https://doi.org/10.31489/2025Ec4/1200

JEL: C53, I21, J11
ORIGINAL RESEARCH
Received: 19.06.2025 | Accepted: 07.10.2025

Comparative Analysis of Forecasting Models for Student Enrollment in Kazakhstan's General Secondary Education System

Duman Tleuzhan[1], **Alexandr Pilipenko**[2], **Kuandyk Tleuzhanuly**[3]*

Abstract

We compared seven forecasting models to predict student enrollment in Kazakhstan's schools using data from 2020-2024. We tested cohort component models, cohort survival models, trend regression with demographic factors, linear trend models, exponential smoothing, multi-factor regression, and weighted moving averages across 20 regions (17 regions and 3 cities) with about 3.9 million students. We measured how accurate each model was using Mean Absolute Percentage Error (MAPE). We trained models on 2020-2023 data and tested them on 2024 numbers. The linear trend model worked best, with 0.70 % MAPE nationally and 0.77 % MAPE across regions. Demographic models didn't work as well—cohort models did poorly at the regional level even though they have good theory behind them. Our forecasts for 2025-2027 show national enrollment growing from 3.9 million to 4.2 million students, but growth varies a lot by region. Big cities like Astana will grow 24.05 % and Almaty 12.81 %, while some regions will barely grow at all. Our results help educational planners pick the right forecasting methods for different areas. This study fills a research gap for post-Soviet countries where detailed forecasting evaluations are hard to find.

Keywords: student enrollment forecasting, educational planning, demographic modeling, time series analysis, Kazakhstan education system, regional analysis.

Introduction

Accurate forecasting of student enrollment is one of the basic challenges facing educational systems worldwide. Educational planners need reliable projections to make sure that infrastructure, teaching staff, and financial resources can meet future demand. This challenge is especially difficult in developing countries where demographic transitions, internal migration, and rapid urbanization create complex planning environments. Kazakhstan's educational system shows these complexities well. The country currently serves about 3.9 million students across 21 major administrative units in a vast and diverse territory. Recent demographic trends have created different pressures: birth rates are declining in some regions while major urban centers see population growth. Meanwhile, large-scale internal migration from rural to urban areas—particularly toward cities like Astana, Almaty, and Shymkent—makes enrollment projections even harder.

The stakes of accurate enrollment forecasting go well beyond simple planning exercises. Enrollment projections directly affect decisions about teacher recruitment and deployment, school construction and renovation programs, budget allocations, and broader educational policy. When forecasts are wrong, the consequences can be serious. Underestimating future enrollment leads to overcrowding and lower educational quality, while overestimating wastes public funds and creates inefficient resource allocation. Despite how important enrollment forecasting is, surprisingly little research has carefully evaluated different forecasting methods using complete data from Kazakhstan's educational system. Most existing studies either focus narrowly on single methods or lack enough historical data to properly test model performance. Moreover, Kazakhstan's unique characteristics—its linguistic diversity, varied settlement patterns, and recent administrative reforms—require specialized analytical approaches that may not transfer directly from other contexts.

This study aims to compare the performance of seven different forecasting models for predicting total student enrollment across Kazakhstan's regional educational system. We work with a complete dataset covering 21 territorial units over five years from 2020 to 2024, including demographic indicators, enrollment statistics, migration patterns, and educational infrastructure characteristics. Our specific goals are fourfold: first, we check whether demographic-based models or statistical time series approaches give more accurate enrollment forecasts; second, we identify which methods work best at different territorial levels from national to regional scales; third, we generate validated forecasts for 2025-2027 to support practical educational

[1] Narxoz University, Almaty, Kazakhstan, duman.tleuzhan@narxoz.kz

[2] Taldau, National Centre for Education Research and Evaluation named after A. Baitursynuly, Astana, Kazakhstan, a.pilipenko@taldau.edu.kz

[3] Narxoz University, Almaty, Kazakhstan, kuandyk.tleuzhanuly@narxoz.kz (corresponding author)

planning; and fourth, we analyze regional variations in enrollment patterns and consider what they mean for resource allocation strategies.

This research contributes through systematic application of multiple forecasting methods to a complete regional dataset, offering evidence-based guidance for educational planners and policymakers. The study addresses a significant research gap by providing real-world validation of model performance within Kazakhstan's distinctive educational and demographic context.

Literature Review

School enrollment forecasting has become essential for educational planning worldwide. It helps policymakers anticipate infrastructure needs, allocate resources efficiently, and ensure access to education. Forecasting methods have evolved with technological advances, but developing countries have shown creativity in adapting these methods to environments where data is limited. Despite growing research in this field, certain regions—particularly post-Soviet countries—remain underrepresented in the literature, pointing toward significant opportunities for investigation.

Time Series Models: Foundation of Enrollment Forecasting

Time series models are the foundational approach to enrollment forecasting, offering solid methods for capturing temporal patterns in educational data. Tang and Yin (2012) established exponential smoothing's relevance in the United States by comparing it with grey prediction models for forecasting education expenditure and enrollment. They found that grey models showed higher accuracy, though exponential smoothing remained valuable as a baseline. Chen (2022) found different results in China, where exponential smoothing actually outperformed grey prediction, ARIMA, and neural network models in forecasting enrollment proportions. These contrasting results show how model performance varies significantly across different national contexts.

The ARIMA methodology has proven versatile in addressing forecasting challenges across both developed and developing nations. Marinoiu (2014) successfully applied the Box-Jenkins methodology to develop an ARIMA model for forecasting gross enrollment ratio in Romanian primary schools, projecting concerning declines. Kornelio et al. (2024) used ARIMA models in Tanzania to project significant enrollment increases in government primary schools. Chen et al. (2021) extended to ARIMAX models incorporating exogenous variables to analyze student-teacher ratios in China's primary education system. Yan (2024) applied Vector Autoregression models to analyze the decline of rural primary schools in China, revealing how urbanization patterns impact educational infrastructure. The comparative effectiveness of these time series approaches shows a crucial finding: no single method dominates across all contexts, making careful consideration of local conditions essential.

Machine Learning and Artificial Intelligence Approaches

The integration of machine learning techniques into enrollment forecasting represents a shift from traditional statistical methods, offering better capabilities for capturing complex, non-linear patterns. James (2021) pioneered the use of Long Short-Term Memory networks at Kansas State University, finding that these deep learning models significantly outperformed both traditional exponential smoothing and standard deep neural networks. Support Vector Machines have emerged as powerful tools when multiple influencing factors must be considered simultaneously. Aksenova et al. (2006) demonstrated this at California State University, Sacramento, where SVM models incorporating demographic variables, income levels, tuition fees, and unemployment rates achieved remarkable accuracy.

Hybrid intelligent systems mark the current frontier in machine learning applications. The Adaptive Neuro-Fuzzy Inference System, as applied by Aji et al. (2023) in Indonesia, combines the learning capabilities of neural networks with the interpretability of fuzzy logic systems. Shafii et al. (2021) applied fuzzy time series models in Malaysia, achieving the highest accuracy for primary school enrollment. The democratization of machine learning tools has also facilitated innovative applications in resource-constrained environments. Sahane et al. (2014) used data mining techniques with accessible tools in India's Aurangabad district, demonstrating that sophisticated analytical capabilities can be deployed even in contexts with limited technical infrastructure.

Spatial and Geographic Information Systems Models

The incorporation of spatial dimensions into enrollment forecasting acknowledges the fundamentally geographic nature of educational access and demand. Geographic Information Systems have changed the field by enabling planners to visualize and analyze the spatial distribution of student populations alongside school infrastructure. Langley's (1997) pioneering work in Leeds, United Kingdom, established the foundation for spatial modeling in educational planning. Haynes (2014) applied spatial Bayesian modeling to enhance small-area enrollment projections in Iowa, improving the accuracy of grade progression rates while accounting for geographic dependencies and migration patterns.

Miller (2008) implemented the Integrated Planning for School and Community model in Wake County, North Carolina, demonstrating how GIS-based approaches can align educational infrastructure development with urban planning. The Greater London Authority's (GLA Intelligence, 2018) school place demand projections illustrated how spatial modeling can inform strategic infrastructure decisions in complex urban environments. Wang et al. (2023) extended spatial methodologies for China's new urban districts by incorporating micro-level factors such as residential location and dwelling type, addressing overcrowding through optimized space utilization.

Demographic and Cohort-Based Approaches

Cohort-survival methods represent the traditional backbone of enrollment forecasting, particularly in contexts with stable demographic patterns. Braden et al. (1972) established the cohort-survival technique as the gold standard for communities with predictable population dynamics. Pajankar and Srivastava's (2019) Reconstructive Cohort Approach for India demonstrates how traditional cohort concepts can be modified for contexts where detailed enrollment data is unavailable, requiring only population figures, repetition rates, and transition rates.

Fabricant and Weinman (1972) applied least squares regression to forecast first-grade enrollment in New York neighborhoods, incorporating variables such as new housing developments, busing policies, and ethnic composition. Grip and Grip's (2019) comparative analysis of confidence intervals and stochastic forecasting using Monte Carlo simulations for New Jersey school districts represents methodological advancement, emphasizing customized approaches based on district size and growth patterns.

System Dynamics and Integrated Modeling Approaches

System dynamics modeling offers a holistic perspective on enrollment forecasting by capturing the complex feedback loops and policy interactions that influence educational participation. Pedamallu et al. (2010) applied system dynamics with cross-impact analysis in developing country contexts, revealing how infrastructural improvements create reinforcing cycles that boost enrollment while reducing dropout rates. The comparative studies by Rynerson and colleagues (2018, 2021, 2022) for various Oregon school districts exemplify integrated approaches, combining demographic analysis, cohort progression, and scenario planning to provide 15-year enrollment forecasts.

National and Comparative Perspectives

The scale of forecasting efforts ranges from local district projections to national policy planning. Hussar and Bailey's (2016) national and state-level enrollment projections for the United States through 2024 illustrate how macro-level forecasting informs federal educational policies and funding allocations. Huynh Van et al. (2019) analyzed Vietnam's six geographical regions and revealed significant disparities in enrollment trends. Chen's (2022) work on China's general and vocational education enrollment ratios highlighted how forecasting must account for regional economic structures and labor market needs.

The evolution of school enrollment forecasting methodologies reflects both technological advancement and adaptive innovation across diverse global contexts. Several key insights emerge from this review. First, no single forecasting method dominates across all contexts; method selection must carefully consider local data availability, demographic patterns, and institutional characteristics. Second, the trend toward hybrid and integrated approaches suggests that future advances may come from creative combinations of existing techniques. Third, the incorporation of spatial dimensions and system dynamics perspectives enriches traditional temporal forecasting.

However, significant gaps remain in the literature. The conspicuous absence of studies from post-Soviet regions and many developing countries suggests that current methodological knowledge may not fully capture the diversity of global educational contexts. Kazakhstan and other Central Asian countries, with their unique demographic transitions and educational system transformations, represent particularly important yet understudied contexts.

Methodology

Data Description

This study utilizes a comprehensive dataset covering Kazakhstan's general secondary education system using the National Education Database (NEDB) for the years 2020-2024. The dataset encompasses 21 territorial units including the national level, 17 regional administrations, and 3 cities of republican significance

(Astana, Almaty, and Shymkent). The primary dataset integrates multiple data sources including demographic indicators, educational enrollment statistics, migration flows, school infrastructure characteristics, and human resource metrics. Demographic variables include birth rates from 2013-2024, enabling cohort tracking with appropriate time lags for school entry age. Educational variables encompass total student enrollment by year, grade-level distributions, and language of instruction categories. Migration data captures both incoming and outgoing student flows between regions, providing insights into population mobility patterns affecting enrollment. Infrastructure variables include school capacity, number of schools, shift arrangements, and facility conditions. Human resource indicators cover teacher numbers, qualifications, and turnover rates. All data sources maintain consistent territorial classifications and temporal coverage, ensuring compatibility across different analytical approaches. Missing values were addressed through interpolation techniques where appropriate, with sensitivity analysis confirming minimal impact on model outcomes.

Forecasting Models

We tested seven different forecasting methods to look at enrollment dynamics from different angles. Each model represents a specific way of predicting enrollment, from demographic-based methods to statistical time series techniques.

Cohort Component Model

The cohort component model is one of the more sophisticated approaches to educational enrollment forecasting. It comes directly from demographic methodology used in population projections. This model works on the idea that future first-grade enrollment can be predicted by tracking specific birth cohorts as they move toward school age. The method starts with historical birth data, usually requiring a six-year lag for standard school entry age. The basic formula is:

$$E_t = B_{t-6} \times S_t \times P_t \times M_t$$

where E_t is enrollment at time t, B_{t-6} is births six years earlier, S_t is the enrollment rate from birth to school age, P_t represents the participation rate in formal education, and M_t captures net migration effects. The model includes survival rates from birth to school age, and participation rates that show what proportion of eligible children actually enter formal education. Migration effects are included through net migration coefficients. For later grades, the cohort progression follows:

$$E_{g,t} = E_{g-1,t-1} \times R_{g,t}$$

where $E_{g,t}$ is enrollment in grade g at time t, and $R_{g,t}$ is the grade progression rate accounting for retention, dropout, and migration. This approach works well in stable demographic settings where birth patterns, migration flows, and educational participation rates follow predictable paths over time.

Cohort Survival Model

The cohort survival model builds on the cohort component framework by adding more sophisticated survival analysis techniques borrowed from actuarial science and epidemiology. Unlike the basic cohort model that uses simple survival rates, this approach recognizes that survival probabilities can vary across different demographic groups, geographic regions, and socioeconomic levels. The model extends the basic formula to include differential survival rates:

$$E_{g,t} = E_{g-1,t-1} \times S_{g,i,t} \times C_{g,i,t}$$

where $S_{g,i,t}$ is differential survival rates for demographic subgroup i in grade g at time t, and $C_{g,i,t}$ denotes continuation probabilities. The model includes differential survival rates that can account for variations in healthcare access, economic conditions, and other factors that influence child survival from birth to school entry age. The cohort survival model adds the concept of educational continuation probabilities, recognizing that not all children who survive to school age will participate in formal education. These continuation probabilities can be modeled as functions of various socioeconomic indicators:

$$C_{g,i,t} = f(X_{i,t})$$

where $X_{i,t}$ is a set of socioeconomic predictors including family characteristics, regional educational infrastructure quality, and economic conditions. This enhanced approach gives more accurate projections where demographic and socioeconomic differences significantly affect educational participation patterns.

Trend Regression with Demographic Factors

Trend regression models with demographic factors combine statistical trend analysis with demographic change theory. These models recognize that enrollment patterns often show both systematic trends over time and responses to underlying demographic drivers. The general form is:

$$E_t = \beta_0 + \beta_1 t + \beta_2 B_{t-k} + \beta_3 N_t + \beta_4 D_t + \epsilon_t$$

where E_t is enrollment at time t, t is the time trend, B_{t-k} is lagged birth rates, N_t captures net migration, D_t represents population density or other demographic variables, and ϵ_t is the error term. The statistical part involves fitting regression equations to historical enrollment data, using time as the main independent variable to capture long-term trends in educational participation. The demographic part adds predictor variables like birth rates, migration flows, age structure changes, and population density variations. More advanced versions may include interaction effects:

$$E_t = \beta_0 + \beta_1 t + \beta_2 B_{t-k} + \beta_3 (t \times B_{t-k}) + \beta_4 N_t + \epsilon_t$$

allowing for non-linear relationships and changing coefficients over time. The methodology often uses multiple regression techniques with various forms including polynomial trends, logarithmic transformations, and piecewise linear functions to capture complex enrollment dynamics. Model specification usually involves testing for autocorrelation using the Durbin-Watson statistic, heteroscedasticity through Breusch-Pagan tests, and structural breaks using Chow tests. These models work well where enrollment patterns are influenced by both long-term demographic transitions and shorter-term policy or economic changes.

Linear Trend Model

The linear trend model, despite its simplicity, is a solid and widely-used forecasting technique that often beats more complex alternatives in practical use. This approach assumes that enrollment changes follow a consistent linear pattern over time, with a constant absolute change per time period. The math involves estimating a simple linear regression equation:

$$E_t = \alpha + \beta t + \epsilon_t$$

where E_t is enrollment at time t, α is the intercept representing baseline enrollment, β trend coefficient, showing average annual change, and ϵ_t is the random error term. Model estimation usually uses ordinary least squares regression, with parameter estimates given by:

$$\hat{\beta} = \frac{\sum_{t=1}^{n}(t - \bar{t})(E_t - \bar{E})}{\sum_{t=1}^{n}(t - \bar{t})^2}$$

$$\hat{\alpha} = \bar{E} - \widehat{\beta \bar{t}}$$

More sophisticated techniques like generalized least squares may be used when autocorrelation is present in the residuals. The forecasting process involves extending the fitted trend line into future periods:

$$\hat{E}_{t+h} = \hat{\alpha} + \hat{\beta}(t + h)$$

where h is the forecast horizon. Prediction intervals are calculated based on the standard error of the forecast:

$$PI_{t+h} = \hat{E}_{t+h} \pm z_{\alpha/2} \times SE(\hat{E}_{t+h})$$

where $SE\left(\hat{E}t + h\right) = s\sqrt{1 + \frac{1}{n} + \frac{(t+h-\bar{t})^2}{\sum t=1^n (t-\bar{t})^2}}$ and s is the residual standard error. While critics often dismiss linear models as too simplistic, research in forecasting accuracy often shows that linear trends give surprisingly accurate predictions for many educational systems, especially over short to medium-term horizons.

Exponential Smoothing (Holt's Method)

Exponential smoothing using Holt's method is a sophisticated time series forecasting technique that handles both level and trend components in enrollment data. Unlike simple exponential smoothing that assumes no systematic trend, Holt's method recognizes that many enrollment series show persistent upward or downward movements that must be explicitly modeled. The method uses two smoothing equations: a level equation that updates the estimated current enrollment level and a trend equation that updates the estimated rate of change. The level equation is:

$$L_t = \alpha E_t + (1 - \alpha)(L_{t-1} + T_{t-1})$$

and the trend equation is:

$$T_t = \beta(L_t - L_{t-1}) + (1 - \beta)T_{t-1}$$

where L_t is the level at time t, T_t is the trend at time t, α is the level smoothing parameter $(0 < \alpha < 1)$, and β is the trend smoothing parameter $(0 < \beta < 1)$. Parameter estimation often uses optimization techniques to minimize forecast errors over historical data, usually minimizing the sum of squared errors:

$$\min_{\alpha,\beta} \sum_{t=1}^{n} \left(E_t - \widehat{E_t}\right)^2$$

The forecasting process generates predictions by adding trend projections to the current level estimate:

$$\hat{E}_{t+h} = L_t + h \times T_t$$

where h is the forecast horizon. Advanced implementations may include damped trend modifications:

$$\hat{E}_{t+h} = L_t + (\phi + \phi^2 + \ldots + \phi^h)T_t$$

where ϕ is the damping parameter $(0 < \phi < 1)$, which assumes trend effects diminish over longer forecast horizons, addressing the common problem of exponential trend extrapolation producing unrealistic long-term projections.

Multi-Factor Regression Model

Multi-factor regression models are the most comprehensive statistical approach to enrollment forecasting, trying to capture the complex mix of demographic, economic, social, and institutional factors that influence educational participation. The general form is:

$$E_t = \beta_0 + \beta_1 B_{t-k} + \beta_2 P_t + \beta_3 I_t + \beta_4 C_t + \beta_5 T_t + \epsilon_t$$

where E_t is enrollment at time t, B_{t-k} represents lagged birth rates, P_t denotes population density or demographic structure, I_t captures income or socioeconomic indicators, C_t represents school capacity or infrastructure measures, T_t is teacher availability, and ϵ_t is the error term. These models usually include demographic variables like birth rates, age structure, and population density; socioeconomic indicators including income levels, parental education, and employment rates; infrastructure measures like school capacity and teacher availability; and policy variables reflecting changes in educational requirements or funding levels. Model specification requires attention to multicollinearity among predictor variables, checked through variance inflation factors:

$$VIF_j = \frac{1}{1 - R_j^2}$$

where R_j^2 is the coefficient of determination from regressing predictor j on all other predictors. Advanced versions may use stepwise regression for variable selection, ridge regression to address multicollinearity with penalty term:

$$\hat{\beta}ridge = arg\,min\,\beta\left[\sum_{t=1}^{n}(E_t - X_t\beta)^2 + \lambda\sum_{j=1}^{p}\beta_j^2\right]$$

or instrumental variables to handle endogeneity problems. The forecasting process requires projecting values for all predictor variables, which can add extra uncertainty.

Weighted Moving Average Model

Weighted moving average models provide a flexible smoothing approach that emphasizes recent observations while including historical patterns to reduce the influence of random fluctuations in enrollment data. Unlike simple moving averages that give equal weights to all observations within the averaging window, weighted moving averages use a declining weight structure that gives more influence to more recent data points. The basic form is:

$$\hat{E}t = \sum i = 0^{k-1}\, w_i E_{t-i}$$

subject to the constraint $\sum_{i=0}^{k-1} w_i = 1$ and typically $w_0 > w_1 > \ldots > w_{k-1}$. The weighting scheme typically follows exponential decay patterns:

$$w_i = \frac{(1-\lambda)\lambda^i}{1-\lambda^k}$$

where λ is a decay parameter ($0 < \lambda < 1$), or geometric progressions, or custom weight distributions designed to match the specific characteristics of the enrollment series. The methodology involves calculating weighted averages of recent enrollment observations and then projecting these smoothed values into future periods. Trend components can be incorporated by calculating the weighted average of first differences:

$$\hat{T}t = \sum i = 0^{k-2}\, w_i(E_{t-i} - E_{t-i-1})$$

and adding these trend estimates to the level projections:

$$\hat{E}_{t+h} = \hat{E}_t + h \times \hat{T}_t$$

The forecasting process often includes automatic weight optimization procedures that minimize historical forecast errors to determine optimal weight parameters, typically by minimizing mean squared error:

$$\min_{w_0,\ldots,w_{k-1}} \sum_{t=k}^{n}\left(E_t - \widehat{E_t}\right)^2$$

Advanced versions may use adaptive weighting schemes that adjust the weight distribution based on how stable or volatile recent enrollment patterns are.

Model Validation and Performance Assessment

We evaluated model performance using out-of-sample validation for accurate assessment. All models were trained only on 2020-2023 data, with 2024 as the testing period. This approach prevents overfitting and gives a realistic assessment of how models perform in actual forecasting situations.

We measured accuracy using Mean Absolute Percentage Error as the main metric. MAPE is easy to interpret because it shows forecast errors as percentages of actual values, making comparison across different areas and enrollment scales straightforward. The MAPE is calculated as:

$$MAPE = \frac{100\%}{n}\sum_{t=1}^{n} | \frac{E_t - \hat{E}_t}{E_t} |$$

where E_t is the actual enrollment at time t, $\hat{E}_t$is the forecasted enrollment, and $n$is the number of observations. We calculated MAPE both at the national level and as average regional performance to check model consistency across different geographic scales. The validation process included sensitivity analysis to test model stability under different parameter settings and data changes. We used cross-validation techniques where possible to make sure performance assessments were solid.

Statistical Implementation

All statistical analyses were conducted using Stata version 18. For the cohort component and cohort survival models, we employed the gen and replace commands to create lagged birth variables and calculate survival rates, with cohort progressions tracked using conditional statements and the bysort prefix for regional groupings. The trend regression with demographic factors utilized the regress command with robust standard errors (vce(robust)) to estimate coefficients, followed by post-estimation diagnostics including the estat vif command to assess multicollinearity, estat hettest for heteroscedasticity testing (Breusch-Pagan test), and estat dwatson for autocorrelation detection (Durbin-Watson statistic). Linear trend models were fitted using regress enrollment year with region-specific estimations executed through statsby loops. Exponential smoothing forecasts implementing Holt’s method were generated using the tssmooth exponential command with optimal smoothing parameters identified through the optimize() function, minimizing mean squared error. Multi-factor regression specifications employed stepwise regression procedures for variable selection, with variance inflation factors computed via vif to identify problematic multicollinearity (threshold VIF > 10), and, when necessary, ridge regression was implemented using the ridgereg command to address collinearity issues. Weighted moving average calculations utilized the ma() function within time series operators, with custom weight specifications created through mata programming for non-standard weighting schemes. Model performance evaluation was conducted using out-of-sample validation, where training-test splits were created using if conditions (2020-2023 for training, 2024 for testing), and Mean Absolute Percentage Error was calculated using egen functions combined with summarize commands. Forecasts for 2025-2027 were generated using the predict command with the dynamic() option for multi-step-ahead predictions, and confidence intervals were constructed using the forecast suite of commands with standard error estimates from predict, stdp. All data management, including merging demographic indicators with enrollment statistics, was performed using merge commands with m:1 relationships, and missing values were handled through linear interpolation using the ipolate command where appropriate.

Results

The comparative analysis shows substantial differences in how well various forecasting methods worked across different territorial levels. Table 1 presents MAPE results for all seven models we tested on total student enrollment.

Table 1. Model Performance for Total Student Enrollment Forecasting (MAPE %, 2024)

Model	National Level MAPE	Regional Average MAPE	Overall Rank
Linear Trend Model	0.70 %	0.77 %	1
Exponential Smoothing (Holt’s)	1.00 %	1.40 %	2
Multi-factor Regression	1.21 %	4.32 %	3
Trend Regression with Demographics	1.84 %	1.63 %	4
Weighted Moving Average	4.16 %	3.53 %	5
Cohort Survival Model	4.96 %	12.57 %	6
Cohort Component Model	7.92 %	18.28 %	7

Note — compiled by authors based on the sources (estimates based on National Education Database 2020-2024)

The linear trend model turned out to be remarkably accurate, posting the best scores both nationally (0.70 % MAPE) and across regions (0.77 % MAPE). What makes this particularly useful is how consistently well it performed regardless of whether we looked at the whole country or individual regions. The model’s straightforward nature combined with this level of accuracy makes it an obvious choice for actual planning work. Exponential smoothing came in second, doing quite well at the national level (1.00 % MAPE) and maintaining decent regional accuracy (1.40 % MAPE). This method’s ability to adapt to recent changes while staying stable explains why it performed reasonably well across the board. What really surprised us was how poorly the demographic models did compare to simpler statistical methods. The cohort component model, which has solid theoretical backing in population dynamics, only managed 7.92 % MAPE nationally and 18.28 % regionally. The cohort survival model wasn't much better at 4.96 % and 12.57 % respectively. This suggests that straightforward demographic relationships aren't capturing what's actually happening in Kazakhstan's education system. The multi-factor regression showed an interesting split—it did well nationally (1.21 % MAPE) but struggled with regional predictions (4.32 % MAPE), which hints at possible overfitting or trouble handling how different regions behave. Since the linear trend model worked so well, we used it to forecast enrollment for 2025-2027. Table 2 shows what we're projecting for different parts of the country.

Table 2. Regional Forecasts Using Linear Trend Model (2025-2027)

Region	Actual 2024	Predicted 2024	MAPE 2024 (%)	Prediction 2025	Prediction 2026	Prediction 2027	Predicted Average Annual Growth Rate 2025-2027 (%)	Predicted Total Growth Rate 2025-2027 (%)
Kazakhstan Total	3,904,496	3,931,649	0.70	4,010,283	4,116,070	4,221,856	2.64	8.13
Abay region	102,494	102,732	0.23	102,876	103,259	103,641	0.37	1.12
Akmola region	142,117	143,257	0.80	144,565	147,013	149,461	1.69	5.17
Aktobe region	185,947	187,020	0.58	192,094	198,241	204,387	3.20	9.92
Almaty region	368,688	371,026	0.63	383,977	399,267	414,556	3.99	12.44
Atyrau region	153,554	153,659	0.07	157,760	161,966	166,171	2.67	8.22
West-Kazakhstan region	126,034	127,571	1.22	129,380	132,725	136,071	2.59	7.96
Zhambyl region	247,648	248,835	0.48	248,972	250,296	251,620	0.53	1.60
Zhetisu region	133,352	133,391	0.03	134,658	135,963	137,269	0.97	2.94
Karaganda region	175,974	177,034	0.60	177,310	178,646	179,982	0.75	2.28
Kostanay region	115,480	116,432	0.82	116,382	117,284	118,185	0.77	2.34
Kyzylorda region	184,917	187,528	1.41	188,260	191,602	194,945	1.78	5.42
Mangystau region	190,643	191,426	0.41	198,277	205,911	213,545	3.85	12.01
Pavlodar region	118,413	119,975	1.32	119,640	120,866	122,093	1.03	3.11
North-Kazakhstan region	74,952	76,229	1.70	74,993	75,034	75,074	0.05	0.16
South-Kazakhstan region	523,618	528,495	0.93	529,168	534,717	540,267	1.05	3.18
Ulytau region	41,218	41,372	0.37	41,923	42,629	43,334	1.68	5.13
East-Kazakhstan region	101,805	102,435	0.62	102,997	104,188	105,380	1.16	3.51
Astana City	286,913	284,716	0.77	309,909	332,906	355,902	7.45	24.05
Almaty City	360,008	366,076	1.69	375,380	390,752	406,123	4.10	12.81
Shymkent City	270,721	272,442	0.64	281,765	292,809	303,853	3.92	12.24

Note — compiled by authors based on the sources (estimates based on National Education Database 2020-2024)

Nationally, we're looking at steady growth from 3.9 million students in 2024 to 4.2 million by 2027—an 8.13 % jump. This reflects ongoing demographic momentum and continuing educational system expansion. The really striking pattern is how unevenly this growth is distributed. Astana leads the pack with a projected 24.05 % increase over three years. This dramatic growth stems from continued urbanization and people moving to the capital for economic opportunities and government jobs. Almaty City is projected to grow 12.81 %, and Shymkent 12.24 %, showing how these major commercial centers keep attracting families. Among regular regions, Almaty region shows 12.44 % growth, Mangystau hits 12.01 %, and Aktobe reaches 9.92 %. These align pretty well with where economic development is concentrated and where resource industries are pulling in workers and their families. The flip side is what's happening in rural and peripheral areas. North-Kazakhstan region barely budges, with just 0.16 % growth over three years. Abay region shows 1.12 % growth, and Zhambyl comes in at 1.60 %. These modest numbers reflect the rural-to-urban migration wave and falling birth rates in these areas, creating entirely different planning headaches for local authorities.Looking at how accurate our 2024 forecasts were across regions, most came in under 1.5 % MAPE. A few outliers include West-Kazakhstan (1.22 %), Kyzylorda (1.41 %), Pavlodar (1.32 %), and Almaty City (1.69 %), though even these are acceptable for planning purposes. North-Kazakhstan region had the highest error at 1.70 %, possibly because of more erratic demographic shifts or data quality issues there.

Discussion

The fact that simple statistical models beat theoretically sophisticated demographic approaches has significant implications for how educational forecasting should work in Kazakhstan. The linear trend model’s accuracy tells us that enrollment patterns over the short term follow fairly predictable growth paths that don’t require complex demographic modeling to capture. This goes against what demographers usually recommend for education forecasting.

Why did demographic models perform so poorly? Several Kazakhstan-specific factors probably explain this. First, the link between birth rates and school enrollment becomes complicated when you have the kind of migration Kazakhstan’s been experiencing. The country has seen major internal population movements over the past decade—people leaving rural areas for cities, moving from smaller towns to Astana, Almaty, and Shymkent. Traditional cohort models assume populations remain stable geographically, which clearly doesn’t match Kazakhstan’s reality. Beyond that, educational participation rates and policy changes have likely disrupted old patterns. Recent reforms—changes to school entry age requirements, curriculum overhauls—could have broken historical relationships between birth cohorts and enrollment. Furthermore, five years of data just is not enough for demographic models to establish reliable parameters. These models typically need much longer time series to estimate survival and progression rates properly.

These findings fit with what forecasting research has found more broadly: simple models often beat complex ones in practice, especially when you don’t have much historical data. The principle here is that unnecessary complexity introduces new sources of error rather than improving accuracy. Our results support this principle specifically for educational forecasting in developing countries. The enrollment patterns we're projecting show significant regional gaps that need different policy responses. High-growth urban centers like Astana and Almaty face serious infrastructure challenges. That 24 % growth in Astana over three years means they must act quickly on school construction, teacher hiring, and budget allocations. Based on typical class sizes and teacher-student ratios, Astana probably needs 15-20 new schools and around 3,000 new teachers by 2027 just to maintain current standards. Almaty and Shymkent face similar pressures, though not quite as extreme.

Meanwhile, regions with flat enrollment like North-Kazakhstan have different problems concerning efficiency and quality. Maintaining the existing school network becomes harder to justify when student numbers aren’t growing. These places might need to think about consolidating schools—fewer institutions with better facilities, more specialized staff, stronger programs—rather than spreading resources thinly across many small schools. That kind of consolidation could actually improve quality while cutting costs.

The urban-rural split in growth patterns mirrors broader trends in Kazakhstan’s development. Educational planners must balance infrastructure investment between booming urban areas that need expansion and slower-growth rural areas that need efficiency improvements. This involves difficult trade-offs between equity (keeping rural schools open) and efficiency (concentrating resources where students actually are). Policymakers need better frameworks for making these allocation decisions that consider both immediate enrollment pressures and long-term regional development goals. Regional variations also affect teacher deployment. High-growth regions need many more teachers, creating recruitment and training challenges. Kazakhstan might need incentive programs to attract qualified teachers to move to rapidly growing areas—maybe housing assistance, salary increases, or faster career advancement. At the same time, slow-growth regions might have too many teachers, requiring retraining programs or helping people transition to other education sectors or regions.

Our analysis has several limitations pointing toward future research. First, the short time series (2020-2024) doesn’t capture long-term cycles or structural breaks in enrollment patterns. The data includes the COVID period, which may have temporarily distorted enrollment in ways that don’t reflect normal trends. Longer time series would allow us to validate models more effectively and understand how enrollment responds to major disruptions. Second, regional aggregation hides important local variations. Regions in Kazakhstan contain diverse territories with different urban-rural mixes, economic bases, and demographic profiles. District-level and school-level analysis would provide much more useful insights for local planning. Future work should examine forecasting at these finer scales, perhaps using spatial models to capture geographic dependencies. Third, we only examined total enrollment without breaking it down by grade. Age-specific forecasts by grade would help with curriculum planning, facility design, and teacher specialization. Different grades might grow at different rates because of varying cohort sizes and changing dropout patterns. Understanding these grade-specific dynamics would sharpen operational planning. Fourth, we did not explicitly model external factors like policy changes, economic shocks, or social shifts. Major policy initiatives or economic disruptions could change enrollment in ways historical trends miss. Future research could incorporate these through scenario analysis or econometric models linking enrollment to external drivers. The system dynamics approaches we discussed in the literature review offer robust frameworks for exploring these complex interactions. Fifth, we did not address quality or outcomes, only quantity. Educational planning ultimately is concerned with quality and results. Future work could link enrollment projections to the resources needed to maintain or improve quality, developing optimization models balancing enrollment accommodation with quality standards.

Conclusion

This study offers the first thorough evaluation of enrollment forecasting models using Kazakhstan's full regional educational dataset. The analysis shows that simple statistical approaches, especially linear trend models, outperform complex demographic models for forecasting total student enrollment in Kazakhstan. This has practical implications for educational planning in Kazakhstan and likely other post-Soviet countries facing similar demographic and institutional situations.

The linear trend model's performance—0.70 % national MAPE and 0.77 % regional MAPE—makes it the obvious choice for operational planning in Kazakhstan. Its simplicity means educational authorities at different levels can use it without the need for sophisticated demographic modeling infrastructure or extensive data collection beyond basic enrollment numbers.

Our 2025-2027 forecasts show continued enrollment growth with significant regional differences reflecting Kazakhstan's ongoing urbanization and economic development. National enrollment should grow from 3.9 million to 4.2 million students, an 8 % jump requiring substantial system-wide capacity expansion. However, this growth is concentrated. Urban centers, especially Astana with 24 % projected growth, face major infrastructure needs demanding substantial investments in school construction, teacher recruitment, and educational resources. Planning authorities in these fast-growing areas need aggressive capacity expansion to prevent overcrowding and maintain quality.

Rural regions with modest growth need different approaches focused on efficiency and quality rather than expansion. These areas might benefit from consolidating schools, concentrating resources in fewer institutions, and implementing new delivery methods like distance learning or shared specialized teachers. Policy frameworks should recognize these different regional needs rather than applying blanket national strategies.

This research contributes to practice by providing evidence-based guidance on model selection and validated projections supporting better resource allocation. The methodological comparison provides educational planners in Kazakhstan and similar contexts practical insights into which forecasting approaches are effective given their data and institutions. Finding that simple models beat complex demographic approaches challenges standard assumptions in educational forecasting and suggests practitioners should not automatically assume theoretical sophistication guarantees forecasting accuracy.

Future research should extend the time series to capture longer-term patterns and validate models across different periods. The current analysis covers a brief period including the unusual COVID disruption; longer datasets would enable stronger conclusions about model performance under various conditions. Increasing geographic resolution to districts and schools would support more detailed planning and allow us to examine local factors influencing enrollment. Adding grade-specific forecasts would improve operational planning for curriculum, facilities, and teacher specialization. Exploring how to incorporate policy variables, economic indicators, and social factors could improve accuracy and enable scenario analysis of how different development paths might affect enrollment.

The methodological framework we've established provides a foundation for improving forecasting practices in Kazakhstan and potentially other Central Asian countries with similar characteristics. The systematic comparison of multiple approaches using comprehensive administrative data offers a model for how educational planning organizations can develop evidence-based forecasting capabilities. Regular updates with new data and periodic reassessment of model performance will ensure planning processes use the most accurate available projections.

The practical implications go beyond academic contribution to directly supporting educational policy and resource planning across Kazakhstan's system. Accurate enrollment forecasts enable proactive rather than reactive planning, allowing authorities to anticipate needs and prepare responses before problems arise. This forward-looking approach supports more efficient resource use, better infrastructure development, and ultimately better educational outcomes through more effective resource management. As Kazakhstan continues to develop and its population changes, we need accurate enrollment forecasting to ensure schools have enough space and all children can receive quality education wherever they live.

Acknowledgment: This research was funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. BR24993019)

References

Aji, B.W., Septiani, N.Z., Putri, W.M., Irawanto, B., Surarso, B., Farikhin, & Dasril, Y. (2023). Prediction of Indonesia school enrollment rate by using adaptive neuro fuzzy inference system. *Indonesian Journal of Artificial Intelligence and Data Mining*, *6*(1), 40–46.

Aksenova, S.S., Zhang, D., & Lu, M. (2006). Enrollment prediction through data mining. In *Proceedings of the 2006 International Conference on Data Mining* (pp. 510–515). IEEE.

Bernhardt, V.L., Pullum, T.W., & Graham, S.N. (1983, April). *Seattle's small-area approach to forecasting enrollments at the school level*. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada.

Braden, B., & Others. (1972). *Enrollment forecasting handbook introducing confidence limit computations for a cohort-survival technique*. New England School Development Council; Educational Facilities Laboratories, Inc.

Chen, C.-C., Chen, Y., & Kang, C.-Y. (2021). Estimating student-teacher ratio with ARIMA for primary education in fluctuating enrollment. *ICIC Express Letters, Part B: Applications*, *12*(6), 515–523.

Chen, Q. (2022). A comparative study on the forecast models of the enrollment proportion of general education and vocational education. *International Education Studies*, *15*(6), 109–126. https://doi.org/10.5539/ies.v15n6p109

Fabricant, R., & Weinman, J. (1972). Forecasting first grade public school enrollment by neighborhood. *Demography*, *9*(4), 625–632.

GLA Intelligence. (2018). *2018 GLA school place demand projections*. Greater London Authority.

Grip, R.S., & Grip, M.L. (2019). Using multiple methods to provide prediction bands of K-12 enrollment projections. *Population Research and Policy Review*, *38*(5), 635–650. https://doi.org/10.1007/s11113-019-09533-2

Haynes, D.A. II. (2014). *Improving enrollment projections through the application of geographic principles: Iowa 1999–2011* (Doctoral dissertation). University of Iowa.

Huynh Van, S., Nguyen Thi, H., Nguyen Vinh, K., Sam Vinh, L., & Giang Thien, V. (2019). Forecasting the results of students attending school in Vietnam by geographical area. *International Journal of Education and Practice*, *7*(3), 274–285. https://doi.org/10.18488/journal.61.2019.73.274.285

Hussar, W.J., & Bailey, T.M. (2016). *Projections of education statistics to 2024* (NCES 2016-013). U.S. Department of Education, National Center for Education Statistics.

James, F. (2021). *Neural network-based time series forecasting of student enrollment with exponential smoothing baseline and statistical analysis of performance* (Master's report). Kansas State University.

Kornelio, S., Balan, R.T., & Deogratias, E. (2024). Forecasting students' enrolment in Tanzania government primary schools from 2021 to 2035 using ARIMA model. *International Journal of Curriculum and Instruction*, *16*(1), 162–174.

Langley, R.J. (1997). *The use and development of geographical information systems (GIS) and spatial modelling for educational planning* (Doctoral dissertation). University of Leeds.

Lavilles, R.Q., & Arcilla, M.J.B. (2012). Enrollment forecasting for school management system. *International Journal of Modeling and Optimization*, *2*(5), 563–566.

Marinoiu, C. (2014). Modeling and forecasting the gross enrollment ratio in Romanian primary school. *Annals of the Constantin Brâncuși University of Târgu Jiu, Economy Series*, *3*, 17–22.

Miller, M.A. (2008). Planning for enrollment growth using land use data to determine future school sites. *Transportation Research Record*, *2074*(1), 12–20. https://doi.org/10.3141/2074-02

Pajankar, V.D., & Srivastava, S. (2019). An approach of estimating school enrolment with reconstructive cohort approach. *Journal of Physics: Conference Series*, *1366*, 012116. https://doi.org/10.1088/1742-6596/1366/1/012116

Pedamallu, C.S., Ozdamar, L., Ganesh, L.S., Weber, G.-W., & Kropat, E. (2010). A system dynamics model for improving primary education enrollment in a developing country. *Organizacija*, *43*(3), 90–100. https://doi.org/10.2478/v10051-010-0010-5

Proehl, R.S. (2000). *Enrollment forecast for Horizon Christian School 2005–2015*. Population Research Center, Portland State University. https://pdxscholar.library.pdx.edu/enrollmentforecasts/71

Rynerson, C., & Ollinger, J. (2018). *Portland Public Schools enrollment forecasts 2018–19 to 2032–33*. Population Research Center, Portland State University. https://pdxscholar.library.pdx.edu/enrollmentforecasts/123

Rynerson, C., & Wei, C. (2021). *Centennial School District enrollment forecasts 2021–22 to 2030–31*. Population Research Center, Portland State University. https://pdxscholar.library.pdx.edu/enrollmentforecasts/148

Rynerson, C., & Wei, C. (2022). *Salem-Keizer Public Schools enrollment forecast 2022–23 to 2041–42*. Population Research Center, Portland State University. https://pdxscholar.library.pdx.edu/enrollmentforecasts/139

Sahane, M., Sirsat, S., Khan, R., & Aglave, B. (2014). Prediction of primary pupil enrollment in government school using data mining forecasting technique. *International Journal of Advanced Research in Computer Science and Software Engineering*, *4*(9), 656–661.

Shafii, N.H., Alias, R., Shamsudin, S.R., & Nasir, D.S.M. (2021). Fuzzy time series for projecting school enrolment in Malaysia. *Journal of Computing Research and Innovation*, *6*(1), 11–21. https://doi.org/10.24191/jcrinn.v6i1.180

Stronge, W.B., & Schultz, R.R. (1981). Models for projecting school enrollment. *Educational Evaluation and Policy Analysis*, *3*(5), 75–81.

Sweeney, S.H., & Middleton, E.J. (2005). Multiregional cohort enrollment projections: Matching methods to enrollment policies. *Population, Space and Place*, *11*(5), 361–379. https://doi.org/10.1002/psp.379

Tang, H.-W.V., & Yin, M.-S. (2012). Forecasting performance of grey prediction for education expenditure and school enrollment. *Economics of Education Review*, *31*(4), 452–462. https://doi.org/10.1016/j.econedurev.2011.12.007

Wang, Q., Wang, Y., & Wang, G. (2023). Research on space potential excavating of existing primary school under the base of schooling demand forecasting model: A case study in new urban districts. *Ain Shams Engineering Journal*, *14*, 101871. https://doi.org/10.1016/j.asej.2022.101871

Yan, B. (2024). Dynamic analysis of influencing factors and forecast of development trend of “disappearance of rural primary schools.” *Journal of Service Science and Management*, *17*(1), 75–88. https://doi.org/10.4236/jssm.2024.171003