Development Economics of China
2025, VOL. 9, NO. 8, 117-121
DOI: 10.47297/wspdecWSP2515-797321.20250908

Research on the Scale Prediction of the Digital Economy Integrating Statistical Models and Machine Learning

Xueying Han[1], Yan Wang[1], Guoqing Chen[2]*, Tianwen Zhao[3]

[1]School of Finance, Chengdu Jincheng College, Chengdu, Sichuan, 611731, China

[2]Mathematical Modeling Research Center, Chengdu Jincheng College, Chengdu, Sichuan, 611731, China

[3]Department of Trade and Logistics, Daegu Catholic University, Gyeongsan, 38430, Republic of Korea

ABSTRACT
The digital economy has become an important component of our country's modern economic system, but its operating conditions are influenced by multiple factors and face significant uncertainties during its development. This paper addresses the issue of forecasting the scale of the digital economy by proposing a combinatorial forecasting method that integrates statistical models and machine learning. The study first constructs an index system of influencing factors suitable for the digital economy in China, addressing issues such as missing data, inconsistent dimensions, and feature redundancy; furthermore, it proposes a combinatorial model framework based on nonlinear error correction strategy and meta-heuristic algorithm optimization. Empirical research shows that the combinatorial model significantly outperforms individual models in terms of convergence speed, parameter optimization, and predictive accuracy, with the PCA-BPNN model's predictive accuracy improving by 30% to 73.8% compared to traditional methods. The research findings provide a scientific basis for government departments in formulating digital economy policies and offer theoretical support for regional digital economy development strategies.

KEYWORDS
Digital economy; Forecasting models; Machine learning; Statistical models; Integration models; ARIMA; BP neural networks; LSTM

1 Introduction

In recent years, China's digital economy has continued to grow at a high speed. However, due to the influence of multiple internal and external factors, its operating pattern has become more complex and uncertainty has increased significantly. Carrying out scientific and accurate predictions of the scale of the digital economy will not only help to reveal its internal development laws, but also provide important references for the government to formulate macroeconomic policies and optimize its structure.

At present, there is still a lack of mature and reliable methods for predicting the scale of China's digital economy. Although traditional statistical models such as time series analysis and regression analysis have a solid theoretical basis and strong interpretability, they have limitations in dealing with nonlinear relationships and high-dimensional data [1]. Although machine learning methods can effectively capture complex nonlinear patterns, they face the dilemma of "high-dimensional small samples" in economic forecasting and are prone to overfitting. Based on this, this study proposes a combined prediction method that integrates statistical models and machine learning from the dual dimensions of economics and management, based on the idea of cross-integration of systems engineering and multidisciplinary studies [2]. The main innovations include: (1) constructing an efficient and concise indicator system of factors affecting the scale of the digital economy; (2) proposing a combined modeling method based on a nonlinear error correction strategy; (3) introducing a meta-heuristic algorithm to optimize the selection of neural network parameters. By comparing and analyzing various models, we determined the optimal prediction path that fits the characteristics of China's digital economy, providing new methodological support for related research.

2 Introduction to the prediction model for the scale of the digital economy

2.1 Construction of the predictive indicator system

This study addresses the lack of an official system of influencing factors and publicly available datasets for predicting the scale of my country's digital economy. By integrating data aggregation, standardization, and correlation analysis into the design process, this study effectively mitigates the impact of missing data, inconsistent dimensions, and redundant features on forecast results. Regarding indicator selection, based on factors influencing regional digital economic development and qualitative evaluation models, this study comprehensively considers dimensions such as digital infrastructure, digital industrialization, industrial digitization, and digital innovation capabilities [3]. Furthermore, this study incorporates Pearson correlation coefficient analysis to identify indicators highly correlated with the scale of the digital

* **Corresponding Author:** Guoqing Chen, chenguoqing@cdjcc.edu.cn

economy.

$$\rho_{X,Y}=\frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y}=\frac{\mathrm{E}\left[\left(X-\mathrm{E}[X]\right)\left(Y-\mathrm{E}[Y]\right)\right]}{\sigma_X\sigma_Y}\#(1)$$

Here, $\rho_{X,Y}$ represents the correlation coefficient between variables X and Y, and σ_X and σ_Y represent the standard deviations of X and Y, respectively. When $\rho_{X,Y}$ is closer to 1, it indicates a higher degree of correlation between the two variables.

2.2 Single prediction model

2.2.1 Time series model

The ARIMA model is a classic method for time series forecasting, which captures linear relationships in the time series through autoregression and moving average on differenced stationary sequences. Its basic form is:

$$\phi(B)(1-B)^d X_t=\theta(B)\varepsilon_t\#(2)$$

Among them, $\varphi(B)$ and $\theta(B)$ represent the autoregressive and moving average polynomials, respectively, d is the order of differencing, and εt is the white noise sequence.

2.2.2 Machine learning model

The BP neural network is a multi-layer feedforward network utilizing error backpropagation, which achieves complex nonlinear mappings through its multi-layer structure consisting of input layers, hidden layers, and output layers. The training process of the network encompasses two processes: forward signal propagation and backward error propagation, wherein the network weights and thresholds are adjusted using the gradient descent method.

The LSTM model is a specialized type of recurrent neural network that effectively addresses the problems of vanishing and exploding gradients in the training of long sequences through a gating mechanism. Its core lies in the control of information flow via forget gates, input gates, and output gates, capturing long-term dependencies in time series data.

2.3 Combined prediction model

2.3.1 Nonlinear error correction combined model

To address the limited effectiveness of single time series models and neural networks in predicting the scale of the digital economy, this study proposes a combined approach based on a nonlinear error correction strategy. This method initially employs time series models for preliminary predictions, and then utilizes neural networks to further extract nonlinear features from the prediction residuals for error correction. The specific steps are as follows:

The preliminary prediction results are obtained by using ARIMA model is $\widehat{Y}_t^{(\mathrm{ARIMA})}$.

Calculate the prediction residual:

$$e_t=Y_t-\widehat{Y}_t^{(\mathrm{ARIMA})}\#(3)$$

The residual prediction model is trained using neural network:

$$\widehat{e}_t=f\left(X_t\right)\#(4)$$

Get the combined prediction results:

$$\widehat{Y}_t^{(\mathrm{comb})}=\widehat{Y}_t^{(\mathrm{ARIMA})}+\widehat{e}_t\#(5)$$

Among them, X_t represents the factor vector affecting the residual.

2.3.2 Original heuristic algorithm optimization model

To address the limitations of neural networks in prediction accuracy and the issue of blind parameter selection, this study proposes a hybrid approach combining meta-heuristics with neural networks. By introducing Particle Swarm Optimization (PSO) and Quantum Particle Swarm Optimization (QPSO), we optimize critical parameters within the neural network framework.

The QPSO algorithm is a quantum version of the PSO algorithm, which implements global search through qubit encoding and quantum gate update, with better convergence performance and optimization ability. Its particle update formula is:

$$m_{\mathrm{best}}=\frac{1}{M}\sum_{i=1}^{M}P_i\#(6)$$

$$X_{id}(t+1)=P_{id}\pm\alpha\left|m_{\mathrm{best},d}-X_{id}(t)\right|\cdot\ln\left(\frac{1}{u}\right)\#(7)$$

Among them, P_i represents the historical optimal position of particle i, mbest denotes the average optimal position, α

is the contraction-expansion coefficient, and u is a random number uniformly distributed between (0,1).

2.4 Model optimization and feature engineering strategy

2.4.1 Principal component analysis and feature dimension reduction

In the prediction of the scale of the digital economy, the feature dimensions are often high, with a large number of indicators exhibiting strong correlations, which can easily lead to model overfitting and increased computational complexity. This study employs Principal Component Analysis (PCA) to reduce the dimensionality of the original features. PCA transforms the original variables, which may exhibit correlations, into a set of linearly independent new variables through orthogonal transformation, arranged in order of variance contribution from largest to smallest. The top k principal components are selected as new feature inputs when the cumulative contribution rate reaches over 85%.

Let the original data matrix X be of $n \times p$ dimensions. After standardization, calculate the covariance matrix $S = \frac{1}{n-1} X^{\mathrm{T}} X$, Then perform eigenvalue decomposition on S.

$$S = V \Lambda V^{\mathrm{T}} \# (8)$$

Among them, Λ is the diagonal matrix of eigenvalues, and V is the corresponding matrix of eigenvectors. By selecting the eigenvectors corresponding to the k largest eigenvalues, we construct the projection matrix W, resulting in the reduced-dimensional data as follows:

$$T = XW \# (9)$$

Through PCA processing, not only the problem of multicollinearity is eliminated, but also the feature dimension is greatly reduced, which improves the training efficiency and generalization ability of the model.

2.4.2 Particle swarm optimization algorithm parameter tuning

The performance of neural network models largely depends on the selection of hyperparameters, including learning rate, number of hidden layer nodes, and number of iterations. Traditional grid search methods are cost-intensive and inefficient. This study introduces the PSO algorithm for automatic parameter optimization.

The PSO algorithm simulates the foraging behavior of bird flocks, where each particle represents a potential solution. It updates its position and velocity by tracking both the individual's historical best position and the group's global best position. The update formula for the i-th particle in the t-th iteration is given by:

$$v_i(t+1) = wv_i(t) + c_1 r_1 \left(p_i - x_i(t) \right) + c_2 r_2 \left(p_g - x_i(t) \right) \# (10)$$

$$x_i(t+1) = x_i(t) + v_i(t+1) \# (11)$$

In this, w is the inertia weight, which controls the balance between global and local search capabilities; c_1 and c_2 are learning factors, usually set to c_1=c_2=2; r_1 and r_2 are random numbers in the range [0, 1]; p_i is the historical optimal position of particle i; p_g is the global optimal position of the group.

2.4.3 Improvement of quantum particle swarm algorithm

To further enhance optimization efficiency and global search capabilities, this study introduces the Quantum Particle Swarm Optimization (QPSO) algorithm based on the PSO. By employing quantum bit encoding and quantum gate update mechanisms, QPSO enables particles to exhibit quantum behavior, allowing them to probabilistically occupy any position in the search space. This innovative approach effectively prevents premature convergence while maintaining optimal search performance. In QPSO, the particle position update formula is:

$$m_{\text{best}} = \frac{1}{M} \sum_{i=1}^{M} P_i \# (12)$$

$$X_{id}(t+1) = P_{id} \pm \alpha \left| m_{\text{best},d} - X_{id}(t) \right| \cdot \ln\left(\frac{1}{u} \right) \# (13)$$

In this framework, m_{best} represents the average of all particles' historical optimal positions, α denotes the contraction-expansion coefficient controlling convergence speed, and u is a uniformly distributed random number in the interval (0, 1). The QPSO algorithm eliminates the need for velocity vectors, features fewer parameters, and demonstrates superior global search capabilities, making it particularly suitable for high-dimensional complex optimization problems.

2.4.4 Cross validation and model evaluation

To objectively evaluate model performance and prevent overfitting, this study employs k-fold cross-validation. The dataset is randomly divided into k mutually exclusive subsets, with each iteration using k-1 subsets as training sets and the remaining 1 subset as a test set. This process is repeated k times, and the final evaluation of model performance is determined by averaging the results from all k tests.

To comprehensively evaluate the predictive performance of the model, a multi-index evaluation system is adopted, including error-based and goodness-of-fit metrics [4-6]:

Mean Absolute Error (MAE):

$$MAE=\frac{1}{n}\sum_{i=1}^{n}\left|y_i-\hat{y}_i\right|\#(14)$$

It measures the average magnitude of prediction errors without considering their direction.

Mean Absolute Percentage Error (MAPE):

$$MAPE=\frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}_i}{y_i}\right|\#(15)$$

It represents the prediction error as a percentage of the actual values, making it easy to interpret across datasets with different scales.

Root Mean Square Error (RMSE):

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}\#(16)$$

It penalizes larger errors more heavily due to the squared term, providing a sensitive measure of prediction accuracy.

Coefficient of Determination (R^2):

$$R^2=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}\#(17)$$

Where $\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$. It evaluates the proportion of variance in the dependent variable that is explained by the model, with values closer to 1 indicating better performance.

Through comprehensive evaluation of multiple indicators, the model is ensured to not only perform well in point prediction accuracy, but also have advantages in trend capture and stability.

3 Analysis of statistical models and machine learning models for digital economy scale prediction

3.1 Data sources and preprocessing

This study collected digital economy-related indicator data from Jiangxi Province between 2011 and 2020. To address missing data in the sample, we employed linear regression for monthly multiple imputation to ensure data integrity. Standardization procedures were implemented to eliminate the impact of inconsistent units across datasets:

$$Z=\frac{X-\mu}{\sigma}\#(18)$$

Among them, μ is the sample mean and σ is the sample standard deviation. In addition, redundant indicators are eliminated through multiple collinearity diagnosis and correlation analysis, reducing the data dimension and improving the efficiency of model training.

3.2 Model performance comparison analysis

To evaluate the predictive performance of different models, this study used Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and the coefficient of determination (R^2) as evaluation metrics. Table 1 presents the performance comparison results of different models in predicting the scale of digital economy in Jiangxi Province.

Table 1 Comparison of performance of different prediction models

Model Type	MAE	MAPE	R^2	Training speed	Interpretability
ARIMA	458.732	0.0452	0.872	Fast	strongly
BP neural network	362.415	0.0358	0.913	Medium	strongly
LSTM	318.627	0.0314	0.935	Slow	Weak
PSO-BP	286.194	0.0283	0.951	Slow medium	Medium
QPSO-LSTM	252.736	0.0249	0.968	Slow medium	Medium
PCA-BPNN	240.018	0.0233	0.976	Medium	Medium

As shown in Table 1, the fusion model demonstrates significantly better prediction accuracy than the single model. The PCA-BPNN model outperforms all others, achieving a Mean Absolute Error (MAE) of 240.018 and a Mean Absolute Percentage Error (MAPE) of 0.0233. This represents an improvement of over 30% compared to the single model, with the highest prediction accuracy reaching 73.8%. These results indicate that principal component analysis effectively eliminates redundant information among features, thereby enhancing both the training efficiency and generalization capability of the neural network。

3.3 Prediction of Digital Economy Development

Through a constructed composite forecasting model, this study analyzes global digital economy trends. The findings indicate that worldwide digital trade is projected to reach $7.39 trillion in 2024, setting a new record and accounting for 22.22% of total global goods and services exports. Digital delivery trade and digital order trade maintain a "dual-engine" dynamic, with respective scales of $4.78 trillion and $2.62 trillion, representing 64.63% and 35.37% of the total.

In terms of national distribution, the United States, the United Kingdom, and Ireland consistently rank as the top three in global digital trade volume, while China ranks fourth with $765.898 billion, demonstrating strong competitiveness. This outcome aligns with the rapid development status of China's digital economy and also confirms the effectiveness of the combined forecasting model.

Table 2 Ranking of digital trade scale of major countries in the world in 2024

Rank	Country	Scale of digital trade (USD 1 billion)	Proportion	Main Features
1	United States	12,458.32	16.85%	Technology leads, all-round development
2	United Kingdom	9,327.16	12.62%	Financial digital services stand out
3	Ireland	8,163.45	11.05%	Multinational enterprises have centralized their European headquarters
4	China	7,658.98	10.36%	Rapid growth, huge scale
5	Germany	6,328.67	8.56%	High degree of industrial digitization
6	Japan	5,237.89	7.09%	Innovation in digital consumption is active
7	France	4,865.44	6.58%	The digital economy is strongly supported by policies

4 Conclusion

This study proposes a combined forecasting method that integrates statistical models and machine learning to predict the scale of China's digital economy. By constructing an indicator system for factors influencing the digital economy, it effectively alleviates issues such as missing data, inconsistent dimensions, and redundant features. Combining correlation analysis with principal component dimensionality reduction improves modeling efficiency. The introduction of a nonlinear error correction mechanism and a meta-heuristic optimization algorithm significantly enhances forecasting accuracy. Experiments show that the PCA-BPNN model achieves mean absolute error (MAE) and mean absolute percentage error (MAPE) of 240.018 and 0.0233, respectively, with an accuracy improvement of up to 73.8%. The QPSO-LSTM model outperforms traditional models in both convergence speed and forecasting accuracy. Forecast results indicate that China's digital trade volume is expected to reach US $765.898 billion in 2024, ranking fourth globally. Artificial intelligence will become a core driver of digital economic growth and is expected to boost global digital services trade growth by 18 percentage points by 2040 compared to the baseline.

Based on the research results, the following policy recommendations are proposed: (1) Improve the digital economy statistical system, establish a unified indicator framework, and enhance data quality and accessibility; (2) Increase investment in artificial intelligence (AI) research and development to promote its deep integration with the real economy; (3) Promote the use of statistical-machine learning fusion models to enhance the scientific nature and foresight of policymaking. This study still has several limitations: First, the influencing factor indicator system needs to be dynamically updated as the digital economy develops; second, the combined model is complex and computationally expensive, requiring future optimization of algorithm efficiency to support real-time prediction; third, insufficient data accessibility, particularly the lack of detailed regional data, hinders refined modeling. With increased data openness, more refined predictive models can be developed.

Funding

This research was financially supported by 2025 Chengdu Jincheng College Campus level Project (NO: 2025JCKY0013)

References

[1] Jiang Y. (2022). Prediction model of the impact of innovation and entrepreneurship on China's digital economy based on neural network integration systems. Neural Computing and Applications, 34(4), 2661-2675.

[2] Wang L., & Zhao L. (2022). Digital economy meets artificial intelligence: forecasting economic conditions based on big data analytics. Mobile information systems, 2022(1), 7014874.

[3] Sun C., Xu M., & Wang B. (2024). Deep learning: spatiotemporal impact of digital economy on energy productivity. Renewable and Sustainable Energy Reviews, 199, 114501.

[4] Zhao T., Chen G., Pang C., & Busababodhin P. (2025). Application and Performance Optimization of SLHS-TCN-XGBoost Model in Power Demand Forecasting. CMES-Computer Modeling in Engineering and Sciences, 143(3), 2883-2917.

[5] Zhao T., Chen G., Suraphee S., Phoophiwfa T., & Busababodhin P. (2025). A hybrid TCN-XGBoost model for agricultural product market price forecasting. PLoS One, 20(5), e0322496.

[6] Zhao T., Chen G., Gatewongsa T., & Busababodhin P. (2025). Forecasting Agricultural Trade Based on TCN-LightGBM Models: A Data-Driven Decision. Research on World Agricultural Economy, 207-221.