British Journal of Applied Science & Technology
16(1): 1-11, 2016, Article no.BJAST.25901
ISSN: 2231-0843, NLM ID: 101664541

SCIENCEDOMAIN *international*
www.sciencedomain.org

Development of a Risk Assessment Mathematical Model to Evaluate Invasion Risk of Invasive Alien Species Using Interval Multivariate Linear Regression

H. O. W. Peiris[1*], S. Chakraverty[2], S. S. N. Perera[1] and S. M. W. Ranwala[3]

[1]Department of Mathematics, Research and Development Centre for Mathematical Modelling, University of Colombo, 94, Cumaratunga Munidasa Mawatha, Colombo 00300, Western Province, Colombo 00700, Sri Lanka.
[2]Department of Mathematics, National Institute of Technology Rourkela, Rourkela - 769 008, Odisha, India.
[3]Department of Plant Sciences, University of Colombo, 94,Cumaratunga Munidasa Mawatha, Colombo 00300, Western Province, Colombo 00700, Sri Lanka.

Authors' contributions

The work was carried out in collaboration between all authors. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/BJAST/2016/25901
Editor(s):
(1) Qing-Wen Wang, Department of Mathematics, Shanghai University, P.R. China.
Reviewers:
(1) N. V. Krishna Prasad, GITAM Univeristy, India.
(2) Pedro Pablo Cárdenas Alzate, Universidad Tecnológica de Pereira, Colombia.
Complete Peer review History: http://sciencedomain.org/review-history/14527

Original Research Article

Received 25th March 2016
Accepted 26th April 2016
Published 9th May 2016

ABSTRACT

Evaluation of risk of Invasive Alien Species (IAS) with uncertain and imprcise data is a challenging task. In the present work, mathematical model for risk assessment is developed by using interval multiple linear regression analysis in which mimic unceratin and imprecise data. Here both dependent and independent variables are interval-valued.
12 invasive attributes selected as model parameters. Proposed a new method find the solution of design matrix using interval least square method. Here obtained a dataset of 28 invasive plant species which contains single-valued observations of 12 parameters and invasion risk scores which

**Corresponding author: E-mail: oshivida@yahoo.com;*

are obtained from National Risk Assessment. Using the dataset formed four interval input datasets. New method is proposed to find the estimates for interval regression coefficient using interval least suqare method. The interval regression coefficents are estimated using four different interval input data set. The quality of the approximated model is evaluted by average accuracy ratio and the models are validated using well known six invasive and four non invasive species.
The approximated model gives average accuracy ratio of 0.730852 along with data set 3 which is the highest among all data sets. Validation results show that the expected risk score of each plant species from National Risk Assessment is within the approximated risk interval.
Comparing the quality and the validation results, it is found that the approximated model along with data set 3 gives better predictions of risks of invasive alien species if its invasion is dominated by biological traits.

Keywords: Interval multiple linear regression; interval least square; invasive alien species; biological traits.

1. INTRODUCTION

Numbers of Invasive Alien Species (IAS) are increasing rapidly worldwide, causing both environmental and economic damage. Many countries have highlighted the urgent need for rigorous and comprehensive risk assessment protocols for invasive alien species for prevention and control strategies. Screening procedures like risk assessments based on questionnaires have been developed in several parts of the world. Such assessments for alien plants need wide range of information of risk factors which define invasion risk of IAS. But most of the risk factors which affect the invasiveness of species are accompanied with imprecision and uncertainty. Therefore it is very important to incorporate mathematical modeling techniques to risk assessments for handling the impression and uncertainty of data. This may give a better prediction of invasion risk of IAS.

Linear regression is one of the fundamental models used to determine the relationship between dependent and independent variables. An extension of this model, namely multiple linear regression, is used to represent the relationship between a dependent variable and several independent variables [1]. These variables usually are single-valued. The need of interval-valued data may arise to mimic the imprecision and uncertainty for obtaining reliable approximations. The lower and upper bounds provide the boundaries of the interval-valued data. Therefore, the interval-valued data x can be written by the pair of values $\underline{x}$ and $\overline{x}$ with $\overline{x} \geq \underline{x}$ where $\underline{x}$ and $\overline{x}$ denote the lower and upper bound respectively. This study is focused on developing a mathematical model to evaluate satisfactory interval approximations for risk of IAS using interval multiple linear regression. Here we propose a new estimation method using interval least square to estimate boundaries of interval regression coefficients. The approximated model is validated, to see whether the predictions are within a satisfactory level.

2. INTERVAL MULTIVARIATE LINEAR REGRESSION

2.1 Least Square

Let us briefly discuss the least square method to approximate the multiple linear regression model for crisp input-output data.

The multiple linear regression equation for p variables is as follows:

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

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 estimated regression coefficients. The regression coefficients $\left(a_j, j = 1, 2, ..., p\right)$ are determined by the normal equations which are obtained, in such a way that $\sum_{i=1}^{N} \left(a_0 + a_1 x_{i1} + a_2 x_{i2} + ... + a_p x_{ip} - \hat{y}_i\right)^2$ is minimal. The least square algorithm to find regression coefficient vector a is defined in [2] as follows:

Step 1: Differentiate $\sum_{i=1}^{N}\left(a_0 + a_1x_{i1} + a_2x_{i2} + \ldots + a_px_{ip} - \hat{y}_i\right)^2$ with respect to a_j for $j = 1, 2, \ldots, p$ where $a_j \in a$.

Step 2. Form the matrix

$$A = \begin{pmatrix} 1 & x_{12} & \cdot & \cdot & \cdot & x_{1p} \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & & & & \cdot \\ 1 & x_{n2} & \cdot & \cdot & \cdot & x_{np} \end{pmatrix} \quad (2.2)$$

and the column vector,

$$b = \left[y_1\ y_2\ \ldots y_n\right]. \quad (2.3)$$

Step 3. Suppose A has full column rank, that is no column in A can be written as a linear combination of other columns. Then the least square estimator a is given by

$$a = \left(A^T A\right)^{-1} A^T b \quad (2.4)$$

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

2.2 Interval Multiple Linear Regression Model

Here we consider the case of response variable, predictor variables and unknown model parameters as intervals.

The functional form of the interval multivariate linear regression model is as follows:

$$\tilde{Y}_i = \tilde{a}_0 + \tilde{a}_1\tilde{x}_{i1} + \tilde{a}_2\tilde{x}_{i2} + \cdots + \tilde{a}_m\tilde{x}_{im} \quad i = 1, 2, \ldots, n. \quad (2.5)$$

Where $\tilde{Y}_i$ is the interval response variable, $\tilde{x}_{ij}$, $j = 0, 1, \ldots, m$ are the interval predictor variables, and $\tilde{a}_j$'s are unknown interval regression coefficients. Below we define the concept of interval least square to estimate the $\tilde{a}_j$ in Eq. (2.5).

2.3 Interval Least Square

Let us denote $\tilde{\hat{y}}_i$ be the estimation of interval response variable $\tilde{y}_i$. We need to obtain $\tilde{a}_j$, $j = 0, 1, \ldots, m$ which minimize the sum of squared distance $\sum_{i=1}^{n} d^2\left(\tilde{y}_i, \tilde{\hat{y}}_i\right)$. First of all let us define the absolute error of interval estimation as in [3].

Definition 2.1 [3]: *Let interval* $\hat{y} = \left[\underline{\hat{y}}, \overline{\hat{y}}\right]$ *be an estimation of an interval* $y = \left[\underline{y}, \overline{y}\right]$. *The left and right absolute errors are* $E_L = \left|\underline{\hat{y}} - \underline{y}\right|$ *and* $E_R = \left|\overline{\hat{y}} - \overline{y}\right|$, *respectively. The absolute error of the estimation is the sum of left and right absolute errors, that is,* $E = E_L + E_R = \left|\underline{\hat{y}} - \underline{y}\right| + \left|\overline{\hat{y}} - \overline{y}\right|$, *respectively.*

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

Definition 2.2 [4]: *Let* U *be the set of* n *interval valued observations of an interval linear regression function* $y = h(x)$ *i.e.* $U = \{(x, y): x \subset R^n, y \subset R \text{ both } x \text{ and } y \text{ are compact}\}$.

According to the definition 1, we say that $\sum_{0 \le j \le m} \tilde{a}_j\tilde{x}_{ij}$ is an interval least square estimation of $\tilde{y}_i$ if the linear combination minimizes:

$$\sum_{i=1}^{n} {E^2}_L + \sum_{i=1}^{n} {E^2}_R,$$

where

$$\sum_{i=1}^{n} {E^2}_L = \sum_{i=1}^{n}\left(\underline{y}_i - \left(\sum_{0 \le j \le m} \underline{a_j x_{ij}}\right)\right), \quad (2.6)$$

and

$$\sum_{i=1}^{n} {E^2}_R = \sum_{i=1}^{n}\left(\overline{y}_i - \left(\sum_{0 \le j \le m} \overline{a_j x_{ij}}\right)\right). \quad (2.7)$$

2.4 Interval Arithmetic (ILS)

While using ILS method, it is required to find the estimates for interval regression coefficients that minimize the sum of Eqs. (2.6) and (2.7). Therefore we need to use interval arithmetic to work with interval-valued data [5].

Let $[a]=[\underline{a},\bar{a}]$, $[b]=[\underline{b},\bar{b}]$ be real compact intervals and $\circ$ is one of the basic operations 'additions', 'subtraction', '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:

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

2.5 Proposed Estimation Method

In this section, we are going to present a new estimation method to approximate the interval regression coefficients in interval multiple linear regression model. Here we consider the case of response variable, predictor variables and unknown regression coefficients intervals.

The interval multiple linear regression model may be written as follows:

$$\tilde{Y}_i=\tilde{a}_0+\tilde{a}_1\tilde{x}_{i1}+\tilde{a}_2\tilde{x}_{i2}+\cdots+\tilde{a}_m\tilde{x}_{im}\quad i=1,2,...,n \tag{2.8}$$

where $\tilde{Y}_i$ is the interval response variable, $\tilde{x}_{ij}$, $j=0,1,...,m$ are the interval predictor variables, and $\tilde{a}_j$'s are unknown interval regression coefficients.

Let us denote the center values of interval regression coefficients $[\underline{a}_j,\bar{a}_j]$ for $j=0,1,...,m$ as a'_m. To find a'_m , first of all we need to construct the matrix $\tilde{A}$ and $\tilde{b}$ as in Eqs. (2.2) and (2.3).

Let us take A_{mid} and b_{mid} as the midpoint matrices of $\tilde{A}$, $\tilde{b}$ respectively. One may note that all components in matrices A_{mid} and b_{mid} are crisp values that is not intervals. If A_{mid} is a full column rank matrix we can evaluate a'_m as below.

$$a'_m=\left(A^T{}_{\text{mid}}A_{\text{mid}}\right)^{-1}b_{\text{mid}}. \tag{2.9}$$

Now we consider $\varepsilon_j>0$ a value which satisfies:

$$\underline{a}_j=a'_j-\varepsilon_j, \tag{2.10}$$

$$\bar{a}_j=a'_j+\varepsilon_j,\ \text{for}\ j=1,2,...,m. \tag{2.11}$$

If boundary values of $\tilde{Y}_i$, $\tilde{x}_{ij}$, and center values a'_m are non-negative then the lower and upper observed responses are satisfying the equations (2.12), (2.13) respectively:

$$\underline{y}_i=\underline{a}_0+\min\left(\underline{a}_1\underline{x}_{i1},\underline{a}_1\bar{x}_{i1}\right)+\min\left(\underline{a}_2\underline{x}_{i2},\underline{a}_2\bar{x}_{i2}\right)+\cdots+\min\left(\underline{a}_n\underline{x}_{in},\underline{a}_n\bar{x}_{in}\right)\quad i=1,2,...,n \tag{2.12}$$

$$\bar{y}_i=\bar{a}_0+\bar{a}_1\bar{x}_{i1}+\bar{a}_2\bar{x}_{i2}+\cdots+\bar{a}_n\bar{x}_{in}\quad i=1,2,...,n \tag{2.13}$$

where $\underline{y}_i,\underline{x}_{ij}$, and $\underline{a}_j$ indicate lower bounds of response variables, predictor variables and model parameters respectively. Similarly $\bar{y}_i,\bar{x}_{ij}$, and $\bar{a}_j$ indicate upper bounds of response variables, predictor variables and regression coefficients respectively. The sign of $\underline{a}_j$ can be changed depending on the values of ε_j. Therefore without knowing the exact ε_j values, Eq. (2.12) cannot be computed as in Eq. (2.13). Let us denote $[\hat{y}_L,\hat{y}_U]$ as the estimated interval

output, where $\hat{y}_L = \left(\underline{a}_0 + \sum_{1 \le j \le m} \min\left(\underline{a}_j \underline{x}_{ij}, \underline{a}_j \overline{x}_{ij} \right) \right)$ and $\hat{y}_U = \left(\sum_{0 \le j \le m} \overline{a}_j \overline{x}_{ij} \right)$.

Here our aim is to find a subset of the $\left[\hat{y}_L, \hat{y}_U \right]$ due to the complexity of finding the lower boundaries for regression coefficients in Eq. (2.12). Let us define $\left[\hat{y}'_L, \hat{y}_U \right]$ as the subset of $\left[\hat{y}_L, \hat{y}_U \right]$ where $\hat{y}'_L = \left(\underline{a}_0 + \sum_{1 \le j \le m} \underline{a}_j \underline{x}_{ij} \right)$. One may note that the defined subset is differ only from the lower bound of set $\left[\hat{y}_L, \hat{y}_U \right]$.

Now let us show that $\left[\hat{y}'_L, \hat{y}_U \right] \subseteq \left[\hat{y}_L, \hat{y}_U \right]$ if boundary values of $\tilde{Y}_i$, $\tilde{x}_{ij}$, and center values of a'_m are non-negative.

For the case when $\underline{a}_j \ge 0$, $\hat{y}_L = \hat{y}'_L = \left(\underline{a}_0 + \sum_{1 \le j \le m} \underline{a}_j \underline{x}_{ij} \right)$ since $\underline{x}_{ij}, \overline{x}_{ij}$ are non-negative values.

Hence $\left[\hat{y}'_L, \hat{y}_U \right] \subseteq \left[\hat{y}_L, \hat{y}_U \right]$.

Now consider the case when $\underline{a}_j < 0$. Without loss of generality we may assume that $\underline{a}_1 < 0$ in Eq. (2.12).

Then it is clear that $\underline{a}_1 \overline{x}_{ij} < \underline{a}_1 \underline{x}_{ij}$ since $\underline{x}_{ij}, \overline{x}_{ij}$ are non-negative values and $\underline{x}_{ij} < \overline{x}_{ij}$.

By adding the remaining right hand terms for both sides of $\underline{a}_1 \overline{x}_{ij} < \underline{a}_1 \underline{x}_{ij}$ we have

$$\begin{aligned} & a_0 + \\ & \underline{a}_1 \overline{x}_{i1} + \underline{a}_2 \underline{x}_{i2} + \cdots + \underline{a}_n \underline{x}_{in} < \underline{a}_0 + \underline{a}_1 \underline{x}_{i1} \\ & + \underline{a}_2 \underline{x}_{i2} + \cdots + \underline{a}_n \underline{x}_{in}, \; i = 1, 2, ..., n \end{aligned} \tag{2.14}$$

Hence $\hat{y}_L < \hat{y}'$. It is to be noted that values of $\hat{y}_U$ in these two cases are same. Therefore it is clear that $\left[\hat{y}'_L, \hat{y}_U \right] \subseteq \left[\hat{y}_L, \hat{y}_U \right]$. Similarly we can prove $\left[\hat{y}'_L, \hat{y}_U \right] \subseteq \left[\hat{y}_L, \hat{y}_U \right]$ for more than one $\underline{a}_j < 0$. One may note that the values of $\forall \underline{a}_j$, $j = 0, 1, ..., m$ cannot be negative since the boundary values of $\tilde{Y}_i$ and $\tilde{x}_{ij}$ are non-negative.

Now we propose our new method to find the boundaries of interval regression coefficient of Eq. (2.8) as below.

Using least square method defined in section 2.1 we find estimates for $\underline{c}_j$'s and $\overline{c}_j$'s of

$$\underline{y}_i = \underline{c}_0 + \underline{c}_1 \underline{x}_{i1} + ... + \underline{c}_m \underline{x}_{im}, \tag{2.15}$$

$$\overline{y}_i = \overline{c}_0 + \overline{c}_1 \overline{x}_{i1} + ... + \overline{c}_m \overline{x}_{im}, \tag{2.16}$$

where $\underline{c}_j = \underline{a}_j - \varepsilon_j$, $\overline{c}_j = \overline{a}_j - \varepsilon_j$ for $j = 1, 2, ..., m$ which minimizes the following

$$\sum_{i=1}^{N} E^2{}_L = \sum_{i=1}^{N} \left(\underline{y}_i - \left(\sum_{0 \le j \le m} \underline{c}_j \underline{x}_{ij} \right) \right), \tag{2.17}$$

$$\sum_{i=1}^{N} E^2{}_R = \sum_{i=1}^{N} \left(\overline{y}_i - \left(\sum_{0 \le j \le m} \overline{c}_j \overline{x}_{ij} \right) \right). \tag{2.18}$$

One may note that, the proposed method can be applied to find the interval estimations for regression coefficients if the boundaries of Interval input-output data and center values of interval regression coefficients of Eq. (2.8) are non negative values.

2.6 Accuracy Assessment of Approximated Model

The quality of the approximation of $\tilde{Y}_i \approx \left(\sum_{0 \le j \le m} \tilde{a}_j \tilde{x}_{ij} \right)$, 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 [3] as below.

Definition 2.3 [3]: *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*

$$Acc(\tilde{Y},\hat{\tilde{Y}}) = \begin{cases} 100\% & \text{if } \tilde{Y} = \hat{\tilde{Y}} \\ \dfrac{w\left(\left[\underline{y},\bar{y}\right]\cap[\underline{\hat{y}},\bar{\hat{y}}]\right)}{w\left(\left[\underline{y},\bar{y}\right]\cup[\underline{\hat{y}},\bar{\hat{y}}]\right)} & \text{if } \left(\tilde{Y}\cap\hat{\tilde{Y}}\right)\neq\varnothing \\ 0 & \text{otherwise.} \end{cases} \quad (2.19)$$

where the function $w()$ returns the width of an interval by taking the difference of upper and lower boundary points of the given 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.

For a set of *N* interval pairs (x_i, y_i), the average accuracy ratio [4] of the approximation is defined as

$$Acc^{*} = \frac{\sum_{i=1}^{N} Acc(\tilde{Y}_i, \hat{\tilde{Y}}(x_i))}{N}. \quad (2.20)$$

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.

3. METHODOLOGY

3.1 Selection of Parameters

According to the view of biologists several factors affect the risk of invasive plant species such as its ecology, establishment, invasive potential, management aspects etc [6,7,8]. In conventional risk assessments these factors are usually considered. In this work we are mainly concerned about the biological traits related to invasive potential. The most important 12 biological traits are selected as the parameters of the model 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 prepared by the Ministry of Environment and Renewable Resources, Sri Lanka. In the process of interval data formulation several interval input data sets have been formed by performing width adjustments in crisp data. Here, the nature of each parameter is assumed to form interval-valued data and keeping the essence of experts' opinions for risk scores. Table 1 presents some interval input data sets which contain interval-valued data of parameters and interval-valued risk scores (output observations). It may be noted that the lower and upper boundaries of interval-valued data are all non negative values.

Table 1. Interval input data sets

Data set	Spread from center of interval-valued input data of parameters	Spread from center of interval- valued risk scores	
		Left	Right
1	±0.4	3	3
2	±0.5	4	4
3	±0.4	5	5
4	±0.5	4	2

3.2 Model Formulation

Here, the invasion risks score Inv_R of a particular alien plant species is assumed to be linearly determined by the 12 biological traits: *SF*, *ASR*, *VS*, *LDD*, *VRS*, *SGR*, *PDS*, *FCS*, *HA*, *NMD*, *AP* and *IR* as

$$\tilde{Inv}_R = \tilde{\theta}_0 + \tilde{\theta}_1(SF) + \tilde{\theta}_2(ASR) + \tilde{\theta}_3(VS) + \tilde{\theta}_4(LDD) + \tilde{\theta}_5(VRS) + \tilde{\theta}_6(SGR) + \tilde{\theta}_7(PDS) + \tilde{\theta}_8(FCS) + \tilde{\theta}_9(HA) + \tilde{\theta}_{10}(NMD) + \tilde{\theta}_{11}(AP) + \tilde{\theta}_{12}(IR) \quad (3.1)$$

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

First of all we have evaluated the center values of regression coefficients of (3.1) by following the procedures given in section 2. The results have shown that all the center values are non negative.

Therefore, the interval coefficient parameters of model (3.1) have been estimated for each interval input data set by following the proposed estimation method defined in section 2.

4. COMPUTATIONAL RESULTS

4.1 Estimation of Regression Coefficients

Tables 2 - 5 summarize the interval estimations of regression coefficients of model along with interval input data sets represented in Table 1.

Table 2. Interval estimates for coefficients from data set 1

Coefficient	Estimates from data set 1
θ_0	[10.344, 11.814]
θ_1	[0.0183830220805138, 0.0183830220805161]
θ_2	[$5.40300982218432\times10^{-6}$, $5.40300982218483\times10^{-6}$]
θ_3	[0.015545382605203, 0.015545382605205]
θ_4	[0.03859443996525, 0.03859443996527]
θ_5	[2.1499, 2.1499]
θ_6	[2.69472395393732, 2.69472395393747]
θ_7	[2.43003869540741, 2.43003869540748]
θ_8	[1.9773, 1.9773]
θ_9	[3.07114393755477, 3.07114393755487]
θ_{10}	[1.34943726439752, 1.34943726439766]
θ_{11}	[2.42486246924464, 2.42486246924466]
θ_{12}	[2.51326594838487, 2.51326594838489]

Table 3. Interval estimates for coefficients from data set 2

Coefficient	Estimates from data set 2
θ_0	[10.3943, 11.745]
θ_1	[0.0183106933740867, 0.0183798813271578]
θ_2	[$5.40142448033662\times10^{-6}$, $5.40644416509869\times10^{-6}$]
θ_3	[0.0155469931613369, 0.0155539824393268]
θ_4	[0.0391974418649532, 0.0414273985753972]
θ_5	[2.15044615933592, 2.15203670731178]
θ_6	[2.69230683186782, 2.69508724936171]
θ_7	[2.42696169733635, 2.43008821070789]
θ_8	[1.97708136819981, 1.97821672074128]
θ_9	[3.0650823787639, 3.07073314364203]
θ_{10}	[1.348870752105111.35368206335038]
θ_{11}	[2.42463624793798, 2.42513167844029]
θ_{12}	[2.51340075543958, 2.51512964449058]

Table 4. Interval estimates for coefficients from data set 3

Coefficient	Estimates from data set 3
θ_0	[9.65394, 12.34399]
θ_1	[0.0183830220805151, 0.01886]
θ_2	[$5.40300982218465\times10^{-6}$, 5.81×10^{-6}]
θ_3	[0.0155453826052045, 0.01555]
θ_4	[0.0385944399652945, 0.07717]
θ_5	[2.14988781032716, 2.19395]
θ_6	[2.69472395393735, 2.71626]
θ_7	[2.43003869540741, 2.44489]
θ_8	[1.93169, 1.9773208835343]
θ_9	[3.0212, 3.07114393755476]
θ_{10}	[1.27264, 1.34943726439764]
θ_{11}	[2.42486246924465, 2.47308]
θ_{12}	[2.51326594838486, 2.5336]

Table 5. Interval estimates for coefficients from data set 4

Coefficient	Estimates from data set 4
θ_0	[8.3943, 11.745]
θ_1	[0.0183106933740868, 0.0183798813271578]
θ_2	[$5.40142448033665\times10^{-6}$, $5.40644416509869\times10^{-6}$]
θ_3	[0.0155469931613369, 0.0155539824393267]
θ_4	[0.0391974418649532, 0.0414273985753901]
θ_5	[2.15044615933592, 2.15203670731174]
θ_6	[2.69230683186785, 2.69508724936171]
θ_7	[2.42696169733638 2.43008821070789]
θ_8	[1.97708136819981, 1.97821672074126]
θ_9	[3.06508237876395, 3.07073314364203]
θ_{10}	[1.34887075210511, 1.3536820633504]
θ_{11}	[2.42463624793801 , 2.42513167844029]
θ_{12}	[2.51340075543958, 2.51512964449059]

4.2 Accuracy Assessment

To measure the overall quality of the model, we have used average accuracy ratio as defined in (2.20). Table 6 summarizes the average accuracy ratios of the model incorporated each of data set given in Table 1. Figs. 1 and 2 show graphical comparison with expected and approximated lower and upper risk boundaries for 28 invasive plant species in the dataset. In these figures, symbol '*', '×', '' and 'Δ' represent expected lower, expected upper, approximated lower and approximated upper boundary of risk respectively.

Table 6. Quality comparison

Data set	Average accuracy ratio
1	0.60306
2	0.679851
3	0.730852
4	0.603521

4.3 Model Validation

The model along with data sets in Table 1 has been validated by well known invasive and non invasive species of Sri Lanka. The data for these species have been gathered from the same source as we mentioned in section 3.1. The validation results are summarized in Tables 7 and 8.

5. DISCUSSION

From Tables 2 - 5, one may observe that width spread of estimated interval regression coefficients are changing with respect to interval input data set. It can be seen that we get larger width spread in data set 3. Table 6 illustrates quality assessment of the approximated model with each data set. It is seen that approximated model gives average accuracy ratio of 0.730852 along with data set 3 which is the highest among all data sets. The variations among average accuracy ratios can be seen from Figs. 1 - 4 which depict the overlaps between estimated and

expected risk intervals. From validation results in Tables 7 and 8, one may see that the expected risk score of each plant species from National Risk Assessment is within the approximated risk interval. It is clear that present model gives better predictions from each data set. The species *Hedychium gardnerianum* with NRA score of 32 is out of boundaries of estimated risk intervals in data set 1 to 3. However, lower boundaries of estimated risk intervals from input data set 2 and 3 are more close to NRA score than the lower boundary of input data set 1. Therefore comparing the quality and validation results, the data set 3 should be incorporated to the approximated model for better prediction of risk of Invasive alien species.

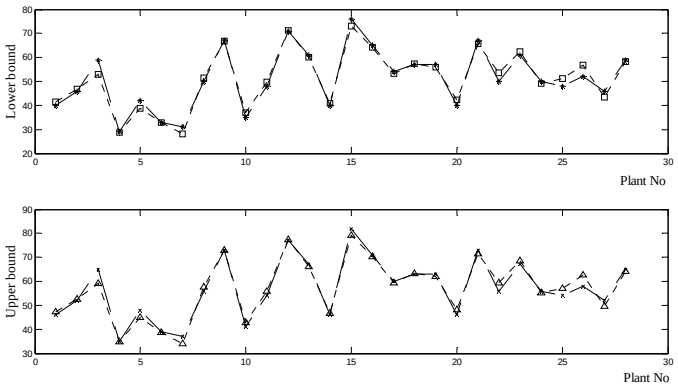

Fig. 1. Interval estimations for risk of IAS from data set 1

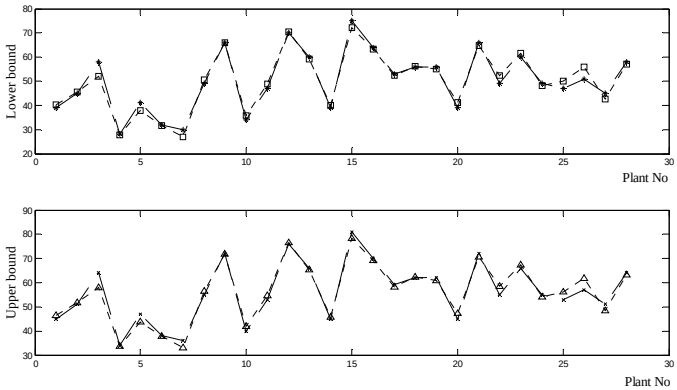

Fig. 2. Interval estimations for risk of IAS from data set 2

Table 7. Validation results

Category of species	Name of species	NRA score (%)	Approximated risk (Data set 1)	Approximated risk (Data set 2)
Invasive	*Austroeupatorium inulifolium*	62	[55.64757,64.58792]	[54.75459, 65.50676]
	Panicum maximum	66	[63.8192, 72.75955]	[62.9152, 73.66792]
	Cuscuta campestris	60	[54.81983, 63.76019]	[53.92563, 64.67496]
	Pueraria montana	55	[54.0905, 63.03085]	[53.17714, 63.93692]
	Acacia mearnsii	64	[57.09867, 66.03903]	[56.1979, 66.94356]
	Magnfera indica	36	[33.57519, 42.51555]	[32.67957, 43.40609]
Non invasive	*Cassia fistula*	32	[31.53293, 40.47328]	[30.62525, 41.36012]
	Cissus rotundi	32	[30.55005, 39.49041]	[29.65617, 40.38129]
	Hedychium gardnerianum	32	[33.18063, 42.12099]	[32.28433, 43.01214]
	Magnfera indica	32	[31.57978, 40.52013]	[30.67473, 41.41821]

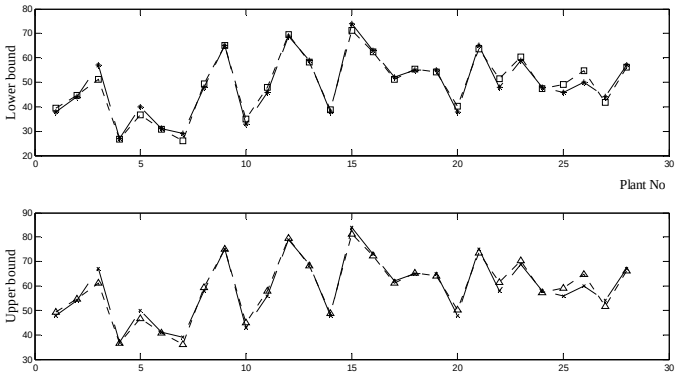

Fig. 3. Interval estimations for risk of IAS from data set 3

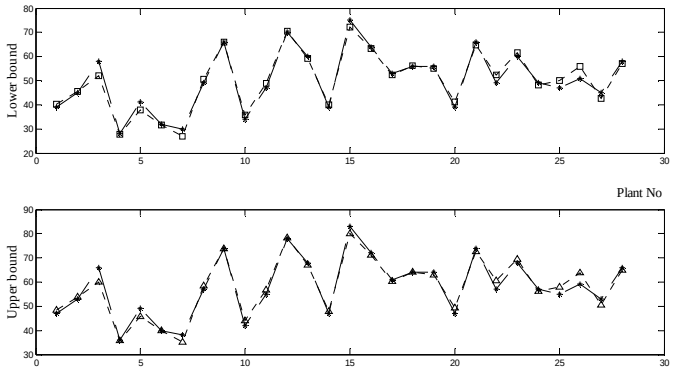

Fig. 4. Interval estimations for risk of IAS from data set 4

Table 8. Validation results

Category of species	Name of species	NRA score (%)	Approximated risk (Data set 3)	Approximated risk (Data set 4)
Invasive	*Austroeupatorium inulifolium*	62	[55.11774, 65.11774]	[52.75459, 65.50676]
	Panicum maximum	66	[63.28938, 73.28938]	[60.9152, 73.66792]
	Cuscuta campestris	60	[54.29001, 64.29001]	[51.92563, 64.67496]
	Pueraria montana	55	[53.56068, 63.56068]	[51.17714, 63.93692]
	Acacia mearnsii	64	[56.56885, 66.56885]	[54.1979, 66.94356]
	Magnfera indica	36	[33.04537, 43.04537]	[30.67957, 43.40609]
Non invasive	*Cassia fistula*	32	[31.00311, 41.00311]	[28.62525, 41.36012]
	Cissus rotundi	32	[30.02023, 40.02023]	[27.65617, 40.38129]
	Hedychium gardnerianum	32	[32.65081, 42.65081]	[30.28433, 43.01214]
	Magnfera indica	32	[31.04996, 41.04996]	[28.67473, 41.41821]

6. CONCLUSION

The interval multiple linear regression method has been applied for the first time to evaluate the risk of invasive alien species. New method to find the estimates for interval regression coefficient along with interval least square method is proposed. In order to apply the proposed estimation method, interval input-output data and center values of regression coefficients need to satisfy certain conditions. Four different interval input data sets have been incorporated to

approximate the model. To see whether the approximated model is within a satisfactory quality level, average accuracy ratio has been used. The model has been validated with each data set using well known invasive and non invasive species of Sri Lanka. Comparing the quality and the validation results, it is found that the approximated model along with data set 3 gives better predictions of risks of invasive alien species if its invasion is dominated by biological traits. However we should explore to extend the proposed method to estimate the interval regression coefficients without considering the sign of boundaries of input-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 evaluation of overall risk of IAS.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Stockburger WD. Multivariate statistics: Concepts, models, and applications. Missouri State University; 1997.
   (Accessed 4th September 2015)
   Available:http://www.psychstat.missouristate.edu/multibook/mlt08m.html
2. Van De Geer SA. Least squares estimation. In: Everitt BS, Howell DC, editors. Encyclopedia of statistics in behavioral science. Chichester: John Wiley & Sons, Ltd; 2005.
3. Available:http://www.csd.uwo.ca/~moreno/SNC-11-file-for-ACM/p16-Hu.pdf
4. Hu C. Interval function and its linear least-squares approximation. Computer Science Department, University of Central Arkansas. 2011;16-23.
   (Accessed 13th September 2015)
5. Alefeld G, Mayer G. Interval analysis: theory and applications. J Comput Appl Math. 2000;121:421–64.
6. Convention on Biological Diversity. Alien species that threaten ecosystems, habitats and species, Article 8[h]. Secretariat of the Convention on Biological Diversity, United Nations; 2008.
7. Hu C, De Korvin RBKA, Kreinovich V. Knowledge processing with interval and soft computing: Advanced information and knowledge processing (ed. Jain L, Wu X.). Springer; 2008. Ranwala SMW, Risk Assessment for Invasive Alien Species, In Invasive Alien Species - Strengthening capacity to control Introduction and Spread in Sri Lanka (Eds. Marambe B, Silva P, Wijesundera S, Attapattu N), Biodiversity Secretariat, Ministry of Environment and Natural Resources, Sri Lanka; 2010.
8. Rejmanek M, Richardson MD. What attributes make some plants species more invasive. J Ecol. 1996;77:1655-61.

© 2016 Peiris et al.; This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Peer-review history:
The peer review history for this paper can be accessed here:
http://sciencedomain.org/review-history/14527