JOURNAL OF GEOSPACE SCIENCE

Vol. 6 No. 1 (2026)

E-ISSN : 2413-6093
https://jogss.org

Forecasting China's GDP Growth with Lagged Investment and Household Consumption

Liangyan Cai

School of Digital Intelligence and Management, Guangzhou Xinhua University, 510520, Guangzhou, China.

Hanfu Liang

Satbayev University,050013,Almaty, Kazakhstan.

ABSTRACT

This study examines a narrow forecasting task: predicting China's annual real GDP growth with one-year lagged macroeconomic indicators from the World Bank World Development Indicators [1]. The predictors include lagged GDP growth, gross capital formation growth, household final consumption growth, industrial value-added growth, and CPI inflation. The training sample covers 1996-2018, and the test sample covers 2019-2024. A historical-mean benchmark, a linear regression model, and a compact random-forest model are compared. The results show that the random forest improves on the historical mean in this small test period, but it still misses the sharp movements around 2020 and 2021. The evidence supports a cautious use of machine-learning forecasts as a complement to conventional macroeconomic interpretation [4, 6].

KEYWORDS

China GDP growth; investment; household consumption; random forest; World Bank WDI

1. Research Question and Data

The research question is intentionally narrow: can a few lagged, publicly available annual indicators provide a usable forecast of China's next-year real GDP growth rate? The design follows the macroeconomic forecasting tradition in which compact indicator sets are evaluated against explicit out-of-sample benchmarks [4, 5]. The data source is transparent and the model can be replicated. Annual observations are limited, so the empirical design emphasizes a clear target, a small predictor set, and a held-out test period rather than a high-dimensional forecasting system. All variables are downloaded from the World Bank WDI database [1]. The target variable is China's annual real GDP growth rate. The explanatory variables are lagged GDP growth, lagged gross capital formation growth, lagged household final consumption growth, lagged industrial value-added growth, and lagged CPI inflation. Every predictor is lagged by one year so that the model does not use information from the year being forecast.

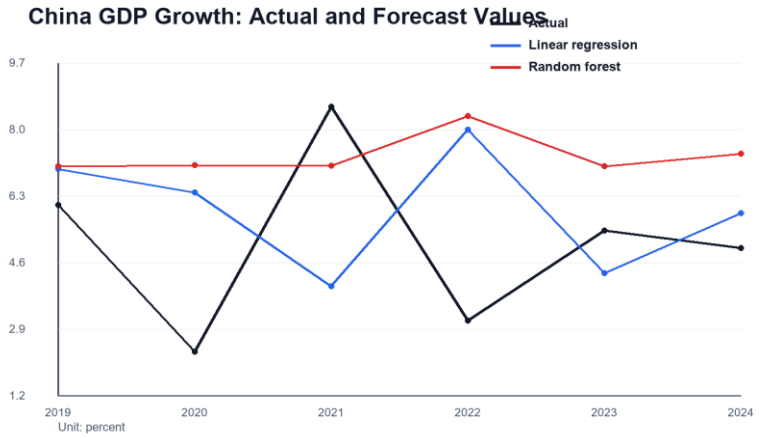


Figure 1. Actual and forecast GDP growth, 2019-2024

Source: World Bank WDI; author's calculation using one-year lagged predictors.

2. Variable Construction

The sample begins in 1996 after the one-year lag is created. Observations from 1996 to 2018 are used for model training, while 2019 to 2024 are held out for testing. This testing window includes the pandemic shock, the rebound in 2021, and the subsequent normalization of growth. It is therefore a difficult but informative test period. The selected variables have direct economic meanings. Investment and consumption represent two important demand-side channels. Industrial value added reflects production-side momentum. CPI inflation captures the price environment. Lagged GDP growth captures persistence in annual growth. This design treats annual GDP growth as a business-cycle signal rather than as a full national-accounting system. Investment, household consumption, and industrial value added are not included because they are the only possible predictors; they are included because they give a compact demand-production view that can be read without private data.

No complex transformation is used beyond the one-year lag. This choice is intentional. With fewer than thirty annual training observations, adding many interaction terms would make the model look more sophisticated but less credible. The compact variable set also makes the table and figure interpretation easier.

Table 1. Recent data used in the GDP growth forecast

Year	GDP	Investment	Consumption	Industry	CPI
2019	6.07	4.13	5.69	4.55	2.90
2020	2.34	4.20	-2.71	2.13	2.42
2021	8.57	3.97	13.14	8.21	0.98
2022	3.13	2.85	1.49	2.30	1.97
2023	5.41	3.44	9.65	4.37	0.23
2024	4.98	3.14	4.98	5.32	0.22

3. Model Design

Three models are estimated. The first is a historical-mean benchmark, which predicts each test-year value using the average GDP growth rate in the training sample. The second is ordinary linear regression. The third is a small random-forest regressor composed of shallow regression trees [6]. The random forest is used because it can capture simple nonlinear splits among lagged indicators. The tree depth is restricted, and the minimum leaf size is set to avoid fitting individual historical years too closely. This conservative specification is suitable for a short annual macroeconomic sample [8, 9]. Model accuracy is evaluated with MAE and RMSE. MAE is easier to interpret as the average absolute forecast error, while RMSE gives more weight to large mistakes [11]. Because there are only six test years, the two indicators should

be read together with the forecast plot and with formal caution about predictive comparison [13].

4. Forecasting Results

The forecast plot shows that all models struggle with the sudden decline in 2020 and the rebound in 2021. This is expected. The one-year lagged variables do not fully contain the information about public-health restrictions, emergency policy responses, or the speed of reopening. The random forest tracks the general direction somewhat better than the historical mean. However, it is still conservative in years with sharp turning points. This pattern is a useful reminder that annual lagged variables can describe persistence, but they cannot fully anticipate exceptional shocks.

Table 2. Forecast errors for GDP growth, test period 2019-2024

Model	MAE	RMSE
Historical mean	3.973	4.458
Linear regression	2.729	3.258
Random forest	2.750	3.199

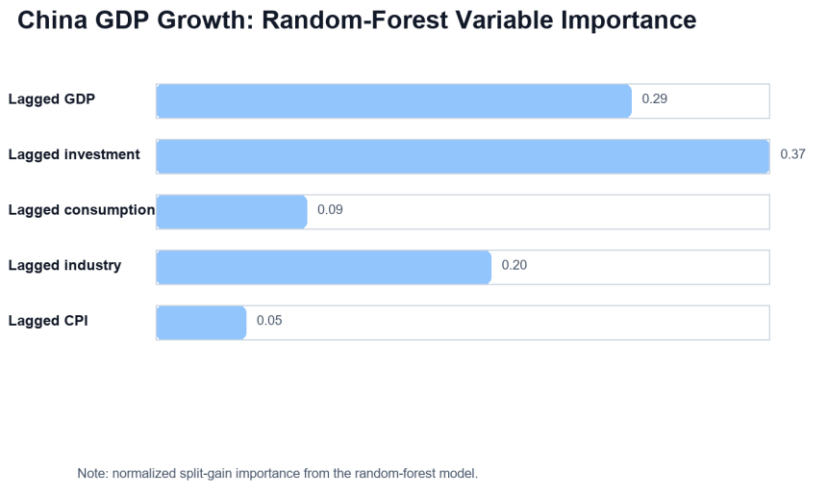


Figure 2. Variable importance in the GDP growth forecast

Source: World Bank WDI; author's calculation using one-year lagged predictors.

The variable-importance figure indicates that lagged investment, lagged GDP growth, and lagged household consumption are used more often in the random-forest splits. This result is economically plausible because these variables are directly connected to aggregate demand and growth persistence.

CPI inflation receives a smaller contribution in this sample. That does not mean prices are irrelevant to macroeconomic performance. It only means that, in this particular annual forecasting setup, lagged CPI provides less direct forecasting information for next-year real GDP growth than the activity variables.

The error table and the plot should be read together. A model can have a lower average error while still missing a turning point. For annual GDP growth, this distinction matters because policymakers and firms care not only about average accuracy but also about whether the model recognizes downshifts and recoveries.

5. Discussion

The main value of the exercise is transparency. Each variable has a public source, each forecast uses only lagged information, and the comparison includes a simple benchmark. This structure is consistent with applied econometric guidance that model performance should be judged against explicit alternatives [8, 9]. The results also show why macroeconomic forecasting should combine model output with economic judgment. A shallow random forest can organize lagged information, but it cannot read unexpected shocks that are absent from the training variables. The model is therefore a decision aid, not a substitute for interpretation.

For GDP growth, the model is most useful when it is read as a bridge between familiar macro indicators and a formal out-of-sample forecast. If investment and industrial output weaken together, the fitted trees tend to move the forecast downward. If consumption and production indicators send mixed signals, the forecast becomes more conservative. For a more practical forecasting system, quarterly data or monthly indicators such as PMI, electricity use, credit growth, and freight indicators would be needed. The present paper is better understood as a small reproducible demonstration using annual open data.

6. Limitations and Replication

The replication files include the raw WDI data, the model-ready dataset, the prediction table, and the generated figures. The WDI indicator codes are NY.GDP.MKTP.KD.ZG, NE.GDI.TOTL.KD.ZG, NE.CON.PRVT.KD.ZG, NV.IND.TOTL.KD.ZG, and FP.CPI.TOTL.ZG. The random seed is fixed at 20260620. The random forest uses 150 trees, a maximum depth of three, and a minimum leaf size of three. These settings are intentionally modest because the annual sample is small. Two limitations should be kept clear. First, WDI annual data are not suitable for genuine nowcasting. Second, the variables move together over time, so the paper compares predictive performance rather than estimating causal effects. Another limitation is the absence of policy and expectation variables. Fiscal support, credit conditions, and confidence changes can alter GDP growth within the forecast year, but they are not represented in the small WDI variable set. The model therefore tends to extrapolate from previous real-activity conditions. A stronger version of the study could use quarterly data and a rolling-origin evaluation design. That would allow the forecast to be updated several times within a year and would make it possible to test whether the random forest adds value beyond conventional bridge equations. Even with these limits, the exercise is useful because every step is inspectable. The downloaded series, lag construction, train-test split, error metrics, and variable-importance chart can all be checked without relying on private data or undocumented software.

7. Conclusion

The results show that a small random-forest model can extract some information from lagged investment and consumption indicators when forecasting China's annual GDP growth. The improvement is modest and should be interpreted within the limits of the annual WDI sample. The main contribution is methodological discipline. A narrow target, public data, a benchmark model, and explicit test-period errors make the forecast evidence easier to evaluate. For applied forecasting work, the result suggests a balanced position. Machine-learning models can help organize lagged macroeconomic information, especially when relationships are nonlinear. However, they still require economic interpretation, careful benchmarks, and honest discussion of shock years. A limited target, a reproducible dataset, two simple figures, and two tables are sufficient for a focused forecasting note when the interpretation is kept proportional to the evidence. A natural next step is a timelier data design, such as quarterly or monthly growth monitoring, where the same modelling logic can be tested under a more realistic information schedule.

References

World Bank. (2026). World Development Indicators database. https://api.worldbank.org/

Zhang, Q., Ni, H., & Xu, H. (2023). Nowcasting Chinese GDP in a data-rich environment: Lessons from machine learning algorithms. Economic Modelling, 122, 106204. https://doi.org/10.1016/j.econmod.2023.106204

Yang, Y., Xu, X., Ge, J., & Xu, Y. (2024). Machine Learning for Economic Forecasting: An Application to China's GDP Growth. arXiv:2407.03595. https://doi.org/10.48550/arXiv.2407.03595

Stock, J. H., & Watson, M. W. (2002). Macroeconomic forecasting using diffusion indexes. Journal of Business & Economic Statistics, 20(2), 147-162. https://doi.org/10.1198/073500102317351921

Giannone, D., Reichlin, L., & Small, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. Journal of Monetary Economics, 55(4), 665-676. https://doi.org/10.1016/j.jmoneco.2008.05.010

Breiman, L. (2001). Random Forests. Machine Learning, 45, 5-32. https://doi.org/10.1023/A:1010933404324

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189-1232. https://doi.org/10.1214/aos/1013203451

Varian, H. R. (2014). Big Data: New Tricks for Econometrics. Journal of Economic Perspectives, 28(2), 3-28. https://doi.org/10.1257/jep.28.2.3

Mullainathan, S., & Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. Journal of Economic Perspectives, 31(2), 87-106. https://doi.org/10.1257/jep.31.2.87

Athey, S., & Imbens, G. W. (2019). Machine Learning Methods That Economists Should Know About. Annual Review of Economics, 11, 685-725. https://doi.org/10.1146/annurev-economics-080217-053433

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. International Journal of Forecasting, 22(4), 679-688. https://doi.org/10.1016/j.ijforecast.2006.03.001

Hyndman, R. J., & Athanasopoulos, G. (2021). Forecasting: Principles and Practice (3rd ed.). OTexts.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. Journal of Business & Economic Statistics, 13(3), 253-263. https://doi.org/10.1080/07350015.1995.10524599

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning: With Applications in R (2nd ed.). Springer. https://doi.org/10.1007/978-1-0716-1418-1

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7