

Cluster Forecasting of Corruption Using Nonlinear Autoregressive with Exogenous Variables (NARX) – An Artificial Neural Network Analysis

S.A. Ghahari^{1*}, C. Queiroz², S. Labi¹, S. McNeil³

¹Lyles School of Civil Engineering, Purdue University, West Lafayette, IN 47907

*Email: sghahari@purdue.edu

²Consultant, World Bank, Washington D.C. 20433

³Department of Civil and Environmental Engineering, University of Delaware, 301 DuPont Hall, Newark, DE 19716 and Visiting Professorial Fellow, School of Civil and Environmental Engineering, University of New South Wales

ABSTRACT

Any effort to combat corruption can benefit from an examination of past and projected worldwide trends. In this paper, we forecast the level of corruption in countries by integrating an artificial neural network modeling and time series analysis. The data were obtained from 113 countries from 2007 to 2017. The study is carried out at two levels: (a) global level where all countries are considered as a monolithic group; and (b) cluster level, where countries are placed into groups based on their development-related attributes. For each cluster, we use the findings from our previous study on the cluster analysis of global corruption using machine learning methods that identified the four most influential corruption factors, and we use those as independent variables. Then, using the identified influential factors, we forecast the level of corruption in each cluster using a nonlinear autoregressive recurrent neural network with exogenous inputs (NARX), an artificial neural network technique. The NARX models were developed for each cluster, with the objective function in terms of the Corruption Perceptions Index (CPI). For each model, the optimal neural network is determined by finetuning the hyperparameters. The analysis was repeated for all countries as a single group. The accuracy of the models is assessed by comparing the mean square errors (MSE) of the time series models. The results suggest that the NARX artificial neural network technique yields reliable future values of CPI globally or for each cluster of countries. This can assist policymakers and organizations in assessing the expected efficacies of their current or future corruption control policies from a global perspective as well as for groups of countries.

Keywords: Policy, corruption, artificial neural networks (ANN), nonlinear autoregressive exogenous models (NARX).

INTRODUCTION

Transparency International (2017) defines corruption as “the abuse of public power for private benefit (or profit).” Fraudulent practice, according to the World Bank (2009) guidelines, is “any act or omission, including a misrepresentation, that knowingly or recklessly misleads, or attempts to mislead, a party to obtain a financial or other benefit or to avoid an obligation”; a collusive practice is “an arrangement between two or more parties designed to achieve an improper purpose, including influencing improperly the actions of another party.”; and a corrupt practice is defined as “the offering, giving, receiving or soliciting, directly or indirectly, of anything of value to influence improperly the actions of another party” (Integrity Vice Presidency, 2009).

Corruption, and fraudulent, collusive and corrupt practices related to development result in inefficiencies, inequities and the diversion of resources. Corruption is a multifaceted phenomenon that ranges from a minor infraction or small act of a forbidden compensation to a pervasive mass looting by public officials. Hence, it has considerable detrimental effects on sustainable development (Loosemore & Lim, 2015; Tabish & Jha, 2012). Sustainable development is defined as “Development that meets the needs of the present without compromising the ability of future generations to meet their own needs” (ASCE, 2010). In other words, sustainable development is conservation of resources and the minimization of waste and pollution (Brundtland, 1987).

The principles of sustainability include enhancing or maximizing the quality and quantity of natural resources through reduction of use, reuse, and recycling, which minimizes the damage to the physical environment (Brundtland, 1987). A corrupt society, however, fails to take the constructive steps toward a sustainable development, such as (a) avoiding adverse institutional effects; (b) maintaining or enhancing the current and future quality of life; (c) providing flexibility for changes in stakeholder requirements; (d) basing policy and business on values such as fairness, duty, knowledge-based solutions, and efficient production; and sharing responsibility for decision making, planning, and results. Corruption causes short-term economic inefficiency (specifically in the private market), and in the long-term, dynamic inefficiency and instability in economic growth and sustainability.

An accurate picture of how global corruption is evolving is need to develop effective policies and corruption-control measures, not only from a monitoring standpoint, but also from the perspective of being able to assess the long-term effectiveness of programs, policies, and initiatives targeted towards corruption mitigation. The main objective of this study is to forecast corruption levels globally and also in clusters of like countries using artificial neural network (ANN) techniques. The data are from 113 countries, and span the time period 2007 to 2017. The study considers two levels of analysis. The first is the global level (all countries considered together as a single group). Then to ensure model flexibility by avoiding making the same predictions for countries that are very dissimilar in terms of development-related attributes, cluster-level analysis was carried out using techniques established in the literature (Ghahari et al., 2021). In this previous study, we identified the four most influential factors of corruption (measured in terms of Corruption Perceptions Index (CPI)) for each cluster, and we use those as independent variables in the model in this paper. The model type used in this study is the nonlinear autoregressive recurrent neural network with exogenous inputs (NARX) technique.

In the next section we review the literature on related studies in the area of corruption. Data collection and the methodology of the research follow. The results of this research are thoroughly discussed afterward, and the final section presents conclusions and recommendations for future work.

LITERATURE REVIEW

Many factors affect the levels of corruption in countries, some exacerbate and others inhibit corruption. Multiple linear regression models or other similar methods are insufficient to model such complex systems that exhibit non-linear relationships among crucial attributes and outcomes. Therefore, a method that can handle time series in complex systems is required to model and predict corruption (Ghahari et al., 2021). Artificial neural network (ANN) techniques are applied in this study due to their potential for solving problems of this nature (Khalil et al., 2019; López-Iturriaga & Sanz, 2018; Woldemariam et al., 2016). ANNs also possess superior predictive accuracy compared to multi-linear regression, support vector machine (SVM), and multivariate adaptive regression splines (Economou, 2010; Yin et al., 2018). Of the various well-known ANN approaches and reliable training algorithms, the nonlinear autoregressive recurrent neural network with exogenous inputs (NARX) forecasting method (a feed backward approach) with the Bayesian regularization training algorithm has been proven to be efficacious in various applications and disciplines (Al-Sbou & Alawasa, 2017; Cicceri et al., 2020; Kayri, 2016; Khan et al., 2014; Yu et al., 2019b). In addition, NARX is an effective instrument to forecast time series (Chen et al., 1990), and in the non-linear time series projection it can utilize its memory capability to recollect the preceding values of the predicted time series. NARX provides more accurate results compared with other neural network techniques and time series models, such as autoregressive integrated moving average (ARIMA) and seasonal ARIMA (SARIMA) approaches (Yu et al., 2019a).

The NARX neural network method has been used in various research studies, for example, forecasting heating and cooling electrical loads (Buitrago & Asfour, 2017; Powell et al., 2014), network traffic flows (Alfred, 2015), rainfall (Benevides et al., 2019; Peña et al., 2020), and crop yield and price (Khamis & Abdullah, 2014; Paul & Sinha, 2016). Peña et al. (2020) found that NARX provides significantly more accurate results for rainfall predictions compared with nonlinear regression models and the SVM techniques, and Paul and Sinha (2016) determined that NARX outperforms ARIMA time series models in forecasting crop yield. NARX has also been applied in macroeconomic modeling. For example, recognizing the episodic and non-linear nature of the gross domestic product (GDP) of a country, researchers have espoused the use of machine learning (ML) techniques such as NARX to improve forecast accuracy of that variable. An example is Cicceri et al. (2020) who showed that the great recession in Italy in 2008-2009 could have been forecasted by NARX neural network methods (Cicceri et al., 2020). Tang (2020) assessed the feasibility of applying NARX for macroeconomic forecasting, national goal setting, and global competitiveness assessment, and carried out case studies using data from countries including China, U.S., and Russia and demonstrated the capability of NARX in forecasting macroeconomic indicators. Khan et al. (2014) performed a performance evaluation of NARX in the foreign exchange market (Khan et al., 2014). With regard to corruption forecasts, the NARX technique seems to be a promising technique. In this research study, we seek to provide some insights in this regard.

DATA

There is rather limited data that can be used in studies of this nature. The data were from the following databases (Ghahari et al., 2021): the World Bank Group (WBG) (WBG, 2017), the United Nations Department of Economic and Social Affairs (UNDESA) (UNDESA, 2017), the United Nations Development Program (UNDP) (UNDP, 2017), the World Economic Forum (WEF) (WEF, 2017), and Transparency International (TI) (TI, 2017) (see Table 1).

Table 1. Data used for the study and sources

	Source (Database)				
	WBG	UNDESA	UNDP	WEF	TI
Variables	Gross National Income per capita (GNI)	E-governance index (EGI)	Human Development Index (HDI)	Global Competitiveness Index (GCI): undue influence; public-sector performance; security; transport infrastructure; goods market efficiency; labor market efficiency; Financial market development; Technological readiness; Market size; Business sophistication	Corruption Perceptions Index (CPI)
Code	C1	C2	C3	C4-C13	C0

The Gross National Income (GNI) is the dollar value of a country's annual income and data on GNI are from the World Bank national accounts database (World Bank, 2017). UNDESA publishes the E-Governance Index (EGI) data, which indicates the consistency of being able to supervise all scales and levels of government authority, and digital interaction of governments and citizens (UNDESA, 2017). According to UNDP, people and their capabilities is the fundamental benchmark to evaluate the development of a country (HDI), and not its economic growth alone (UNDP, 2017).” In this study, the Human Development Index (HDI) is taken from the UNDP database (UNDP, 2017). The Global Competitiveness Index (GCI) shows the competitiveness landscape of economies, offering exceptional vision into the contributors to the productivity and prosperity of countries (WEF, 2018). In this study, we use the following attributes from GCI: undue influence, public sector performance, security, transport infrastructure, goods market efficiency, labor market efficiency, financial market development, technological readiness, market size, and business sophistication. Finally, the corruption Perceptions Index (CPI) from TI is a ranking indicator that indicates the perceived levels of public sector corruption (TI, 2017); the CPI is our dependent variable.

More discussion on the reasoning behind choosing these attributes can be found in our recent research study on the cluster analysis of corruption level in continents using principal component analysis and machine learning techniques (Ghahari et al., 2021). In this paper, a principal component analysis (PCA), cluster analysis, and a random forest technique are used to determine CPI values for countries. By performing PCA, we were able to deal with the potential correlations between the thirteen attributes (C1 to C13 shown in Table 1), and we were able to condense the original potentially correlated attributes into principal components – with a minimum potential loss of data information. Then, we used the top three selected principal components (PC1, PC2, and PC3) to measure the Euclidean distance between the components for each of the 113 countries to form the clusters. We verified the optimum number of clusters using the K-means machine learning technique, and we categorized the countries into four clusters. Table 2 shows the list of the countries within their corresponding clusters.

Table 2. The cluster analysis results

Cluster	Countries	No. of Countries in Each Continent	
1	Albania, Armenia, Azerbaijan, Bahrain, Barbados, Chile, China, Costa Rica, Cyprus, Czech, Estonia, France, Georgia, Hungary, Iceland, India, Indonesia, Italy, Jordan, Kazakhstan, Korea (Rep.), Latvia, Lithuania, Mauritius, Mexico, Montenegro, Morocco, Oman, Panama, Poland, Portugal, Russia, KSA, Slovakia, Slovenia, South Africa, Spain, Thailand, Turkey, Uruguay, Viet Nam	Africa	3
		Asia & Oceania	13
		Europe	19
		North America	4
		South America	2
2	Algeria, Argentina, Bangladesh, Bolivia, Bosnia-Herzegovina, Brazil, Bulgaria, Croatia, Dominican, Egypt, El Salvador, Greece, Guatemala, Honduras, Iran, Lebanon, Moldova, Nicaragua, Nigeria, Pakistan, Paraguay, Peru, Philippines, Romania, Serbia, Trinidad Tobago, Tunisia, Ukraine	Africa	4
		Asia & Oceania	5
		Europe	8
		North America	6
		South America	5
3	Australia, Austria, Belgium, Canada, Denmark, Finland, Germany, Ireland, Israel, Japan, Luxembourg, Malaysia, Netherlands, New Zealand, Norway, Singapore, Sweden, Switzerland, UK, USA	Africa	0
		Asia & Oceania	6
		Europe	11
		North America	3
		South America	0
4	Benin, Burkina Faso, Burundi, Cameroon, Côte d'Ivoire, Ethiopia, Gabon, Gambia, Ghana, Guyana, Jamaica, Kenya, Lesotho, Libya, Madagascar, Malawi, Mauritania, Mozambique, Namibia, Nepal, Rwanda, Senegal, Uganda, Zambia	Africa	21
		Asia & Oceania	1
		Europe	0
		North America	1
		South America	1

Furthermore, using a random forest algorithm, we found the marginal effects of the given variables on the outcome, and we identified and ranked the most important attributes in determining the CPI values using Gini charts. In the time series analysis that is presented in this paper, we use the top four important attributes corresponding to each cluster to predict the CPI values. These attributes are shown in Table 3 in descending order of importance/influence.

Table 3. Attributes corresponding to the world level and cluster level, 2007 to 2017

Level	Top four influential attributes
World	C11 - technological readiness, C3 - Human Development Index, C2 - E-Governance Index, C4 - undue influence
Cluster1	C11 - technological readiness, C1 - Gross National Income, C6 - security, and C4 - undue influence
Cluster 2	C3 - Human Development Index, C4 - undue influence, C2 - E-Governance Index, C5 - public sector performance
Cluster 3	C5 - public sector performance, C9, C2 - E-Governance Index, C11- technological readiness
Cluster 4	C4 - undue influence, C5 - public sector performance, C6 - security, C1 - Gross National Income

METHODOLOGY

Artificial Neural Network Techniques

An ANN model has an input layer, hidden layers, and an output layer, uses neurons to find a pattern within a dataset and expands the pattern to the other or future events. The model is established on a nonlinear relationship between the input layers and the output layers (Muyeen et al., 2014). ANN accuracy varies with the network structure. Therefore, different training/learning algorithms, and changes in the number of hidden layers, neurons, lags, iterations, hyperparameters, etc. can change the output (Beyca et al., 2019).

ANN techniques can be categorized as follows: feed-forward and feed-backward. As shown in Figure 1, each category consists of different training algorithms (Poznyak et al., 2018). Feed-forward NN training algorithms include: single-layer perceptron, multi-layer perceptron, and radial-based function network. On the other hand, recurrent or feed-backward NN algorithms include: Bayesian regularization NNs, Hopfield networks, competitive networks, art models, and Kohonen's self-organizing map. One well-known ANN approach and reliable training algorithm for nonlinear complex time series analysis is the nonlinear autoregressive with exogenous variables (NARX) NN time series forecasting method (a feed-backward approach) with the Bayesian regularization training algorithm (Al-Sbou & Alawasa, 2017; Cicceri et al., 2020; Kayri, 2016; Khan et al., 2014; Murat & Ceylan, 2006; Yu et al., 2019b).

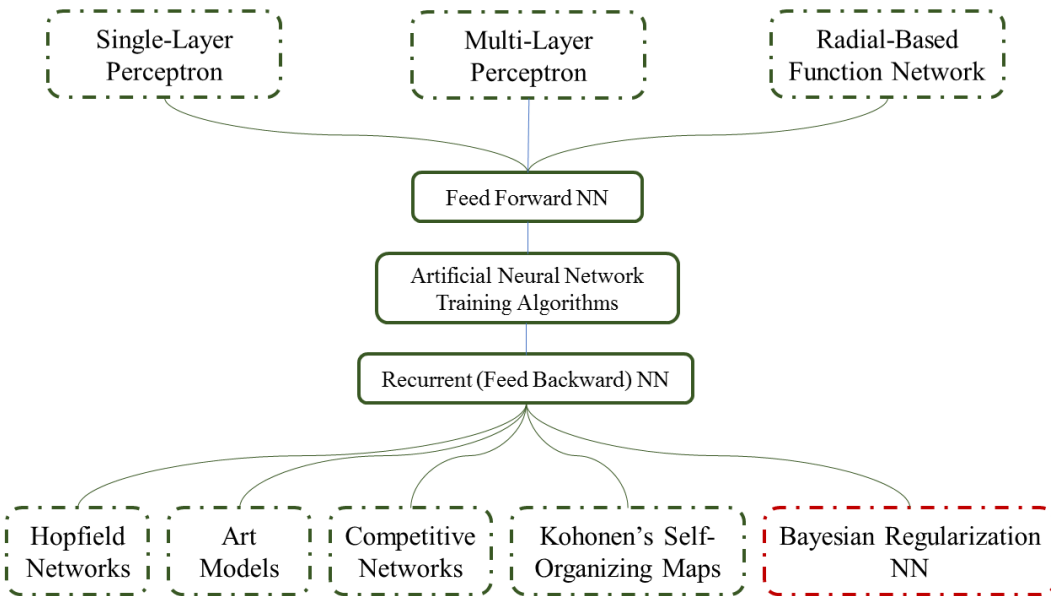


Figure 1. Training algorithm classification for artificial neural networks

Nonlinear Autoregressive Recurrent Neural Network with Exogenous Inputs (NARX) Models

The nonlinear autoregressive recurrent neural network with exogenous inputs (NARX) technique is a time series modeling technique that relates the current value of a time series to both past values of the time series and the current and past values of the exogenous inputs time series (Taqvi et al., 2020). In fact, this characteristic of NARX, which accepts dynamic variables from different time series sets, makes it superior over other feed-forward backpropagation through-time algorithm (BPTT) neural networks (Diaconescu, 2008; Jaeger, 2002). The recurrent NNs, including NARX, are cyclic in nature. Time lag connections, which transfer values between successive activations, form the cycles that include exogenous inputs and endogenous inputs (Paul & Sinha, 2016). NARX NN performs this procedure via autonomous learning (Yu et al., 2019a). The NARX technique builds complex interconnections amongst the exogenous variables and ultimately creates a function, and this renders NARX as a reliable approach for time series forecast analysis (Boussaada et al., 2018; Ruiz et al., 2016). The architecture of the NARX neural network applied to the world-level corruption data is illustrated in Figure 2. Under this architecture, the output is forecasted from the past values of CPI as well as the past and present values of the exogenous variables. The NARX technique is defined according to the following equation:

$$y(t+n) = f(y(t+1), \dots, y(t+n-1), x(t-l+1), \dots, x(t-l+n-1)) + \varepsilon \quad (1)$$

where, n is the discrete time step, $y(t+n)$ is the predicted value of CPI, $f(\cdot)$ is the neural network mapping function, $y(t+1), \dots, y(t+n-1)$ are the past predicted values for CPI, n is the number of lags, $x(t-l+1), \dots, x(t-l+n-1)$ are the past values for the exogenous variables (including l number of lags), and ε is the error term. The variable $y(t)$ (Figure 2) is defined as follows:

$$y(t) = \sum_i \alpha_i \phi \left(\sum_{j=1}^k (\beta_{ji} x_{t-j} + \gamma_{ji} y_{t-1}) \right) \quad (2)$$

where, ϕ is the hidden layer activation function, β_{ji} and γ_{ji} are the hidden layer input weights at the neuron j , α_i is the hidden layer output weights, and k is the number of input nodes.

In the NARX technique, a recurrent multi-layer perceptron (RMLP) is utilized to estimate the mapping function of $f(\cdot)$, which consists of input layers, hidden layers, and output layers. RMLP also includes neurons, activation functions, and weights. Within the hidden layer, neural network functions are operated through the interior neurons. The neurons multiply the previous layers' input vectors by the weight vectors, and they provide the scalar output. The connection weights are tuned using the Bayesian regularization algorithm. Afterward, the activation function maps each output layer to generate the neuron output to be forwarded to the next layer. In other words, to compute the output, the weighted sum of the inputs is being applied to the activation function. When the generalization improvement (in the training period) ends, and the changes in the mean square error values (MSE) become stable, the training process automatically stops. MSE is a crucial performance evaluation criterion that assists with determining the optimum initial hyperparameters for the neural network. MSE can be obtained according to Equation (3):

$$MSE = SSE/df \quad (3)$$

where, SSE is the sum of square errors, and df is the degree of freedom. Obviously, the lowest MSE value for the neural network models lead to the optimum model (Hagan et al., 1997). After the first model is fitted through the series-parallel architecture, more time steps ahead can be forecast in a closed-loop parallel architecture, where each predicted output (in the previous step) is fed into the model to predict a future output.

In NARX, the number of hidden layers, lags, and neurons, as the main hyperparameters, influence the accuracy of the results. Therefore, we investigate several different numbers of hidden layers, lags, and neurons to find the optimum model. The variation of the number of hidden layers, lags, and the number of neurons selected in this study were one to seven for the number of hidden layers, one to three for the number of lags, and one to 20 for the number of neurons, respectively. Two other hyperparameters are important: the number of epochs and the learning rate. One epoch is one cycle that the full training dataset is performed and is initially set to 100 and then varied from 100 to 1000. The learning rate is the step size at each iteration while moving toward a minimum of a loss function and is initially set to 0.1 and then varied from 0.0001 to 0.1.

We evaluate the precision of the models by comparing the mean square error values (MSE), accordingly. Our exogenous variables (inputs), as shown in Table 4, include GNI, E-governance index (EGI), human development index (HDI), undue influence, public-sector performance, security, labor market efficiency, and technological readiness. CPI is used as the dependent variable (output). Data from 2007 to 2017 for 113 countries are assembled for all variables with 70% of the data used for training the

model, 15% for validation, and 15% of the data used to test the model. In the next section we present the results of the NARX analysis outcome.

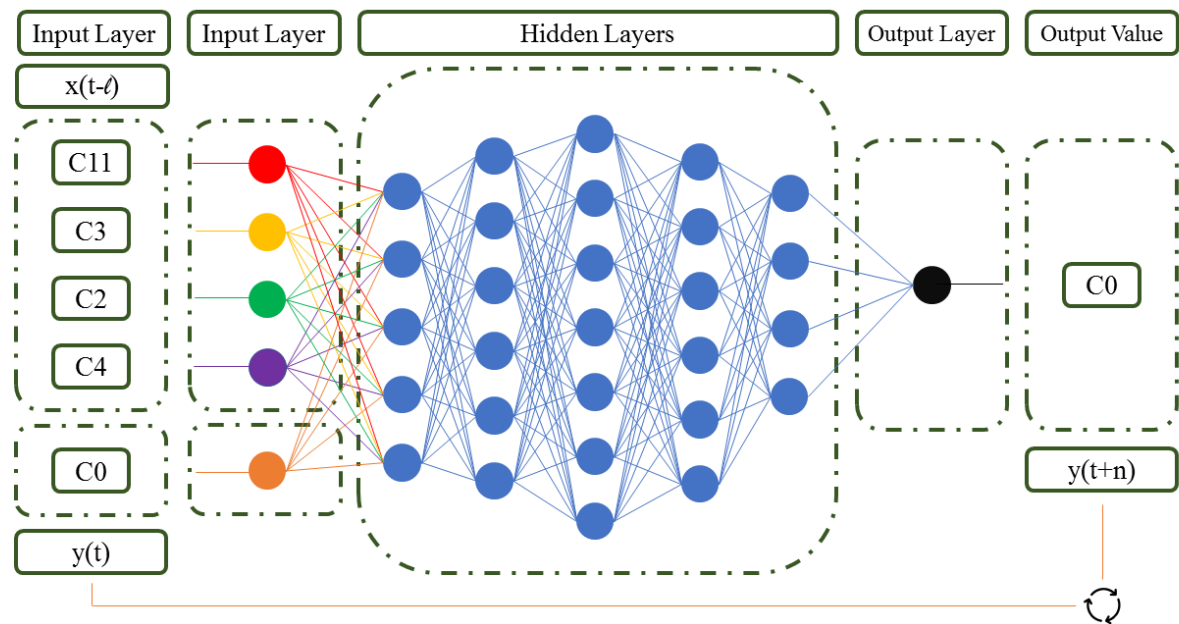


Figure 2. The architecture of NARX neural network applied to the World-level data

RESULTS AND DISCUSSION

Hyperparameters play a critical role in the accuracy of the NARX analysis or any neural network analysis. (Al-Sbou & Alawasa, 2017). The hyperparameters, which need to be tuned to give models with higher accuracies, in the NARX analysis are the number of hidden layers, lags, neurons, and epochs and the learning rate. In many cases, a higher number of hidden layers causes overfitting in the model, and lower prediction accuracy (Kim et al., 2019; Liu & Kim, 2018). In this study, we investigate different numbers of hidden layers and lags to initiate the neural network analysis, and we choose the least error associated with a hidden layer and a lag. Table 4 presents the errors associated with each hyperparameter (number of lags and number of hidden layers) for the world-level data. The data shows that four hidden layers with one lag gives the least possible error among the other number of hidden layers and lags. The training MSE is calculated 0.261, the error for the validation phase is 0.180, and the testing error is 0.243. The 4 hidden layer training MSE for two lags and three lags are 26.82% and 15.71% higher than that of the four hidden layer MSE for one lag, respectively. When it comes to the testing MSE, two lags and three lags show 93.49% and 181.61% higher MSE compared to one lag, respectively. While the training MSE for four layers and one lag is 1.91% higher that the training MSE for six layers and one lag, the validation and testing MSE are 4.76% and 64.68% lower, respectively. This also shows that when the number of hidden layers increases, lower prediction accuracy is obtained.

Table 4. NARX errors associated with the number of hidden layers and the number of lags (world-level category) (neuron = 1, epochs = 100, and learning rate = 0.1)

Lag	No. of Hidden Layers	Training MSE	Validation MSE	Testing MSE
1	1	0.278	0.175	1.073
2		0.494	0.385	0.253
3		0.315	1.484	0.715
1	2	0.351	0.228	0.561
2		0.375	0.669	0.309
3		0.619	1.164	0.675
1	3	0.267	0.188	0.280
2		0.341	0.485	0.540
3		0.527	0.793	0.466
1	4	0.261	0.180	0.243
2		0.331	0.505	0.553
3		0.302	0.735	0.487
1	5	0.379	0.287	0.413
2		0.271	0.374	0.499
3		1.128	0.652	0.425
1	6	0.256	0.189	0.688
2		1.388	0.372	0.375
3		0.488	0.502	0.814
1	7	0.368	0.360	1.507
2		0.327	0.851	0.966
3		0.278	0.175	1.073

To fine-tune another crucial NARX NN hyperparameter, we focus on the number of neurons at each hidden layer. Table 5 presents the errors associated with the number of hidden layers (H) and the number of neurons (N) for the world-level data. The data indicates that four hidden layers with five neurons give the least possible error among the other number of hidden layers and neurons. The training MSE for H4|N5 is calculated 0.236 and the testing error is 0.209. H3, H5, and H6 with five neurons show 2.48%, 4.45%, and 0.84% higher training MSE compared with H4|N5, respectively. Likewise, testing MSE for H4|N5 is 15.72%, 36.08%, and 53.76% lower than that of the H3, H5, and H6 testing MSE with five neurons. When the number of neurons exceeds 10 neurons, the errors significantly increase. The H4|N5 training and testing MSE values are 43.13% and 49.64% lower than that of the training and testing MSE for H4 with 10 neurons. This also confirms the importance of fine-tuning hyperparameters for the NARX NN.

Table 5. Hyperparameter fine tuning for the world-level data – NARX errors associated with the number of hidden layers (H3-H6) and number of neurons (N1-N10, N15, and N 20) (lag =1, epochs = 100, and learning rate = 0.1)

	Training MSE				Testing MSE			
	H3	H4	H5	H6	H3	H4	H5	H6
N1	0.267	0.261	0.279	0.256	0.280	0.243	0.413	0.688
N2	0.257	0.253	0.271	0.251	0.265	0.235	0.399	0.612
N3	0.250	0.244	0.265	0.246	0.264	0.228	0.379	0.545
N4	0.242	0.239	0.251	0.238	0.254	0.224	0.352	0.462
N5	0.242	0.236	0.247	0.238	0.248	0.209	0.327	0.452
N6	0.252	0.249	0.254	0.240	0.264	0.233	0.340	0.480
N7	0.271	0.267	0.286	0.278	0.297	0.249	0.371	0.509
N8	0.295	0.282	0.304	0.315	0.329	0.259	0.400	0.691
N9	0.349	0.334	0.361	0.403	0.374	0.347	0.478	0.922
N10	0.431	0.415	0.435	0.615	0.460	0.401	0.563	1.085
N15	1.034	1.008	1.175	1.456	1.199	0.972	1.140	2.266
N20	1.471	1.456	1.640	2.649	1.656	1.370	1.578	3.646

The final hyperparameter tuning is related to epochs and learning rates. The epochs range from 100 to 1000 and learning rates (LRs) from 0.0001 to 0.1 are investigated to determine the optimum MSE. Table 6 presents the results for the training and testing MSE values associated with different ranges of epochs and LR for the NARX NN with four hidden layers, five neurons, and one lag. According to the results, the differences between the epochs and LR are insignificant, demonstrating that the number of hidden layers, neurons, and lags have been selected properly. We keep the epochs and the learning rates at 100 epochs and 0.1 learning rate, our starting point, to keep the calculation costs as lowest as possible.

Table 6. Hyperparameter fine tuning for the world-level data – NARX errors associated with the epochs and learning rates (LR) (lag =1, hidden layers = 4, and neurons =5)

Epoch	Learning Rate							
	Training MSE				Testing MSE			
	0.0001	0.001	0.01	0.1	0.0001	0.001	0.01	0.1
100	0.240	0.235	0.238	0.236	0.213	0.206	0.211	0.209
200	0.241	0.239	0.240	0.237	0.215	0.207	0.215	0.211
400	0.240	0.234	0.237	0.236	0.213	0.205	0.210	0.209
600	0.250	0.235	0.236	0.236	0.213	0.207	0.211	0.209
800	0.241	0.238	0.238	0.238	0.214	0.209	0.215	0.212
1000	0.241	0.239	0.240	0.240	0.215	0.210	0.216	0.215

The data at each level (world-level data and cluster level data) is distinct, therefore, different errors are likely for each cluster. This means that the number of lags, hidden layers, and neurons can vary for the NARX analysis for each cluster. The analysis of training, validation and testing MSE for ranges of the various hyperparameters for each cluster was conducted following the process described for the world-level data. The hyperparameter values resulting in optimal performance and MSEs for each the world-level and each cluster are presented in Table 7. According to the results, cluster 2 and cluster 3 have a lower MSE value compared with the MSE values for Cluster 1 and Cluster 4, which could be due to fact that the CPI variance among the countries in Cluster 2 and Cluster is significantly less (as shown in Figure 3). Likewise, Cluster 1 and Cluster 4 exhibits higher errors of 0.254 and 0.259 at the testing phase. Considering the error values, four hidden layers are selected for Cluster 1 and Cluster 4, whereas three hidden layers is found to be optimum for Cluster 2 and Cluster 3.

Table 7. NARX model hyperparameters and performance values for the world-level and the cluster-level analysis

Category	Lag	Nr. of hidden layers	Nr. Of neurons	Training MSE	Validation MSE	Testing MSE
World level	1	4	5	0.236	0.161	0.209
Cluster 1		4	6	0.324	0.267	0.254
Cluster 2		3	6	0.280	0.189	0.210
Cluster 3		3	5	0.208	0.140	0.150
Cluster 4		4	6	0.350	0.294	0.259

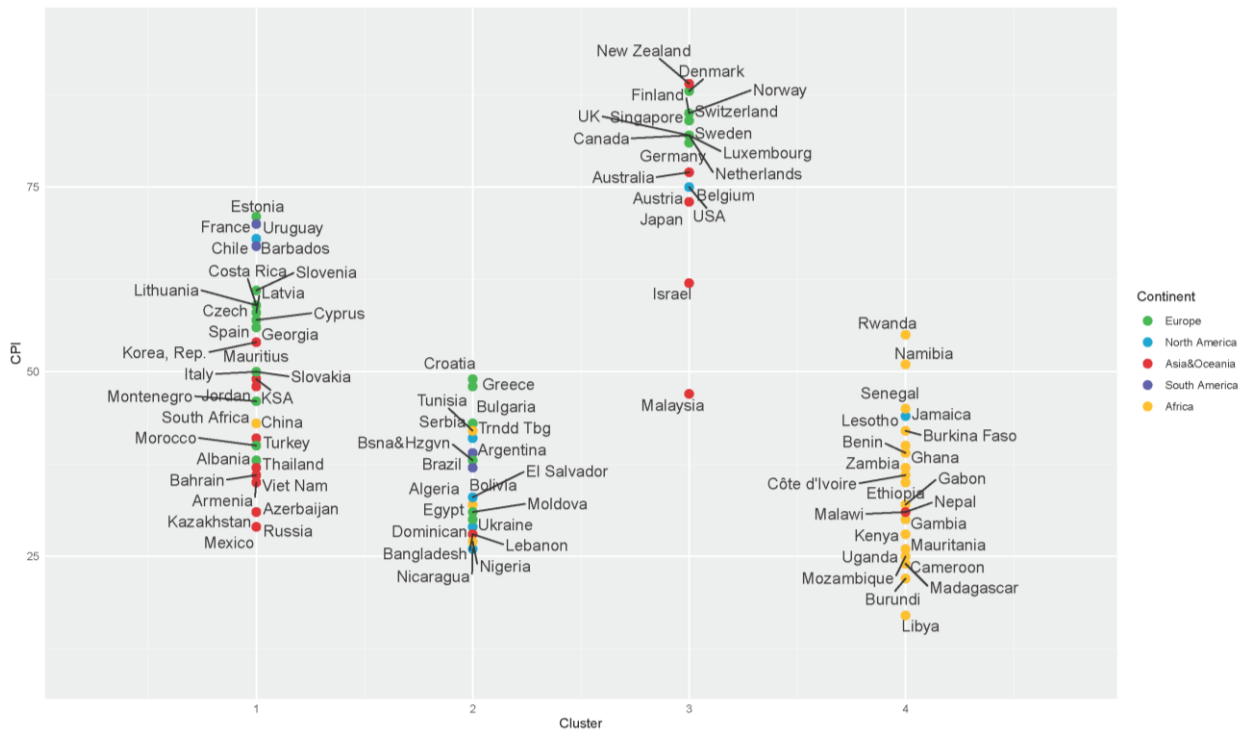


Figure 3. Position of countries in each cluster considering CPI values (Ghahari et al., 2021)

Finally, the results of the NARX analysis using the world-level data and the clusters are discussed in this section of this manuscript. The NARX ANN time series response for the world-level data is presented in Figure 4. Figure 4.a indicates the training, target, and predicted outputs and the corresponding errors (target output – training output) with a 97.5% confidence band. This figure also presents the predicted values for the 2017–2020 period. Figure 4.b presents the optimum ultimate epoch that is selected for obtaining the optimum results regarding the world-level data. The best training performance is identified as occurring at 0.10020 MSE and at the epoch 298 with no observable overfitting. This means that after the initial training of the first neural network model, it retrained the network for 298 epochs until it reached a near-zero change in MSE. According to the results, the highest difference between the training target and training outputs is calculated for 2012 with a value of -0.999 , and the second-highest error is achieved for 2011 with a value of -0.619 due to the significant change in the average CPI values from 2011 to 2013. The results show that the predicted CPI values for 2018, 2019, and 2020 (shown in black triangles connected with a dashed line) are comparatively close to the real values reported by Transparency International for those specific years (TI, 2020). The CPI actual and forecast values are presented in Table 8 showing generally insignificant error between the two; the

differences between the CPI forecasts and actual values in 2017, 2018, 2019, and 2020 are calculated at 0.25, 0.04, −0.07, and −0.08, respectively. Figure 4a. also indicates that the overall CPI value of the world is increasing. Although a 0.18% decrease in CPI value is seen from 2007 to 2010, the general trend is positive with a 6.71% increase in the CPI value from 2010 to 2020.

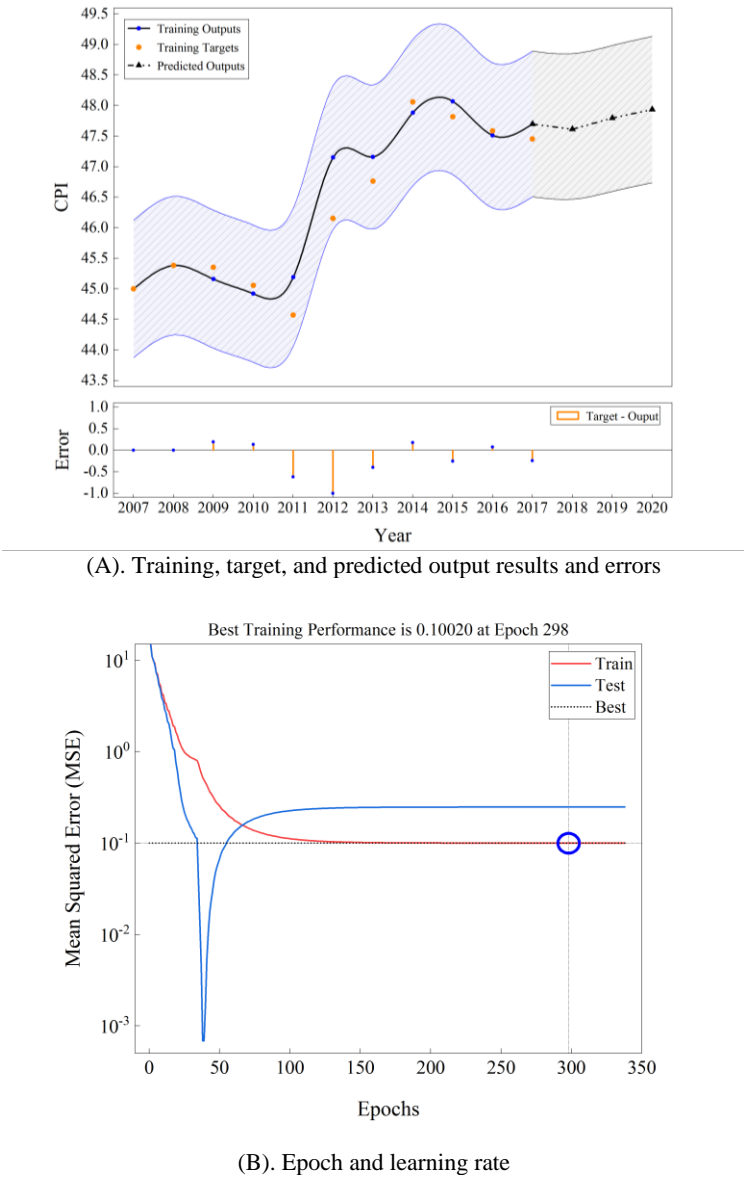


Figure 4. NARX ANN time series response for the world-level data

Figure 5 presents the NARX ANN time series response for Cluster 1. Figure 5.a illustrates training, target, and predicted outputs and the corresponding errors (target output – training output) with a 97.5% confidence band. Also, this figure indicates the predicted values for the 2017–2020 period. Figure 5.b illustrates the optimum epoch chosen for calculating the optimum results for Cluster 1; the best training performance is set at 0.2135 MSE and epoch 151 with no observable overfitting. This denotes the fact that after the initial training of the first neural network model, it retrained the network for 151 epochs until it reached a near-zero change in MSE. Based on the results, the maximum difference between the training target and training outputs is in 2013 with a value of 0.835, and the second-highest error is in

2011 with a value of -0.690. This could be due to the considerable change in the average CPI values from 2011 to 2013 for this cluster. The predicted CPI value results for 2017, 2018, 2019, and 2020 are close to the real values reported by Transparency International for those specific years (TI, 2020). Table 8 presents the actual and forecast CPI values. The results show a 0.01, 0.26, 0.21, and 0.39 differences between the forecast and actual CPI values in 2017, 2018, 2019, and 2020, respectively. Furthermore, Figure 5.a shows that the overall CPI value for Cluster 1 is increasing. Despite a 1.67% decrease in CPI value from 2008 to 2010, the general trend is positive with a 7.41% increase in the CPI value from 2010 to 2020.

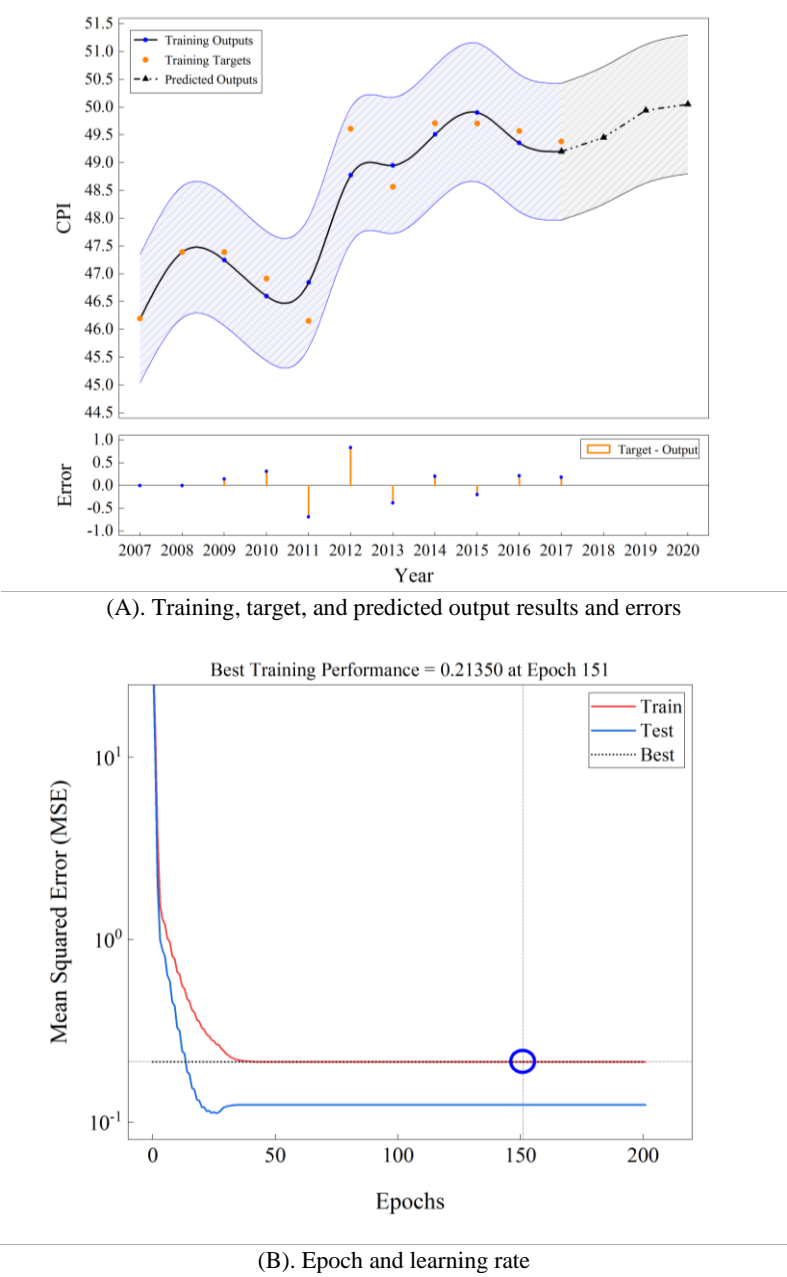
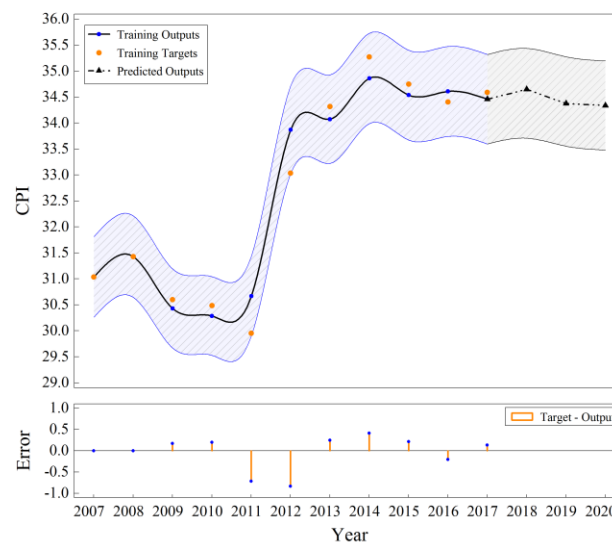


Figure 5. NARX ANN time series response for Cluster 1

The NARX ANN time series analysis results for Cluster 2 is presented in Figure 6. Figure 6.a shows training, target, and predicted outputs and the corresponding errors (target output – training output) with a 97.5% confidence band. Furthermore, this figure shows the predicted values for the 2017–2020 period. Figure 6.b denotes the optimum ultimate epoch selected in this analysis in order to obtain the optimum results. The best training performance is caught at epoch 248 and 0.20484 MSE with no observable overfitting. This shows that after the initial training of the first neural network model, it retrained the network for 248 epochs until it reached a near-zero change in MSE. Results show that the highest training target and training outputs difference is in 2012 and 2011 with values of -0.833 and -0.714 , respectively, which could be due to the significant change in the average CPI values from 2011 to 2013. The results indicate that the predicted CPI values for 2017 to 2020 are comparatively close to the real values reported by Transparency International (TI, 2020). The values presented in Table 8 indicate a minor error between the real CPI values and the predicted CPI values. Differences between the predicted and real CPI values for 2017 to 2020 are calculated as 0.23, -0.01 , -0.35 , and -0.21 , respectively. Moreover, Figure 6a. shows an overall increase in the CPI values in this cluster. Although a 3.64% decrease in the CPI value is seen from 2008 to 2010, the general trend is upward with a 13.37% increase in the CPI value from 2010 to 2020.



(A). Training, target, and predicted output results and errors

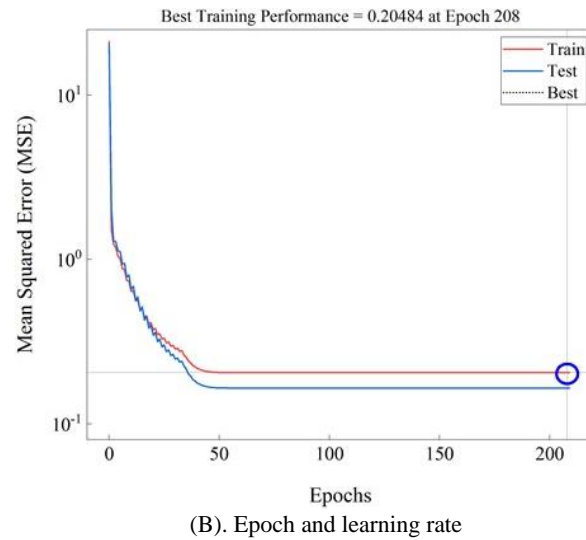
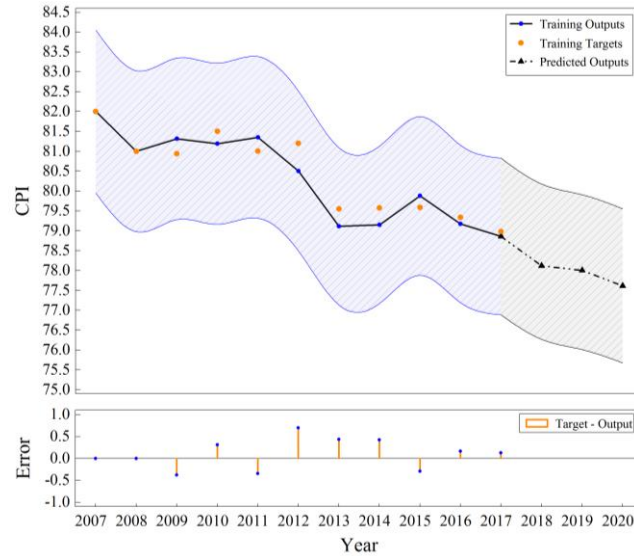
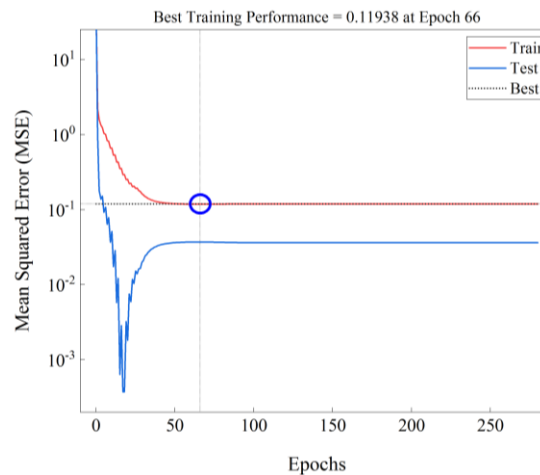


Figure 6. NARX ANN time series response for Cluster 2

The NARX ANN time series response for the third cluster is illustrated in Figure 7. Figure 7.a shows training, target, and predicted outputs and the corresponding errors (target output – training output) with a 97.5% confidence band. This figure also illustrates the predicted values for 2017, 2018, 2019, and 2020. Figure 7.b shows the optimum epoch chosen for calculating the optimum results for Cluster 2; the best training performance is obtained at 0.11938 MSE and epoch 66 with no observable overfitting. This indicates that after the initial training of the first neural network model, it retrained the network for 66 epochs until it reached a near zero change in MSE. Based on the results, the difference between the training target and training outputs is at its maximum value of 0.696 in 2012, and the second-highest error at 0.436 in 2013, which could be due to the significant change in the average CPI values in 2012 and 2013 for this cluster. The predicted CPI value results for 2017 to 2020 are comparatively close to the real values reported by the Transparency International for those specific years (TI, 2020). The actual and forecast CPI values are illustrated in Table 8. The results show differences of -0.63 , -0.32 , -0.35 , and -0.16 between the forecast and actual CPI values in 2017 to 2020, respectively. Figure 7.a also illustrates that the overall CPI value for Cluster 3 is decreasing. The general CPI trend in this cluster is negative, with a 5.35% decrease in the CPI value from 2007 to 2020.



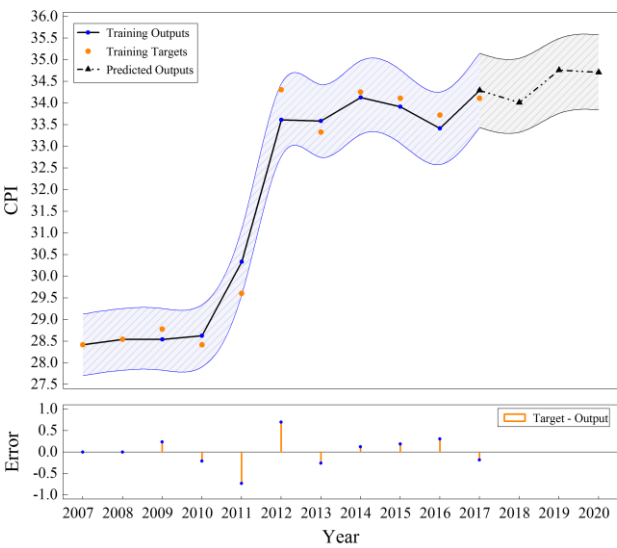
(A). Training, target, and predicted output results and errors



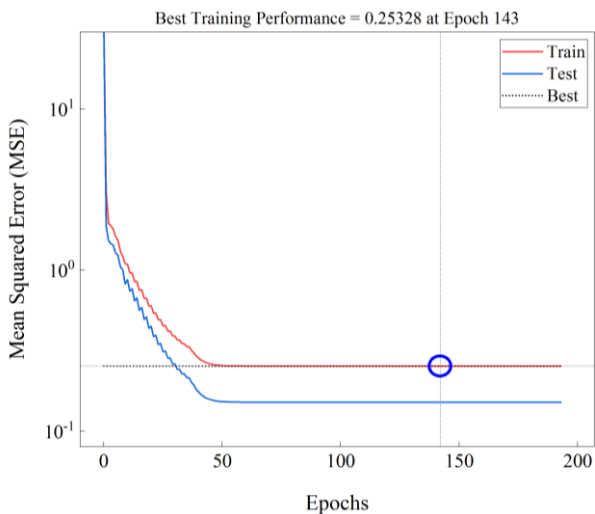
(B). Epoch and learning rate

Figure 7. NARX ANN time series response for cluster 3

The NARX ANN time series analysis results for Cluster 4 is presented in Figure 8. Figure 8.a shows training, target, and predicted outputs and the corresponding errors (target output – training output) with a 97.5% confidence band. This figure also illustrates the predicted values for 2017 – 2020. Figure 8.b shows the optimum ultimate epoch selected in this analysis to obtain the optimum results; the best training performance is caught at epoch 143 and 0.25328 MSE with no observable overfitting. This means that after the initial training of the first neural network model, it retrained the network for 143 epochs until it reached a near-zero change in MSE. Results indicate that the highest training target and training outputs difference are in 2011 and 2012 with values of -0.733 and 0.696 , respectively, which could be due to the significant change in the average CPI values from 2010 to 2012 in this cluster. The results denote that the predicted CPI values for 2017 - 2020 are comparatively close to the real values reported by Transparency International (TI, 2020). The values are presented in Table 8 indicating a minor difference between the real CPI values and the predicted CPI values. The differences between the predicted and real CPI values for 2017 - 2020 are calculated at 0.18 , 0.17 , -0.17 , and 0.08 , respectively. Furthermore, Figure 8.a indicates an overall increase in the CPI values. The general trend in this cluster is upward with a 21.25% increase in the CPI value from 2010 to 2020.



(A). Training, target, and predicted output results and errors



(B). Epoch and learning rate

Figure 8. NARX ANN time series response for Cluster 4

Table 8. CPI actual and forecast values

Year	Category	Actual CPI	CPI Forecast	Error (Forecast - Actual)
2017	World level	47.70	47.45	0.25
	Cluster 1	49.39	49.38	0.01
	Cluster 2	34.82	34.59	0.23
	Cluster 3	78.35	78.98	-0.63
	Cluster 4	34.29	34.11	0.18
2018	World level	47.65	47.61	0.04
	Cluster 1	49.71	49.45	0.26
	Cluster 2	34.64	34.65	-0.01
	Cluster 3	77.80	78.12	-0.32
	Cluster 4	34.17	34.00	0.17
2019	World level	47.73	47.80	-0.07
	Cluster 1	50.15	49.94	0.21
	Cluster 2	34.11	34.37	-0.26
	Cluster 3	77.65	78.00	-0.35
	Cluster 4	34.58	34.75	-0.17
2020	World level	47.86	47.94	-0.08
	Cluster 1	50.44	50.05	0.39
	Cluster 2	34.14	34.35	-0.21
	Cluster 3	77.45	77.61	-0.16
	Cluster 4	34.79	34.71	0.08

CONCLUDING REMARKS

Artificial neural networks (ANNs) are effective tools for non-linear mapping of multiple variables on one or more outputs. In this study, we use a well-known neural network method, the nonlinear autoregressive recurrent neural network with exogenous inputs (NARX), to model and forecast corruption in countries. The analysis was carried out using the data on 113 countries from 2007 to 2017. The development-related attributes that have significant influence on the levels of corruption in countries, as measured by the CPI, were identified from the literature. We split the countries into four clusters based on their development-related attributes, and developed corruption forecasting models in each cluster. NARX NN training was performed on 70% of the data, 15% of the data was used for validation, and the rest of the data was used for testing the output.

Any reliable neural network model needs precise hyperparameter fine-tuning before training. The variation of the number of hidden layers, lags, and the number of neurons were selected as: 1–7, 1–3, and 1–20, respectively. Considering MSE as a baseline for the hyperparameter tuning process showed that one lag, four hidden layers, and five neurons would give an optimum NARX model for forecasting CPI values for the world-level data. For Cluster 1 and Cluster 4 the number of hidden layers were found to be four, versus three for Cluster 2 and Cluster 3. At the same time, the number of neurons for Cluster 1, Cluster 2, and Cluster 4 were chosen to be six, versus five for Cluster 3. Epochs and learning rates were found to have no significant influence on the initial hyperparameter MSE values for the NARX models. It was observed that when the number of neurons and hidden layers increased, a comparatively lower prediction accuracy was obtained, due to the models' overfitting.

As expected, the NARX NN prediction models showed different results for the world-level data analysis and the cluster-level data analysis. For the world-level data, it was found that there is a general uptrend momentum in the value of CPI showing a 6.71% increase in CPI from 2010 to the predicted value of CPI in 2020. Cluster 1, Cluster 2, and Cluster 4 showed the same uptrend with 7.41%, 13.37%, and 21.25% increase in CPI from 2010, despite having a comparatively minor downtrend in CPI from 2007 to 2010. However, Cluster 3 – despite containing developed countries mostly – showed a 5.35% decrease in

CPI from 2007. For countries within the clusters introduced in this paper, the study results can be valuable to policymakers, governments, and NGOs as they continue to assess the efficacy of their current or prospective future corruption-mitigation policies, programs, and initiatives.

In this paper, the lack of adequate data on development-related attributes was one of the main limitations. In future studies, access to other data will be helpful to develop more confident conclusions. Another limitation was the reliance on only one attribute (CPI) as the indicator of corruption. Suggested directions for future research include (a) a detailed investigation of the causes of the uptrend and downtrend momentum in CPI values in each cluster; (b) adequate and explicit assessment of corruption-mitigation initiatives implemented in countries in each cluster, and identifying solutions that have worked well as those that failed, and an overall assessment of the extent to which these solutions succeeded or failed. Further, future studies could investigate not only project-level data (instead of country-level in this study). In this regard, researchers could examine the effect of corruption on infrastructure delivery quality, time delay, and cost overruns, and thereby measure, for example, the portion of overrun cost that could be attributed to corruption and the portion that could be attributed to inefficiency.

FUNDING

Publication of this article was funded in part by *Purdue University Libraries Open Access Publishing Fund*.

AUTHOR CONTRIBUTIONS

Conceptualization, SeyedAli Ghahari; Data curation, SeyedAli Ghahari and Cesar Quieroz; Formal analysis, SeyedAli Ghahari, Cesar Quieroz and Sue McNeil; Funding acquisition, SeyedAli Ghahari; Investigation, SeyedAli Ghahari and Samuel Labi; Methodology, SeyedAli Ghahari, Cesar Quieroz and Sue McNeil; Resources, SeyedAli Ghahari, Cesar Quieroz and Samuel Labi; Software, SeyedAli Ghahari; Supervision, SeyedAli Ghahari, Cesar Quieroz, Samuel Labi and Sue McNeil; Validation, SeyedAli Ghahari; Visualization, SeyedAli Ghahari; Writing – original draft, SeyedAli Ghahari; Writing – review & editing, SeyedAli Ghahari, Cesar Quieroz, Samuel Labi and Sue McNeil.

DATA AVAILABILITY STATEMENT

The datasets used in this study is available upon request.

REFERENCES

- Al-Sbou, Y.A., & Alawasa, K.M. (2017). Nonlinear Autoregressive Recurrent Neural Network Model for Solar Radiation Prediction. *International Journal of Applied Engineering Research*, 12(14), 4518-4527.
- Alfred, R. (2015). *Performance of Modeling Time Series Using Nonlinear Autoregressive with Exogenous Input (Narx) in the Network Traffic Forecasting*. Paper presented at the 2015 International Conference on Science in Information Technology (ICSITech), Yogyakarta, Indonesia.
- ASCE. (2010). Policy Statement 418 – the Role of the Civil Engineer in Sustainable Development. . *American Society of Civil Engineers, Virginia*.
- Benevides, P., Catalao, J., & Nico, G. (2019). Neural Network Approach to Forecast Hourly Intense Rainfall Using Gns Precipitable Water Vapor and Meteorological Sensors. *Remote Sensing*, 11(8), 966.
- Beyca, O.F., Ervural, B.C., Tatoglu, E., Ozuyar, P.G., & Zaim, S. (2019). Using Machine Learning Tools for Forecasting Natural Gas Consumption in the Province of Istanbul. *Energy Economics*, 80, 937-949.

- Boussaada, Z., Curea, O., Remaci, A., Camblong, H., & Mrabet Bellaaj, N. (2018). A Nonlinear Autoregressive Exogenous (Narx) Neural Network Model for the Prediction of the Daily Direct Solar Radiation. *Energies*, 11(3), 620.
- Brundtland, G.H. (1987). *Report of the World Commission on Environment and Development: "Our Common Future."*: United Nations.
- Buitrago, J., & Asfour, S. (2017). Short-Term Forecasting of Electric Loads Using Nonlinear Autoregressive Artificial Neural Networks with Exogenous Vector Inputs. *Energies*, 10(1), 40.
- Chen, S., Billings, S., & Grant, P. (1990). Non-Linear System Identification Using Neural Networks. *International Journal of Control*, 51(6), 1191-1214.
- Cicceri, G., Inserra, G., & Limosani, M. (2020). A Machine Learning Approach to Forecast Economic Recessions—an Italian Case Study. *Mathematics*, 8(2), 241.
- Diaconescu, E. (2008). The Use of Narx Neural Networks to Predict Chaotic Time Series. *Wseas Transactions on computer research*, 3(3), 182-191.
- Ekonomou, L. (2010). Greek Long-Term Energy Consumption Prediction Using Artificial Neural Networks. *Energy*, 35(2), 512-517.
- Ghahari, S., Assi, L., Carter, K., & Ghotbi, S. (2019). *The Future of Hydrogen Fueling Systems for Fully Automated Vehicles*. Paper presented at the ASCE International Conference on Transportation and Development 2019: Innovation and Sustainability in Smart Mobility and Smart Cities, Alexandria, VA.
- Ghahari, S.A., Queiroz, C., Labi, S., & McNeil, S. (2021). Impact of E-Governance on National Corruption Indexes: New Evidence Using Panel Vector Auto Regression Analysis. *Preprint*.
- Hagan, M.T., Demuth, H.B., & Beale, M. (1997). *Neural Network Design*: PWS Publishing Co.
- Integrity Vice Presidency. (2009). *Fraud and Corruption Awareness Handbook: How It Works and What to Look For*. Retrieved from
- Jaeger, H. (2002). *Tutorial on Training Recurrent Neural Networks, Covering Bppt, Rtrl, Ekf and the "Echo State Network" Approach* (Vol. 5): GMD-Forschungszentrum Informationstechnik Bonn.
- Kayri, M. (2016). Predictive Abilities of Bayesian Regularization and Levenberg–Marquardt Algorithms in Artificial Neural Networks: A Comparative Empirical Study on Social Data. *Mathematical and Computational Applications*, 21(2), 20.
- Khalil, A.J., Barhoom, A.M., Abu-Nasser, B.S., Musleh, M.M., & Abu-Naser, S.S. (2019). Energy Efficiency Prediction Using Artificial Neural Network.
- Khamis, A., & Abdullah, S. (2014). Forecasting Wheat Price Using Backpropagation and Narx Neural Network. *The International Journal of Engineering and Science*, 3(11), 19-26.
- Khan, Z., Pathak, D.K., Pandey, A., & Kumar, S. (2014). *Performance Evaluation of Nonlinear Auto-Regressive with Exogenous Input (Narx) in the Foreign Exchange Market*. Paper presented at the Proceedings of 10th IRF International Conference.
- Kim, J.-H., Seong, N.-C., & Choi, W. (2019). Cooling Load Forecasting Via Predictive Optimization of a Nonlinear Autoregressive Exogenous (Narx) Neural Network Model. *Sustainability*, 11(23), 6535.
- Labi, S. (2014). *Introduction to Civil Engineering Systems: A Systems Perspective to the Development of Civil Engineering Facilities*: Wiley & Sons, Hoboken, NJ.
- Liu, H., & Kim, H. (2018). Ecological Footprint, Foreign Direct Investment, and Gross Domestic Production: Evidence of Belt & Road Initiative Countries. *Sustainability*, 10(10), 3527.
- Loosemore, M., & Lim, B. (2015). Inter-Organizational Unfairness in the Construction Industry. *Construction Management and Economics*, 33(4), 310-326.
- López-Iturriaga, F.J., & Sanz, I.P. (2018). Predicting Public Corruption with Neural Networks: An Analysis of Spanish Provinces. *Social Indicators Research*, 140(3), 975-998.
- Murat, Y.S., & Ceylan, H. (2006). Use of Artificial Neural Networks for Transport Energy Demand Modeling. *Energy Policy*, 34(17), 3165-3172.

- Muyeen, S., Hasanien, H.M., & Al-Durra, A. (2014). Transient Stability Enhancement of Wind Farms Connected to a Multi-Machine Power System by Using an Adaptive Ann-Controlled Smes. *Energy Conversion and Management*, 78, 412-420.
- Paul, R.K., & Sinha, K. (2016). Forecasting Crop Yield: Arimax and Narx Model. *RASHI*, 1(1), 77-85.
- Peña, M., Vázquez-Patiño, A., Zhiña, D., Montenegro, M., & Avilés, A. (2020). Improved Rainfall Prediction through Nonlinear Autoregressive Network with Exogenous Variables: A Case Study in Andes High Mountain Region. *Advances in Meteorology*, 2020.
- Powell, K.M., Sriprasad, A., Cole, W.J., & Edgar, T.F. (2014). Heating, Cooling, and Electrical Load Forecasting for a Large-Scale District Energy System. *Energy*, 74, 877-885.
- Poznyak, T., Oria, J.I.C., & Poznyak, A. (2018). *Ozonation and Biodegradation in Environmental Engineering: Dynamic Neural Network Approach*: Elsevier.
- Ruiz, L.G.B., Cuéllar, M.P., Calvo-Flores, M.D., & Jiménez, M.D.C.P. (2016). An Application of Non-Linear Autoregressive Neural Networks to Predict Energy Consumption in Public Buildings. *Energies*, 9(9), 684.
- Tabish, S., & Jha, K.N. (2012). The Impact of Anti-Corruption Strategies on Corruption Free Performance in Public Construction Projects. *Construction Management and Economics*, 30(1), 21-35.
- Tang, L. (2020). Application of Nonlinear Autoregressive with Exogenous Input (Narx) Neural Network in Macroeconomic Forecasting, National Goal Setting and Global Competitiveness Assessment. *National Goal Setting and Global Competitiveness Assessment (May 15, 2020)*.
- Taqvi, S.A., Tufa, L.D., Zabiri, H., Maulud, A.S., & Uddin, F. (2020). Fault Detection in Distillation Column Using Narx Neural Network. *Neural Computing and Applications*, 32(8), 3503-3519.
- TI. (2017). Corruption Perception Index. Retrieved from https://www.transparency.org/news/feature/corruption_perceptions_index_2017
- TI. (2020). Corruption Perception Index. Retrieved from <https://www.transparency.org/en/cpi/2020/index/nzl>
- UNDESA. (2017). E-Government Development Index. Retrieved from <https://publicadministration.un.org/egovkb/en-us/Reports/UN-E-Government-Survey-2018>
- UNDP. (2017). Human Development Reports. Retrieved from <http://hdr.undp.org/en/content/human-development-index-hdi>
- WBG. (2017). Worldwide Governance Indicators. Retrieved from <https://datacatalog.worldbank.org/dataset/worldwide-governance-indicators>
- WEF. (2017). *The Global Competitiveness Report*. Retrieved from <http://www3.weforum.org/docs/GCR2017-2018/05FullReport/TheGlobalCompetitivenessReport2017%E2%80%932018.pdf>
- WEF. (2018). *The Global Competitiveness Report 2018*. Paper presented at the World Economic Forum.
- Woldemariam, W., Murillo-Hoyos, J., & Labi, S. (2016). Estimating Annual Maintenance Expenditures for Infrastructure: Artificial Neural Network Approach. *Journal of Infrastructure Systems*, 22(2), 04015025.
- World Bank. (2017). Gni Per Capita. Retrieved from <https://data.worldbank.org/indicator/NY.GNP.PCAP.CD>
- Yin, Z., Jia, B., Wu, S., Dai, J., & Tang, D. (2018). Comprehensive Forecast of Urban Water-Energy Demand Based on a Neural Network Model. *Water*, 10(4), 385.
- Yu, X., Chen, Z., & Qi, L. (2019a). Comparative Study of Sarima and Narx Models in Predicting the Incidence of Schistosomiasis in China. *Math. Biosci. Eng*, 16(4), 2266-2276.
- Yu, X., Chen, Z., & Qi, L. (2019b). Comparative Study of Sarima and Narx Models in Predicting the Incidence of Schistosomiasis in China. *Mathematical biosciences and engineering: MBE*, 16(4), 2266-2276.