Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 4
ICCIA 2024

Using Deep Learning to Predict Global Population Dynamics and Construction Risks

Kai Jin, Xingxin Yang, Yonghao Zhou, Yongzheng Wu
Guilin university of technology at nanning, Guangxi, Chongzuo, 541006, China

Abstract. In this study, a long short-term memory network (LSTM) model is used to predict the population size and risk level of each country in 2030 by combining global population data from 1950 to 2021 and country risk rating indices from 1954 to 2023.The core of the LSTM model is its three gating units: forgetting gates, input gates, and output gates, which enable the model to effectively deal with the long-term dependence problems and avoid the problems of gradient vanishing and gradient explosion in traditional recurrent neural networks. Through correlation analysis of the input data and prediction errors, as well as an autocorrelation study of the errors, the model shows prediction accuracy at certain time delays, although the prediction errors of the model are not observed to be completely random on certain delay terms, implying that the model may need further optimisation in these areas. In addition, this study applies the entropy method to calculate the construction suitability evaluation index for each country in 2030, which provides a scientific basis for the development of future construction strategies. By analysing the mean square error (MSE) during the training process, the study shows that the performance of the model on the training and validation sets gradually improves as the number of training rounds increases, although the MSE on the test set increases at some points, showing signs of overfitting. Finally, countries were scored and ranked for construction suitability based on predicted population and risk class scores. Based on the composite suitability score index, we categorised the countries into three building suitability classes, each corresponding to a different building strategy. In areas of low suitability, robust materials and designs that are resistant to natural hazards are recommended; in areas of moderate suitability, sustainable building design and green building materials are recommended; and in areas of high suitability, investment in high-quality and high-end design building projects should be emphasised to meet the needs of higher-income groups. These strategies will contribute to the needs of different regions and populations and improve the accuracy and efficiency of real estate decisions. Through the application of this deep learning model, we have not only improved our ability to predict global population dynamics and construction risks, but also provided important data support for urban planning and development strategies on a global scale.

Keywords: Long Short-Term Memory Networks (LSTM); Global Population Projections; Country Risk Ratings; Time Series Analysis; Error Autocorrelation.

1. Introduction

As globalisation accelerates and urbanisation continues, accurate prediction of population dynamics and construction risks is crucial for the sustainable development of countries and regions. In order to address the challenges faced by traditional prediction methods when dealing with complex time series data, such as gradient vanishing or gradient explosion, this study adopts a state-of-the-art deep learning model, the Long Short-Term Memory Network (LSTM).The LSTM, through its unique internal structure including forgetting gates, input gates, and output gates, is able to effectively capture long-term dependencies, thereby improving the accuracy of predictions of changes in populations and their risk levels [1].

In this study, the LSTM model was used to predict the population size and risk level in 2030 based on the global population data of each country from 1950 to 2021 and the composite risk assessment index from 1954 to 2023. By further applying the entropy method, we assessed the construction suitability index of each country in 2030, which provides a scientific basis for future construction strategies. In addition, the analysis of the model prediction results reveals the correlation between the input variables and the prediction errors and the autocorrelation of the errors, pointing out the room that the model may need to be further optimised in some aspects.

Through this study, we have not only improved our understanding of global population dynamics, but also provided strong data support for reducing construction risks, optimising resource allocation and developing effective urban planning strategies. This study demonstrates the potential of deep learning techniques for application in global development strategies, and provides new perspectives and methodological foundations for future research and practice in related fields [2].

2. Related work

In recent years, with the rapid development of big data and machine learning technologies, it has become possible to use deep learning models to predict global population dynamics and construction risks. In this study, a deep learning framework, the Long Short-Term Memory (LSTM) network, is used to predict the long-term trend of the population data of each country from 1950 to 2021, and to predict the population and the risk level in 2030 by combining the comprehensive risk evaluation index from 1954 to 2023.

The LSTM network, as a special recurrent neural network (RNN) architecture, effectively solves the problems of gradient vanishing and gradient explosion faced by traditional RNNs by introducing three key control units, namely, forgetting gate, input gate and output gate. These gating units enable the model to selectively retain or forget information to better capture long-term dependencies in time series data [3]. In this study, the accuracy of the LSTM model predictions was verified by several statistical graphs, including correlation plots between inputs and errors and autocorrelation plots of errors. These graphs show that despite the positive correlation of the model on certain time delay terms, the overall prediction error shows random and no significant pattern, indicating that the model has some room for optimisation.

In addition, this study uses the entropy value method to assess the construction suitability evaluation index in 2030, so as to develop a more scientific construction strategy. Through the steps of constructing the decision matrix, standardising the indicators, calculating the characteristic weights, entropy values, and coefficients of variation, the construction suitability scores of each country were finally formed, and the ultimate priority ranking was carried out accordingly [4].

Based on the prediction results, countries were classified into three different construction suitability classes, with countries in each class showing different construction strategy needs. For example, in regions with low suitability, robust and durable building materials and designs that can withstand natural disasters are recommended; in regions with medium suitability, it is recommended that sustainable building design and green building materials be prioritised to reduce environmental impacts and increase energy efficiency; and in regions with high suitability, investment in high-quality and high-end design construction projects can be prioritised to cater to the needs of high-income groups.

With this approach, not only can future demographic changes and risk levels be predicted, but data support and decision-making can be provided for construction projects globally, thus improving the accuracy and efficiency of real estate decisions. This work demonstrates the potential of deep learning in the application of global development strategy planning, which is an important reference value for future urban planning and development.

3. Modelling

Population data were collected for each country for the years 1950-2021, and the data were subjected to missing value tests, and outliers were treated. (Not specifically developed here for space reasons). Combined with the comprehensive evaluation index of risk level of each country from 1954 to 2023 obtained from Problem 1, the LSTM prediction model was used to predict the population and the comprehensive evaluation index of risk level of each country in 2030. The entropy value method is then used to find out the evaluation index of suitable construction in 2030 to formulate the construction strategy [5].

LSTM (Long Short-Term Memory) is a deep learning neural network architecture particularly suited for processing and predicting time-series data, natural language processing tasks, and other problems that require the capture of long-term dependencies. LSTM networks are a variant of Recurrent Neural Networks (RNNs), which are Designed to solve problems such as gradient vanishing and gradient explosion in traditional RNNs. LSTM introduces a special internal structure consisting of three key gating units that help the model to remember or forget previous information and selectively pass information to the next time step. These three gating units are the forgetting gate, the input gate, and the output gate. The forgetting gate determines how much of the previously memorized information is retained at the current time step [6]. By learning the weights, it can decide to forget or retain the previous unit states. Its formula is as follows:

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{1}$$

Where f_t is the output of the forgetting gate, σ is the sigmoid function, W_f is the weight matrix, h_{t-1} is the previous hidden state, x_t is the current input, and b_f is the bias. The input gate determines how much new information to memorize at the current time step. It determines which information to add to the unit state by learning the weights. Its formula is as follows:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2}$$

$$\tilde{C}_t = \tan h(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{3}$$

Where i_t is the output of the input gate and $\tilde{C}_t$ is the new vector of candidate values. The output gate decides how much output to generate based on the current cell state and input information. This gating unit filters out unnecessary information. Its formula is as follows:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{4}$$

$$h_t = o_t * \tan h(C_t) \tag{5}$$

Where o_t is the output of the output gate and h_t is the hidden state of the current time step. the internal structure of the LSTM allows the model to better capture long-term dependencies when dealing with time-series data without being affected by gradient vanishing.

Table 1. EWM

Step 1	Constructing a decision matrix $X=\left(x_{ij}\right)$
Step 2	Standardization of indicators: homogenization of heterogeneous indicators 1. For positive indicators $x_{ij}^{'} = \left[\frac{x_{ij} - \min(x_{1j}, x_{2j}, \cdots, x_{nj})}{\max(x_{1j}, x_{2j}, \cdots, x_{nj}) - \min(x_{1j}, x_{2j}, \cdots, x_{nj})}\right] \times 100\%$ 2. For negative indicators $x_{ij}^{'} = \left[\frac{\max(x_{1j}, x_{2j}, \cdots, x_{nj}) - x_{ij}}{\max(x_{1j}, x_{2j}, \cdots, x_{nj}) - \min(x_{1j}, x_{2j}, \cdots, x_{nj})}\right] \times 100\%$
Step 3	Calculate the proportion of characteristics of i evaluation objects under indicator j
Step 4	Calculate the entropy value of indicator j $e_j = -\frac{1}{\ln n}\sum_{i=1}^{n} p_{ij} \ln p_{ij}, (j = 1, \cdots, n)$
Step 5	Calculation of the coefficient of variation in indicator j $g_j = 1 - e_j, (j = 1, \cdots, n)$
Step 6	Calculation of the weighting factor for indicator j $w_j = 1 - E_j / \sum_{k=1}^{n}(1 - E_k), (j = 1, \cdots, n)$

The population and composite evaluation scores were predicted using the LSTM network model, and some of the parameters and results are shown below:

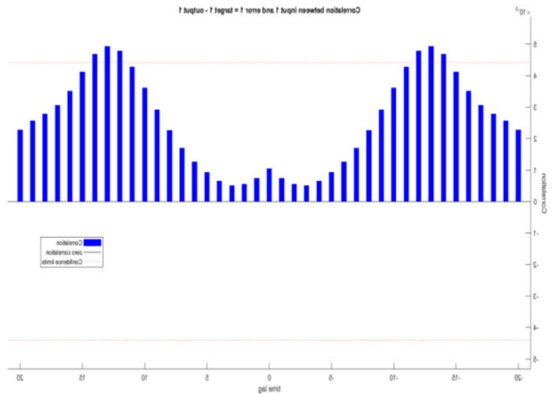

Figure 1. Correlation between input and error

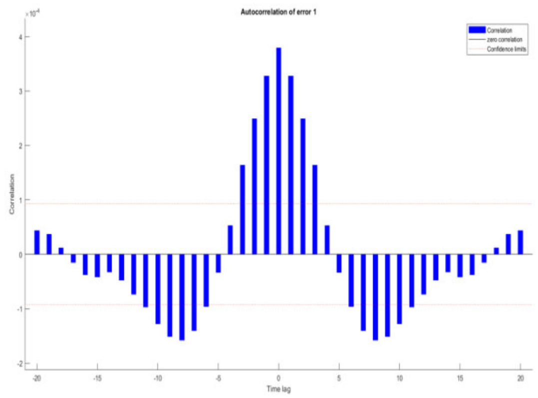

Figure 2. Autocorrelation of errors

The first graph shows the correlation between input 1 and the error (target 1 - output 1), where the error is the difference between the model prediction and the actual value. The time lag (time lag) ranges from -20 to +20 and can be thought of as a lag term in time series analysis. Positive numbers indicate that future inputs are correlated with current errors, and negative numbers indicate that past inputs are correlated with current errors. Most of the correlation values in the graph are positive, which implies that there is some positive correlation between the inputs and the errors on some of the lagged terms. This may indicate that there may be room for further optimization of the model, and the confidence limits shown by the red dashed lines indicate the statistical significance thresholds for the correlation statistics, above which the correlation is considered statistically significant. The second graph shows the autocorrelation of Error 1 [7]. Autocorrelation measures the correlation between a time series and itself at different time lags. For a good forecasting model, we want no autocorrelation between errors, i.e., random and patternless. In this graph, we see significant autocorrelation in some of the lagged terms, especially when the time lag is zero, which is where the autocorrelation is strongest [8]. This means that the model's prediction errors are not completely random and there may be some kind of pattern or trend, which is something the model may need to improve.

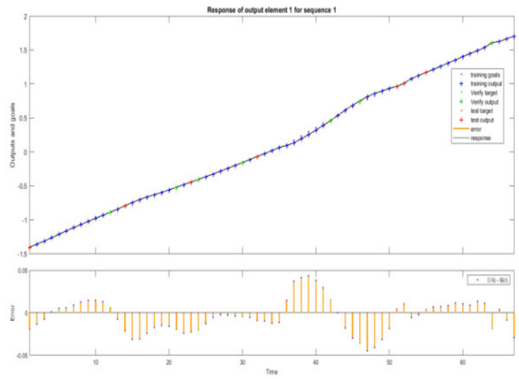

Figure 3. Response of the output elements

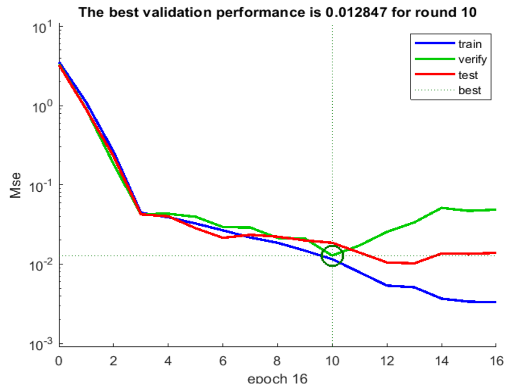

Figure 4. MSE of the training process

The training output (blue line), validation output (green line) and test output (red line) in the response plot of the output element closely follow the target value (black dots), while the model's response (orange line) shows the approximation to the target value. The bottom half shows the error bars at different points in time, and it can be seen that most of the errors are clustered around zero, indicating that the model's predictions match the actual data quite well.

This graph shows how the mean square error (MSE) of the model varies with the number of training rounds (epoch) on the training, validation, and test sets. The mean square error is the average of the squares of the prediction errors and is a common metric for assessing prediction accuracy [9]. In this case, we can see that as the number of training rounds increases, the MSE decreases for both the training and validation sets, indicating that the model is learning and improving its performance on the training set. The MSE of the test set increases slightly at some points, which may indicate that the model is overfitting at those points [10]. The circled portion may point out where the best validation performance occurs, i.e., at round 10 when the MSE reaches its lowest point of 0.012847.

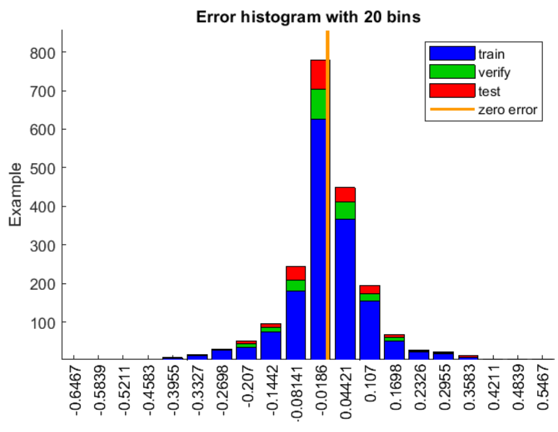

Figure 5. Histogram of errors

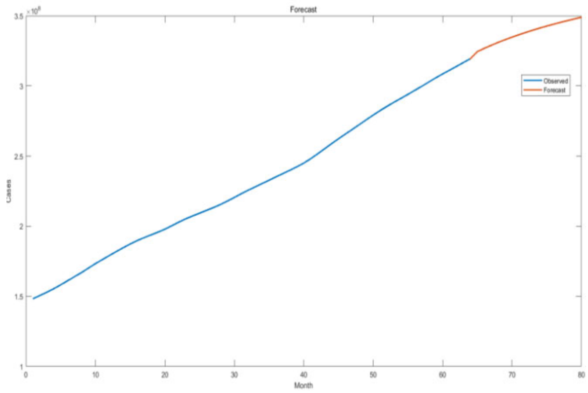

Figure 6. Population projections

This figure is an error histogram with 20 bins to show the error distribution of the training, validation and test data. The distribution of errors should be around the zero error line, and ideally it should be close to a normal distribution, which indicates that there is no systematic bias in the model's predictions. In this graph, most of the errors are indeed centered around zero, but there seems to be some bias. The distributions of training error (blue), validation error (green), and testing error (red) look relatively consistent, which suggests consistency in the model's performance across different datasets. This shows that the model works well, and the table below shows the predictions of the population and risk level scores for each country in 2030.

Table 2. Projections of population and risk level scores for countries in 2030

Country	Populations	Risk level score
United States	19757183024	0.008655003
Antigua and Barbuda	5701685.43	0.003862827
Azerbaijan	612098552.6	0.000461548
……	……	……
Georgia	347000174.2	0.002974218
Uzbekistan	1745597975	1.61466E-06
Serbia	616700856.6	0.001227105

The countries were then given a suitability score for population and risk level to obtain a final prioritization. Finally, we obtained a composite suitability index and ranking as shown in the figure below:

The center of cluster 1: 0.20, the center of cluster 2: 0.44, and the center of cluster 3: 0.72. represents the location of the center of each cluster. Based on the value of the composite score index, these three clusters represent different levels of suitability for construction. In these areas of Cluster 1, we labeled them as having a medium degree of suitability for construction, and recommended prioritizing building designs and materials that are resistant and resilient to mitigate the potential impacts of natural disasters. For areas in Cluster 2, we have labeled them as medium suitable for construction. In these areas, we recommend prioritizing sustainable building designs and green building materials to reduce environmental impacts and improve energy efficiency. It is also key to invest in infrastructure upgrades to ensure that areas have good transportation, water, power, and communication facilities to improve residential and business accessibility.

Finally, for areas in Cluster 3, we labeled them as having a high degree of suitability for construction. In these areas, there could be a focus on investment in construction with an emphasis on building quality and design, luxury and high-end real estate products to meet the needs of high-income groups, and high-end community services and facilities. These strategies will help to better meet the needs of different regions and populations, and improve the accuracy and effectiveness of real estate decisions.

4. Conclusions

By collecting and processing the population data of countries around the world from 1950 to 2021, and combining with the comprehensive evaluation index of the country risk level from 1954 to 2023, we predicted the population and risk level of each country in 2030 by using the Long Short-Term Memory (LSTM) model, which, by virtue of its unique internal structure, effectively deals with the problem of gradient disappearance and explosion faced by the traditional recurrent neural network in the long-term data processing, thus improving the accuracy and stability. The LSTM model, with its unique internal structure, effectively deals with the problem of gradient vanishing and explosion faced by traditional recurrent neural networks in long-term dependent data processing, thus improving the accuracy and stability of the prediction.

In addition, we further analyse the construction suitability evaluation index of each country in 2030 by using the entropy method, which provides a scientific basis for the development of future construction strategies. The prediction results show that the model is able to capture the correlation between the input data and the prediction error better, while revealing the autocorrelation of the error, indicating that the model still has room for improvement in some aspects.

By analysing the mean square error (MSE) of the model during training, validation and testing, we observe that the model's performance on the training and validation sets is gradually optimised as the number of training rounds increases, but the MSE on the testing set slightly increases at some points, suggesting that there may be overfitting. Nevertheless, the overall performance of the model shows good consistency and high prediction accuracy.

In summary, the LSTM model shows strong potential for the prediction of global population dynamics and construction risk. By accurately predicting the population and risk level of each country, as well as evaluating the suitability of future construction, the model not only provides important decision support for policy makers, but also provides a scientific basis for future urban planning and development. The findings will help to guide construction activities on a global scale, optimise resource allocation and enhance resilience to future challenges.

5. Discussion

Globally, accurate projections of national population dynamics and construction risks are essential for the development of effective policies and strategies. In this study, a Long Short-Term Memory (LSTM) network model is used to successfully predict the population of countries and their risk ratings in 2030 by combining population data from 1950 to 2021 and a composite index of risk ratings from 1954 to 2023.The LSTM model efficiently solves the gradient vanishing and exploding in the traditional recurrent neural network through its built-in forgetting gate, input gate and output gate problems, thus enabling the capture of long-term dependencies and improving the accuracy and reliability of predictions.

Through further data analysis, the entropy method was used to calculate the appropriate construction evaluation index in 2030, which provides a scientific basis for the development of specific construction strategies. The prediction results of the model show that there is a certain positive correlation between the input data and the prediction error, while the autocorrelation analysis of the error reveals that there is still a certain systematic bias in the prediction of the model, which points out the potential direction of model optimisation.

In addition, the analysis of the mean square error (MSE) during model training shows that the model's performance on the training and validation sets gradually improves as the number of training rounds increases, but there is a slight increase in the MSE on the test set, suggesting that overfitting may have occurred. Despite these challenges, the model as a whole showed consistency across datasets, proving the reliability of its predictive ability.

Ultimately, based on the population and risk class data predicted by the model, we scored and ranked countries for suitability to build, which in turn resulted in a suitability to build grading for each country. The results of these ratings provide guidance for construction activities in different regions of the world, recommending the prioritisation of disaster-resistant building designs and materials in regions with high risk of natural disasters, and the promotion of sustainable building designs and green building materials in regions with high environmental impacts. In addition, areas with a high concentration of high-income groups can focus on investing in construction projects with high quality and high-end design.

The application of this deep learning model not only enhances the accuracy of predicting future population and construction risks, but also provides important data support for urban planning and development strategies globally, helping to optimise resource allocation and improve the effectiveness of decision-making.

References

[1] Lin, S. S., Shen, S. L., Zhou, A., & Xu, Y. S. (2021). Risk assessment and management of excavation system based on fuzzy set theory and machine learning methods. Automation in Construction, 122, 103490.

[2] Yakovlev, S., Bazilevych, K., Chumachenko, D., Chumachenko, T., Hulianytskyi, L., Meniailov, I., & Tkachenko, A. (2020). The concept of developing a decision support system for the epidemic morbidity control.

[3] Pourhomayoun, M., & Shakibi, M. (2021). Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision-making. Smart health, 20, 100178.

[4] Zhang, L., Wang, Y., Niu, M., Wang, C., & Wang, Z. (2020). Machine learning for characterizing risk of type 2 diabetes mellitus in a rural Chinese population: The Henan Rural Cohort Study. Scientific reports, 10(1), 4406.

[5] Berrang-Ford, L., Sietsma, A. J., Callaghan, M., Minx, J. C., Scheelbeek, P. F., Haddaway, N. R., ... & Dangour, A. D. (2021). Systematic mapping of global research on climate and health: a machine learning review. The Lancet Planetary Health, 5(8), e514-e525.

[6] Quiroz-Juárez, M. A., Torres-Gómez, A., Hoyo-Ulloa, I., León-Montiel, R. D. J., & U'Ren, A. B. (2021). Identification of high-risk COVID-19 patients using machine learning. Plos one, 16(9), e0257234.

[7] Zhao, H., Zhang, X., Xu, Y., Gao, L., Ma, Z., Sun, Y., & Wang, W. (2021). Predicting the risk of hypertension based on several easy-to-collect risk factors: a machine learning method. Frontiers in Public Health, 9, 619429.

[8] Pang, L., Wang, K., Tao, Y., Zhi, Q., Zhang, J., & Lin, H. (2021). A new model for caries risk prediction in teenagers using a machine learning algorithm based on environmental and genetic factors. Frontiers in Genetics, 12, 636867.

[9] Yang, J., Ju, J., Guo, L., Ji, B., Shi, S., Yang, Z., ... & Yuan, P. (2022). Prediction of HER2-positive breast cancer recurrence and metastasis risk from histopathological images and clinical information via multimodal deep learning. Computational and structural biotechnology journal, 20, 333-342.

[10] Pham, B. T., Luu, C., Van Dao, D., Van Phong, T., Nguyen, H. D., Van Le, H., ... & Prakash, I. (2021). Flood risk assessment using deep learning integrated with multi-criteria decision analysis. Knowledge-based systems, 219, 106899.