Journal of Theoretical and Applied Information Technology
30th April 2019. Vol.97. No 8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

HANDLING OF MISSING DATA FOR E-READINESS INDICATORS

[1]RABII LAMRIQ, [2]ABDELAZIZ DOUKKALI, [3]NAJIB BELKHATAT

[1]Ph. D Student., TIES team, ENSIAS, MedV Souissi University, RABAT

[2]Assoc Prof., TIES team, ENSIAS, MedV Souissi University, RABAT

[3]Assoc Prof., TIES team, ENSIAS, MedV Souissi University, RABAT

E-mail: [1]lamriq.rabii@gmail.com, [2]doukkali@ensias.ma, [3]najib.belkhayat@gmail.com

ABSTRACT

Missing data is one of the main problems associated with composite indicators of electronic readiness (e-readiness), but the way in which these missing values are processed can have a serious impact on the results of e-readiness assessments. The complexity of this problem increases with the number of missing values. However, despite the known limitations on the performance of some missing data processing methods, such as imputation based on the following year's values or the average of previous years' values, many composite indices of e-readiness continue to use these methods. The main objective of this article is to improve the estimation of missing data in a dataset used by the Networked Readiness Index (NRI) organisation. In order to improve existing estimates, we establish a predictive model based on multiple linear regressions for each indicator containing missing values. We also use variable selection techniques to choose the best input variables for each model.

Keywords: *E-Readiness, Missing Data, Imputation, Variable Selection, Linear Regression, Composite Indicators.*

1. INTRODUCTION

Information and communication technology (ICT) play a significant role in many areas of society around the world. In particular, ICT has penetrated social and economic activities and has changed their traditional methods. Thus, these transformations positively impacted communication, the financial sectors and contributed to the growth of the economy. This strong correlation between ICT use and economic development was the subject of numerous studies that proved this reality .

Thus, several organizations were interested in collecting data and measuring indicators that reflect the use of ICT in different sectors related to the economy, or in assessing the relationship between the degree of ICT use and economic progress. Other studies called: e-readiness assessment, gauge the readiness of each country and its ability to benefit from ICT and innovation for its economic development.

However, most indicators used in statistical studies suffer from a lack of data. In particular, many datasets collected by e-readiness assessment organisations experience a significant absence of certain measures, and especially measures collected for developing and least developed countries.

This issue of missing data may be due to several reasons:

(a) Data collection tools: Since most of the information on the use of information and communication technologies (ICT) comes from surveys, this collection method can generate many missing values, arising from a lack of answers for certain items [1], [2].

(b) Environmental instability: When the conditions within a region or country are unfavourable due to internal conflicts, wars or natural disasters, the gathering of information or measurements becomes very difficult [3], [4];

(c) Creation of new indicators: Due to the rapid development of ICT technologies, e-readiness

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

assessment organisations frequently modify indicators or define new ones, thus making data collection difficult for some countries [5].

The purpose of this study is to explore the use of machine learning methods to improve the estimation of missing data in e-readiness datasets. The method proposed in this paper is based on multiple linear regression and variable selection algorithms. This article is organised as follows: Section 2 provides an overview of composite indicators of e-readiness. Section 3 examines the problem of missing data in the evaluation of e-readiness. Subsequently, a new estimate of the missing data and its application is described in Sections 4 and 5. Finally, some conclusions are drawn in Section 6.

2. RELATED WORK

2.1 E-readiness composite indicators

It appears that there is no general definition of e-readiness in the literature; however, e-readiness generally refers to the use of ICT in the economy and in industry [5], [6], [7]. Similarly, e-readiness assessment has been defined in many studies as a new tool that is designed to give an overall picture, represented as a composite index based on measurable variables related to economic and social development. This index is calculated using a composite indicator approach involving the following steps [8]: data selection, data processing, data normalisation, data weighting and finally data aggregation.

This study focuses on the data processing stage, in which the issue of missing data is handled. In the literature, several processing methods for missing data are reviewed. However, some studies of e-readiness assessment do not describe how they handle missing values [5], and few have solved the problem of missing data. In addition, the ways in which missing values are processed can have a significant impact on the reliability of the e-readiness index, and can give a misleading result in terms of the ranking of countries. The use of advanced techniques to deal with missing data is strongly recommended in any calculation of composite indicators, and we therefore aim in this paper to re-examine all the methods used by e-readiness indices and to propose a new estimate that is capable of providing plausible values for missing data.

2.2 Missing data handling methods used in e-readiness

In this section, we analyse several methods that are frequently used by the most well-known e-readiness composites in the literature [8].

(i) **Hot-deck imputation:** This is a very commonly used technique for missing data problems. Its principle consists of using two similar units from the same set to replace the missing value of the destination unit, based on the corresponding value from a donor unit [9]. In the case of an e-readiness assessment, this method requires a strong similarity between the destination country with missing values and the donor country. However, given the limit on the number of countries, the existence of a strong similarity is not always possible.

(ii) **Nearest year** (or average of last two years): This method is simple to use when data are available from previous years. In general, missing data are estimated based on the average of two or more values from previous years. One limitation of this method is that the estimate depends on the availability of the values of the nearest years, meaning that imputation from a distant year (last forest year for example) is not a good estimate. In addition, the estimated value depends solely on the variation in the variable itself, and does not take into account the variation in the other variables in the dataset.

(iii) **Growth rate:** This method uses the time series concept to estimate and predict missing data. The use of time series typically requires a large number of observations in order to give an accurate estimate.

Other methods have been proposed in the literature, such as the Kalman filter used by Belkhayat [5] to produce an estimate and a prediction of missing data. In fact, he modelled the evolution of layered e-readiness indicators based on the concept of inter-indicator impactability, leading to a dynamic system state model. Belkhayat applied the Kalman filter to the "i2010 Initiative" dataset to test the convergence of the state model and to estimate the missing data [5].

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

2.3 Missing data in NRI

NRI is a composite indicator developed in 2002 by World Economic Forum in collaboration with INSEAD. the NRI index is designed to reflect the readiness of countries and their ability to benefit from the technological revolution. the data used in the calculation of the NRI index is derived from 53 indicators divided into 4 categories. Table 1 shows the categories of the NRI indicators [10].

Table 1: Description of NRI variables

Xi	Categories	Description
x_1	Environment	Effectiveness of law-making bodies, 1-7 (best)
x_2		Laws relating to ICTs, 1-7 (best)
x_3		Judicial independence, 1-7 (best)
x_4		Efficiency of legal system in settling disputes, 1-7 (best)
x_5		Efficiency of legal system in challenging regs, 1-7 (best)
x_6		Intellectual property protection, 1-7 (best)
x_7		Software piracy rate, % software installed
x_8		No. of procedures to enforce a contract
x_9		No. of days to enforce a contract
x_{10}		Availability of latest technologies, 1-7 (best)
x_{11}		Venture capital availability, 1-7 (best)
x_{12}		Total tax rate, % profits
x_{13}		No. of days to start a business
x_{14}		No. of procedures to start a business
x_{15}		Intensity of local competition, 1-7 (best)
x_{16}		Tertiary education gross enrolment rate, %
x_{17}		Quality of management schools, 1-7 (best)
x_{18}		Gov't procurement of advanced tech, 1-7 (best)
x_{19}	Readiness	Electricity production, kWh/capita
x_{20}		Mobile network coverage, % pop.
x_{21}		Int'l Internet bandwidth, kb/s per user
x_{22}		Secure Internet servers/million pop.
x_{23}		Prepaid mobile cellular tariffs, PPP $/min.
x_{24}		Fixed broadband Internet tariffs, PPP $/month
x_{25}		Internet & telephony competition, 0–2 (best)
x_{26}		Quality of educational system, 1-7 (best)
x_{27}		Quality of math & science education, 1-7 (best)
x_{28}		Secondary education gross enrolment rate, %
x_{29}		Adult literacy rate, %
x_{30}	Usage	Mobile phone subscriptions/100 pop.
x_{31}		Individuals using Internet, %
x_{32}		Households w/ personal computer, %
x_{33}		Households w/ Internet access, %
x_{34}		Fixed broadband Internet subs/100 pop.
x_{35}		Mobile broadband subs/100 pop.
x_{36}		Use of virtual social networks, 1-7 (best)
x_{37}		Firm-level technology absorption, 1-7 (best)
x_{38}		Capacity for innovation, 1-7 (best)
x_{39}		PCT patents, applications/million pop.
x_{40}		ICT use for business-to-business transactions, 1-7 (best)
x_{41}		Business-to-consumer Internet use, 1-7 (best)
x_{42}		Extent of staff training, 1-7 (best)
x_{43}		Importance of ICTs to gov't vision, 1-7 (best)
x_{44}		Government Online Service Index, 0–1 (best)
x_{45}		Gov't success in ICT promotion, 1-7 (best)
x_{46}	Impact	Impact of ICTs on business models, 1-7 (best)
x_{47}		ICT PCT patents, applications/million pop.
x_{48}		Impact of ICTs on new organisational models, 1-7 (best)
x_{49}		Knowledge-intensive jobs, % workforce
x_{50}		Impact of ICTs on access to basic services, 1-7 (best)
x_{51}		Internet access in schools, 1-7 (best)
x_{52}		ICT use & gov't efficiency, 1-7 (best)
x_{53}		E-Participation Index, 0-1 (best)

In the next section, we will describe the proposed approach for processing and estimating all missing data in the Networked Readiness Index (NRI) database for the year 2016. This database contains 104 missing values affecting 13 indicators from 75 countries. Table 2 gives details of all indicators with missing data.

The proposed approach is based on the application of machine learning algorithms to the NRI dataset in order to develop a predictive model that is capable of modelling a linear relationship between the dependent output variable (which contains missing values) and the independent input variables of the dataset.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 1: List of variables with missing values and countries impacted

Indicator	Description	Number of missing data	Impacted countries
x_7	Software piracy rate, % software installed	35	Benin, Bhutan, Burundi, Cambodia, Cape Verde, Chad, Ethiopia, Gabon, Gambia, Ghana, Guinea, Guyana, Haiti, Iran Islamic Rep, Jamaica, Kyrgyz Republic, Lao PDR, Lesotho, Liberia, Madagascar, Malawi, Mali, Mauritania, Mongolia, Mozambique, Myanmar, Namibia, Nepal, Rwanda, Seychelles, Swaziland, Tajikistan, Tanzania, Trinidad and Tobago, Uganda
x_{49}	Knowledge-intensive jobs, % workforce	29	Bahrain, Benin, Bosnia and Herzegovina, Burundi, Cameroon, Cape Verde, Chad, China, Côte d'Ivoire, Gabon, Gambia, Guyana, Haiti, Honduras, India, Jordan, Kenya, Kuwait, Lao PDR, Malawi, Mali, Mauritania, Mozambique, Myanmar, Nigeria, Oman, Senegal, Swaziland, Tajikistan
x_{29}	Adult literacy rate, %	23	Australia, Austria, Belgium, Canada, Czech Republic, Denmark, Finland, France, Germany, Hong Kong SAR, Iceland, Ireland, Israel, Japan, Korea, Luxembourg, Netherlands, New Zealand, Norway, Sweden, Switzerland, United Kingdom, United States
x_{16}	Tertiary education gross enrolment rate, %	2	Canada, Zambia
x_{24}	Fixed broadband Internet tariffs, PPP $/month	2	Argentina, Guinea
x_{25}	Internet & telephony competition, 0–2 (best)	2	Mongolia, Venezuela
x_{39}	PCT patents, applications/ million pop.	2	Hong Kong SAR, Taiwan China
x_{44}	Government online service index, 0–1 (best)	2	Hong Kong SAR, Taiwan China
x_{47}	ICT PCT patents, applications/ million pop.	2	Hong Kong SAR, Taiwan China
x_{53}	E-participation index, 0–1 (best)	2	Hong Kong SAR, Taiwan China
x_{20}	Mobile network coverage, % pop.	1	Tajikistan
x_{23}	Prepaid mobile cellular tariffs, PPP $/min.	1	Argentina
x_{28}	Secondary education gross enrolment rate, %	1	Zambia

3. METHODOLOGY

For each variable x_i containing missing values, a predictive model is developed using the multiple linear regression method. Each model represents a linear relationship between the dependent variable x_i and the independent variables in the dataset. This predictive model will be then used to produce estimates of missing values.

3.1 Multiple linear regression model

Linear regression is a powerful method that is used to analyse the relationship between variable and build predictive models. The multiple linear

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

regression model used in this mathematical analysis is given in Equation (1) below [11]:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \quad (1)$$

where $\{x_1, x_2, \ldots, x_p\}$ are independent variables; y is the output or dependent variable; β_{ij} are the unknown regression coefficients of the multiple linear regression model; and ε is a random error. The parameters β_{ij} are estimated using the ordinary least squares (OLS) method.

In general, a better multiple linear regression model can be obtained by finding the subset of the best independent predictor variables, i.e. the subset that minimises the error ε, in order to obtain very good accuracy of prediction for the output variable y. The classical method involves performing all possible combinations if p is not large. However, the number of possible models is 2^pcombinations; in the case of a large number of predictors, and particularly when p ≥ 30, this is a very large number of calculations that cannot be fully performed due to the computational complexity involved.

This complexity has been identified in many datasets used by e-readiness assessments with more than 30 indicators: for example, 52 indicators have been identified in the National E-Readiness Measurement Framework [6], 53 in the NRI [10], and 44 in the E-Trade Readiness Index [8]. In addition, the number of indicators is constantly increasing due to the rapid development of ICT. For this reason, the use of a simple technique to select the best variables or perform all combinations is not feasible in e-readiness datasets.

To overcome this complexity, we propose the use of other methods of advanced variable selection or subset selection, as described in the next section, to find the best subset that can provide a good regression model.

3.2 Variable selection

Variable selection techniques are used when the number of predictor variables is very important in a dataset. The main reasons for using them are to increase the prediction accuracy of the regression model and reduce computational complexity in order to find the best model [12]. To facilitate the operation of these regression models, there are three methodologies for under-defining the feature space and selecting the best predictor variables. The first uses filtering methods in which variables are selected based on their importance using statistical tests, such as Pearson's correlation coefficient, information gain, random forest, chi-square test, and so on. The second is the wrapper method, in which the selection of a subset is treated as an optimisation problem that involves searching for the best predictor variables, using a predefined precision factor to evaluate the generated subset. Some common examples of wrapper methods are backward selection, forward elimination, elimination of recursive variables, genetic algorithm and colony ants. These methods are usually very expensive. The third method is a combination of filtering and wrapper methods, called the hybrid method. Two very popular examples of this method are: (i) a ranked forward search (RFS), which searches for the best subset of the ranked variables using forward variable selection; and (ii) a refined exhaustive search (RES), which works in the same way as RFS except that it examines all possible combinations of ranked variable space to find the best subset [13].

In this paper, we propose the use of RES, the second of these methods, with three statistical criteria to classify variables by importance: (a) correlation ranking; (b) ranking based on information gain; and (c) ranking based on the random forest algorithm. We then compare these methods with a random selection of all variables. This comparison was made using validation criteria.

3.3 Validation of the selection model

Several criteria are used to evaluate the selected subset, for example the multiple coefficient of determination R^2, adjusted R^2, C mean residual square, Mallow's Cp statistics, the Akaike information criterion (AIC), Bayesian analogies (BIC) and statistical prediction residual sum of squares (PRESS) [11]. The criterion chosen for the validation of the subset depends on the intended use of the model. Since our regression model will be used to predict or estimate the missing value of the output, the appropriate criterion is the PRESS statistic [11].

The PRESS statistic is given in Equation (3), as the sum of squared PRESS residues defined in Equation (2) [9].

$$e_{(i)} = y_i - \hat{y}_{(i)} \quad (2)$$

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

where $\hat{y}_{(i)}$ is the predicted value adjusted by the model used, and y_i is the value of the test data.

$$PRESS = \sum_{i=1}^{n} [y_i - \hat{y}_{(i)}]^2 \quad (3)$$

The algorithm below illustrates the steps followed to find the best subset corresponding to the minimum value of PRESS, using the cross-validation method.

Algorithm 1: Embedded feature selection

1. Define the training data and test data for the dependent variable X_i.
2. Calculate the importance of variables or their ranking.
3. Select the m best ranked variables ($20 \leq m \leq 25$).
4. Generate all possible combinations of predictor variables.
5. **For** each combination S_i, $i = 1\ to\ 2^m$ **do**
6. Train the linear regression model using the training data.
7. Generate predictions using test data.
8. Calculate the PRESS statistical coefficient.
9. Retain the combination of predictor variables that minimises the PRESS coefficient.
10. **End**.
11. Retain the subset of predictor variables S_i that minimises the PRESS coefficient.

The same algorithm was used for the three statistical tests: Pearson correlation, information gain and random forest. For the selection of random variables, we proceeded with the same steps, except for steps (3) and (4).

Algorithm 2: Feature random selection

1. Define the training data and test data for the dependent variable X_i.
2. **Do**
3. Randomly select a subset of the predictor variables.
4. Train the linear regression model using the training data.
5. Generate predictions using test data.
6. Calculate the PRESS statistical coefficient.
7. Retain the combination of predictor variables that minimises the PRESS coefficient.
8. **Until** A convincing minimal value of PRESS is reached.
9. **End**.
10. Retain the subset of predictor variables S_i that minimises the PRESS coefficient.

3.4 Software

The two algorithms mentioned in subsection 3.3 are implemented in the R statistic software. For the calculation of important variables, we used the library "randomForest" for Random Forest, "FSelector" for Information Gain and "Stats" for Pearson Correlation. these packages are available on the R libraries website [14].

4. DATA

4.1 Description of the NRI indicator database

The NRI dataset was published by the World Economic Forum (WEF) for the computation of indices for e-readiness [15]. It contains data for 53 individual indicators, represented by four main categories (environment, readiness, use and impact), and covers more than 150 countries [10]. The overall dataset used for processing missing data is a concatenation of annual data from 2012 to 2016, and contains a total of 775 lines. Each line is an observation that contains 54 columns, where the first column represents the name of the country and the last 53 represent the individual indicators.

4.2 Data preparation

The database is divided into three sets of data. The first is the training data; this contains data from 2012 to 2015 and is intended for learning and constitution of the regression model for each target variable. Rows in this training set are cleaned if they contain one or more missing values. The second set is the test data, which contains only data from 2016 with no missing values, and is used for validation of the regression model based on the PRESS criterion.

The third set is the missing data, and this contains only data from 2016 that contain missing values.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

5. RESULTS AND DISCUSSION

We used Algorithm 1 with the three integrated methods using a filter (where Method 1 was based on correlation, Method 2 on random forest and Method 3 on information gain), and Algorithm 2 with the random selection method. Both algorithms were applied to the training dataset. Each model set in the training set was tested using the test dataset to calculate the PRESS value. The goal was to find the model that gave the minimum value of PRESS.

Table 3 shows the results of the cross-validation used to validate the best model, using the PRESS statistic as a criterion.

Table 3: Comparison between predictive models using PRESS statistic

Variables	PRESS statistic			
	COR	I.G	R.F	R.S
x_7	22.28	21.51	29.73	15.62
x_{16}	34.09	33.96	32.47	21.40
x_{20}	4.88	5.24	4.04	4.31
x_{23}	46.99	51.28	47.59	38.27
x_{24}	0.33	0.27	0.28	0.07
x_{25}	66.58	54.37	49.76	44.19
x_{28}	12.26	12.72	12.84	9.78
x_{29}	10.66	11.34	11.74	10.53
x_{39}	3.79	4.11	4.16	0.70
x_{44}	46.74	42.16	39.78	25.87
x_{47}	2.53	2.95	2.82	0.51
x_{49}	19.93	19.02	20.7	18.06
x_{53}	72.59	99.69	62.63	40.78

To evaluate the accuracy of the predictive model, we used the Pearson correlation (PC) between the real value and the estimated value. The result is figured in table 4.

Table 4: Comparison between predictive models using Pearson Correlation

Variables	Pearson Correlation			
	COR	I.G	R.F	R.S
x_7	0.9	0.9	0.86	0.93
x_{16}	0.78	0.78	0.8	0.85
x_{20}	0.46	0.37	0.56	0.56
x_{23}	0.51	0.43	0.5	0.67
x_{24}	0.44	0.59	0.56	0.73
x_{25}	0.42	0.58	0.63	0.75
x_{28}	0.79	0.79	0.78	0.87
x_{29}	0.85	0.83	0.82	0.89
x_{39}	0.75	0.73	0.72	0.8
x_{44}	0.71	0.74	0.77	0.81
x_{47}	0.75	0.7	0.71	0.79
x_{49}	0.84	0.84	0.86	0.89
x_{53}	0.56	0.59	0.66	0.73

The results show that the random selection method outperforms the three integrated methods in terms of PRESS statistics for all variables except X_{20}, where the RF-based RES method gives a better value for PRESS than the random selection method.

The table 5 shows in ascending order the residual between the real values of the variable X_{23} and the values predicted using the final model chosen for all countries in the NRI dataset for 2016. Appendix I shows a comparison graph of the real values and the predicted values for all 13 variables.

Table 5: Comparison between predictive models using PRESS statistic

Country	Real value	Predicted value	Residual
BRA	5.54	5.53	0.01
BWA	5.09	5.05	0.04
PRY	5.44	5.40	0.04
LTU	5.81	5.86	0.05
TUR	6.54	6.46	0.08
CHL	5.48	5.39	0.09
LKA	6.78	6.65	0.13
KAZ	6.45	6.59	0.14
EST	5.43	5.57	0.14
ECU	5.31	5.15	0.16
IDN	6.10	6.27	0.17
ITA	5.76	5.58	0.18
PAN	5.94	6.12	0.18
ZWE	4.96	4.77	0.19
VNM	6.28	6.49	0.21
EGY	6.71	6.94	0.23
LVA	6.16	5.93	0.23
SVN	5.60	5.35	0.25
SLV	5.53	5.80	0.27
PHL	5.13	4.82	0.31
MDA	5.94	5.61	0.33
LBN	4.75	5.09	0.34
MKD	6.05	5.69	0.36
BOL	4.83	5.19	0.36
COL	5.61	5.25	0.36
MNE	5.79	5.41	0.38
ZAF	5.96	5.57	0.39
ARM	5.97	6.36	0.39
POL	6.26	5.83	0.43
MYS	6.23	5.79	0.44
AZE	5.37	5.83	0.46
BGD	6.83	6.36	0.47
PER	5.51	5.03	0.48
SGP	6.13	5.63	0.50
DOM	4.77	5.28	0.51
HRV	5.73	5.21	0.52
MUS	6.17	5.63	0.54
PAK	6.71	7.25	0.54
PRT	6.37	5.83	0.54
SVK	5.88	5.34	0.54
URY	5.18	5.73	0.55
MLT	5.05	4.46	0.59
MAR	6.34	5.74	0.60
SAU	5.52	6.14	0.62
UKR	6.21	5.59	0.62

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

CYP	6.42	5.79	0.63
ARE	6.30	5.63	0.67
DZA	5.68	4.95	0.73
RUS	6.86	6.13	0.73
SRB	5.92	5.18	0.74
GEO	6.58	5.81	0.77
CRI	6.59	5.81	0.78
TUN	6.71	5.87	0.84
MEX	6.43	5.46	0.97
QAT	5.97	5.00	0.97
HUN	5.74	4.75	0.99
THA	6.60	5.57	1.03
GTM	4.08	5.21	1.13
ESP	6.31	5.13	1.18
ROU	4.28	5.81	1.53
ALB	3.66	5.46	1.80
BGR	3.33	5.47	2.14
GRC	3.34	5.59	2.25
NIC	1.48	3.86	2.38

Note that the real and estimated values are very close for the majority of countries; the exceptions are Bulgaria, Greece, and Nicaragua, which show a marked difference between the real values and the estimated values. Furthermore, the Pearson correlation, which is used as a precision indicator to validate the regression model of the 13 variables, contains two small decreases of 0.56 and 0.67 for the two variables X_{20} and X_{23}, respectively, although these values remain acceptable. This shows that these models can plausibly estimate missing data in the missing NRI 2016 dataset.

The missing values of each variable in the NRI 2016 dataset were estimated by replacing the input variables in each final predictive model, which were chosen according to the PRESS criteria for each of the variables. Table 6 (Appendix II) gives the 104 estimated values for the 13 variables for each of the 75 countries.

6. CONCLUSION

This article began by discussing the importance of processing missing data for electronic composite indicators, many approaches to which still use conventional methods that do not provide a plausible estimate. To solve this problem, we proposed the use of multiple linear regression as a method for estimating the missing values of the e-readiness dataset, with the help of variable selection techniques to handle the large number of independent variables in the model. A comparison of four methods of variable selection was then given. Three of these are RES methods using filtering of important variables: the first is based on the Pearson correlation, the second on the random forest method and the third on the information gain. The final method is random selection. The results show that the best method of selecting predictor variables is random selection.

REFRENCES:

[1] Y. Dong et C. Y. J. Peng, *Principled missing data methods for researchers. SpringerPlus, 2, 222*. DOI 10.1186/2193-1801-2-222, 2013.

[2] G. L. Schlomer, S. Bauman, et N. A. Card, « Best practices for missing data management in counseling psychology. », *J. Couns. Psychol.*, vol. 57, nº 1, p. 1, 2010.

[3] J. Honaker et G. King, « What to do about missing values in time-series cross-section data », *Am. J. Polit. Sci.*, vol. 54, nº 2, p. 561–581, 2010.

[4] J.-M. Pasteels, « Review of best practice methodologies for imputing and harmonising data in cross-country datasets », *Available Natlex Ilo Chwcmsp5groupspublic—dgreports—statdocumentsgenericdocumentwcms389375 Pdf*, 2013.

[5] N. Belkhayat, A. D. A. Doukkali, et R. B. Regragui, « e-readiness: a novel approach for indicators measurements estimation and prediction », *JATIT, vol. 69, n°3*, p. 617-631, nov-2014.

[6] T. X. Bui, S. Sankaran, et I. M. Sebastian, « A framework for measuring national e-readiness », *Int. J. Electron. Bus.*, vol. 1, nº 1, p. 3–22, 2003.

[7] P. Hanafizadeh, M. R. Hanafizadeh, et M. Khodabakhshi, « Taxonomy of e-readiness assessment measures », *Int. J. Inf. Manag.*, vol. 29, nº 3, p. 189–195, 2009.

[8] L. Rabii et D. Abdelaziz, « Comparison of e-readiness composite indicators », in *Intelligent Systems Design and Applications (ISDA), 2015 15th International Conference on*, 2015, p. 93–97.

[9] R. R. Andridge et R. J. Little, « A review of hot deck imputation for survey non-response », *Int. Stat. Rev.*, vol. 78, nº 1, p. 40–64, 2010.

[10] S. Baller, S. Dutta, et B. Lanvin, *Global information technology report 2016*. Ouranos Geneva, 2016.

[11] D. C. Montgomery, E. A. Peck, et G. G. Vining, *Introduction to linear regression analysis*. John Wiley & Sons, 2015.

[12] P. S. Kumar et D. Lopez, « A review on feature selection methods for high

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

dimensional data », *Int. J. Eng. Technol.*, vol. 8, nº 2, p. 669–672, 2016.

[13] M. Shardlow, « An analysis of feature selection techniques », *Univ. Manch.*, 2016, p. 1–7.

[14] R. C. Team, « R: A language and environment for statistical computing », 2013.

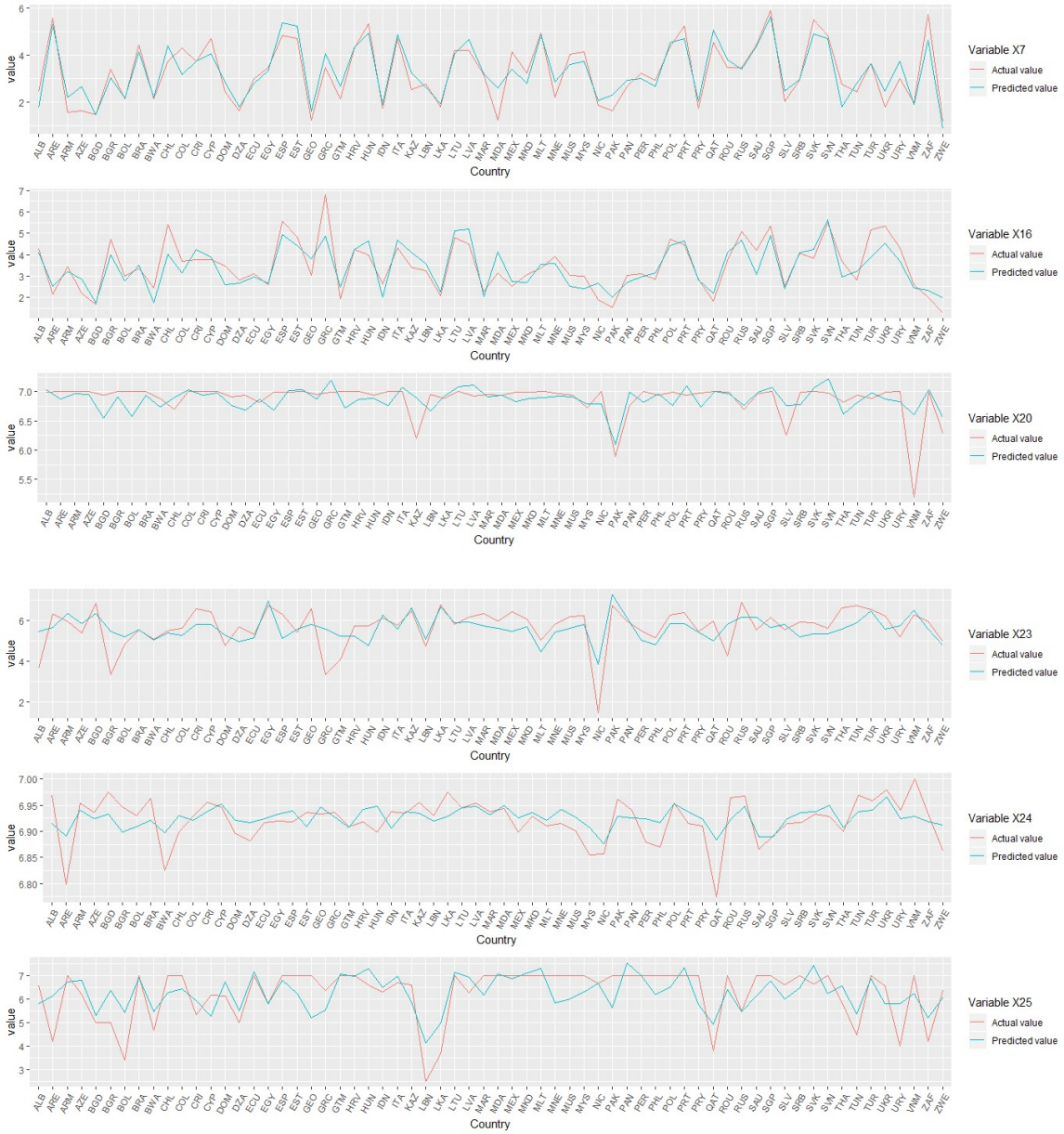
[15] World Economic Forum, INSEAD, « Network Readiness Dataset (xls) 2016 ». 2016.

Journal of Theoretical and Applied Information Technology
30th April 2019. Vol.97. No 8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

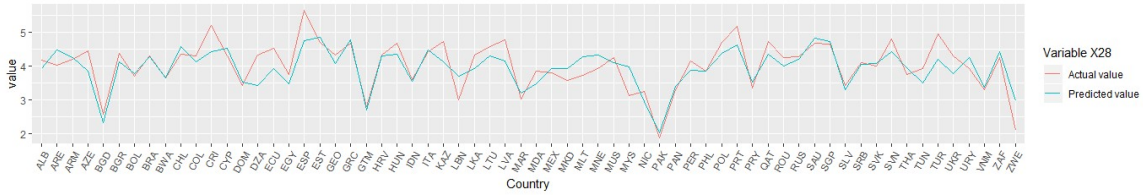
APPENDIX I

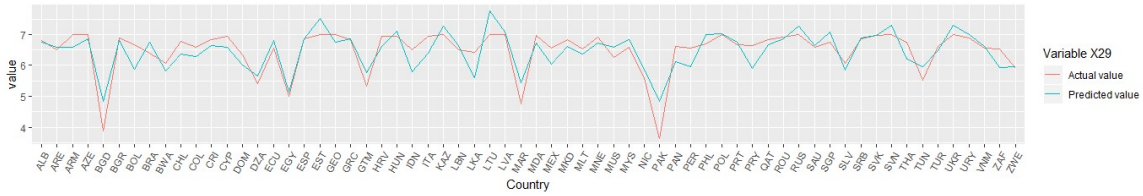
Graph comparison of actual value and predicted value for the 13 variables with missing data

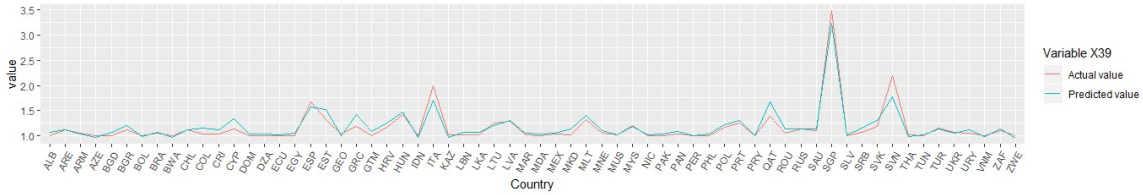


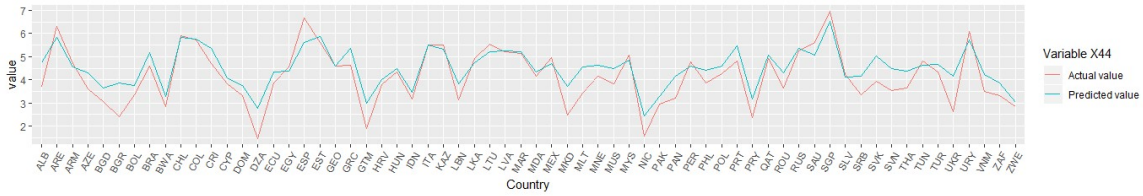
Journal of Theoretical and Applied Information Technology
30th April 2019. Vol.97. No 8

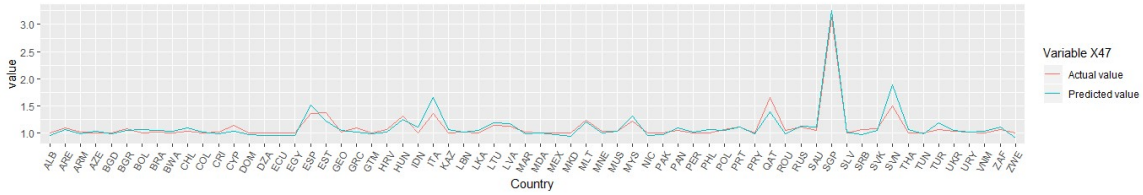
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

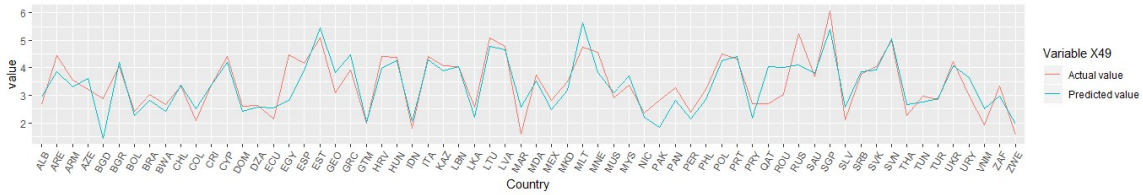


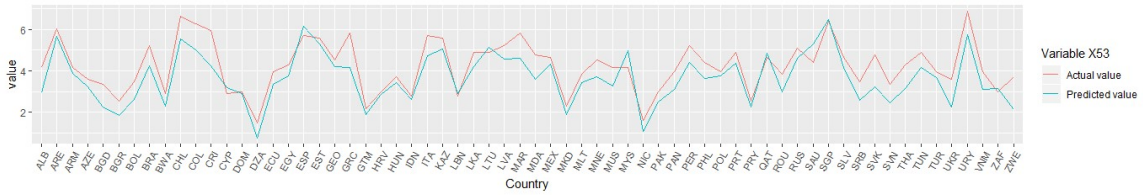












ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

APPENDIX II

Table 6: Estimation of 104 missing values

Country	X7	X16	X20	X23	X24	X25	X28	X29	X39	X44	X47	X49	X53
Argentina				0.56	22.3								
Australia								109.23					
Austria								93.85					
Bahrain												32.93	
Belgium								120.62					
Benin	72.7											11.82	
Bhutan	70.39												
Bosnia and Herzegovina												15.03	
Burundi	107.53											9.11	
Côte d'Ivoire												10.21	
Cambodia	91.7												
Cameroon												13.96	
Canada		80.16						101.69					
Cape Verde	66.99											15.6	
Chad	90.47											9.58	
China												17.87	
Czech Republic								101.28					
Denmark								104.28					
Ethiopia	85.43												
Finland								108.63					
France								103.3					
Gabon	86.98											16.13	
Gambia. The	70.78											8.56	
Germany								94.3					
Ghana	71.48												
Guinea	85.2				34.74								
Guyana	74.56											12.46	
Haiti	80.96											8.51	
Honduras												10.23	
Hong Kong SAR								101.74	111.38	0.72	25.94		0.83
Iceland								98.16					
India												9.36	
Iran. Islamic Rep.	72.54												
Ireland								103.66					
Israel								99.37					
Jamaica	63.9												
Japan								96.43					

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Jordan												21.35	
Kenya												7.57	
Korea. Rep.								91.92					
Kuwait												31.45	
Kyrgyz Republic	83.24												
Lao PDR	84.5											7.12	
Lesotho	88.67												
Liberia	82.51												
Luembourg								89.5					
Madagascar	88.28												
Malawi	88.21											5.13	
Mali	73.53											15.03	
Mauritania	88.94											12.75	
Mongolia	70.81					1.93							
Mozambique	85.27											5.12	
Myanmar	105.52											8.93	
Namibia	72.46												
Nepal	79.59												
Netherlands								100.78					
New Zealand								105.56					
Nigeria												9.47	
Norway								103.01					
Oman												25.17	
Rwanda	72.42												
Senegal												11.26	
Seychelles	67.02												
Swaziland	86.09											8.88	
Sweden								108.06					
Switzerland								89.44					
Taiwan. China									27.5	0.6	18.23		0.43
Tajikistan	84.11		93.64									10.57	
Tanzania	96.26												
Trinidad and Tobago	63.6												
Uganda	89.89												
United Kingdom								106.05					
United States								95.05					
Venezuela						1.68							
Zambia		8.18					53.11						