

Forecasting the Digital Divide: A Comparative Machine Learning Study of ICT Access Trajectories

Liana Rakhimzhanova

`liana.rakhimzhanova@nu.edu.kz`

Nazarbayev University <https://orcid.org/0009-0005-0575-0788>

Agzam Shamsadinov

Kazakh-British Technical University <https://orcid.org/0009-0007-7792-8666>

Research Article

Keywords: digital divide, digital inequality, connectivity index, machine learning, time-series forecasting, Kazakhstan, Central Asia

Posted Date: October 8th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11283628/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Forecasting the Digital Divide: A Comparative Machine Learning Study of ICT Access Trajectories

Liana Rakhimzhanova & Agzam Shamsadinov

Abstract—Digital connectivity is an important dimension of digital inclusion, but connectivity should not be treated as equivalent to digital skills, quality of use, downstream benefits, or cybersecurity-governance capacity. This study constructs a composite Connectivity Index from Internet use, mobile subscriptions, fixed broadband subscriptions, and electricity access and evaluates its short-term predictability using persistence and multivariate supervised models. World Bank World Development Indicators are used after a reporting-coverage filter. The baseline index assigns weights of 0.40, 0.20, 0.20, and 0.20, respectively. Because these weights are theory-informed rather than estimated from data, alternative weighting schemes are evaluated as robustness checks. In the final rerun, the index is available for 2002–2024, while the lagged modeling sample covers 2004–2024 and contains 3,732 country-year observations from 196 countries. On the common test set of 537 observations, linear regression has the lowest RMSE (4.300), followed by random forest (4.353), XGBoost (4.454), and naive persistence (4.582). Thus, the final run supports a modest predictive improvement from multivariate regression over persistence, but not a general machine-learning advantage. The attempted ARIMA and Holt/ETS benchmarks produced zero valid test observations in the supplied rerun and are therefore not used for substantive model-ranking claims. Weight sensitivity shows that the index is highly correlated with the baseline specification, but country-level changes are non-negligible and predictive performance varies across weighting schemes. For Kazakhstan, long-horizon analysis is presented as scenario extrapolation rather than a point forecast, with 2049 scenario values of 96.72 for a recent-trend linear path, 77.35 for a full-history logistic path, and 77.70 for an accelerated pre-2018 logistic calibration. These values are relative index positions, not percentages of connected people or measures of digital-system scale. The study does not estimate cybersecurity-governance capacity or a causal connectivity-cyber-risk relationship; connectivity trajectories are interpreted only as planning inputs for stress-testing digital-service and cybersecurity requirements.

Index Terms—digital divide, digital inequality, connectivity index, machine learning, time-series forecasting, Kazakhstan, Central Asia

I. INTRODUCTION

Digital connectivity is an important condition for participation in contemporary social, economic, educational, and public-service systems. However, a single penetration statistic does not capture the full access environment. Internet use, mobile subscriptions, fixed broadband, and electricity access represent different aspects of the conditions under which digital participation can occur. This study therefore forecasts a composite Connectivity Index rather than treating Internet penetration alone as a complete measure of the digital divide.

The paper addresses four empirical questions. First, how accurately can short-term Connectivity Index values be predicted using persistence and multivariate socioeconomic fea-

tures? Second, do multivariate models improve on a simple persistence benchmark? Third, how sensitive are short-term results and focal-country trajectories to the weights used to construct the composite index? Fourth, how should long-horizon trajectories for Kazakhstan be interpreted when recent observations may represent either structural saturation or a temporary plateau?

The study also clarifies the boundary between connectivity forecasting and cybersecurity governance. A connectivity trajectory can identify a changing access environment, but the Connectivity Index does not measure the number of digital systems, transactions, dependencies, cyber incidents, institutional coordination mechanisms, or incident-response capacity. The present study therefore does not claim that connectivity growth causes cyber risk or that governance capacity is insufficient. Cybersecurity governance is treated as an application context for scenario planning, not as an outcome estimated by the forecasting model.

The contribution is consequently fourfold: (1) construction of a multidimensional, explicitly relative Connectivity Index; (2) evaluation of persistence and multivariate forecasting models on a common chronological test set; (3) sensitivity analysis of alternative index weights, including focal-country trajectory changes; and (4) scenario-based long-horizon analysis that makes the uncertainty of Kazakhstan’s recent plateau explicit.

II. LITERATURE REVIEW

A. Digital divide as a multidimensional concept

Digital-divide research distinguishes physical access from differences in skills, quality and patterns of use, and unequal outcomes from digital participation [1], [2]. The present Connectivity Index primarily captures first-level access and infrastructure prerequisites. It must therefore not be interpreted as a direct measure of digital skills, quality of use, or downstream educational, economic, civic, or welfare outcomes. Secondary-school enrollment, GDP per capita, and urbanization are contextual predictors rather than direct measures of these later levels of inequality.

B. Composite measurement

Composite indicators are useful when a construct cannot be represented adequately by one variable. The Digital Divide Index literature illustrates the value of multidimensional policy measurement [3]. Composite construction nevertheless requires decisions about normalization, weighting, and missing data. The present index uses within-year min-max normalization. Its value therefore represents a country’s relative position

among reporting countries in a given year; it is not an absolute percentage of connectivity.

C. Technology diffusion and long-horizon extrapolation

Logistic diffusion models represent an S-shaped process in which adoption accelerates, reaches an inflection region, and approaches a ceiling [4]. In this study, the logistic form is not assumed to be a law of digital development. It is used only as one conditional scenario generator. Long-horizon results are therefore presented as calibration scenarios rather than expected outcomes.

D. Cybersecurity governance as a separate construct

Cybersecurity governance emphasizes institutional responsibilities, coordination, incentives, risk assessment, and alignment between technical systems and institutional arrangements [5], [6], [7]. These concepts provide a planning context for connectivity trajectories but do not justify inferring governance capacity from connectivity alone. A genuine connectivity-governance analysis would require independent governance indicators and outcomes.

III. METHODS

A. Data and panel construction

The panel uses World Bank World Development Indicators: individuals using the Internet (% of population), mobile subscriptions per 100 people, fixed broadband subscriptions per 100 people, electricity access (% of population), GDP per capita, secondary-school enrollment, and urban population share. Aggregate entities are removed. A year is retained for index construction only if each of the four index components has at least 100 reporting countries.

The final rerun retained index years 2002-2024. The excluded years were 1990-2001 and 2025. The 2025 exclusion is particularly relevant because electricity-access coverage ended in 2024 in the retrieved data. The reporting threshold is a reproducibility and data-quality rule intended to prevent a small number of early reporters from determining annual min-max normalization.

After construction of one- and two-year target lags and removal of rows with missing required predictors, the modeling sample contains 3,732 country-year observations from 196 countries. The chronological split is training through 2017, validation from 2018 through 2021, and test after 2021. The lagged modeling period is therefore 2004-2024.

B. Composite Connectivity Index

The baseline index is a weighted average of four within-year normalized components:

$$CI_{c,t} = 0.40I_{c,t} + 0.20M_{c,t} + 0.20B_{c,t} + 0.20E_{c,t}, \quad (1)$$

where I , M , B , and E denote normalized Internet use, mobile subscriptions, fixed broadband subscriptions, and electricity access, respectively.

For each component x , within-year normalization is

$$N_{c,t}(x) = 100 \frac{x_{c,t} - \min(x_t)}{\max(x_t) - \min(x_t)}. \quad (2)$$

Mobile subscriptions are capped at 150 subscriptions per 100 people before normalization to limit the influence of multiple-SIM behavior. When a component is missing, the available component weights are renormalized rather than assigning the missing component a value of zero.

The index is explicitly relative. Future values are not recomputed using unknown future cross-country minima and maxima. Instead, the historical normalized index is the forecasting target and long-horizon models extrapolate that historical relative-position series.

C. Weight sensitivity

Four specifications are evaluated:

- baseline: 0.40/0.20/0.20/0.20;
- equal: 0.25/0.25/0.25/0.25;
- Internet-dominant: 0.50/0.1667/0.1667/0.1667;
- infrastructure-balanced: 0.30/0.15/0.25/0.30.

Sensitivity is assessed using the correlation with the baseline index, linear-model test error, and mean absolute changes in the country-level index trajectory for Kazakhstan and Uzbekistan. The country-level comparison is important because a high aggregate correlation can conceal meaningful changes in an individual trajectory.

D. Features and temporal separation

Short-term models use GDP per capita, electricity access, mobile subscriptions, secondary-school enrollment, urban population, and one- and two-year lags of the Connectivity Index. Missing socioeconomic predictors are forward-filled within country only; no backward fill or interpolation is used. The notebook also records the filled observations for an imputation audit.

The target is persistent, so lagged predictors should not be interpreted as independent causal drivers. Their role is predictive: they summarize recent trajectory information available at the forecast origin.

E. Forecasting models

The final common-test comparison contains four models:

- 1) naive persistence, using the previous observed Connectivity Index;
- 2) linear regression;
- 3) random forest regression;
- 4) XGBoost regression.

The train, validation, and test partitions are chronological. Hyperparameter/model selection is performed without using future test observations.

The notebook also attempts ARIMA and Holt/exponential smoothing benchmarks with validation-based selection. In the final supplied run, however, both methods have zero valid test observations. They are therefore retained as methodological extensions in the code but are excluded from substantive claims about comparative predictive performance.

TABLE I
FINAL COMMON-TEST FORECASTING RESULTS.

Model	N	RMSE	MAE
Naive persistence	537	4.582	2.371
Linear regression	537	4.300	2.234
Random forest	537	4.353	2.468
XGBoost	537	4.454	2.523
ARIMA	0	-	-
Holt/ETS	0	-	-

F. Long-horizon scenarios

A single 25-year point forecast is inappropriate when a short recent plateau may be mistaken for a permanent ceiling. Kazakhstan is therefore evaluated under three calibration scenarios. The conservative scenario extrapolates the recent 2018-2024 pattern with a bounded linear trend. The baseline scenario fits a logistic curve to the full historical trajectory. The accelerated scenario fits a logistic curve to the pre-stabilization period through 2017.

These are calibration scenarios, not probabilities or policy forecasts. A logistic fit does not establish that the true future process is S-shaped.

IV. RESULTS

A. Short-term predictive performance

Table I reports the final common-test results. Linear regression has the lowest RMSE and MAE among the four models. Relative to naive persistence, its RMSE decreases from 4.582 to 4.300, a reduction of approximately 6.1%. Random forest is slightly worse than linear regression, while XGBoost does not improve on either of them.

The result does not support a blanket claim that more complex machine-learning models outperform simple benchmarks. Instead, the final run supports a narrower conclusion: multivariate linear regression provides a modest improvement over persistence for this index and sample, whereas the nonlinear ensemble models do not improve on it.

B. Horizon-specific persistence

The horizon diagnostics show increasing difficulty as the persistence horizon extends. For the feasible one-year origins, RMSE ranges from 2.922 in 2021 to 5.529 in 2022 and 3.822 in 2023. At the two-year horizon, RMSE is 6.642 for the 2021 origin and 6.922 for the 2022 origin. The three-year diagnostic from the 2021 origin has RMSE 8.110.

These values reinforce that a lag-1 persistence result should not be described as a general three-year forecasting result. Forecast uncertainty grows with horizon even for the simplest benchmark.

C. Weight sensitivity

The final weight sensitivity results are shown in Table II. All alternative indices remain highly correlated with the baseline, but predictive performance is not invariant to the weighting choice.

The Internet-dominant specification produces the lowest linear-model RMSE (3.791), lower than the baseline 4.300. Equal weighting produces the highest RMSE (5.465), while the infrastructure-balanced specification produces RMSE 4.772. Thus, the baseline weighting is not empirically unique. At the same time, the high correlations with the baseline indicate that the broad ordering and shape of the index are relatively preserved.

Country-level sensitivity is more visible than the aggregate correlations alone suggest. Kazakhstan's mean absolute index change is 2.953 under equal weights, 1.969 under Internet-dominant weights, and 2.390 under infrastructure-balanced weights. The corresponding Uzbekistan values are 2.844, 1.896, and 4.743. The latter result demonstrates why focal-country trajectory checks are necessary: a high global correlation does not imply negligible country-specific changes.

D. Feature importance and model interpretation

The final pipeline evaluates standard XGBoost importance, permutation importance, predictor correlations, variance inflation factors, and lag ablation. These diagnostics are used to distinguish predictive persistence from claims about substantive causal importance. In particular, lag variables should be interpreted as information about the recent state of the index rather than as independent causal determinants of connectivity.

Because the final notebook summary does not print numerical values for the permutation/VIF/ablation tables, this manuscript does not report unsupported numerical importance percentages. The corresponding CSV outputs should be used if a numerical feature-importance table is required in a subsequent revision.

E. Long-horizon Kazakhstan scenarios

The final Kazakhstan scenario endpoints for 2049 are:

- conservative recent linear: 96.72;
- baseline full-history logistic: 77.35;
- accelerated pre-stabilization logistic: 77.70.

The spread is substantively important. The conservative scenario rises substantially above the two logistic scenarios, showing that the interpretation of the 2018-2024 plateau has a large effect on long-run extrapolation. The two logistic calibrations are much closer to one another, suggesting that the historical diffusion structure places the modeled relative position near the high-70s under those assumptions.

None of these values is a percentage of connected people. A value of 77.35 means a modeled position on the relative 0-100 Connectivity Index. It is not an estimate of a physical saturation ceiling, digital-system scale, or cybersecurity capacity.

V. DISCUSSION

The final rerun provides a more conservative empirical conclusion than the earlier version of the study. The Connectivity Index is sufficiently persistent that naive persistence is a strong benchmark, but multivariate linear regression improves its common-test RMSE from 4.582 to 4.300. The improvement is modest, and the random forest and XGBoost models do not

TABLE II
CONNECTIVITY INDEX WEIGHT SENSITIVITY IN THE FINAL RERUN.

Scenario	w_I	w_M	w_B	w_E	Corr.	RMSE	MAE	KZ mean $ \Delta CI $
Baseline	.40	.20	.20	.20	1.0000	4.300	2.234	0.000
Equal	.25	.25	.25	.25	.9953	5.465	2.782	2.953
Internet-dominant	.50	.1667	.1667	.1667	.9984	3.791	2.002	1.969
Infrastructure-balanced	.30	.15	.25	.30	.9936	4.772	2.371	2.390

improve on linear regression. The appropriate interpretation is therefore not that machine learning solves Connectivity Index forecasting, but that a small amount of multivariate information can improve on pure persistence in this sample.

The horizon-specific diagnostics reinforce this interpretation. One-year persistence errors can be moderate, while two- and three-year persistence errors become substantially larger. Consequently, model performance should always be described with its forecast horizon and test construction.

The weight sensitivity analysis shows both robustness and non-uniqueness. The alternative indices remain highly correlated with the baseline, but their predictive errors and focal-country trajectories differ. In particular, the Internet-dominant specification performs better in the final linear-model sensitivity exercise, whereas equal and infrastructure-balanced weighting perform worse. This means the baseline 0.40 Internet weight should be presented as a theory-informed specification, not as an empirically identified optimum.

The index is also intrinsically relative. If all countries improve simultaneously, a country's index can remain stable because the cross-country distribution changes. Conversely, a country can change its index because other countries' values change. The Connectivity Index should therefore be interpreted as a comparative trajectory rather than a direct measure of the physical scale of digital systems.

The digital-divide interpretation is similarly bounded. The index primarily captures first-level access and infrastructure prerequisites. It cannot demonstrate whether people possess the skills to use digital technologies effectively or whether connectivity produces equal economic, educational, social, or civic benefits. The contextual predictors used for forecasting do not transform the index into a direct measure of second- or third-level inequality.

The cybersecurity-governance implication is consequently a planning application rather than a dependent variable. A higher-connectivity scenario can be used to stress-test whether incident-response procedures, interagency coordination, regulatory coverage, authentication and security requirements, and staffing arrangements can scale with expanding digital services. A plateau scenario can instead direct attention toward persistent access barriers. These are planning applications, not estimated governance effects.

VI. LIMITATIONS

First, the index uses theory-informed rather than empirically estimated weights. The sensitivity results demonstrate that

weighting can affect predictive performance and focal-country trajectories.

Second, within-year min-max normalization makes the index relative to the reporting-country distribution. It therefore cannot be interpreted as an absolute connectivity percentage.

Third, the 100-country reporting threshold excludes thin-coverage years and can affect historical trajectories. The final run explicitly excludes 1990-2001 and 2025.

Fourth, forward filling of socioeconomic predictors is past-only, but filled values can still influence model predictions. The rerun therefore records an imputation audit.

Fifth, country-year observations are not independent. The model-comparison significance procedure uses country-cluster resampling rather than treating every country-year row as an independent cluster.

Sixth, the attempted ARIMA and Holt/ETS models have zero valid test observations in the final supplied run. They should therefore not be presented as successful competing forecasts. A future rerun should diagnose the coverage/selection failure before making claims about univariate time-series superiority or inferiority.

Finally, the long horizon is intrinsically uncertain. Structural shocks, policy changes, infrastructure investment, geopolitical events, and revisions to international statistics can invalidate historical extrapolation. The Kazakhstan 2018-2024 stabilization is too short to establish a permanent saturation ceiling.

VII. CONCLUSION

This study develops a composite Connectivity Index and evaluates its short-term predictability while explicitly separating connectivity measurement from broader claims about digital inequality and cybersecurity governance. The final rerun shows that linear regression achieves the lowest common-test error among the evaluated persistence and multivariate models, with RMSE 4.300 compared with 4.582 for naive persistence. The improvement is modest, and nonlinear ensemble models do not outperform linear regression.

The weight analysis shows that the index is highly correlated across specifications but that weighting still matters for predictive performance and country-level trajectories. The baseline 0.40 Internet weight should therefore be understood as a transparent theory-informed specification rather than an empirically established optimum.

For long-horizon analysis, the study replaces a single expected 2049 value with scenarios. Kazakhstan's 2049 values range from 77.35 under the full-history logistic calibration to 96.72 under the recent-trend linear scenario. This spread

demonstrates why the recent plateau should not be interpreted as a confirmed saturation ceiling.

Finally, connectivity forecasts can support preparedness and stress-testing, but they do not measure cybersecurity governance capacity or establish a causal relationship between connectivity and cyber risk. Future integrated research should combine connectivity trajectories with independent measures of institutional coordination, cybersecurity staffing, incident-response capacity, regulatory implementation, and documented cyber outcomes.

VIII. CODE AND DATA AVAILABILITY

The code used to construct the Connectivity Index, perform the forecasting analysis, conduct sensitivity checks, and generate the reported tables and figures will be made available in a public GitHub repository (<https://github.com/lianaphoenix/digital-divide-forecasting>) at the time of publication. The analysis uses publicly available World Bank World Development Indicators; the repository will contain the executable notebook and scripts required to reproduce the analysis, together with documentation of the computational environment and generated intermediate outputs.

REFERENCES

- [1] M. Ragnedda and G. W. Muschert, *The Digital Divide: The Internet and Social Inequality in International Perspective*. Taylor & Francis, 2013.
- [2] E. Hustad, J. L. Hansen, A. Skaiaa, and P. Vassilakopoulou, "Digital inequalities: A review of contributing factors and measures for crossing the divide," in *Lecture Notes in Computer Science*, pp. 505-519, 2019.
- [3] S. Park and G. Kim, "Lessons from South Korea's digital divide index (DDI)," *Information*, vol. 16, no. 3, pp. 72-84, 2014.
- [4] J. Kaminski, "Diffusion of innovation theory," *Canadian Journal of Nursing Informatics*, vol. 6, no. 2, 2011.
- [5] K. R. Agbodo-Falschau and B. H. Ravaonoroahanta, "Investigating the influence of governance determinants on reporting cybersecurity incidents to police," *Technology in Society*, vol. 74, 102309, 2023.
- [6] H. Zhang, Z. Tang, and K. Jayakar, "A socio-technical analysis of China's cybersecurity policy," *Telecommunications Policy*, vol. 42, no. 5, pp. 409-420, 2018.
- [7] M. Kamariotou and F. Kitsios, "Information systems strategy and security policy: A conceptual framework," *Electronics*, vol. 12, no. 2, 382, 2023.
- [8] World Bank, "World Development Indicators," World Bank Open Data.