JSc EUSL(2018) Vol. 9 No. 1, p 12-30,

Print- ISSN: 1391-586X, E- ISSN: 2602-9030,

NOVEL INTERVAL MULTIPLE LINEAR REGRESSION MODEL TO ASSESS THE RISK OF INVASIVE ALIEN PLANT SPECIES

Peiris, H.O.W[1]., Chakraverty, S[2]., Perera, SSN[3*]., and Ranwala, SMW[4]

[1]Department of Mathematics, Faculty of Natural Sciences, The Open University of Sri Lanka, Sri Lanka.

[2]Department of Mathematics,National Institue of Technology, Rourkela, India

[3]Research & Development centre for Mathematical Modelling, Department of Mathematic, University of Colombo, Sri Lanka

[4]Department of Plant Sciences, University of Colombo, Sri Lanka

ABSTRACT

Invasive Alien Species (IAS) can be considered as a serious threat to the existence of the environment as they alter physical, chemical and biological components of the environment. Invasive potential of species can be recognized by their biological traits. Therefore, it is very important to model the risk of species using biological traits before going into a new environment. The purpose of this study is to build interval multiple linear regression models with interval input-output data to evaluate invasion risk of IAS related to biological traits. A new method has been proposed to estimate the interval regression coefficients. Two different regression models are developed using interval least square algorithm. In the first model we use the method developed by Chenyi Hu and the second model is newly developed. The estimated accuracy of the model that is developed by the proposed method is higher in comparison to the model with Chenyi Hu method. These two models are validated using known invasive and non-invasive species. The model that incorporates the proposed method provides better prediction of risk of IAS.

Keywords: Invasive Alien Species; Interval mulitiple linear regression; Interval least square.

[*]*Corresponding author*: ssnp@maths.cmb.ac.lk

https://orcid.org/0000-0002-6484-6000

1. INTRODUCTION

Invasive Alien Species (IAS) can be considered as a serious threat to the environment as they alter physical, chemical and biological components of the environment [3]. Risk assessment is a key tool to identify the potential Invasive Alien Species (IAS). Many countries all over the world are now practicing this tool as a vital part of comprehensive IAS prevention strategy. These risk assessments are in the form of questionnaires based on risk factors of IAS with predefined answers. The outcome of the assessment will be the risk value of a particular IAS, which is the sum of scores that have been given to each question by the domain expert. Nevertheless, most of the risk factors which affect the invasiveness of species are accompanied by imprecision and uncertainty. Data is usually collected by groups of experts rather than the individuals, due to the unavailability or lack of proper mechanism to directly measure biological traits [7]. Therefore it is very important to develop a mathematical model that incorporates the uncertainty and imprecision of input data to evaluate the risk of IAS efficiently.

In Multivariate Linear Regression (MLR), the relationship between the response variable and several explanatory variables is investigated. These variables usually are single-valued. The need of interval-valued data may arise in an effort to deal with the imprecision and uncertainty in obtaining reliable approximations. In fact, the minimum and maximum recorded values offer a fair insight into the phenomenon rather than considering the average values. This study is focused on developing a mathematical model to evaluate satisfactory interval approximations for risk of IAS using multivariate linear regression. The model can be used to predict reliable risk range of potential IAS rather a single-valued risk score. The model parameters can be found by applying interval least square method. There are some methods to solve interval least square with complex computational methods [2]. In this work, we have developed two different interval multivariate regression models. The difference between these two models is the approach used to estimate the unknown regression coefficients. We have used interval least square algorithm proposed by Chenyi Hu [4] as the model I, and proposed a new method, viz., model II to estimate the interval-valued regression coefficients.

Biological traits play a vital role in identifying invasive characteristics of plant species. This article provides mathematical risk assessment model for the first time with the application of interval multivariable regression analysis. As we mentioned above, two different models have been developed to study the relationship among the biological

traits subject to invasion risk. The models have been validated by testing a set of well-known invasive plant species and non- invasive species in Sri Lanka.

2. INTERVAL MULTIVARIATE LINEAR REGRESSION

2.1 Basic concepts of multivariate linear regression

Multivariate regression is an extension of simple linear regression in which more than one independent variable (x) is used to predict a single dependent variable (y) [9]. The predicted value of y is a linear combination of the x variables such that the sum of squared deviations of the observed and predicted y is a minimum. The multiple linear regression equation for p variables is as follows:

$$y_i = a_0 + a_1 x_{i1} + a_2 x_{i2} + ... + a_p x_{ip}\ , \quad i = 1, 2, ..., n. \qquad (2.1)$$

where y_i is the predicted or expected value of the dependent variable, x_{i1} through x_{ip} are p distinct independent or predictor variables, and a_0 through a_p are the regression coefficients to be estimated utilizing the input and output data. The regression coefficients $\left(a_i, i = 1, 2, ..., p\right)$ can be determined by using least squares estimation.

2.2 Least square estimation

This method can be used to estimate the regression coefficients in the linear regression model by minimizing the squared discrepancies between observed data and their expected values [8]. To find the least square estimators, it will be more convenient to use matrix notation of Eq. (2.1).

Let Y be a $p-$dimensional vector consisting of expected values, X is an $n \times p$ data matrix of n observations of the p variables, then we have

$$X = \begin{pmatrix} 1 & x_{12} & \cdot & \cdot & \cdot & x_{1p} \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \mathrm{i} & x_{n2} & \cdot & \cdot & \cdot & x_{np} \end{pmatrix} \qquad (2.2)$$

a is $p-$dimensional vector of unknown regression coefficients.

The least square estimator, denoted by $\hat{a}$ is that the value of a that minimizes the

$$\|Y - Xa\|^2 \tag{2.3}$$

Suppose X has full column rank, that is no column in X can be written as a linear combination of other columns. Then the least square estimator $\hat{a}$ is given by

$$\hat{a} = \left(X^T X\right)^{-1} X^T y \tag{2.4}$$

Next we define the interval multiple linear regression model for interval input-output data.

2.3 Interval multivariate linear regression model

The interval MLR model can be defined as follows:

For the situation of a linear dependence of a variable $\tilde{Y}$ on m variables $\tilde{\mu}_i = 1, 2, ..., m$ is of form

$$\tilde{Y} = \theta_0 + \theta_1 \tilde{X}_{i1} + \theta_2 \tilde{X}_{i2} + ... + \theta_m \tilde{X}_{im}, \quad i = 1, 2, ..., n. \tag{2.5}$$

Assume that for the variables $\tilde{X}_{i1}, \tilde{X}_{i2}, ..., \tilde{X}_{im}$, we are given interval observations $\left[\underline{X}_{mi}, \bar{X}_{mi}\right]$, $i = 1, 2, ..., N > m.$ Similarly the dependent variable $\tilde{Y}$ are considered as $\left[\underline{y}_i, \bar{y}_i\right]$, $i = 1, 2, ..., N.$ The interval vector $\theta = \left(\theta_0, \theta_1, \theta_2, ..., \theta_m\right)^T$ represents the unknown regression coefficients.

2.4 Interval least square

The concept of interval least squares can be applied to the interval-valued linear system in Eq. (2.5) to obtain interval estimates of θ. In the present study, we basically concern on minimizing the overall absolute error of approximation.

Let us define the absolute error of interval estimation as in [5].

Definition 2.1 [5]: *Let interval* $\hat{y} = \left[\underline{\hat{y}}, \bar{\hat{y}}\right]$ *be an estimation of an interval* $y = \left[\underline{y}, \bar{y}\right]$. *The left and right absolute errors are* $E_L = \left|\underline{\hat{y}} - \underline{y}\right|$ *and* $E_R = \left|\bar{\hat{y}} - \bar{y}\right|$, *respectively.*

The absolute error of the estimation is the sum of left and right absolute errors, that is, $E = E_L + E_R = \left|\hat{\underline{y}} - \underline{y}\right| + \left|\hat{\bar{y}} - \bar{y}\right|$, *respectively.*

Using the Definition 1, we define sum of squares error (SSE) of interval-valued multiple linear regression system as in [4]:

Definition 2 [4]: *Let* U *be the set of* n *interval valued observations of an interval linear regression function* $y = h(x)$ *i.e.*

$U = \left\{(x, y) : x \subset \mathbb{R}^n,\ y \subset \mathbb{R},\ \text{both } x \text{ and } y \text{ are compact}\right\}$. *We say that* $\sum_{0 \le j \le m} \tilde{\theta}_j \tilde{X}_{ij}$ *is an interval least squares approximation of* $h(x)$ *if the linear combination minimizes:* $\sum_{i=1}^{N} E^2{}_L + \sum_{i=1}^{N} E^2{}_R$,

where

$$\sum_{i=1}^{N} E^2{}_L = \sum_{i=1}^{N} \left(\underline{y}_i - \left(\sum_{0 \le j \le m} \underline{\theta_j X_{ij}} \right) \right), \tag{2.6}$$

and

$$\sum_{i=1}^{N} E^2{}_R = \sum_{i=1}^{N} \left(\bar{y}_i - \left(\sum_{0 \le j \le m} \overline{\theta_j X_{ij}} \right) \right). \tag{2.7}$$

2.5 Interval arithmetic

While using ILS method, it is required to find the estimates for interval regression coefficients that minimize the sum of equations. (2.6) and (2.7). Therefore we need to use interval arithmetic to work with interval-valued data [1]. Let $[a] = [\underline{a}, \bar{a}]$, $[b] = [\underline{b}, \bar{b}]$ be real compact intervals and $\circ$ is one of the basic operations 'addition','substraction','multiplication' and 'division', respectively (for real numbers) that is $\circ \in \{+, -, \cdot, \div\}$.

Then, $[a] \circ [b] = \{a \circ b \mid a \in [a], b \in [b]\}$. if $\circ$ is $\div$ then $0 \notin [b]$.

for the corresponding operations:

$$[a]+[b]=[\underline{a}+\underline{b},\bar{a}+\bar{b}],$$

$$[a]-[b]=[\underline{a}-\underline{b},\bar{a}-\bar{b}],$$

$$[a]\cdot[b]=\left[\min\left\{\underline{a}\cdot\underline{b},\underline{a}\cdot\bar{b},\bar{a}\cdot\underline{b},\bar{a}\cdot\bar{b}\right\},\max\left\{\underline{a}\cdot\underline{b},\underline{a}\cdot\bar{b},\bar{a}\cdot\underline{b},\bar{a}\cdot\bar{b}\right\}\right],$$

$$[a]\div[b]=\left[\min\left\{\underline{a}\div\underline{b},\underline{a}\div\bar{b},\bar{a}\div\underline{b},\bar{a}\div\bar{b}\right\},\max\left\{\underline{a}\div\underline{b},\underline{a}\div\bar{b},\bar{a}\div\underline{b},\bar{a}\div\bar{b}\right\}\right],\text{ provided } 0\notin[b].$$

In the subsequent heads we now incorporate the two models as mentioned earlier.

3. MODEL I

In this section, we approximate the interval regression coefficients using ILS Algorithm given in [4].

3.1 Initial solution

Let $\tilde{X}\theta=\tilde{Y}$ be the interval linear equations. It is a computational challenge to find θ such that $\tilde{X}\theta\approx\tilde{Y}$. Also it is very difficult to solve $\tilde{X}^T\tilde{X}\theta=\tilde{Y}$ directly or performing the QR factorization on interval-valued matrix X. Therefore, this work mainly aims to find approximated solution to meet our needs. First we find the midpoint solution for θ in the interval normal equation $\tilde{X}\theta=\tilde{Y}$ as the initial point solution. The algorithm for finding midpoint solution is given in [4] as follows:
For a given set of interval-valued pairs (x_i,y_i)

- Evaluate the interval matrix $\tilde{X}$ as defined in (2.2).
- Perform QR factorization on midpoint matrix, X_{mid} of $\tilde{X}$ such that $X_{\text{mid}}=QR$.
- Calculate $c=Q^T\tilde{Y}$ with interval arithmetic.
- Compute midpoint solution of θ by solving $R\theta=c_{\text{mid}}$ with backward substitution.

3.2 Width adjustments for initial point solution

One may note that the initial point solution which is obtained from the above algorithm is a point solution. Therefore we cannot directly apply this solution to the interval linear equations. Accordingly, sequence of $\varepsilon-$inflations of 1%, 2%,..., 10% have been used to form various interval-valued θ.

Here we have considered the concept of ratio of estimation to find a satisfactory approximated solution. Applying the concept of volume of an interval vector $V=\left(q_1,q_2,...,q_n\right)$, which is $\upsilon(q)=\prod_{1\le i\le n}\left(\bar{q}_i-\underline{q}_i\right)$, the notion of ratio of estimation for an approximated solution of interval linear systems $X\theta=Y$ is defined in [5] as:

$$r=\begin{cases} 0\% \quad \text{if} \quad X\theta\cap Y=\varnothing; \\ 100\% \quad \text{if } X\theta=Y; \\ \dfrac{\upsilon(Y)}{\upsilon(X\theta)} \quad \text{if } X\theta\supset Y; \\ \dfrac{\upsilon(X\theta)}{\upsilon(Y)} \quad \text{if } X\theta\subset Y; \\ \dfrac{\upsilon(X\theta\cap Y)}{\upsilon(X\theta\cup Y)} \quad \text{otherwise} \end{cases} \tag{3.1}$$

The ratio estimation has been performed for each of the approximated solution which have been obtained by various interval-valued θ. The approximated solution would be better if the ratio estimation is large. Therefore, we have chosen the interval-valued θ which gives the highest ratio estimation as the approximated interval regression coefficients.

Below we now propose a new method for finding estimates for interval regression coefficients.

4.MODEL II

The problem in the model I is the adjusting the widths of center values of regression coefficients using $\varepsilon-$inflations of 1%, 2%,..., 10%. But, if the center values of coefficients and boundaries of input and output data are positive we can directly find the boundary estimations for regression coefficients.

In this section we propose a new method to estimate the interval regression coefficient of a multivariate regression model which satisfies certain conditions.

Let us consider the interval multivariate regression model as

$$\left[\underline{y}_i, \overline{y}_i\right] = [\underline{a}_0, \overline{a}_0] + [\underline{a}_1, \overline{a}_1]\left[\underline{x}_{i1}, \overline{x}_{i1}\right] + \cdots + [\underline{a}_m, \overline{a}_m]\left[\underline{x}_{im}, \overline{x}_{im}\right] \quad \text{for } i = 1, 2, ..., n. \tag{4.1}$$

The center values of interval regression coefficients $[\underline{a}_j, \overline{a}_j]$ for $j = 0, 1, ..., m$ are considered as a'_m. We consider $\varepsilon_j > 0$ as a value that satisfies:

$$\underline{a}_j = a'_j - \varepsilon_j, \tag{4.2}$$

$$\overline{a}_j = a'_j + \varepsilon_j, \text{ for } j = 1, 2, ..., m. \tag{4.3}$$

If $a'_j > 0$ and $\underline{x}_{ij}, \overline{x}_{ij} \geq 0$ for each $j = 0, 1, ..., m$ and $i = 0, 1, ..., n$ then Eq. (4.1) can be written by using interval arithmetic mentioned in section 2.5 as follow

$$\begin{aligned} \underline{y}_i &= \left(a'_0 - \varepsilon_0\right) + \min\left(\left(a'_1 - \varepsilon_1\right)\underline{x}_{i1}, \left(a'_1 - \varepsilon_1\right)\overline{x}_{i1}\right) \\ &+ \min\left(\left(a'_2 - \varepsilon_2\right)\underline{x}_{i2}, \left(a'_2 - \varepsilon_2\right)\overline{x}_{i2}\right) + \cdots + \min\left(\left(a'_m - \varepsilon_m\right)\underline{x}_{in}, \left(a'_m - \varepsilon_m\right)\overline{x}_{in}\right), \end{aligned} \tag{4.4}$$

$$\overline{y}_i = \left(a'_0 + \varepsilon_0\right) + \left(a'_1 + \varepsilon_1\right)\overline{x}_{i1} + ... + \left(a'_m + \varepsilon_m\right)\overline{x}_{im}. \tag{4.5}$$

One may note that all $\underline{y}_i$, $\overline{y}_i$, $\underline{a}_j$, $\overline{a}_j$, $\underline{x}_{ij}$ and $\overline{x}_{ij}$'s are point values and not intervals. The most important observation here is, without knowing the sign of $\underline{a}_j$ it is not easy to find the estimations for $\underline{a}_j$ in Eq. (4.4). In order to find the estimates for $\underline{a}_j$ first we find estimates for $\overline{a}_j$.

Rearranging Eq. (4.5) we have

$$\overline{y}_i = \overline{c}_0 + \overline{c}_1\overline{x}_{i1} + ... + \overline{c}_m\overline{x}_{im}, \tag{4.6}$$

where $\overline{c}_j = a'_j + \varepsilon_j$ for $j = 1, 2, ..., m$.

Using Least squares method we find estimates for $\overline{c}_j$'s. Solving the equation $\overline{c}_j = a'_m + \varepsilon_j$ for ε_j we estimate lower boundary estimation $\underline{a}_j$ as $\underline{a}_j = a'_j - \varepsilon_j$. Likewise we can estimate the interval regression coefficients of Eq. (4.1).

5. ACCURACY ASSESSMENT OF INTERVAL APPROXIMATION

The quality of the approximation of $\tilde{Y}_i \approx \left(\sum_{0 \le j \le m} \tilde{\theta}_j \tilde{X}_{ij} \right)$, can be examined by considering the overlap between approximated output $\hat{\tilde{Y}}$ and expected output $\tilde{Y}$. The approximation would be better if the overlap between $\hat{\tilde{Y}}$ and $\tilde{Y}$ is considerably large. The accuracy ratio of an interval approximation is defined in [6] as below.

Definition 3 [5]: *Let* $\hat{\tilde{Y}} = [\underline{\hat{y}}, \overline{\hat{y}}]$ *be an approximation for the interval* $\tilde{Y} = \left[\underline{y}, \overline{y}\right]$. *The accuracy ratio of the approximation is*

$$Acc(\tilde{Y}, \hat{\tilde{Y}}) = \begin{cases} 100\% & \text{if } \tilde{Y} = \hat{\tilde{Y}} \\ \dfrac{w\left(\left[\underline{y}, \overline{y}\right] \cap [\underline{\hat{y}}, \overline{\hat{y}}]\right)}{w\left(\left[\underline{y}, \overline{y}\right] \cup [\underline{\hat{y}}, \overline{\hat{y}}]\right)} & \text{if } \left(\tilde{Y} \cap \hat{\tilde{Y}}\right) \neq \varnothing \\ 0 & \text{otherwise.} \end{cases} \tag{5.1}$$

where the function $w()$ *returns the width of an interval.*

In this study, we consider the average accuracy ratio to assess the quality of an interval approximation on a set of N interval pairs (x_i, y_i) qualitatively [4].

For a set of N interval pairs (x_i, y_i), the average accuracy ratio of the approximation is defined as

$$Acc^{*} = \frac{\sum_{i=1}^{N} Acc(\tilde{Y}_i, \hat{\tilde{Y}}(x_i))}{N} \tag{5.2}$$

The average accuracy ratio is a quality measurement in addition to the sum of squares of left and right errors defined in Eqs. (2.6) and (2.7). Maximizing the average accuracy ratio and minimizing the sum of squares are inter-connected. The higher the average accuracy ratio is, smaller the sum of squares and the better the approximation.

6. EVALUATION OF INVASION RISK OF INVASIVE ALIEN PLANT SPECIES

6.1 *Selection of parameters*

One of the major concerns in determining invasiveness of IAS is their biological traits [6].Invasive plants usually have higher ability adaption, reproduction and dispersal, and thus make them turn to a great diversity of habitats. Depending on viable seed

production, strength vegetative reproduction and diversity of dispersal mechanism enable IAS occupying invaded habitat quickly, and disperse with a far range. The interfering competition based on alleopathy and physical defensive structures make invasive pants more invasive. Therefore in this study we are mainly concerned about the biological traits related to invasive potential. The most important 12 biological traits are selected as the parameters from National Risk Assessment (NRA) for alien invaders in Sri Lanka. These parameters may be written as below:

Number of seeds per fruit (*SF*), Annual seed production per m^2 (*ASR*), Viability of seeds (*VS*), Long distance dispersal strength (*LDD*), Vegetative reproduction strength (*VRS*), Seed germination requirements (*SGR*), Presence of physical defensive structures (*PDS*), Formation of climbing or smothering growth habit (*FCS*), Potential to be spread by human activities (*HA*), Role of natural and manmade disturbances (*NMD*), Alleopathic property (*AP*), Existence of invasive races (*IR*).

The dataset of known 28 invasive alien species is provided by the invasive specialists group attached to Ministry of Environment and Renewable Resources, Sri Lanka. It contains single-valued observations of 12 parameters and invasion risk scores which are obtained from NRA procedure using a checklist prepared and accepted by the Ministry of Environment and Renewable Resources, Sri Lanka after a broad stakeholder participation and discussion.

6.2 Formulation of Interval-valued data

As mentioned in section I NRA score is a decision given for the invasion risk of a plant by the group of plant science experts. For each risk factor the experts give a collective opinion over a plant species. For example the collective opinion given for the invasive species *Austroeupatorium inulifolium*'s vegetative reproduction strength is Medium. These linguistic labels are scored before generating the NRA score. For example the linguistic labels for the vegetative reproduction strength are *Very Low*, *Low*, *Medium*, *High*, *Very High* and for each score has been assigned as 1, 2, 3, 4, 5 respectively. The same procedure is followed for the other qualitative risk factors. If it is a quantitative risk factor a score is given to the appropriate output range. For example, if a species' viability of seeds is up to 2 years, then score is 1 or if between 2 to 5 years score is 2 otherwise score is 4. One may see that the NRA score is generated by single values without considering the range of linguistic label or interval. One may question that is it

realistic to represent the linguistic label or an interval. It is clear that the uncertain and imprecision cannot be captured by a single value.

On the other hand the data for quantitative parameters are approximations because it cannot be measured exactly. Therefore, working with interval inputs and interval outputs is a vital role in this kind of a situation. However the NRA scores have been considered as the basis of this study because these are given by using experts' knowledge and experience.

In the process of interval data formulation, scores assigned to the linguistic labels of qualitative parameters and real data of quantitative parameters have been converted into intervals by performing width adjustments considering those values as centers. Here, the nature of each parameter is assumed to form interval-valued data and keeping the essence of experts' opinions for risk scores. One may note that data for the parameters *SGR*, *PDS*, *AP* and *IR* are in the form of yes/no answers. In reality the term yes/or no is exact, no need of converting into intervals. Therefore the scores assigned to the yes/no answers in the NRA have been considered as the data for those parameters. It has been found that the appropriate widths from center to end point of parameters and risk score are ± 0.4, ± 5 respectively.
It may be noted that the lower and upper boundaries of interval-valued data are all non-negative values.

6.3 Model formulation

Here the invasion risk Inv_R of a particular alien plant species is assumed to depend on by the 12 biological traits: *SF*, *ASR*, *VS*, *LDD*, *VRS*, *SGR*, *PDS*, *FCS*, *HA*, *NMD*, *AP* and *IR*. Accordingly a linear relation is assumed for Inv_R as

$$\begin{aligned} \tilde{Inv}_R = {} & \tilde{\theta}_0 + \tilde{\theta}_1(SF) + \tilde{\theta}_2(ASR) + \tilde{\theta}_3(VS) + \tilde{\theta}_4(LDD) + \tilde{\theta}_5(VRS) + \tilde{\theta}_6(SGR) \\ & + \tilde{\theta}_7(PDS) + \tilde{\theta}_8(FCS) + \tilde{\theta}_9(HA) + \tilde{\theta}_{10}(NMD) + \tilde{\theta}_{11}(AP) + \tilde{\theta}_{12}(IR) \end{aligned} \qquad (6.1)$$

Where *SF*, *ASR*, *VS*, *LDD*, *VRS*, *SGR*, *PDS*, *FCS*, *HA*, *NMD*, *AP*, *IR* are all in intervals.

To find the coefficient parameters of (6.1), first the interval-valued input data matrix $\tilde{X}$ has been constructed. Then two approximation models for (6.1) have been obtained by following the procedures mentioned in sections 2 and 3.

7. COMPUTATIONAL RESULTS

7.1 Estimations of coefficients

Tables 1 and 2 summarize the interval estimations of regression coefficients that have been obtained from Models I and II respectively.

Table 1: Interval estimates for coefficients of Model I

Coefficient	Center value of coefficient	Interval estimates- Model I
θ_0	11.07908	[11.07323, 11.08492]
θ_1	0.0183830220805151	[0.01837, 0.01840]
θ_2	$5.40300982218465\times10^{-6}$	$[5.396\times10^{-6}, 5.409\times10^{-6}]$
θ_3	0.0155453826052045	[0.01553, 0.01556]
θ_4	0.0385944399652945	[0.03855, 0.03864]
θ_5	2.14988781032716	[2.14781, 2.15196]
θ_6	2.69472395393735	[2.69187, 2.69758]
θ_7	2.43003869540741	[2.42756, 2.43251]
θ_8	1.9773208835343	[1.97549, 1.97915]
θ_9	3.07114393755476	[3.06775, 3.07454]
θ_{10}	1.34943726439762	[1.34851, 1.35037]
θ_{11}	2.42486246924465	[2.42239, 2.42733]
θ_{12}	2.51326594838486	[2.51067, 2.51586]

Table 2: Interval estimates for coefficients of Model II

Coefficient	Center value of coefficient	Interval estimates-Model II
θ_0	11.07908	[9.81417, 12.34399]
θ_1	0.0183830220805151	[0.0183830220805139, 0.0183830220805163]
θ_2	$5.40300982218465\times10^{-6}$	$[5.40300982218348\times10^{-6}, 5.40300982218582\times10^{-6}]$
θ_3	0.0155453826052045	[0.0155453826052041, 0.0155453826052049]
θ_4	0.0385944399652945	[0.0385944399652919, 0.0385944399652971]
θ_5	2.14988781032716	[2.14988781032714, 2.14988781032718]
θ_6	2.69472395393735	[2.69472395393733, 2.69472395393737]
θ_7	2.43003869540741	[2.43003869540739, 2.43003869540743]
θ_8	1.9773208835343	[1.97732088353429, 1.97732088353431]
θ_9	3.07114393755476	[3.07114393755475, 3.07114393755477]
θ_{10}	1.34943726439762	[1.3494372643976, 1.34943726439764]
θ_{11}	2.42486246924465	[2.42486246924464, 2.42486246924466]
θ_{12}	2.51326594838486	[2.51326594838485, 2.51326594838487]

7.2 Comparison of quality of the models

Average accuracy ratio has been used to compare the overall quality of the approaches that have been used in Models I and II. Table 4 summarizes the accuracy ratios with each of the model. The Figs. 1 and 2 show graphical comparison among expected lower

and upper risk boundaries with approximated lower and upper risk boundaries that have been obtained from both models for 28 invasive plant species in the dataset. In these two figures, symbol '*', '×', '↟' and 'Δ' represent expected lower, expected upper, approximated lower and approximated upper boundary of risk respectively.

Table 3: Quality comparison between Model I and Model II

Model No	Average Accuracy Ratio
Model I	0.65779
Model II	0.727641

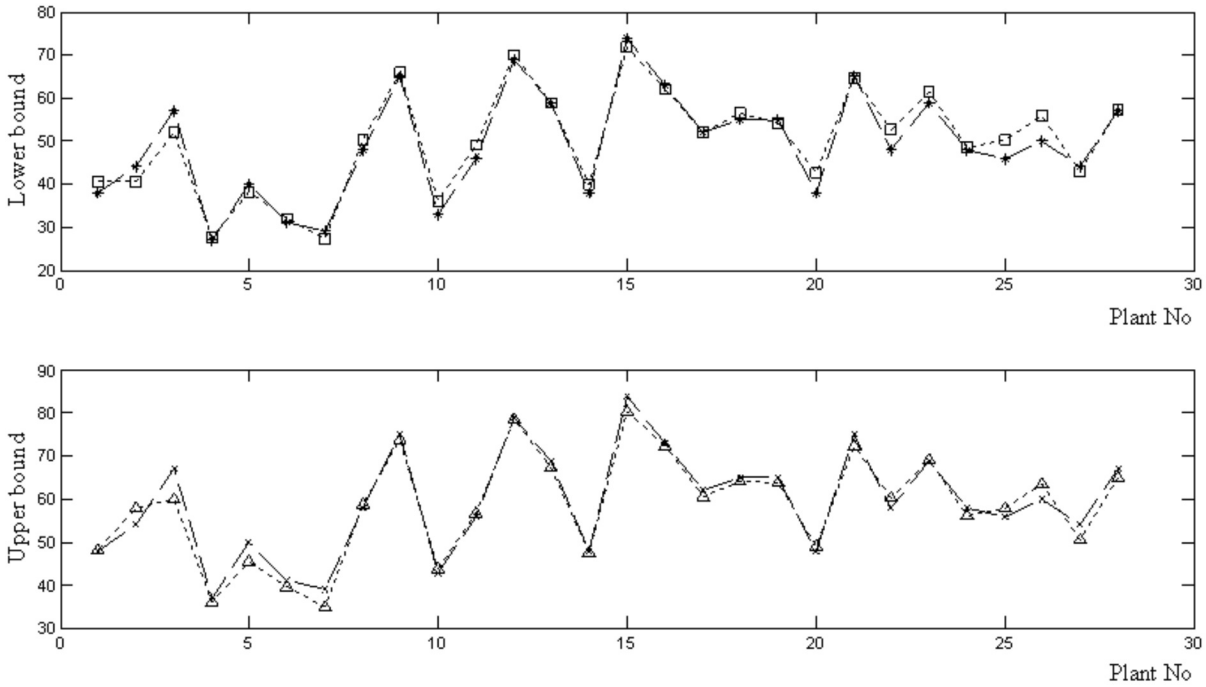

Figure 1: Comparison of expected risk boundaries with approximated risk boundaries from Model I

7.3 Validation results

The Models I and II have been validated using some well-known invasive and non-invasive species in Sri Lanka, to see whether these models provide better predictions. The data for these species have been gathered from the same source as we mentioned in section 6.1.

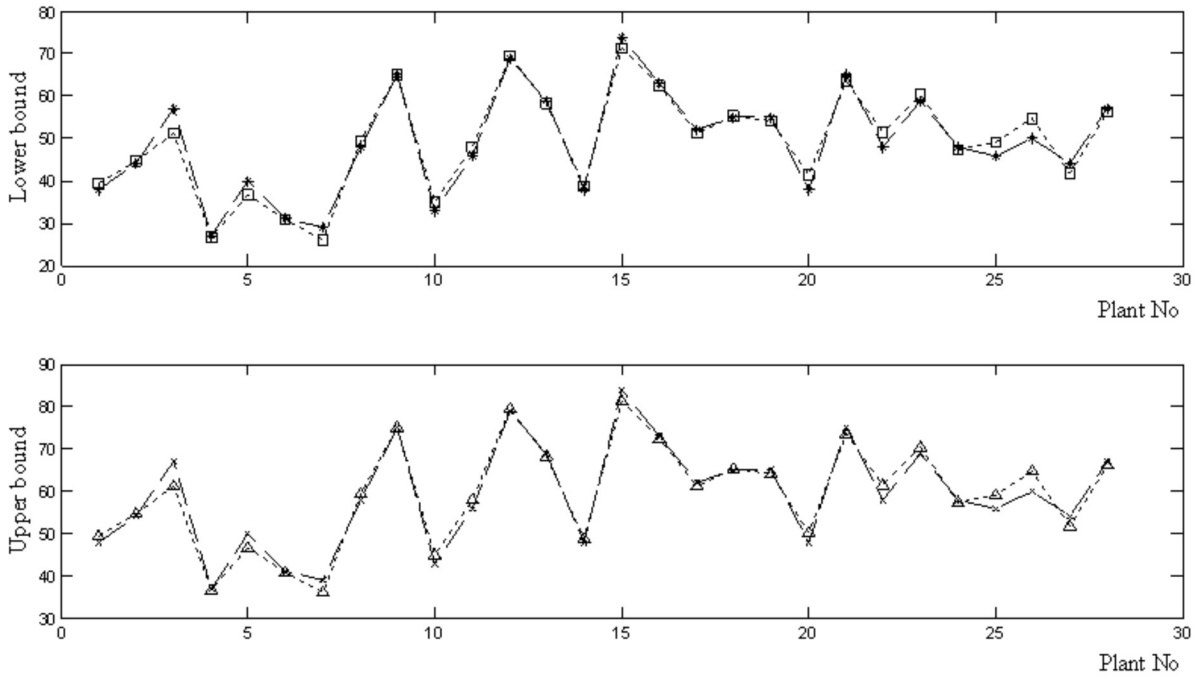

Figure 2: Comparison of expected risk boundaries with approximated risk boundaries from Model II

Table 4: Validation results

Category of species	Name of species	Whether Native (N) or Exotic (E)	NRA score (%)	Approximated Risk Model I	Approximated Risk Model II
Invasive In Sri Lanka	*Austroeupatorium inulifolium*	E	62	[56.333, 63.91]	[55.118, 65.118]
	Panicum maximum	E	66	[64.495, 72.091]	[63.29, 73.289]
	Cuscuta campestris	E	60	[55.506, 63.082]	[54.29, 64.29]
Outside Sri Lanka	*Pueraria montana*	E	55	[54.777, 62.352]	[53.561, 63.561]
	Acacia mearnsii	E	64	[57.781, 65.364]	[56.5689, 66.569]
	Myrica faya	E	36	[34.282, 41.817]	[33.045, 43.045]
Non Invasive	*Cassia fistula*	N	32	[32.24, 39.773]	[31.003, 41.003]
	Cissus rotundifolia	E	32	[31.26, 38.788]	[30.02, 40.02]
	Hedychium gardnerianum	E	32	[33.888, 41.421]	[32.651, 42.650]
	Magnfera indica	N	32	[32.287, 39.82]	[31.05, 41.05]

7.4. DISCUSSION

Validation results show that both models reflect the invasiveness of plant species more or less the same way that a group of experts have given a rank in the national risk assessment procedure. It is obvious that the known invasive species have obtained a higher risk value compared to non-invasive species as shown by species used for the validation process. In addition, it reflects a higher percentage of species that have not been found in Sri Lanka but reported to be invasive in other countries. Therefore, it is clear that the model has the potential to serve as a tool which could screen the invasiveness of species before introduction to the country.

It is clear that the 12 parameters used in this study could be the traits that contribute heavily to the invasiveness of plant species by elevating the risk score. This in another way to confirm that the trait selection for NRA in Sri Lanka has been successfully conducted by picking the most relevant traits to evaluate invasive risks of plants.

However, both models developed in this study satisfy the assessment of invasiveness of a plant species, with respect to mathematical aspects the Model II is considered better over the Model I.

The ε – inflation process computationally finds the interval regression coefficients with reasonable quality. If the decision maker is satisfied with the value of ratio estimation, then the current ε value will be the width of interval coefficient. However, it is not easy to find which ε value would fit better with a particular coefficient. According to Tables. 1 and 2, the boundaries of coefficients of Model I and II are positively related to the invasion risk. In this kind of a situation the Model II can be easily applied. On the other hand accuracy ratios of Models in Table 3 reveal how well the expected outcome intersects with approximated outcome. As mentioned in section 5 the average accuracy ratios have been obtained for each model to measure the intersection between expected and approximated outcomes. It can clearly be seen that the Model II reflect a higher level of intersection compared to the Model I and the Figs. 1 and 2 confirm that the expected outcome and approximated outcome of Model II are closer than the Model I.

On the other hand, as mentioned earlier the NRA scores have been considered as the basis to compare the model outcomes. Here we consider the model outcome is a better interpretation of risk of species if that the NRA score is within that output interval. For example, if we consider non-invasive species *Cassia fistula,* which has NRA score of 32 in Table 4. This risk score is out of the boundaries of the predicted risk interval in Model I, but it is within the boundaries of the risk interval in Model II. Similarly, the results are for non-invasive species *Magnfera indica*. On the other hand, the species *Hedychium gardnerianum* with NRA score of 32 is out of boundaries of estimated risk intervals in both models. But the lower boundary of Model II is closer to NRA score than the lower boundary of Model I. According to the discussion, the Model II can be considered as the tool which gives a significant outcome as the risks of IAS.

8. CONCLUSION

In this paper, two different interval multiple linear regression models with interval inputs and outputs, have been constructed to assess the risk of IAS. We have used interval least square algorithm proposed by Chenyi Hu [7], and proposed a new method to find the interval-valued regression coefficients. The proposed method is a direct and efficient method to find the boundaries of interval estimations of coefficients than

adjusting widths of coefficients using ε – inflation which is used in Chenyi Hu [7]. Proposed method has produced significantly improved results in comparison to Chenyi Hu [7]. The proposed method gives better prediction of risks of invasive alien species if its invasion is dominated by biological traits. However we should explore to extend this method to estimate the interval regression coefficients without considering the sign of boundaries ofinput-output data. Also, the model needs to be modified by incorporating the risk factors other than biological traits, e.g. ecology, establishment, management aspects etc to evaluate overall invasion risk. But the limited amount of available data on those factors sets serious constraints to the evaluation of overall risk of IAS.

Acknowledgment

H.O.W. Peiris would like to acknowledge the University of Colombo research grant (No: AP/3/2012/CG/26) for providing necessary financial support for visiting National Institute of Technology Rourkela, Odisha, India to undertake this collaborative work.

REFERENCES

[1]. Alefeld G., and Mayer G., (2000). Interval analysis: theory and applications, *Journal of Computational and Applied Mathematics*, 121: 421-464.

[2]. Bentbib, A.H. (2002). Solving the full rank interval least squares problem, *Journal of Applied Numerical Mathematics*, 41(2): 283-294,

[3]. Convention on Biological Diversity (2008). Alien species that threaten ecosystems, habitats and species, Article 8[h]. Secretariat of the Convention on Biological Diversity, United Nations Subsidiary Body on Scientific, Technical and Technological advice, Thirteenth meeting FAO, Rome, 18-22, February 2008.

[4]. Hu, Chenyi., (2012). Interval Function and its Linear Least-squares Approximation, *Proceeding*, SNC '11 Proceedings of the 2011 International Workshop on Symbolic-Numeric Computation, 16-23.

[5]. Hu C., De Korvin R.B.K.A., and Kreinovich V., (2008). Knowledge Processing with Interval and Soft Computing: Advanced Information and Knowledge Processing (ed. L. Jain, X. Wu), Springer.

[6]. Ranwala S.M.W., (2010). Risk Assessment for Invasive Alien Species, In Invasive Alien Species- Strengthening capacity to control Introduction and Spread in Sri Lanka (Eds. B. Marambe, P. Silva, S. Wijesundera, N. Attapattu), Biodiversity Secretariat, Ministry of Environment and Natural Resources, Sri Lanka.

[7]. Rejmanek M., and Richardson M. D. (1996). What attributes make some plants species more invasive, *Journal of Ecology*, 77: 1655-1661.

[8]. Sara A Van De Geer (2005). Least Squares Estimation, *Encyclopaedia of Statistics in Behavioral Science*, (eds. Brian S. Everitt, David C. Howell), John Wiley & Sons, Ltd, Chichester, 2: 1041-1045.

[9]. Stockburger W.D., (2001). Multivariate Statistics: Concepts, Models, and Applications. 3rd Web Edition, Retrieved from http://www.psychstat.missouristate.edu/multibook/mlt08m.html.