3,473.2095	0.4184	4438.2772
W_{235}	-24.4780	34.4610	0.4779	3950.4274
W_{245}	-6.3184	7.8736	0.4227	4456.7351
W_{345}	-3.3527	371.7082	0.9928	465.6279

Table 3 shows that the variable W_{15} was removed from the model since it has the highest VIF value among the other variables. The model was again rerun to obtain the new regression output. The process will continue until all the VIF values for the variables are less than 5 which indicated no multicollinearity effects exist in the model. Table 4 illustrated the regression outputs from model M73.17 until M73.19. Similar procedures on removal of multicollinearity source variables were carried out.

Table 4. Removal of multicollinearity source variable from M73.17 until M73.19.

Model	Variables	Coefficient	Std.Error	p-value	VIF	Action
M73.17.	**Constant**	1.0114	0.5545	0.0688		
	W_1	0.0033	0.0191	0.8646	9.618	
	W_2	0.5535	0.2286	0.0159	1.625	
	W_5	0.0236	0.0632	0.7090	9.259	
	W_{24}	-2.1040	2.8119	0.4547	9.785	
	W_{134}	-7.5157	16.6528	0.6520	12.703	
	W_{135}	-0.0148	0.1593	0.9262	11.456	
	W_{145}	0.0434	0.0388	0.2642	15.150	**Removed**
	W_{345}	-48.7001	66.4007	0.4637	14.877	
M73.18.	**Constant**	0.6729	0.4647	0.1483		
	W_1	0.0186	0.0132	0.1614	4.643	
	W_2	0.4469	0.2078	0.0320	1.342	
	W_5	0.0716	0.0463	0.1228	4.976	
	W_{24}	0.0234	2.0708	0.9910	5.303	
	W_{134}	-8.2980	16.6428	0.6183	12.681	
	W_{135}	-0.0358	0.1582	0.8213	11.297	
	W_{345}	-35.9082	65.4259	0.5834	14.435	**Removed**

	Constant	0.6127	0.4512	0.1752		
	W_1	0.0245	0.0077	0.0016	1.569	
	W_2	0.4775	0.2000	0.0174	1.246	
M73.19.	W_5	0.0589	0.0400	0.1421	3.721	
	W_{24}	-0.5007	1.8360	0.7852	4.175	
	W_{134}	-13.3159	13.8956	0.3384	8.854	**Removed**
	W_{135}	-0.0773	0.1389	0.5782	8.715	

Table 5. Model M73.20.0 free from multicollinearity effects and insignificant variables.

Variables	Coefficient	Std.Error	p-value	VIF
Constant	0.4871	0.4317	.2598	
W_1	0.0232	0.0076	.0023	1.522
W_2	0.5477	0.1861	.0034	1.079
W_5	0.0873	0.0269	.0013	1.680
W_{24}	-2.0155	0.9338	.0314	1.080
W_{135}	-0.1918	0.0706	.0068	2.254

Table 5 showed the model M73.20. which was free from multicollinearity effects, had twenty of its independent variables removed from the parent model M73, due to multicollinearity and no insignificant variables were eliminated. Model M73.20.0 can thus be given as: $\hat{Y}_2 = f(W_1, W_2, W_5, W_{24}, W_{135})$ ……(5)

It could also be seen in Table 5 that model M73.20 was affected by single independent variables (W_1, W_2 and W_5), first order interact variable (W_{24}) and second order interaction variable (W_{135}). The negative coefficients of variables (W_{24}) and (W_{135}) respectively implied negative contribution of the heavy metal (Copper) concentration, while the others implied positive impacts. Table 5 thus implicated that best model M73.20.0 was free from multicollinearity effects when all the p-values were more than 0.05 and the VIF were less than 5.0. The regression equation of the best model is given by:-

$$Y_{Cu} = 0.4871 + 0.0232\ W_1 + 0.5477\ W_2 + 0.0873\ W_5 - 2.0155\ W_{24} - 0.1918\ W_{135} \quad \text{.....(6)}$$

The all possible 80 models were reduced to 34 selected models after carrying out the four-phase modelling procedures. Table 6 below showed that the best model chosen for Copper based on the eight selection criteria (8SC) was M73.20.0.

Table 6. The corresponding selection criteria values for selected models for Copper.

Model	k+1	SSE	AIC	FPE	GCV	HQ	RICE	SCHWARZ	SGMASQ	SHIBATA
M2.0.0	2	219.9888	0.4932	0.4932	0.4932	0.4968	0.4932	0.5023	0.4910	0.4932
M3.0.0	2		0.4927	0.4927	0.4927	0.4962	0.4927	0.5017	0.4905	0.4926
↓	:	:	:	:	:	:	:	:	:	:
M25.0.0			0.4866	0.4866	0.4867	0.4937	0.4867	0.5048	0.4824	0.4866
↓	:	:	:	:	:	:	:	:	:	:
M45.3.0			0.4868	0.4868	0.4868	0.4938	0.4868	0.5049	0.4825	0.4867
↓	:	:	:	:	:	:	:	:	:	:
M55.6.1			0.4849	0.4849	0.4850	0.4920	0.4850	0.5030	0.4807	0.4849
↓	:	:	:	:	:	:	:	:	:	:
M73.20.0	6	210.4851	0.4804	0.4804	0.4805	0.4909	0.4806	0.5074	0.4741	0.4802
↓	:	:	:	:	:	:	:	:	:	:
M80.26. 2	4	213.5004	0.4830	0.4830	0.4830	0.4900	0.4830	0.5009	0.4787	0.4829

For model's goodness-of-fit, the runs test for randomness and normality tests were carried out on the standardized residuals of model M73.20.0. The hypothesis statements were shown as follows:

H_0: The standardized residual, u_i are randomly distributed.

H_1: The standardized residual, u_i are not randomly distributed.

The runs test of Table 7 below showed that the |Z| value of standardized residuals of model M73.20.0 was less than the significant value; hence, the null hypothesis was accepted. It could be thus concluded that the standardized residuals for dependent variable, Y_2 was randomly distributed. The scatter plot as shown in Figure 6 was the supporting evidence that the standardized residual was randomly distributed with the upper control limit (UCL) and lower control limit (LCL) within ± 3 standard deviation.

Table 7. Runs test for standardized residual, u_i.

	Standardized
Test Value	-0.269
Cases < Test Value	213
Cases >= Test Value	237
Total Cases	450
Number of Runs	215
Z	-0.981
Asymp. Sig. (2-tailed)	0.327

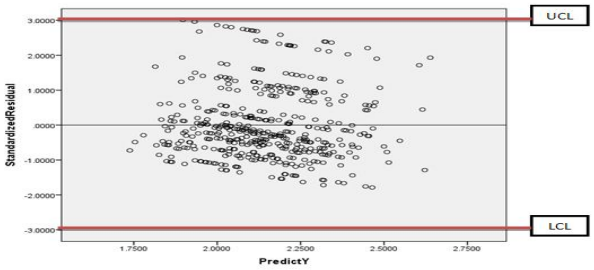

Figure 6. Randomness plot for standardized residuals.

The hypothesis statements for normality test were shown as follows:

H_0: The standardized residual, u_i are normally distributed.

H_1: The standardized residual, u_i are not normally distributed.

Since the sample size was 450, the normality test using Kolmogorov-Smirnov statistics of value (0.123) with a significant value of $p < 0.0001$ was obtained. Since the p-value was less than 0.05, the residuals were not normally distributed. Figure 7 showed that the normality plot was skewed. This could be explained by the non-parametric bootstrapping which did not consider the assumption of normality (Saeid *et al*., 2008).

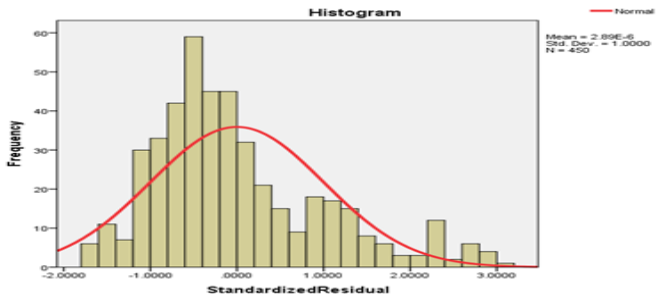

Figure 7. Normality plot for standardized residuals.

The forecasting criteria namely, the Mean Absolute Percentage Error (MAPE) was used for the prediction efficiency. Using equation (3), the value of MAPEs for all the heavy metal concentrations were calculated and comparisons were made as shown in Table 8 below. The heavy metal concentration, Y_2 (Copper) was equal to 22.8096% (since < 25%), hence, was acceptable.

Heavy Metal	Dependent Variable	MAPE	Action
Cadmium	Y_1	108.2176	Not accurate
Copper	Y_2	22.8096	Acceptable
Iron	Y_3	25.0080	Not Acceptable
Lead	Y_4	32.9134	Not Acceptable
Nickel	Y_5	28.1237	Not Acceptable
Zinc	Y_6	38.6099	Not Acceptable

Therefore, the best model for Y_2 (Copper) was thus said to be acceptable and considered as accurate. In other words, the concentration of Copper had the most effect on the heavy metal concentration in the barks of *Cinnamomum iners*.

CONCLUSION

The multiple regression analysis can be used to determine the factors that affect the heavy metal concentration in the cinnamon tree barks. Factors such as, the diameter of breast height (W_1), stem height (W_2) and concentration of copper in soil (W_5) had positive relationships with the concentration of copper in bark, while the first order interaction variable of stem height and average ppm in soil (W_{24}) and second order interaction variable of diameter of breast height with average ppm in bark and concentration of copper in soil (W_{135}) had negative relationships with the concentration of Copper (Y_2) in bark. The regression equation can thus be used to predict the amount of this toxic heavy metal, that is Copper, absorbed by the cinnamon tree, and hence further its toxicity. It is also suggsted that further works can be done using these model building procedures via VIF multicollinearity test on parts of other plants or trees in identifying heavy metals concentrations.

CONFLICTING OF INTEREST

The authors hereby declared that this research is partially funded by the Universiti Malaysia Sabah where experiments had been carried out in the Chemical Analysis Laboratory of the university.

REFERENCES

1. Akaike, H. 1970. Statistical Predictor Identification. *Institute of Statistical Mathematics*, **22**(2): 203-217.
2. Akaike, H. 1974. A New Look at model Statistical Identification. *IEEE Transactions on Automatic Control*, **19**:716-723.
3. Agarwal, K. S. 2009. Heavy Metal Pollution. A.P.H Publishing Corporation. New Delhi.
4. Drupal. 2014. Department of Statistics. The Pennsylvania State University. Available at: https://onlinecourses.science.psu.edu/stat501/node/190. Retrieved: 9 May 2014.
5. Garnham, B. 2008.Data Generation.*The SAGE Encyclopedia of Qualitative Research Methods*. Available at: http://srmo.sagepub.com/view/sage-encyc-qualitative-research-methods/n97.xml. Retrieved : 19 October 2013.
6. Golub, G.H., Heath, M. & Wahba, G. 1979.Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2): 215-223.
7. Google Maps. 2013. Gateway of School of Science and Technology, and School of Engineering and Information Technology Information, University Malaysia Sabah. Street map. Available at: https://maps.google.com/maps?hl=en&authuser=0&q=School+of+science+and+Technology+and+School+of+Engineering+and+Information+gateway+in+Universiti+Malaysia+Sabah&ie=UTF8&ei=qJKHUqKLJMSMrQejrYDADQ&ved=0CAgQ_AUoAg. Retrieved: 13 October 2013.
8. Hannan, E.J. & Quinn, B.G. 1979.The Determination of the Order of an Autoregression.*Journal of Royal Statistics Society*, 41(2):190-195.
9. Hogan, C.M. 2010. Heavy metal. *The Encyclopedia of Earth*. http://www.eoearth.org/view/article/153463.html. Retrieved: 14 October 2013.
10. Huseyin, A. & Mustafa, T. 2011. Comparison of Dry, Wet & Microwave Digestion Methods for the Multi Element Determination in Some Dried Fruit Samples by ICP-OES. *Food & Chemical Toxicology*, 49: 2800–2807.
11. Igor, Y.P., Andrew, R.W., & Julio, C.D. 2010. Resampling Approach for Determination of the Method for Reference Interval Calculation in Clinical Laboratory Pratice. Journal of American Society for Microbiology, 17(8): 1217-1222.
12. John, A.R.J. 2011. Bootstrapping Analysis, Inferential Statistics and EXCEL.Spreadsheets in Education (eJSiE), 4(3):1-23.
13. Kutner, M.C., Nachtsheim, C., & Neter, J. 2008. Applied Linear Regression Models.4th edition. McGraw Hill, Inc, New York, pp:409.
14. Mohammad, A. & Hong, S.N. 2010. Do Instructional Attributes pose Multicollinearity Problems? An Empirical Exploration.*Economic Analysis and Policy*, 40 (3): 351-361.

15. Mitsios, I.K., Golia, E.E. & Tsadilas, C.D. 2005. Heavy metal concentrations in soils and irrigation waters in Thessaly Region, Central Greece. *Communications in Soil Science and Plant Analysis*, 36 (4-6): 487-501.
16. Noraini, A., Zainodin, H.J., & Nigel, J.J.B. 2008. Multiple Regression Models of the Volumetric Stem Biomass. WSEAS Transaction on Mathematics, 7(7): 492-502.
17. Noraini, A., Amran, A., & Zainodin, H.J. 2011.An Improved Volumetri Estimation Using Polynomial Regression. *Journal of Science and Technology*, *UTHM*, 3(2): 29-42.
18. Ramanathan, R. 2002. *Introductory Econometrics with Applications.* 5th Ed. South-Western, Thomson Learning,Ohio.
19. Rice, J. 1984. Bandwidth Choice for Nonparametric Kernel Regression.*The Annals of Statistics*,12(4):1215-1230.
20. Saeid, A., Dietrich, V.R., & Silvelyn, Z. 2008. On the comparison of parametric and nonparametric bootstrap.*Department of Mathematics Uppsala University*, pp:1-13.
21. Schwarz, G. 1978. Estimating the Dimension of a Model.*Annals of Statistics*, **6**(2):461-464.
22. Shibata, R. 1981. An Optimal Selection of Regression Variables.*Biometrika*, **68**(1):45-54.
23. Skoog, D., Jarey, H. & Stanley, C. 2007. Principles of Instrumental Analysis 6th Edition. *Brooks/Cole Cengage Learning*, United States of America.
24. Smical, A.I., Hotea, V., Oros, V., Juhasz, J., and Pop, E. 2008. Studies on transfer and bioaccumulation of heavy metals from soil into lettuce. *Environmental Engineering and Management Journal*, **7**(5): 609-615.
25. Tiina, M., Nieminen, Kirsti, D., Henning, M. & Bruno D.V. 2013. Chapter 16 – Soil Solution: Sampling & Chemical Analyses. *Environmental Science*, **12**: 301–315.
26. Trochim, W.M.K. 2006. Descriptive Statistics.Research Methods, Knowledge Base. Available at: http://www.sosialresearchmethods.net/kb/statdesc.php. Retrieved: 13 October 2013.
27. Vries, G. 2001. *What is Mathematical Modeling*. Department of Mathematical Sciences. University of Alberta. Canada. Available at: http://www.math.ualberta.ca/~devries/erc2001/slides.pdf. Retrieved: 13 October 2013.
28. Zainodin, H.J., Noraini, A., & Yap, S.J. 2011. An Alternative Multicollinearity Approach in Solving Multiple Regression Problem. *Trends in Applied Sciences Research*, **6**(11): 1241-1255.
29. Zainodin, H.J. & Khuneswari, G. 2009. A Case Study on Determination of House Selling Price Model Using Multiple Regressions. *Malaysian Journal of Mathematical Sciences*,3(1):27 -44.
30. Zainodin, H.J., Khuneswari, G., Noraini, A. & Haider, F.A.A. 2015. Selected Model Systematic Sequence via Variance Inflationary Factor. International Journal of Applied Physics and Mathematics, 5(2): 105-114.
31.
32. Zhao, K., Liu, X.M., Xu, J.M. & Selim, H.M. 2010. Heavy Metal Contaminations In a Soil-rice system: Identification of Spatial dependence in relation to soil properties of paddy fields. *Journal of Hazardous Materials*, 181:778-787.