

Original Research Article

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

ABSTRACT

The abstract is written in a concise, clear, and understandable way. Please correct your abstract.

Aims: To develop a mathematical model for assess the invasion risk of Invasive Alien species by using interval multiple linear regression analysis which mimic unceratin and imprecise data. Here both dependent and independent variables are interval-valued.

Study design: 12 invasive attributes selected as model parameters. Proposed a new method find the solution of design matrix using interval least square method.

Place and Duration of Study: Department of Mathematics, National Institute of Technology Rourkela, India, between August 2015 and November 2015.

Methodology: We prepared a dataset of 28 invasive plant species which contains single-valued observations of 12 parameters and invasion risk scores which 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.

Results: 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. The species *Hedychium gardnerianum* with NRA score of 32 is out of boundaries of estimated risk intervals in data set 1 to 3. The 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.

Conclusion: 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. The proposed estimation should be extended 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

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 [7]. 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 $\bar{x}$ with $\bar{x} \geq \underline{x}$ where $\underline{x}$ and $\bar{x}$ 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:

$$(2.1) \quad y_i = a_0 + a_1x_{i1} + a_2x_{i2} + \dots + a_px_{ip}, \quad i = 1, 2, \dots, n.$$

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 $(a_j, j = 1, 2, \dots, p)$ are determined by the normal equations which are obtained, in such a way that

$\sum_{i=1}^N (a_0 + a_1x_{i1} + a_2x_{i2} + \dots + a_px_{ip} - \hat{y}_i)^2$ is minimal. The least square algorithm to find regression coefficient vector a is defined in [8] as follows:

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

where $a_j \in a$.

Step 2. Form the matrix

59

60 (2.2)

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

61

62 and the column vector,

63 (2.3)

$$b = [y_1 \ y_2 \ \dots \ y_n].$$

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

66 (2.4)

$$a = (A^T A)^{-1} A^T b$$

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

68

69

70 2.2 Interval multiple linear regression model

71

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

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

75 (2.5)

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

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

79

80 2.3 Interval least square

81

82 Let us denote $\tilde{\hat{y}}_i$ be the estimation of interval response variable $\tilde{y}_i$. We need to obtain

83 $\tilde{a}_j$, $j = 0, 1, \dots, m$ which minimize the sum of squared distance $\sum_{i=1}^n d^2(\tilde{y}_i, \tilde{\hat{y}}_i)$. First of all let us

84 define the absolute error of interval estimation as in [4].

85

86 **Definition 2.1 [4]:** Let interval $\hat{y} = [\underline{\hat{y}}, \bar{\hat{y}}]$ be an estimation of an interval $y = [\underline{y}, \bar{y}]$.

87 The left and right absolute errors are $E_L = |\underline{\hat{y}} - \underline{y}|$ and $E_R = |\bar{\hat{y}} - \bar{y}|$, respectively. The
88 absolute error of the estimation is the sum of left and right absolute errors, that is,
89 $E = E_L + E_R = |\underline{\hat{y}} - \underline{y}| + |\bar{\hat{y}} - \bar{y}|$, respectively.

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

92

93

94 **Definition 2.2 [3]:** Let U be the set of n interval valued observations of an interval linear
 95 regression function $y = h(x)$ i.e. $U = \{(x, y) : x \in [0, 1], y \in [0, 1], \text{ both } x \text{ and } y \text{ are compact}\}$.

96 According to the definition 1, we say that $\sum_{0 \leq j \leq m} \tilde{a}_j \tilde{x}_{ij}$ is an interval least square estimation of

97 $\tilde{y}_i$ if the linear combination minimizes:

98
$$\sum_{i=1}^n E_L^2 + \sum_{i=1}^n E_R^2,$$

99 where

100 (2.6)
$$\sum_{i=1}^n E_L^2 = \sum_{i=1}^n \left(y_i - \left(\sum_{0 \leq j \leq m} a_j x_{ij} \right) \right)^2,$$

101 and

102 (2.7)
$$\sum_{i=1}^n E_R^2 = \sum_{i=1}^n \left(\bar{y}_i - \left(\sum_{0 \leq j \leq m} \bar{a}_j \bar{x}_{ij} \right) \right)^2.$$

103

104 2.4 Interval arithmetic (ILS)

105

106 While using ILS method, it is required to find the estimates for interval regression coefficients
 107 that minimize the sum of Eqs. (2.6) and (2.7). Therefore we need to use interval arithmetic to
 108 work with interval-valued data [1].

109 Let $[a] = [\underline{a}, \bar{a}]$, $[b] = [\underline{b}, \bar{b}]$ be real compact intervals and $\circ$ is one of the basic operations
 110 'additions', 'subtraction', 'multiplication' and 'division, respectively (for real numbers) that is
 111 $\circ \in \{+, -, \cdot, \div\}$.

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

113 for the corresponding operations:

114 $[a] + [b] = [\underline{a} + \underline{b}, \bar{a} + \bar{b}],$

115 $[a] - [b] = [\underline{a} - \bar{b}, \bar{a} - \underline{b}],$

116 $[a] \cdot [b] = [\min\{\underline{a} \cdot \underline{b}, \underline{a} \cdot \bar{b}, \bar{a} \cdot \underline{b}, \bar{a} \cdot \bar{b}\}, \max\{\underline{a} \cdot \underline{b}, \underline{a} \cdot \bar{b}, \bar{a} \cdot \underline{b}, \bar{a} \cdot \bar{b}\}],$

117 $[a] \div [b] = [\min\{\underline{a} \div \underline{b}, \underline{a} \div \bar{b}, \bar{a} \div \underline{b}, \bar{a} \div \bar{b}\}, \max\{\underline{a} \div \underline{b}, \underline{a} \div \bar{b}, \bar{a} \div \underline{b}, \bar{a} \div \bar{b}\}],$ provided $0 \notin [b]$.

118

119 2.5 Proposed estimation method

120

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

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

126

127 (2.8)
$$\tilde{Y}_i = \tilde{a}_0 + \tilde{a}_1 \tilde{x}_{i1} + \tilde{a}_2 \tilde{x}_{i2} + \cdots + \tilde{a}_m \tilde{x}_{im} \quad i = 1, 2, \dots, n$$

128 where $\tilde{Y}_i$ is the interval response variable, $\tilde{x}_{ij}$, $j = 0, 1, \dots, m$ are the interval predictor
 129 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 $j = 0, 1, \dots, m$ 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.

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

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

$$(2.10) \quad \underline{a}_j = a'_j - \varepsilon_j,$$

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

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:

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

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

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

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

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

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

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

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

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

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

169
 170 (2.14)
$$\underline{a}_0 + \underline{a}_1 \bar{x}_{i1} + \underline{a}_2 \underline{x}_{i2} + \dots + \underline{a}_n \underline{x}_{in} < \underline{a}_0 + \underline{a}_1 \underline{x}_{i1} + \underline{a}_2 \underline{x}_{i2} + \dots + \underline{a}_n \underline{x}_{in}, i = 1, 2, \dots, n$$

 171

172 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
 173 is clear that $[\hat{y}'_L, \hat{y}_U] \subseteq [\hat{y}_L, \hat{y}_U]$. Similarly we can prove $[\hat{y}'_L, \hat{y}_U] \subseteq [\hat{y}_L, \hat{y}_U]$ for more than
 174 one $\underline{a}_j < 0$. One may note that the values of $\forall \underline{a}_j, j = 0, 1, \dots, m$ cannot be negative since the
 175 boundary values of $\tilde{Y}_i$ and $\tilde{x}_{ij}$ are non-negative.

176
 177 Now we propose our new method to find the boundaries of interval regression coefficient of
 178 Eq. (2.8) as below.

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

180 (2.15)
$$\underline{y}_i = \underline{c}_0 + \underline{c}_1 \underline{x}_{i1} + \dots + \underline{c}_m \underline{x}_{im},$$

181
 182 (2.16)
$$\bar{y}_i = \bar{c}_0 + \bar{c}_1 \bar{x}_{i1} + \dots + \bar{c}_m \bar{x}_{im},$$

183
 184 where $\underline{c}_j = \underline{a}_j - \varepsilon_j$, $\bar{c}_j = \bar{a}_j - \varepsilon_j$ for $j = 1, 2, \dots, m$ which minimizes the following
 185

186 (2.17)
$$\sum_{i=1}^N E^2_L = \sum_{i=1}^N \left(\underline{y}_i - \left(\sum_{0 \leq j \leq m} \underline{c}_j \underline{x}_{ij} \right) \right)^2,$$

187 (2.18)
$$\sum_{i=1}^N E^2_R = \sum_{i=1}^N \left(\bar{y}_i - \left(\sum_{0 \leq j \leq m} \bar{c}_j \bar{x}_{ij} \right) \right)^2.$$

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

191 192 **2.6 Accuracy assessment of approximated model**

193 The quality of the approximation of $\tilde{Y}_i \approx \left(\sum_{0 \leq j \leq m} \tilde{a}_j \tilde{x}_{ij} \right)$, can be examined by considering the
 194 overlap between approximated output $\hat{\tilde{Y}}$ and expected output $\tilde{Y}$. The approximation would
 195 be better if the overlap between $\hat{\tilde{Y}}$ and $\tilde{Y}$ is considerably large. The accuracy ratio of an
 196 interval approximation is defined in [4] as below.

197 **Definition 2.3 [4]:** Let $\hat{\tilde{Y}} = [\hat{\underline{y}}, \hat{\bar{y}}]$ be an approximation for the interval $\tilde{Y} = [\underline{y}, \bar{y}]$.
 198 The accuracy ratio of the approximation is

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

200 where the function $w()$ returns the width of an interval. In this study, we consider the
 201 average accuracy ratio to assess the quality of an interval approximation on a set of N
 202 interval pairs (x_i, y_i) qualitatively.

203 For a set of N interval pairs (x_i, y_i) , the average accuracy ratio [3] of the approximation is
 204 defined as

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

207
 208

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

213

214 3. METHODOLOGY

215

216 3.1 Selection of parameters

217

218 According to the view of biologists several factors affect the risk of invasive plant species
 219 such as its ecology, establishment, invasive potential, management aspects etc [2, 5, 6]. In
 220 conventional risk assessments these factors are usually considered. In this work we are
 221 mainly concerned about the biological traits related to invasive potential. The most important
 222 12 biological traits are selected as the parameters of the model from National Risk
 223 Assessment (NRA) for alien invaders in Sri Lanka. These parameters may be written as
 224 below:

225

- 226 • Number of seeds per fruit (SF)
- 227 • Annual seed production per m^2 (ASR)
- 228 • Viability of seeds (VS)
- 229 • Long distance dispersal strength (LDD)
- 230 • Vegetative reproduction strength (VRS)
- 231 • Seed germination requirements (SGR)
- 232 • Presence of physical defensive structures (PDS)
- 233 • Formation of climbing or smothering growth habit (FCS)
- 234 • Potential to be spread by human activities (HA)
- 235 • Role of natural and manmade disturbances (NMD)
- 236 • Alleopathic property (AP)
- 237 • 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

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

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

275 **Table 2. Interval estimates for coefficients from data set 1**
276

Coefficient	Estimates from data set 1
θ_0	[10.344, 11.814]
θ_1	[0.0183830220805138, 0.0183830220805161]
θ_2	$[5.40300982218432 \times 10^{-6}, 5.40300982218483 \times 10^{-6}]$
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]

277
278
279 **Table 3. Interval estimates for coefficients from data set 2**
280

Coefficient	Estimates from data set 2
θ_0	[10.3943, 11.745]
θ_1	[0.0183106933740867, 0.0183798813271578]
θ_2	$[5.40142448033662 \times 10^{-6}, 5.40644416509869 \times 10^{-6}]$
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, 1.35368206335038]
θ_{11}	[2.42463624793798, 2.42513167844029]
θ_{12}	[2.51340075543958, 2.51512964449058]

281
282
283
284
285
286
287
288
289
290
291
292

293 **Table 4. Interval estimates for coefficients from data set 3**

294

Coefficient	Estimates from data set 3
θ_0	[9.65394, 12.34399]
θ_1	[0.0183830220805151, 0.01886]
θ_2	$[5.40300982218465 \times 10^{-6}, 5.81 \times 10^{-6}]$
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]

295

296

297

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	$[5.40142448033665 \times 10^{-6}, 5.40644416509869 \times 10^{-6}]$
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]

298

299

300

4.2 Accuracy assessment

301

302

303

304

305

306

307

308

309

310

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 ‘*’, ‘x’, ‘-’ 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

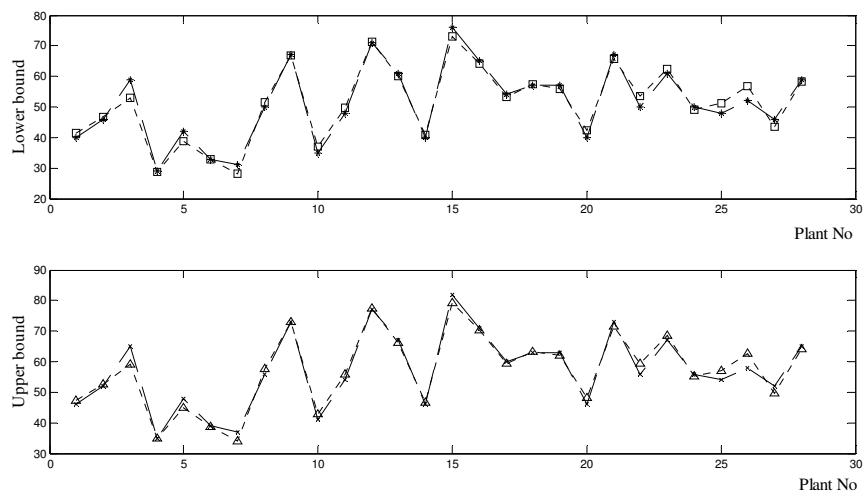


Fig. 1. Interval estimations for risk of IAS from data set 1

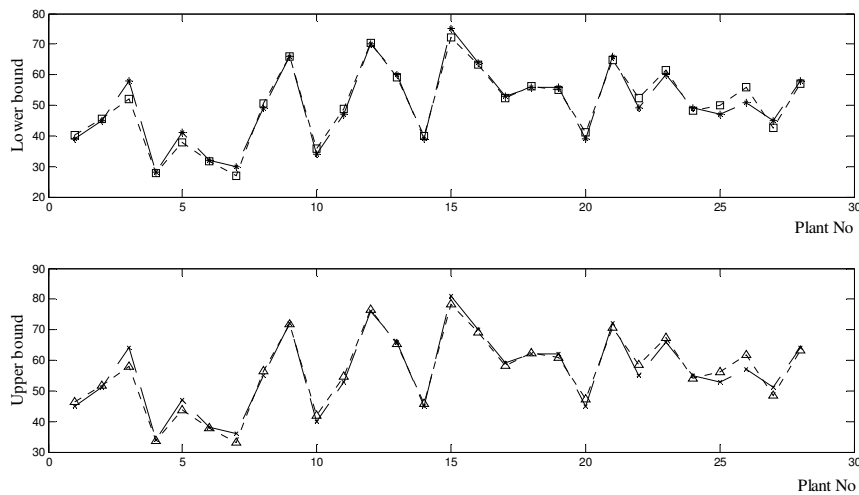


Fig. 2. Interval estimations for risk of IAS from data set 2

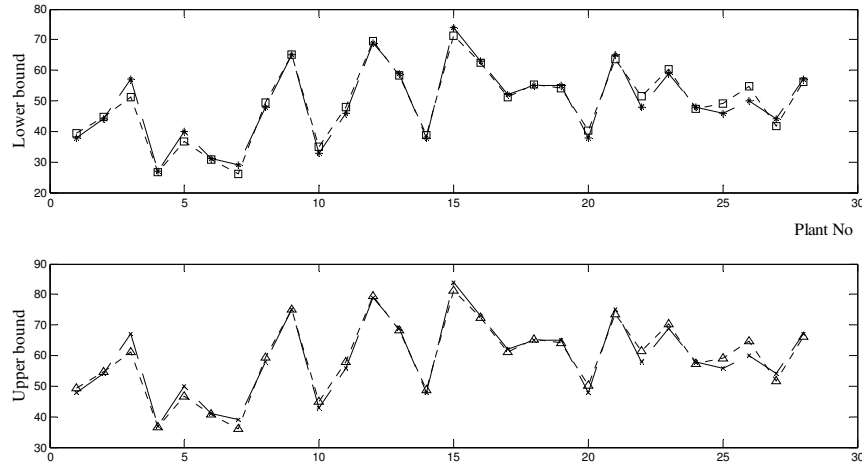


Fig. 3. Interval estimations for risk of IAS from data set 3

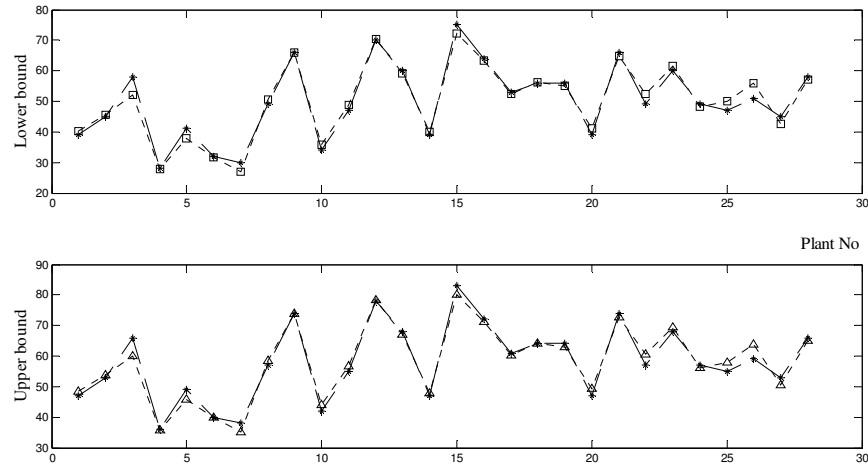


Fig. 4. Interval estimations for risk of IAS from data set 4

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.

417 **Table 7. Validation results**

418

419

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

420

421

Table 8. Validation results

Category of species	Name of species	NRA score (%)	Approximated Risk (Data set 3)	Approximated Risk (Data set 4)
Invasive	<i>Austroeupatorium inulifolium</i>	62	[55.11774, 65.11774]	[52.75459, 65.50676]
	<i>Panicum maximum</i>	66	[63.28938, 73.28938]	[60.9152, 73.66792]
	<i>Cuscuta campestris</i>	60	[54.29001, 64.29001]	[51.92563, 64.67496]
	<i>Pueraria montana</i>	55	[53.56068, 63.56068]	[51.17714, 63.93692]
	<i>Acacia mearnsii</i>	64	[56.56885, 66.56885]	[54.1979, 66.94356]

	<i>Magnifera indica</i>	36	[33.04537, 43.04537]	[30.67957, 43.40609]
Non invasive	<i>Cassia fistula</i>	32	[31.00311, 41.00311]	[28.62525, 41.36012]
	<i>Cissus rotundi</i>	32	[30.02023, 40.02023]	[27.65617, 40.38129]
	<i>Hedychium gardnerianum</i>	32	[32.65081, 42.65081]	[30.28433, 43.01214]
	<i>Magnifera indica</i>	32	[31.04996, 41.04996]	[28.67473, 41.41821]

4.4 Discussion

From Tables 2 to 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 to 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.

5. 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.

REFERENCES

1. Alefeld G and Mayer G. Interval analysis: theory and applications. J Comput Appl Math. 2000; 121:421–64.
2. 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.
3. Hu C, Interval Function and its Linear Least-squares Approximation. Computer Science Department, University of Central Arkansas. 2011; 16-23. Accessed 13th September 2015.
Available: <http://www.csd.uwo.ca/~moreno/SNC-11-file-for-ACM/p16-Hu.pdf>.
4. Hu C, De Korvin RBKA and Kreinovich V. Knowledge Processing with Interval and Soft Computing: Advanced Information and Knowledge Processing (ed. L. Jain, X. Wu). Springer; 2008.
5. Ranwala SMW, 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. 2010.
6. Rejmanek M and Richardson MD, What attributes make some plants species more invasive. J Ecol. 1996; 77:1655-61.
7. 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>.
8. 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.