Research on trade quantity prediction on ARIMA and Long Short-Term Memory Networks

Zihao Guo*, Yuhan Yang, Xinxiong Wu, Zhiyong Zheng, Weiqi Liu

College of Computer and Data Science/College of Software, Fuzhou University, Fujian, China, 350108

*Corresponding author: Gzhao924@163.com

Abstract. In the face of the enormous threat that illegal wildlife trade poses to global ecology and biodiversity. In this paper, two prediction models, ARIMA and LSTM, are used for comparative experiments. By analyzing the data of global illegal wildlife trade from 1995 to 2021, two prediction models, ARIMA and LSTM, are established to obtain the prediction results of global illegal wildlife trade from 2022 to 2029 respectively. The experimental results show that both the ARIMA prediction model and the LSTM prediction model show an upward trend in the prediction of the amount of illegal wildlife trade. Especially noteworthy is that the prediction results of the LSTM model show a more obvious upward trend and larger prediction values compared with the ARIMA model. The projections tell us the grim story of the illegal wildlife trade. If we do not take urgent and effective interventions, the scale of the illegal wildlife trade will continue to show a worrying trend and could rise further. This not only poses a serious threat to global biodiversity, but also has potential negative impacts on ecological balance, economic and social stability. These projections can effectively guide efforts to reduce the amount of illegal wildlife trade.

Keywords: Big Data, ARIMA, LSTM.

1. Introduction

In recent years, LSTM has become the focus of deep learning is increasingly used in various fields. (Karim et al., 2019) [1] conducted a series of ablation tests (3627 experiments) on LSTM-FCN and ALSTM-FCN to better understand the model and each of its sub-modules. Finally, (Karim et. al., 2019) demonstrated the performance of LSTM-FCN when the LSTM block is replaced by GRU, basic RNN and Dense blocks. (Yu et. al., 2019) [2] reviewed LSTM units and their variants and explored the learning ability of LSTM units. LSTM networks are categorized into two main types: the LSTM dominant network and the integrated LSTM network. In order to improve the prediction accuracy (Wei et al., 2019) [3] proposed a novel traffic flow prediction method called autoencoder long short-term memory (AE-LSTM) prediction method. The prediction error and fluctuation of the AE-LSTM method are small for different stations and different dates. A hybrid CNN-LSTM model was established by combining a convolutional neural network (CNN) with a long short-term memory (LSTM) neural network for predicting the PM2.5 concentration in Beijing in the next 24 h. The advantages of the two were fully utilized, i.e., the CNN can efficiently extract the features related to air quality, and the LSTM can reflect the long-term historical process of the input time series data (Li et al. 2020)[4]. Four models were developed for PM2.5 concentration prediction, namely univariate LSTM model, multivariate LSTM model, univariate CNN-LSTM model and multivariate CNN-LSTM model. (Park et al., 2020) [5] proposed a novel RUL prediction technique based on long short-term memory (LSTM). Experimental results show that the mean absolute percentage error (MAPE) of the proposed single-channel LSTM model is improved by 39.2% compared to the baseline LSTM model. (Cui et al., 2020) [6] focused on RNN-based models and attempted to reformulate methods for incorporating RNNs and their variants into traffic prediction models. They proposed a bi-directional and uni-directional superimposed LSTM network structure (SBU-LSTM) to help design neural network structures for traffic state prediction. For time-series forecasting of confirmed cases, deaths, and recoveries in 10 major countries affected by COVID-19 (Shahid et al., 2020) [7], an autoregressive integrated moving average (ARIMA), support vector regression (SVR), long-shot short-term memory

(LSTM), and bi-directional long-short-term memory (Bi-LSTM) consisting of a Predictive modeling. In all scenarios, Bi-LSTM, LSTM, GRU, SVR and ARIMA models are ranked in order of performance from high to low. (Fan et al., 2021) [8] investigated well production prediction based on arima-lSTM model considering manual operations. A novel hybrid model was developed which considered the advantages of linear and nonlinear as well as the effect of manual operations. The coupled models (ARIMA-LSTM, ARIMA-LSTM-DP) show better results than the separate ARIMA, LSTM or LSTM-DP models, where the ARIMA-LSTM-DP model performs better when the production sequence of the wells is affected by frequent manual operations. The illegal wildlife trade is a major driver behind the global loss of wildlife and is pushing many species towards extinction[9]. Therefore, this paper investigates the illegal wildlife trade quantity prediction through LSTM model. By analyzing the illegal wildlife trade data and using the attributes, features and algorithmic principles of the data to make a comprehensive decision, we realize the prediction of illegal wildlife trade quantity based on LSTM algorithm.

2. Quantitative data analysis of illegal wildlife trade

In order to gain an in-depth understanding of the scale and development trend of illegal wildlife trade, the study drew a curve of global illegal wildlife trade from 1975 to 2021(data from https://www.kaggle.com/datasets/azizsalmi/cites-wildlifetrade-database-1975-2022), and conducted a detailed analysis of wildlife trade trends in different protection levels (I, II, III). As Fig 1 shows, the volume of illegal wildlife trade has risen dramatically over the past four decades, revealing this profound global problem. Among these illegal transactions, the trade share of protected species II is the most significant, indicating that this category of wildlife faces more serious threats and poaching activities.

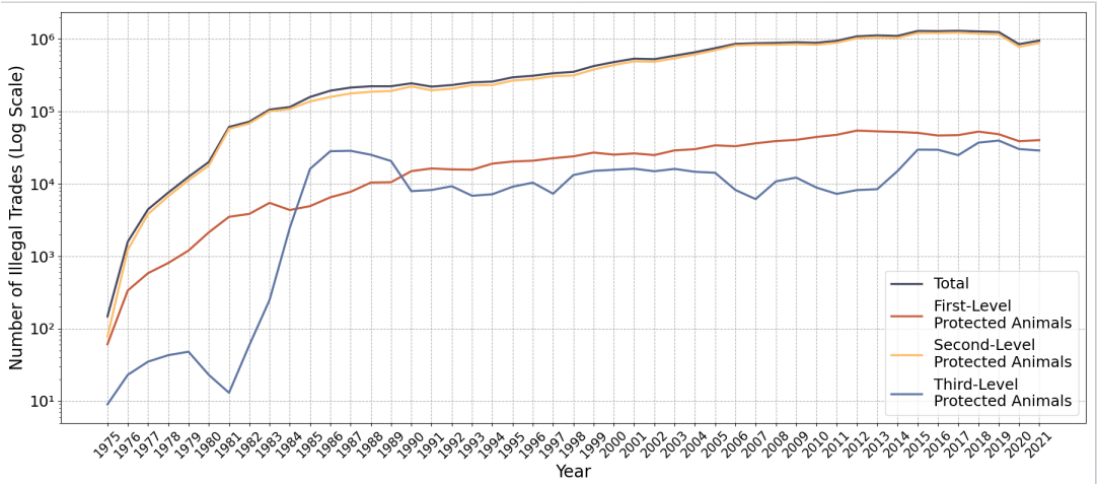

Fig 1. Number of illegal wildlife trades in each protection class

This paper have created a pie chart illustrating the proportion of various wildlife species in illegal wildlife trade, categorized by conservation class. As shown in Fig 2, among the first-level protected animals, Falco hybrid, Crocodylus siamensis, and Loxodonta africana constitute a significant proportion. Within the second-level protected animals, Alligator mississippiensis, Malayopython reticulatus, and Crocodylus niloticus dominate the distribution.

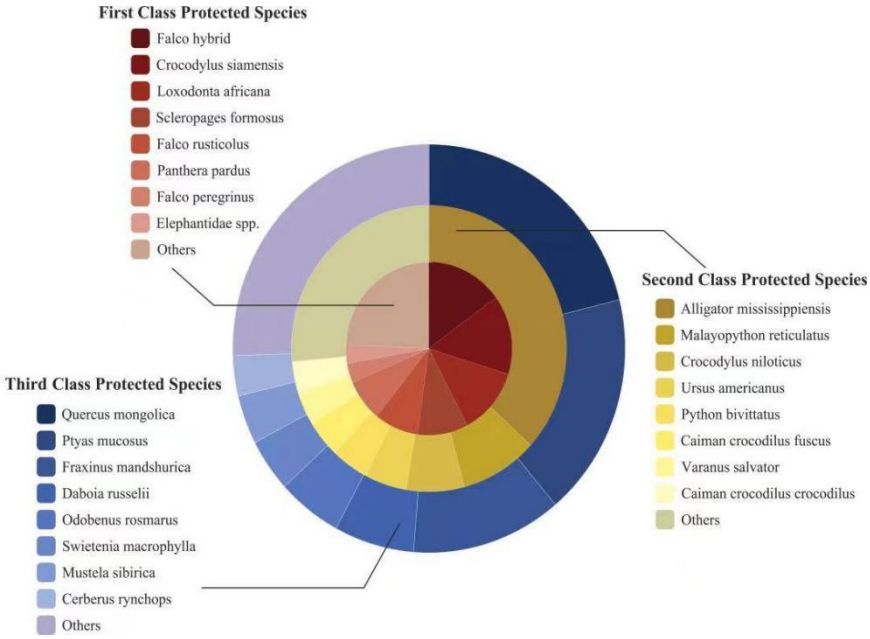

Fig 2. The proportion of animals in trade at all levels of protection

3. The basic funamental of Long-Short Term Memory

The LSTM model has a good long-term dependence relationship, and uses the nonlinear fitting characteristics of the algorithm to process time series, which is composed of forgetting gate, input gate and output gate. From t-1 to t time, the model propagates forward and generates memory cells. A gate is used to control the flow of information in the network, wherein the forgetting gate determines the amount of original information to be retained and the type of unimportant memory to be discarded. The role of the input gate is to determine the amount of current information to be saved by the model, and the output gate determines the amount of output information. The structure of the LSTM model, as shown in Fig 3.

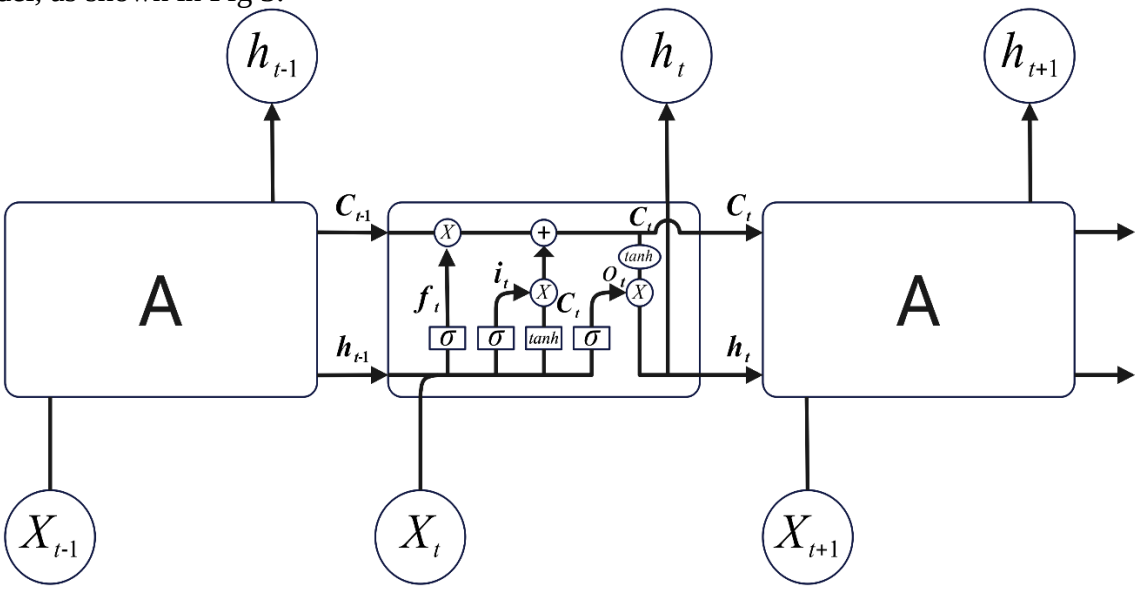

Fig 3. Structural diagram of the LSTM model

The general model of artificial neural network consists of four basic elements, which are:

(1) Forgetting gate: Determines the proportion of information that needs to be forgotten by the neuron, that is, which information needs to be retained or forgotten. The activation function consists of a sigmoid with an output between 0 and 1, where "0" means that this part of the information is completely forgotten by the neuron and "1" means that this part of the information is completely

retained by the neuron. By learning the appropriate forgetting ratio, the model can dynamically decide which information to retain or forget according to the current input and the contents of the memory cells in the previous time step, so as to better deal with the dependency in long sequence data. The calculation formula is as follows:

$$f_t = \sigma(w_f \times [h_{t-1}, x_t] + b_f) \tag{1}$$

where, f_t represents the output of the forgetting gate at time t, that is, the forgetting factor; σ represents the activation function sigmoid; w_f represents the weight matrix of the forgetting gate; $[h_{t-1}, x_t]$ represents the formation of a new vector of short-term memory h_{t-1} and event information x_t; b_f represents the bias term of the forgetting gate

(2) Input gate: Controls what information needs to be stored. The input gate consists of two parts, the first part is to use the previously hidden information and the current input information through the sigmoid layer to determine what values need to be updated. The second part is to use the tanh layer to create a new candidate value vector and generate candidate memory. The calculation formula is as follows:

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

$$\tilde{C}_t = tanh(w_g \times [h_{t-1}, x_t] + b_c) \tag{3}$$

$$tanh = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{4}$$

$$C_t = C_{t-1} \otimes f_t + \tilde{C}_t \otimes i_t \tag{5}$$

Where, i_t represents the output of the input gate, w_i represents the weight vector of the input gate, b_i is the bias term of the input gate, $\tilde{C}_t$ represents the short-term memory input from the learning gate, that is, the current memory, w_g represents the weight matrix of the input node, and b_c is the bias term of the current input,C_t is the memory cells generated during positive propagation from t-1 to t time.

(3) Output gate: Controls the output of new states and determines what information to keep and what to discard. It accepts the input data of the current moment and the hidden state of the previous moment, the new cell state is obtained by sigmoid function, and the value of C_t is scaled to between -1 and 1 by $tanh$ function, and the result is obtained by multiplication.

$$o_t = \sigma(w_o \times [h_{t-1}, x_t] + b_o) \tag{6}$$

$$h_t = o_t \otimes \tanh(C_t) \tag{7}$$

Where, o_t represents the output of the output gate, w_o is the weight vector of the output gate, b_o is the offset term of the output gate, h_t is the new state at time t, and $\otimes$ represents element-by-element multiplication

4. The large data prediction model

4.1. The establishment of simulation model

For the prediction of illegal wildlife trade data, two prediction models, ARIMA and LSTM, are used for comparison experiments, and the prediction models are realized by PyCharm software and Python language. By analyzing the global wildlife illegal trade data from 1995 to 2021, the two prediction models of ARIMA and LSTM were established to obtain the prediction results of the number of global wildlife illegal trade for each year from 2022 to 2029, respectively.

The ARIMA model in machine learning is a commonly used time series forecasting method,which is applicable to short- and medium-term time series without considering external influences[10]. Based on the autocorrelation and trend of historical observations, ARIMA (autoregressive integral moving Average) model captures the intrinsic pattern of time series through the steps of difference,

autoregression and moving average. This method performs well in predicting and analyzing data trends that are not significantly affected by external factors, and is especially widely used in fields such as finance, economics, and natural sciences.

$$y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \ldots - \theta_q \varepsilon_{t-q} \quad (8)$$

4.2. Analysis of experimental results

In Fig 4, two different prediction models, ARIMA (autoregressive integral Moving average) and LSTM (Short term Memory Network), are presented for the global illegal wildlife trade from 2022 to 2029. The ARIMA model, shown on the left, uses a statistical approach based on time series to make predictions using past observations and is well adapted to trends and seasonal changes. The LSTM model, shown on the right, is a deep learning model that captures more complex time series relationships and is suitable for non-linear and dynamic data patterns.

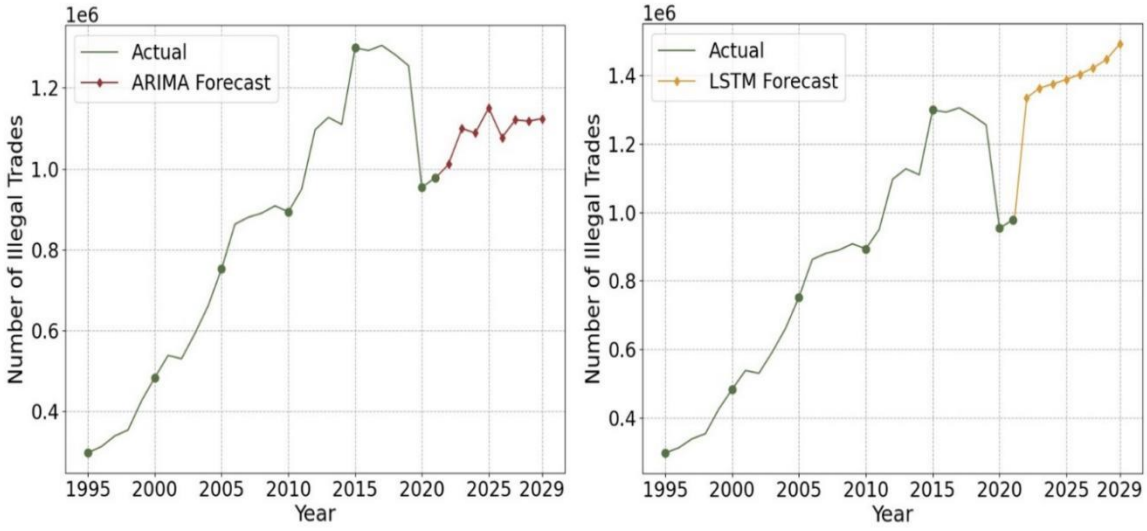

Fig 4: Prediction curves for the volume of illegal wildlife trade

It can be seen that the ARIMA algorithm for a class of models that capture the temporal structure in time series data, however, ARIMA is a forecasting method based on linear regression, and it is difficult to model the nonlinear relationship between variables with ARIMA model alone. Autoregressive Integrated Moving Average (ARIMA) is a generalized autoregressive moving average (ARMA) model that combines an autoregressive (AR) process with a moving average (MA) process to construct a composite model of the time series. AR: Autoregressive. A regression model that uses dependencies between one observation and multiple lagged observations. Smooths the time series by calculating the difference between observations at different times. MA: Moving Average. A method of calculating the dependence between observations and residual terms when a moving average model is used for lagged observations (q). A simple form of the AR model of order p, AR(P), can be written as a linear process. For seasonal time series data, the short-term non-seasonal component is likely to contribute to the model. For this dataset, the ARIMA predictions are steady and the predictions are closer to the historical average.

And Long Short-Term Memory (LSTM) network, as a special recurrent neural network structure, can solve the gradient vanishing problem. Capturing long-term dependencies in time series data. Store the past data into 'memory nerves', which are forgetting gates, input gates, and output gates. Having sensitivity to time and being able to learn patterns and features in time-series data gives LSTM an advantage in tasks such as time-series prediction and signal processing. The prediction results are presented relatively aggressively, the prediction results are presented in the high level of gradual growth, for similar (trading volume) such as the original data fluctuation is more intense and non-linear complex time series data, the performance of the prediction using LSTM is superior, closer to the actual situation of the data.

Based on the experimental results, it can be observed that both the ARIMA and LSTM prediction models show an upward trend in the amount of illegal wildlife trade. It is particularly noteworthy that compared with the ARIMA model, the prediction results of the LSTM model show a more obvious upward trend, and the prediction value is larger.

This may indicate that LSTM models have advantages in capturing more complex nonlinear relationships and dynamic changes in time series. Its greater ability to learn enables it to better adapt to the complex and volatile phenomenon of the illegal wildlife trade. However, it is important to note that the high predictive value of the LSTM model may also reflect the sensitivity of the model, which may be more sensitive to outliers or noise.

5. Conclusions

The changing trend of global wildlife illegal trade data provides the basis for the establishment of prediction models, and the dataset is characterized by large fluctuations, non-linear relationships, and non-seasonal time data. In this paper, by combining the characteristics of two prediction models, ARIMA and LSTM, and establishing two wildlife illegal trade prediction models, and analyzing the samples and prediction results. The experimental results show that for this kind of data, LSTM can solve the problem of gradient vanishing, capture the long-term dependence in time series data, and perform better for prediction.The LSTM recurrent neural network model has good prediction and robustness, and has certain practical application value.

Based on the above conclusions, in the face of the growing illegal wildlife trade, which is a threat to ecological balance and species diversity, global efforts are needed to stop it, and public education should be strengthened to raise people's awareness of the dangers of illegal wildlife trade. At the same time, the government is urged to formulate stricter regulations and enforcement measures to crack down on the sources and channels of trade. The international community should strengthen coordination, jointly develop a framework for transnational cooperation, share information and assist in investigations, and jointly address this global challenge. Support wildlife conservation organizations, advocate for sustainable use, and actively participate in social movements to build a more sustainable and harmonious natural ecosystem. Only through joint efforts can people around the world effectively protect wildlife and safeguard the precious diversity of life on Earth.

References

[1] Karim F, Majumdar S, Darabi H. Insights into LSTM fully convolutional networks for time series classification[J]. IEEE Access, 2019, 7: 67718-67725.

[2] Yu Y, Si X, Hu C, et al. A review of recurrent neural networks: LSTM cells and network architectures[J]. Neural computation, 2019, 31(7): 1235-1270.

[3] Wei W, Wu H, Ma H. An autoencoder and LSTM-based traffic flow prediction method[J]. Sensors, 2019, 19(13): 2946.

[4] Li T, Hua M, Wu X U. A hybrid CNN-LSTM model for forecasting particulate matter (PM2. 5)[J]. Ieee Access, 2020, 8: 26933-26940.

[5] Park K, Choi Y, Choi W J, et al. LSTM-based battery remaining useful life prediction with multi-channel charging profiles[J]. Ieee Access, 2020, 8: 20786-20798.

[6] Cui Z, Ke R, Pu Z, et al. Stacked bidirectional and unidirectional LSTM recurrent neural network for forecasting network-wide traffic state with missing values[J]. Transportation Research Part C: Emerging Technologies, 2020, 118: 102674.

[7] Shahid F, Zameer A, Muneeb M. Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM[J]. Chaos, Solitons & Fractals, 2020, 140: 110212.

[8] Fan D, Sun H, Yao J, et al. Well production forecasting based on ARIMA-LSTM model considering manual operations[J]. Energy, 2021, 220: 119708.

[9] Vincent Nijman, Thais Morcatty, Jaima H. Smith, Sadek Atoussi, Chris R. Shepherd, Penthai Siriwat, K. Anne-Isola Nekaris, and Daniel Bergin. Illegal wildlife trade - surveying open animal markets and online platforms to understand the poaching of wild cats. Biodiversity, 20(1):58–61, 2019.

[10] ZHANG Huifeng CAO Xiaoyu, XU Bo and WEI Xinjiang. Research on price level prediction of antitumor drugs in shandong province based on lstm and arima models. China Hospital Statistics, 29(419-423), 2022.