SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

A Trend-Augmented Multivariate Regression Model for University Enrollment Forecasting: The Case of Thu Dau Mot University, 2011–2025

Vo Van On[1], Phan Van Huan[2], Nguyen Thi Hong Tham[3]

[1]Institute of Pharmaceuticals and Health Foods, Thu Dau Mot University, Ho Chi Minh City, Vietnam.
[2]Faculty of Basic Sciences, Binh Duong University, Ho Chi Minh City, Vietnam.
[3]School of Law and Management, Thu Dau Mot University, Ho Chi Minh City, Vietnam.
*Corresponding authors: Vo Van On (onvv@tdmu.edu.vn); Nguyen Thi Hong Tham (thamnth@tdmu.edu.vn)

Abstract
This study develops a multivariate regression model (M2) to analyze and forecast student enrollment at Thu Dau Mot University from 2011 to 2025, using four explanatory variables: the number of academic programs, two regime-adjustment variables for the COVID-19 and post-COVID periods, and a linear time trend.
The model attains R^2 = 0.8637 and adjusted R^2 = 0.8092, with a regression standard error of 668.63 students. In-sample errors are MAE = 412.62, RMSE = 545.94 and MAPE = 10.04%. The trend coefficient is 140.23 students per year (p = 0.0229), and the COVID coefficient is 1,202.23 (p = 0.0145). The maximum VIF is 1.8326, and diagnostic tests do not reject the OLS assumptions.
One-step-ahead expanding-window validation over 2016–2025 yields MAE = 881.01 and MAPE = 19.30%, with a negative bias of ME = −478.54 students.
The model is benchmarked against seventeen alternative methods within the same validation framework, including AICc-selected ARIMA, ARIMAX, three exponential smoothing variants, the Theta method, Ridge regression, Random Forest, Gradient Boosting, a hybrid specification, and forecast combinations. Some important results emerge. First, the ARIMA order choice converges to (0, 1, 0), hence in this case the ARIMA model becomes exactly equivalent to the naïve forecasting method, and no one of the univariate extrapolation methods (ARIMA, exponential smoothing, Theta, naïve with drift) performs better than the naïve oneSecond, even though M2 is one of the leaders, it is not the best approach, since three modifications of M2 (ARIMAX, MAE 822.09; hybrid, 844.03; and Ridge, 880.42) give better results, although not statistically significantly better based on Diebold-Mariano test. Third, an obvious accuracy-bias trade-off emerges: combining M2 and Random Forest reduces bias from -478.54 to -70.33 students at the cost of only 1.6% higher MAE.
A real-time bias-correction experiment yields a negative outcome: errors worsen, and the bias switches signs, showing that the model's bias should be corrected with asymmetric usage rather than an offset term.
Assuming 51 programs in 2026 and 55 in 2027, the model forecasts 6,754 and 7,064 students, with 95% prediction intervals of [4,684; 8,825] and [4,837; 9,291]. Converted into planning units, the standard error range equates to about 52 full-time faculty positions at a 25:1 student-faculty ratio. We recommend interpreting the forecast as ranges rather than point targets.

Keywords: university enrollment; multivariate regression; time trend; rolling-origin validation; Diebold–Mariano test; forecast combination; Thu Dau Mot University.

1. Introduction

Enrollment numbers play a crucial role in planning staffing, facilities, and budgets in higher education institutions. Enrollment forecasting is usually the first and most critical step in the planning process because it directly affects revenue forecasting, staffing plans, and facilities requirements [1].
The enrollment series of Thu Dau Mot University (TDMU) between 2011 and 2025 possesses four characteristics that any model should address. First, the number of academic programs varies widely, from 22 to 50 per year. Second, 2020 to 2022 show volatility associated with COVID-19. Third, enrollment fell unexpectedly during 2023 to 2024 before recovering in 2025. Fourth, this series shows an increasing trend over time: enrollment in 2025 was almost 2.5 times that in 2011.
The research aims at five objectives: (i) specify and estimate a multivariate regression model that includes a linear time trend; (ii) check for multicollinearity and test the OLS assumptions through a battery of diagnostics; (iii) evaluate the model using both in-sample and out-of-sample metrics; (iv) benchmark the model against a broad set of alternative forecasting methods within the same validation framework; and (v) produce forecasts for 2026–2027 together with 95% prediction intervals and a translation of forecast uncertainty into planning units.

1.1. Relationship to the authors' earlier work

Our previous study [2] also uses the same TDMU registration data for the years 2011–2025. For clarity for readers and the editorial board, we outline the connection between the two studies below.

Reference [2] formulates a univariate regression model in which enrollment is explained using one covariate and accounts for the heterogeneous nature of periods using Phase-aware Weighted Least Squares. That paper emphasizes the estimation method: how to weight observations in each phase to avoid the impact of abnormal phases on parameter estimation.

This paper is distinctive in four respects.

Model structure. M2 is a multi-variable model having four explanatory variables. Phase effects are incorporated directly using two adjustment variables in the regression equation, rather than as a weight outside the equation, as in [2]. These two representations produce different results: [2] estimates the slope coefficient under the phase-based weighting scheme, but this paper estimates a unique adjustment factor for each phase, along with the coefficient on the number of programs and the time trend.

Estimation method. This paper employs OLS, tests inference robustness with HC3 standard errors, and does not use weighted least squares.

Research question. Reference [2] asks how to obtain stable estimates when a series contains anomalous phases. This paper asks two different questions: how much of the variation in enrollment the observed factors explain, and how well a multivariate model forecasts out of sample relative to a range of alternative methods.

Scope of evaluation. This paper adds three elements absent from [2]: a comparison against seventeen alternative methods within a rolling-origin validation framework (§4.6), Diebold–Mariano testing of ranking uncertainty (§4.7), and translating forecast uncertainty into planning units (§6).

The two papers share no tables or figures. Every estimated coefficient, fit statistic, diagnostic result, error metric, and forecast value reported here comes from a different model specification and does not duplicate the figures published in [2]. The two studies share the underlying data, and we disclose this overlap here and in the cover letter accompanying the manuscript.

2. Data and model

2.1. Data

The sample consists of 15 annual observations, from 2011 to 2025. The dependent variable is total enrollment.

The original data also include four other variables: reputation, public-institution status, cost of living, and tuition. None enter the final model for three technical reasons. The variable *public-institution status* takes the same value at every observation, since TDMU was a public institution throughout the study period; a variable with no variation cannot be estimated. The two variables*, cost of living and tuition,* take identical values across all 15 observations, so they are perfectly collinear, and at most one can appear in the model. The variable *reputation* is coded on a coarse ordinal scale with five levels and varies almost monotonically over time, so it duplicates the information carried by the time trend. Given the limited degrees of freedom in a 15-observation sample, we drop all four variables from the final specification.

Table 1 presents the study data together with the coding of the explanatory variables.

Table 1. Study data and variable coding. COVID: 0 for 2011–2019, 1 for 2020, 2 for 2021, 1 for 2022, and 0 from 2023 onward, reflecting the year-by-year magnitude of the disruption. After COVID: 1 in 2023, 0.5 in 2024, and 0 for all other years, representing a gradual reduction adjustment. Time trend: t = year – 2011.

Year	Enroll	Majors	COVID	PostCOVID	t
2011	2,523	31	0	0	0
2012	4,854	32	0	0	1
2013	3,835	32	0	0	2
2014	5,330	50	0	0	3
2015	5,336	28	0	0	4
2016	3,843	22	0	0	5
2017	4,405	28	0	0	6
2018	4,959	30	0	0	7
2019	4,956	36	0	0	8
2020	6,598	47	1	0	9
2021	8,547	48	2	0	10

Year	Enroll	Majors	COVID	PostCOVID	t
2022	7,099	46	1	0	11
2023	3,646	34	0	1	12
2024	4,297	38	0	0.5	13
2025	6,287	40	0	0	14

Enrollment rose from 2,523 in 2011 to 8,547 in 2021, then fell to a low of 3,646 in 2023 before rebounding to 6,287 in 2025. The correlation coefficient between programs and enrollment is 0.707, while that between time trend and enrollment is 0.468.

2.2. Model

The suggested multivariate regression model is of the form:

$$Enroll_t = \beta_0 + \beta_1 \cdot Majors_t + \beta_2 \cdot COVID_t + \beta_3 \cdot PostCOVID_t + \beta_4 \cdot t + \varepsilon_t$$

where Enrollt represents enrollment in year t, Majorst is the number of academic majors offered, COVIDt and PostCOVIDt are the two regime adjustment variables, t is the linear trend, and εt is an error term. The model is estimated using ordinary least squares (OLS). For n = 15 and k = 4 independent variables, the degrees of freedom for residuals are n – k – 1 = 10.

Expected signs: β1 > 0 because an expanded program portfolio provides a wider variety of options for the applicant; β2 > 0 and β3 < 0, corresponding to the higher level of 2020–2022 and the lower level of 2023–2024; β4 > 0, representing the long-term positive trend.

These coefficients on the regime dummy variables should be understood as phase-related statistical controls that account for all other covariates, rather than causal effects. In particular, β2 > 0 does not mean that COVID-19 raised enrollment; it means that during that period TDMU's enrollment was higher than what the relationship with program count and trend alone would predict.

3. Metrics and evaluation design

3.1. Fit metrics

R^2, adjusted R^2 and the regression standard error are computed in the usual way. In a small sample, adjusted R^2 is preferred because it penalizes the number of parameters used. Information criteria are computed from the likelihood as:

$$AIC = -2 \cdot lnL + 2p, BIC = -2 \cdot lnL + p \cdot lnn$$

where p = k + 1 = 5 is the number of estimated regression coefficients, including the intercept; following the statsmodels convention, the error variance is not counted. Counting it (p = k + 2) would add the same constant to AIC and BIC for every model and would not change any comparison.

3.2. Forecast error metrics

The study uses mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error (MAPE) to measure error magnitude; mean error (ME) and mean percentage error (MPE) to measure error direction; and the mean absolute scaled error (MASE) for a scale-free comparison:

$$MASE = \frac{MAE}{MAE(naïve, in-sample)}$$

where the denominator is the MAE of the naïve forecast on the initial 2011–2015 training window, equal to 1,212.75 students.

Hyndman and Koehler [3] warn against using only one class of metrics: The first three measures cannot distinguish between a biased forecast model and a model with purely random errors, but both types of errors have important implications for planning. Random errors offset each other year after year, whereas biases build up in the same direction.

3.3. Multicollinearity

The extent to which the explanatory variables are linearly related is determined by the variance inflation factor:

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

where R^2j is the coefficient of determination obtained when regressing the j-th regressor against the other variables. Typical thresholds are VIF > 5 for concern and VIF > 10 for a serious issue.

3.4. Diagnostic tests

We check residual normality using the Shapiro-Wilk test [4]. We test heteroscedasticity using the Breusch-Pagan test [5]. In addition to OLS standard errors, we report HC3 standard errors to assess whether inference holds in small samples and with high leverage [6]. We test autocorrelation with the Durbin–Watson statistic [7], along with the Ljung–Box [8] and Breusch–Godfrey [9, 10] tests. Because Durbin–Watson tables at n = 15 leave a wide inconclusive region, the study reports an exact p-value obtained by simulating the statistic's distribution under the null, conditional on the model's actual design matrix

(200,000 replications). We measure observation influence using leverage and Cook's distance [11], with reference thresholds hi > 2(k+1)/n = 0.667 and Di > 4/n = 0.267.

3.5. Out-of-sample validation

In-sample metrics measure how well a model describes data it has already seen; they do not measure its ability to forecast data it has not seen [12]. Because the series is short and has a time structure, the study does not use random k-fold splits but instead applies one-step-ahead rolling-origin validation: for each year τ from 2016 to 2025, the model is re-estimated using only data from 2011 through τ − 1, and used to forecast year τ. This procedure ensures that every forecast uses only information available before the forecast date [13]; Bergmeir et al. [14] discuss how time dependence affects the validity of cross-validation procedures more broadly. At origins where a regime variable remains identically zero throughout the training window (COVID before 2020, PostCOVID before 2023), we do not identify its coefficient; we omit the variable at those origins, and it contributes nothing to the corresponding forecast. At the 2016 origin, for example, we estimate three coefficients from five observations.

This design yields ten out-of-sample forecasts without sacrificing any training-set observations—an important consideration for a 15-observation series, where holding out the last three years as a test set would discard 20% of the data.

3.6. Prediction intervals

The prediction interval with a 95% confidence level for a future observation at feature vector x0 is found using the conventional method by taking the t-critical value corresponding to n - k - 1 = 10 degrees of freedom as t(0.975, 10) = 2.2281 instead of the z-value 1.96; using z would underestimate the true width by 12.0%. A prediction interval is used instead of a confidence interval for the mean because the aim is to predict the true enrollment for that year, not just the mean [15, 16]. The value of h0 is mentioned explicitly since it shows how much the forecasted value falls outside the range of observed values:

$$h_0 = x_0{}^T (X^T X) \quad {}^{-1} x_0$$

3.7. Design of the comparison with alternative methods

We consider all seventeen additional techniques using the same rolling-origin framework described in §3.5. No method has access to information beyond the forecast date, and we reselect every tuning choice—including the ARIMA order, smoothing parameters, and the Ridge regularization parameter—at each origin using only the corresponding training data.

The methods fall into four groups. Trivial benchmarks: naïve, naïve with drift, and the historical mean. Univariate time-series methods: ARIMA with order (p, d, q) selected by AICc [17, 18] over a grid p, q ∈ {0, 1, 2} and d ∈ {0, 1}; simple exponential smoothing (SES); linear Holt; damped Holt [19]; and the Theta method [20]. Methods using the same covariates as M2: Ridge [21] with the regularization parameter chosen by LOOCV; Random Forest [22] with 500 trees; Gradient Boosting [23]; a reduced model M1 with only the program count and time trend; and a trend-only model. Combination group: ARIMAX, i.e., regression with ARMA errors; a hybrid model consisting of M2 plus a Random Forest fitted to M2's residuals, in the spirit of the hybrid design of Tapio and Tarepe [24]; a simple average of M2 and Random Forest; and a simple average of M2, Random Forest, and naïve.

Two of the covariate-based benchmarks are nested versions of M2. The reduced model M1 drops the two regime-adjustment variables and keeps only the number of programs and the time trend:

$$Enroll_t = \gamma_0 + \gamma_1 \cdot Majors_t + \gamma_2 \cdot t + \varepsilon_t$$

and the trend-only model further drops the number of programs:

$$Enroll_t = \delta_0 + \delta_1 \cdot t + \varepsilon_t$$

The labels reflect this nesting: M2 is obtained by adding COVID and PostCOVID to M1. Both reduced models are estimated by OLS and evaluated in the same rolling-origin framework as M2. They address a separate question: whether the two regime-adjustment variables genuinely improve out-of-sample forecasting or merely inflate the in-sample fit.

To assess the reliability of the error differences between methods, the study uses the Diebold–Mariano test [25] with an absolute-error loss function, together with the small-sample correction of Harvey, Leybourne and Newbold [26].

4. Results

4.1. Estimated coefficients

Table 2 reports the OLS estimates of model M2, together with OLS and HC3 standard errors and 95% confidence intervals.

Table 2. Estimated coefficients of model M2. Confidence intervals use t(0.975, 10) = 2.2281.

Variable	Coefficient	SE OLS	t	p OLS	SE HC3	p HC3	CI 95%
Intercept	2,490.8670	940.2182	2.649	0.0243	1,132.8333	0.0525	[395.93; 4,585.80]

Variable	Coefficient	SE OLS	t	p OLS	SE HC3	p HC3	CI 95%
Majors	42.3537	28.3412	1.494	0.1659	34.0521	0.2419	[−20.79; 105.50]
COVID	1,202.2340	407.5266	2.950	0.0145	437.1561	0.0205	[294.21; 2,110.26]
PostCOVID	−2,224.6696	775.3595	−2.869	0.0167	1,358.9569	0.1327	[−3,952.28; −497.06]
Trend t	140.2318	52.2225	2.685	0.0229	69.9519	0.0728	[23.87; 256.59]

$$Enroll_t = 2{,}490.87 + 42.35 \cdot Majors_t + 1{,}202.23 \cdot COVID_t - 2{,}224.67 \cdot PostCOVID_t + 140.23 \cdot t$$

All four coefficients have the expected sign. Three of the four explanatory variables are statistically significant at the 5% level under OLS inference.

The trend variable has a coefficient of 140.23, meaning that after controlling for the number of programs and the two phases, enrollment grows by roughly 140 students per year on average, with a 95% confidence interval of [23.87, 256.59].

The COVID variable has a coefficient of 1,202.23 and is the most robust result in the model: significant under both OLS (p = 0.0145) and HC3 (p = 0.0205). Based on the coding in Table 1 above, this suggests that the adjustment is positive for 1,202 students in 2020 and 2022, and 2,404 students in 2021.

The PostCOVID variable has a coefficient of −2,224.67, implying a reduction of about 2,225 students in 2023 and 1,112 students in 2024. The coefficient is statistically significant using OLS (p = 0.0167), but not using HC3 (p = 0.1327). The reason is discussed in §4.3

The variable number-of-programs has a coefficient of 42.35 but is statistically insignificant (p-value = 0.1659). The proper reading is a positive association with the correct sign, but our sample size is insufficient to identify it accurately. New initiatives vary greatly in size and attractiveness, so a single average coefficient cannot capture this heterogeneity.

Comparison of actual enrolment figures and M2 fitted values from 2011 to 2025, with forecasts for 2026 and 2027.

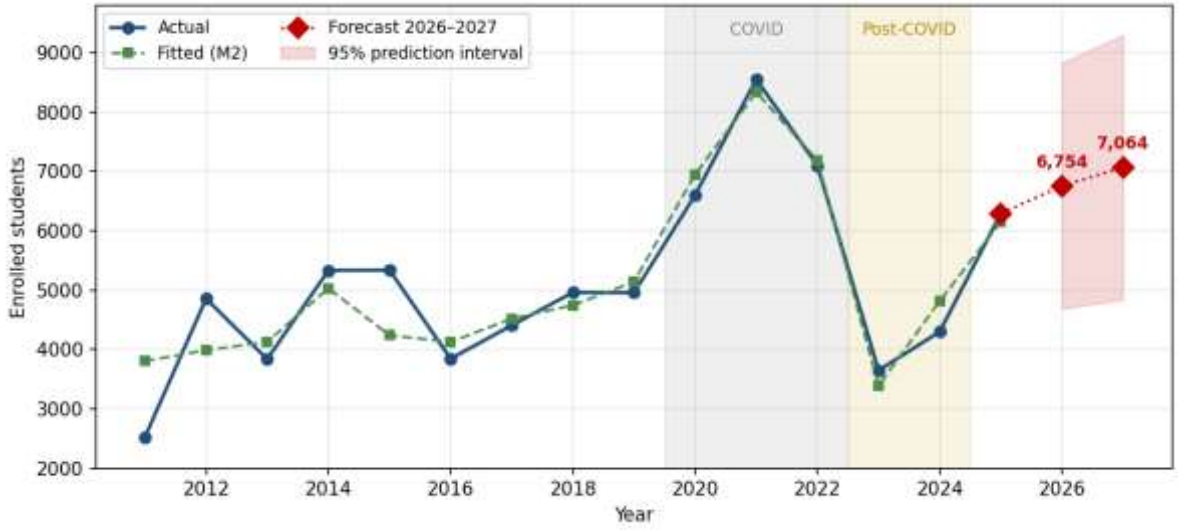

Figure 1. Actual enrollments, M2 model fits, and predictions for 2026-2027. The shaded area represents the 95% prediction interval. The model fits the peaks in 2021 and troughs in 2023 well, but is highly inaccurate at the beginning of the time series.

4.2. Multicollinearity

Table 3 reports the variance inflation factors of the four explanatory variables.

Table 3. Variance inflation factors for the explanatory variables.

Variable	VIF	Assessment
Majors	1.7998	Not a concern
COVID	1.8326	Not a concern
PostCOVID	1.4792	Not a concern

Variable	VIF	Assessment
Trend t	1.7080	Not a concern

The largest VIF is 1.8326, well below both conventional thresholds. This result is worth noting specifically for the time trend, since trend variables are often suspected of causing multicollinearity when included alongside other variables that also rise over time. In this sample, the correlation between the number of programs and the trend is 0.398 — low enough that the VIF reaches only 1.708.

Nonetheless, two distinct issues should be kept apart. A low VIF indicates that the variables do not overlap linearly, but it does not guarantee stable estimation in a sample of only 15 observations. The stability of the estimates is a separate matter, taken up in §7.1.

4.3. Diagnostic tests

Table 4 summarizes the diagnostic tests applied to the residuals of model M2.

Table 4. Diagnostic tests on the residuals of model M2**.** The Durbin–Watson p-value is obtained from 200,000 simulations of the statistic's distribution under the null, conditional on the model's design matrix; the corresponding 5% critical value is 1.5362. The reported p-value is the lower-tail probability (alternative: positive autocorrelation); against negative autocorrelation, the upper-tail p-value is about 0.47, so neither alternative is supported.

Test	Statistic	p	Conclusion at 5%
Shapiro–Wilk (normality)	W = 0.9494	0.5158	Not rejected
Breusch–Pagan (heteroscedasticity)	LM = 5.4854	0.2410	Not rejected
Durbin–Watson (autocorrelation)	DW = 2.3971	0.5263	Not rejected
Breusch–Godfrey lag 1	LM = 2.5787	0.1083	Not rejected
Ljung–Box lag 1	Q = 2.6886	0.1011	Not rejected
Ljung–Box lag 2	Q = 2.7017	0.2590	Not rejected

None of the tests reject the corresponding OLS assumption (normality, homoscedasticity, no autocorrelation), so the statistical inference in Table 2 rests on valid footing. Two points about reading this table deserve mention.

First, with n = 15 and four explanatory variables, these tests have low power. A “not rejected” result should be read as “no evidence of violation was detected,” not as “certainly no violation exists.” This is why we used HC3 standard errors alongside OLS in Table 2: any difference in results suggests some reservations. In fact, we found such a difference for the PostCOVID variable.

Second, the Durbin-Watson test value of 2.3971 is above 2, implying negative autocorrelation. Although this is not statistically significant, the trend is mechanically understandable: a linear trend makes the residuals alternate in sign because the regression line is bound to run through the middle of both the ascending and descending parts of the time series.

Figure 2 shows the residual plot.

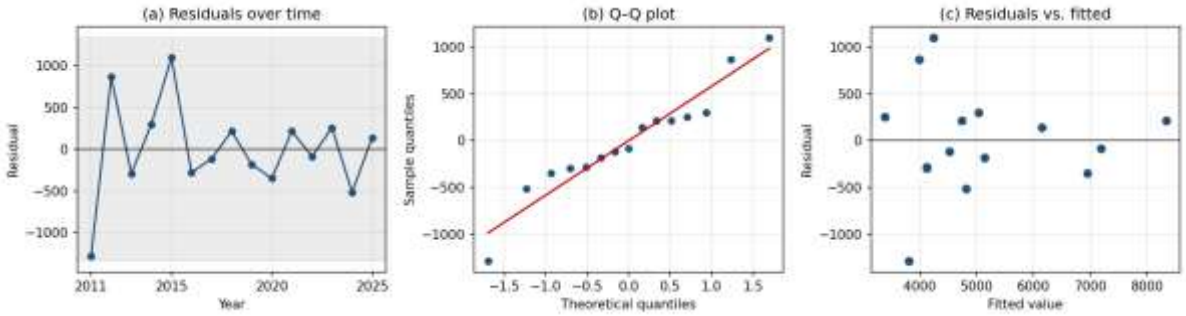

Figure 2. Residual diagnosis for the M2 model. Figure (a): residual plotted over time, where the gray area indicates ±2s; all data points lie within it, with 2011 having the largest residual value (−1,280.8). Panel (b): Q-Q plot indicating that the residuals fall near the normal line. Panel (c): Plot of residuals vs. fitted values, indicating no presence of non-linearity or increasing variance with the forecast level.

On observation influence. Three observations exceed the Cook's distance threshold: 2023 (Di = 0.7701; hi = 0.8224), 2011 (Di = 0.3414), and 2014 (Di = 0.2794; hi = 0.6850). All three are located at the extremes of

the data: 2023 represents the trough of the dataset and the only period when PostCOVID = 1, which means that it essentially determines the value of β3 – that is why HC3 estimation of this coefficient is not stable. 2014 has the most programs in the sample (50). 2011 has the largest residual (−1,280.8). In addition, 2021 — the peak of the series and the only observation with COVID = 2 — exceeds the leverage threshold (hi = 0.7085), although its Cook's distance is small (Di = 0.1746). This is a structural feature of a 15-observation sample and should be flagged as a limitation.

4.4. Whole-sample fit metrics

Table 5 summarizes the whole-sample fit metrics of model M2.

Table 5. Whole-sample fit metrics for model M2.

Metric	Value
Number of observations n	15
Number of explanatory variables k	4
Residual degrees of freedom	10
R^2	0.8637
Adjusted R^2	0.8092
Regression standard error s	668.63 students
MAE	412.62 students
RMSE	545.94 students
MAPE	10.04%
AIC	241.64
BIC	245.18
Whole-model F statistic	F(4, 10) = 15.84
p of the F statistic	0.0003

The model explains 86.37% of the variation in enrollment; after penalizing for the four parameters, adjusted R^2 still reaches 80.92%. The whole-model F statistic has p = 0.0003, so we strongly reject the hypothesis that all coefficients are jointly zero. MAE = 412.62 students against a mean enrollment of 5,101 students per year (= the 76,515-student total in Table 1 divided by 15 observations; see Appendix A.1) corresponds to MAPE = 10.04%. A mean error of zero is a mechanical consequence of OLS with an intercept, so it carries no information about in-sample bias; bias only becomes informative when assessed out of sample.

4.5. Out-of-sample validation

Table 6 contrasts M2's in-sample errors with its leave-one-out (LOOCV) and rolling-origin out-of-sample errors, using the naïve forecast as a benchmark.

Table 6. Error metrics for model M2. In-sample and LOOCV metrics are computed over all 15 observations (2011–2025); rolling-origin and naïve metrics over the 2016–2025 validation period (10 forecasts).

Metric	In-sample	LOOCV	Rolling	Naïve
MAE	412.62	678.15	881.01	1,374.50
RMSE	545.94	840.77	1,091.51	1,665.85
MAPE	10.04%	16.21%	19.30%	27.24%
ME	0.00	+131.05	−478.54	+95.10
MPE	—	—	−14.04%	−3.56%
U-ratio vs. naïve	—	—	0.641	1.000

Out-of-sample error is substantially higher than in-sample error. MAPE rises from 10.04% to 16.21% under LOOCV and to 19.30% under rolling-origin validation. This near-doubling shows that the metrics in Table 5 reflect how well the model describes already observed data more than its forecasting ability. For planning purposes, use 19.30%, not 10.04%.

It is worth flagging a structural limitation of the validation window: since both regime variables only vary from 2020 onward, forecasts for 2016–2019 effectively rely on the program-count and trend variables alone.

The model shows a negative bias. Mean error is −478.54 students, corresponding to MPE = −14.04%, with 7 of the 10 years showing negative errors. In other words, the model tends to systematically over-forecast enrollment. Using the point forecast as a planning target risks over-committing resources. This is precisely the type of information that MAE, RMSE, and MAPE alone cannot reveal [3].

The model clears the trivial benchmark. A U-ratio of 0.641 means the model incurs only 64.1% of the naïve forecast's error. In our view, this metric is more persuasive evidence of the model's usefulness than R^2 = 0.8637, since it answers directly the question a planner cares about.

Figure 3 shows the rolling-origin forecast path of M2 and its year-by-year errors.

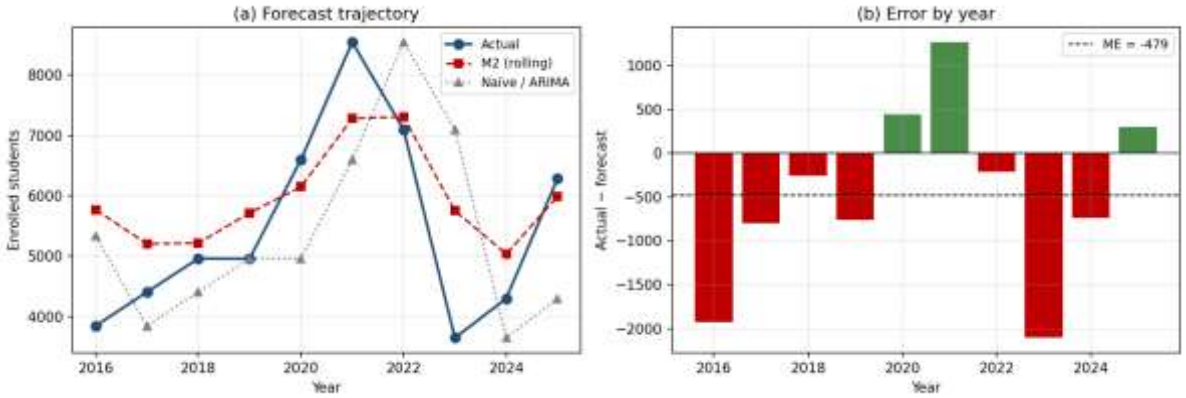

Figure 3. One-step-ahead rolling-origin validation results, 2016–2025. Panel (a): the forecast path of model M2 against actual enrollment and the naïve benchmark. Panel (b): year-by-year error; green bars are positive errors, red bars are negative errors; the horizontal line marks the average bias.

4.6. Comparison with seventeen alternative methods

Table 7 reports the out-of-sample error metrics of M2 and the seventeen alternative methods, sorted by increasing MAE.

Table 7. Comparison of eighteen forecasting methods under the same one-step-ahead rolling-origin validation, 2016–2025 validation period (10 forecasts). Sorted by increasing MAE. U is the ratio of MAE to the naïve method. Error is defined as actual minus forecast, so a negative ME corresponds to over-forecasting.

Method	MAE	RMSE	MAPE (%)	ME	MPE (%)	MASE	U
ARIMAX	822.09	1,113.80	17.83	−336.54	−10.68	0.68	0.598
M2 + RF on residuals	844.03	1,056.21	18.19	−332.30	−11.29	0.70	0.614
Ridge	880.42	1,084.90	18.85	−199.99	−8.48	0.73	0.641
M2	881.01	1,091.51	19.30	−478.54	−14.04	0.73	0.641
Combination of M2 + RF	895.32	1,129.03	18.15	−70.33	−6.66	0.74	0.651
Combination of M2 + RF + naïve	968.80	1,260.71	19.25	−15.19	−5.63	0.80	0.705
Gradient Boosting	1,021.31	1,278.59	19.74	+157.15	−2.38	0.84	0.743
Random Forest	1,069.00	1,334.88	19.96	+337.87	+0.72	0.88	0.778
M1 (programs + trend)	1,175.61	1,482.94	25.64	−647.66	−18.90	0.97	0.855
Historical mean	1,358.04	1,722.91	22.92	+787.71	+8.07	1.12	0.988
Trend-only	1,373.07	1,871.91	29.87	−569.39	−19.47	1.13	0.999

Method	MAE	RMSE	MAPE (%)	ME	MPE (%)	MASE	U
ARIMA (AICc-selected)	1,374.50	1,665.85	27.24	+95.10	−3.56	1.13	1.000
Naïve	1,374.50	1,665.85	27.24	+95.10	−3.56	1.13	1.000
Linear Holt	1,389.17	1,874.53	30.20	−508.41	−18.24	1.15	1.011
Damped Holt	1,401.92	1,765.03	29.53	−422.54	−16.48	1.16	1.020
Naïve with drift	1,425.00	1,790.01	29.11	−268.33	−10.74	1.18	1.037
Simple exponential smoothing	1,548.96	1,917.61	29.27	+192.72	−3.38	1.28	1.127
Theta	1,573.00	1,942.58	30.18	+44.65	−6.44	1.30	1.144

Figure 4 visualizes the MAE ranking of the eighteen methods relative to the naïve benchmark.

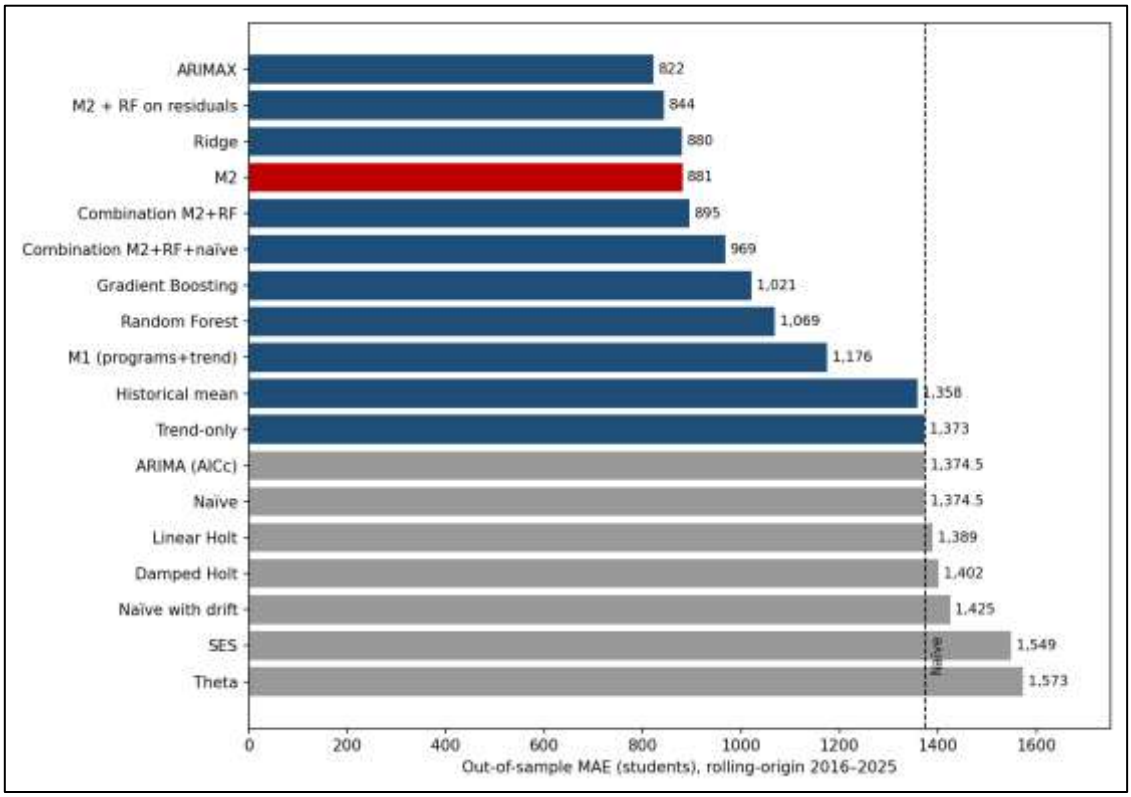

Figure 4. Out-of-sample mean absolute error for the eighteen methods. The dashed line marks the naïve benchmark; the gray bars are methods that fail to beat it.

Result 1: ARIMA degenerates into a random walk. ARIMA and naïve coincide exactly on all seven metrics, because at all ten origins the order selected by AICc is (0, 1, 0) — which is precisely a random walk. With training windows ranging from 5 to 14 observations, AICc penalizes any additional AR or MA structure more heavily than the fit it buys. As a result, the naïve benchmark used in §4.5 is simultaneously the ARIMA benchmark for this dataset. This is consistent with the Ljung–Box and Breusch–Godfrey tests in Table 4, which detect no autocorrelation in the residuals.

Result 2: no extrapolative univariate method beats naïve. All five methods that rely solely on the series' own past — SES, linear Holt, damped Holt, Theta, and naïve with drift — have U > 1; the only univariate method below naïve is the historical mean, which does not extrapolate, and it does so only marginally (U = 0.988). The process is simple because these approaches extrapolate the last level and trend; thus, at the turning point, they simply continue the previous trend in the opposite direction. The largest miss from linear Holt happens in 2023, where the prediction is 7,474 students, but the actual figure is 3,646, a miss of 105%, since the model continues to extrapolate the increase from 2020 through 2022. It implies that M2 outperforms because of the exogenous information used, and not by modeling the endogenous dynamics of the series.

Result 3: M2 belongs to the group of leaders, although not the optimal one. Three methods have smaller MAE errors than M2: ARIMAX (822.09), hybrid M2 + Random Forest for residuals (844.03), and Ridge (880.42). All three methods extend M2: ARIMAX includes an ARMA model for the error terms; the hybrid method incorporates a nonlinear function for the error terms; and Ridge adds regularization to the same four covariates used in M2. No method from another model class performs better than M2. The difference between Ridge and M2 is 0.59 in MAE; for all intents and purposes, the two are equal.

Result 4: The two regime indicators are of some forecasting value. The trend-only model yields MAE = 1,373.07, which is basically the same as the naïve model. Adding the program-count variable (M1) brings this down to 1,175.61. Adding the two regime variables (M2) brings it further down to 881.01, a 25.1% improvement. This directly rebuts the concern that the phase dummies merely soak up residual variation: in the rolling-origin design, a regime variable enters a forecast only after it has been observed in earlier training years, so the gain cannot come from absorbing the residual of the year being forecast. The same contrast holds in-sample: estimated on all 15 observations, M1 attains $R^2 = 0.5409$ (adjusted $R^2 = 0.4644$) and MAPE = 18.41%, against $R^2 = 0.8637$ (adjusted $R^2 = 0.8092$) and MAPE = 10.04% for M2.

Result 5: machine-learning methods are worse on magnitude but better on bias. Random Forest and Gradient Boosting use the same four covariates as M2, so this compares identical information. Both have higher MAE than M2, consistent with the theoretical expectation that tree-based methods cannot extrapolate beyond the observed range. But Random Forest's MPE is only +0.72% and Gradient Boosting's is −2.38%, against M2's −14.04%. In other words, M2 is more accurate but systematically wrong, while the machine-learning methods are less accurate but wrong in a nearly random way.

4.7. The accuracy–bias trade-off

Result 5 above suggests that the two groups of methods make two different kinds of mistakes that could offset each other. Figure 5 plots all eighteen methods on a plane of MAE (error magnitude) against the absolute value of ME (bias).

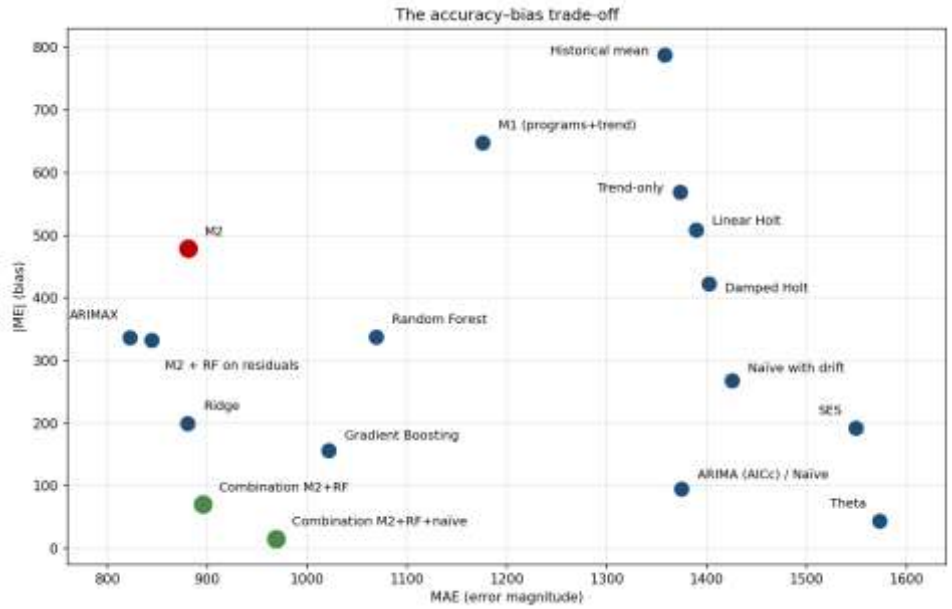

Figure 5. The accuracy–bias trade-off across the eighteen methods. Methods near the origin perform well on both dimensions. M2 is marked in red; the two forecast combinations are marked in green.

A simple forecast combination nearly eliminates the bias at a small cost in accuracy. The simple average of M2 and Random Forest achieves MAE = 895.32, only 1.6% above M2, while bias falls from −478.54 to −70.33 students, an 85.3% reduction. Combining all three of M2, Random Forest, and naïve pushes bias down to −15.19 students — essentially unbiased — at MAE = 968.80.

The mechanism behind this result is straightforward: M2 is biased −478.54 while Random Forest is biased +337.87, so averaging the two forecasts nearly cancels the bias. This is the standard point: forecast combinations tend to outperform any single approach [27].

Practical implications depend on the intended application. If the objective is one-year forecasting accuracy, use M2 or ARIMAX. For long-range resource planning, the typical university scenario, the M2–Random Forest combination is best, since random errors tend to cancel year after year, whereas biases accumulate in one direction [3].

4.8. Diebold–Mariano test

Table 8 presents the Diebold-Mariano test results for comparing M2 against other methods.

Table 8. Diebold- Mariano test on M2 vs. the competing model(s). A negative value means that M2 has the lower mean absolute error. P-values for two-sided tests using the t-statistic with 9 degrees of freedom after correction by Harvey-Leybourne-Newbold.

Comparison	DM	p	Conclusion at 10%
M2 vs. damped Holt	−2.214	0.0541	M2 better
M2 vs. M1 (programs + trend)	−1.955	0.0823	M2 better
M2 vs. Theta	−1.899	0.0900	M2 better
M2 vs. linear Holt	−1.871	0.0942	M2 better
M2 vs. naïve with drift	−1.810	0.1037	Inconclusive
M2 vs. trend-only	−1.767	0.1111	Inconclusive
M2 vs. SES	−1.767	0.1111	Inconclusive
M2 vs. ARIMA / naïve	−1.732	0.1173	Inconclusive
M2 vs. historical mean	−1.115	0.2938	Inconclusive
M2 vs. Gradient Boosting	−0.559	0.5901	Inconclusive
M2 vs. Random Forest	−0.675	0.5169	Inconclusive
M2 vs. combination M2+RF+naïve	−0.416	0.6874	Inconclusive
M2 vs. combination M2+RF	−0.093	0.9280	Inconclusive
M2 vs. Ridge	+0.006	0.9957	Inconclusive
M2 vs. ARIMAX	+0.360	0.7271	Inconclusive
M2 vs. M2 + RF on residuals	+0.628	0.5454	Inconclusive

Table 8 supports two opposite conclusions, and both are significant.

The M2 is superior to the inferior models with confidence. Four results are significant at the 10% level, all compared with univariate techniques or reduced models. Comparison with ARIMA and naïve shows a p-value = 0.1173, which is not significant enough even with a 35.9% difference in MAE.

The three techniques that have MAE lower than M2 are not significantly different from M2. The p-values for the tests conducted against Ridge (0.9957), ARIMAX (0.7271), and the hybrid model (0.5454) are all large. Given ten pairs of forecast differences, the Diebold-Mariano test has limited power; it shows that the differences are within normal sampling variation, not that these three methods are better.

Here we give the complete table rather than just the MAE ranking, since the uncertainty around that ranking is also informative for the user. The conclusion, cautious and consistent with the data presented, is that M2 is part of a prominent class with three extensions, and M2 is the simplest specification among them. For a series of 15 observations, simplicity constitutes an important strength, and not simply a matter of presentation.

4.9. Can the negative bias be corrected?

The estimate of the bias of M2 as −478.54 students in §4.5 clearly points to the solution of simply subtracting the bias from the prediction. We tested this alternative under the condition that bias must be estimated in real time from errors in previous years up to year τ; no forward-looking information is allowed. Because we require at least three years of error data, our evaluation period is 2019-2025. This is shown in Table 9.

Table 9. M2 before and after real-time bias correction, 2019–2025 period (7 forecasts).

Approach	MAE	RMSE	MAPE (%)	ME
Original M2	832.38	1,034.65	17.08	−257.42
Bias-corrected M2	930.52	1,175.02	16.75	+364.71

The bias correction further aggravates the error. The MAE increases by 11.8%, while RMSE increases by 13.6%. Only MAPE slightly improves. More importantly, the bias does not disappear; it changes sign to +364.71 from −257.42.

This is because the M2 bias is not constant. However, this is not a static correction factor but the result of the model's slow reaction at two breakpoints: underestimation during the 2020-2021 growth period and overestimation during the 2023-2024 decline period. Averaging these errors yields a negative value, but it does not forecast next year's error.

Comparing this finding with §4.7 yields an extremely helpful result: the model's bias cannot be corrected by subtracting a constant, but it is corrected well through forecast combination. Combining M2 with a method whose bias is opposite reduces the bias by 85.3%, while direct correction changes the sign of the bias.

5. Forecasts for 2026–2027

Table 10 provides forecasts for 2026-2027, along with the 95% prediction intervals and leverage h_0 at each point forecast.

Table 10. Predictions and 95% Prediction Intervals for 2026–2027. The assumption is that COVID = 0 and PostCOVID = 0 in both.

Year	Programs	Forecast	PI95% low	PI95% high	Half-width	h_0
2026	51	6,754	4,684	8,825	2,070	0.9313
2027	55	7,064	4,837	9,291	2,227	1.2343

The model predicts 6,754 students by 2026 and 7,064 by 2027, an increase of 7.4% and 12.4%, respectively, over 6,287. The increase comes from two sources: growth in the number of programs (+11 and +15 programs relative to 2025, contributing roughly 466 and 635 students) and the time trend (contributing 140 students per year).

Three caveats need to be stated explicitly when using these figures.

First, these are conditional forecasts. The assumption of 51 programs in 2026 and 55 in 2027 is a user-supplied input, not a model output. Check this assumption against the University's approved program-expansion plan.

Second, the forecast points lie outside the observed data region in both directions. The number of programs in the sample ranges only over [22; 50], so the values 51 and 55 both fall outside it; the same is true of the time trend. The quantity h0 quantifies this: the 2027 forecast point has h0 = 1.2343, larger than the leverage of every observation in the sample (the highest is 0.8224). In addition, the assumed increase of 11 programs in a single year equals the second-largest year-on-year increase in the sample (2020, from 36 to 47 programs) and is exceeded only by the increase in 2014 (+18 programs from 2013 to 2014; see Appendix A.5).

Third, the prediction intervals are very wide: ±2,070 students for 2026 and ±2,227 for 2027, roughly ±31% of the point forecast. This width is not a defect of the calculation but an honest reflection of the uncertainty inherent in a 15-observation series with two structural breaks. It is consistent with the out-of-sample error in Table 6: with a rolling-origin MAPE of 19.30%, a narrow prediction interval would be inconsistent with the evidence.

6. Implications for planning

This section converts the above statistics into values the planner can use directly. The overall principle is that a forecast model becomes worthwhile once its error is translated to decision units, such as faculty members, classrooms, and scholarships, rather than in percentages.

6.1. Converting forecast error into resource terms

With an expected number of 6,754 students in 2026 and a MAPE of 19.30% out-of-sample, the average error range is ±1,304 students. The 95% prediction interval covers 4,141 students. Table 11 converts these numbers to part-time faculty needs under varying student-to-faculty ratios.

Table 11. Transforming uncertainty about the 2026 forecast into faculty needs. The table illustrates student-faculty ratios under different scenarios; in practice, however, they must be replaced with the highest ratio set by the Ministry of Education and Training for each academic discipline at the University.

Student–faculty ratio	±MAPE band (faculty)	PI95% width (faculty)	Range of faculty needed
20: 1	± 65	207	234 – 441

Student–faculty ratio	±MAPE band (faculty)	PI95% width (faculty)	Range of faculty needed
25: 1	± 52	166	187 – 353
30: 1	± 44	138	156 – 294
40: 1	± 33	104	117 – 221

Given the 25:1 ratio, the standard error range is 52 faculty members, while the total range of the 95% prediction interval is 166 faculty members. This is one such budget-related gap. This is why 19.30%, and not 10.04%, is the right starting point for planning resources.

6.2. Three principles for use

First principle: make cost-reversibility decisions separately. Low-reversal decisions such as creating class sections, hiring adjunct professors, and renting more classrooms should revolve around the primary estimate of 6,754 students because they may be modified as needed based on actual enrollment throughout the year. Decisions involving large reversal costs, such as recruiting tenure-track employees, building projects, and procurement of expensive equipment, must be based on the lower limit of the prediction interval. Given the negative bias described in §4.5, one solution is to adopt 5,500-6,000 students as the standard for the second set of decisions, with shortfall managed through capacity flexibility.

Second principle: distinguish fixed capacity from flexible capacity. The margin of error of ±1,304 students is inherent to the problem structure and cannot be improved by a more accurate model because, as the results in section 4.6 show, none of the seventeen methods, including machine learning ones, significantly reduces this error. This implies that the University should maintain an adjustable capacity buffer of about 20% of anticipated enrollment through part-time staff, borrowed classrooms, and combination classes, rather than trying to estimate the figures.

Third principle: choose a method that matches the purpose. As discussed in §4.7, if the goal is accurate forecasting for a single year, use M2 or ARIMAX; if the goal is planning resources across several consecutive years, use the combination of M2 with Random Forest, since its bias is roughly 6.8 times lower (478.54 / 70.33 ≈ 6.8; see Appendix A.2) at a cost of only 1.6% in accuracy. In either case, the model should be used only one to two years ahead and re-estimated every year as new enrollment data arrive.

6.3. Conditions under which the forecast holds

The forecast of 6,754 students for 2026 assumes 51 academic programs. The sensitivity is roughly 42 students per program, so if the actual number of programs turns out to be 45, the forecast falls by roughly 254 students; if the number of programs stays at 40, as in 2025, the forecast falls by roughly 466 students and the central estimate reverts to about 6,288 — close to the actual 2025 enrollment.

This needs to be clearly pointed out to users: most of the expected increase for 2026 comes from adding 11 programs rather than anything inherent in the series. If the current program portfolio does not change, the model projects enrollment to be essentially flat compared to 2025.

7. Discussion and limitations

7.1. Overall assessment of the model

M2 has four merits. The coefficients have the expected sign, and three of four are statistically significant. Multicollinearity is not an issue. No diagnostic test violates any OLS assumption. The model outperforms the naïve benchmark by 35.9% in error while ranking fourth among the eighteenth methods under consideration as the least complex specification.

On the other hand, the model has four natural weaknesses that need to be exposed.

The gap between in-sample and out-of-sample error. MAPE increases from 10.04% to 19.30%. The difference between the explanatory model and the predictive model [12] is highly significant in this context; the adjusted R^2 of 0.8092 reflects historical fit, not forecast performance.

A negative bias in forecasting. The model over-forecasts by an average of 478.54 students. As discussed in §4.9, this bias cannot be corrected by subtracting a constant; the feasible remedies are forecast combination or an asymmetric usage rule.

Stability of the trend coefficient. Leaving out one observation at a time shows that this coefficient ranges from 87.8 to 167.6 students per year depending on which observation is dropped (see Appendix A.4). A further check — dropping both 2011 and 2025 at once, the two observations at either end of the series — yields a coefficient of 33.4, which loses statistical significance ($p = 0.575$). Most of the evidence for the trend comes from these two observations.

The data do not resolve the trend's functional form. We tried three alternative specifications — quadratic, log ln(1+t), and damped $1 - 0.85^t$ — and all three fit as well as or better than the linear form. Because these forms diverge sharply when extrapolated further, use the model only for one- to two-year-ahead forecasting.

7.2. Comparison with the literature

Dwiansyah et al. [28] forecast graduate student enrollment at an Indonesian university using linear regression on the previous year's enrollment and report that it outperforms multilayer perceptron, k-nearest neighbors, decision tree, and Random Forest regressors—a pattern consistent with M2's advantage over tree-based methods in §4.6. Tapio and Tarepe [24] compare Random Forest against a hybrid ARIMA–Random Forest model for enrollment forecasting at a higher education institution in the Philippines, using 75 years of data (1949–2024); their hybrid model achieves a MAPE of 14.52%, against 15.62% for Random Forest and 16.95% for ARIMA.

This study adapts Tapio and Tarepe's hybrid design in a short-series setting, with a notably different outcome. In their study, ARIMA contributes substantially to the hybrid model. In the TDMU dataset, ARIMA degenerates into a random walk at every origin, so the corresponding hybrid must be built on a multivariate regression base rather than an ARIMA base. As a result, the M2-plus-Random-Forest hybrid reaches a MAPE of 18.19%, an improvement over M2 alone's 19.30% that does not, however, reach statistical significance.

A direct comparison of MAPE figures across the two studies carries little weight, given differences in series length (15 versus 75 observations), volatility structure, and validation design. Notably, the TDMU series contains two structural breaks within a very short span—a feature that makes forecasting substantially harder for every method—confirmed by the fact that none of the seventeen alternative methods, including three machine-learning methods, bring MAPE below 17.8%.

7.3. Data limitations and directions for future research

Sample size. Fifteen observations severely restrict the number of estimable parameters and the power of any test, such as the Diebold-Mariano test discussed in §4.8. With four independent variables, the residual degrees of freedom are 10.

Re-verification of the program-count series is necessary at the source. The series declines from 50 programs (2014) to 28 (2015) and then to 22 (2016), after another decline. If the swing represents a shift in how the programs were enumerated or even a one-time reorganization of the program portfolio, but not a shutting down and restarting of programs, the coefficient for this indicator is essentially combining two very different things.

Missing variables. The model does not include variables for applicant demand, competition among other institutions, quota admissions, admissions ratios, or regional pool sizes. The trend component of the current model serves as a proxy for whatever else changes over time that is not captured by the model.

In light of this, the most important area for future research is replacing the trend variable with meaningful covariates and constructing a panel data set organized by program or source region. A balanced panel with 15 years and about 40 programs will provide an order of magnitude more efficient observations than the current cumulative sample, allowing estimation of program fixed effects and separating program-mix variance from total variance. It is also the most direct way to overcome all three of the above-mentioned limitations.

8. Conclusion

A regression model called M2 with four independent variables, such as the number of academic programs offered, two regime adjustment variables, and a linear time trend, is estimated for the analysis and forecasting of the TDMU enrollment from 2011 through 2025.

The model attains $R^2 = 0.8637$, adjusted $R^2 = 0.8092$, and MAPE = 10.04% over the whole sample. Of the four independent variables, three are statistically significant at the 5% level. None of the diagnostic tests reject the OLS assumptions.

One-step-ahead rolling-origin validation over 2016–2025 yields a MAPE of 19.30% and reveals a negative bias of 478.54 students, while confirming that the model beats the random-walk benchmark with a 35.9% error reduction. The gap between in-sample and out-of-sample error is the most important practical result: the 19.30% figure, not 10.04%, is the appropriate basis for planning.

An extended comparison against seventeen alternative methods within the same validation framework supports three conclusions. First, the strength of M2 comes from its exogenous information source, not the inherent process of the series: No univariate method of extrapolation, even automatically ordered ARIMA that reduces to a random walk for any origin, beats the naïve forecast. Second, M2 is one of the four best methods and the easiest specification of all of them; three methods with lower error are derived from M2, and the differences are not statistically significant. Third, an obvious trade-off between accuracy and bias can be observed: combining M2 and RF reduces bias by 85.3% at the expense of 1.6% accuracy.

The real-time bias-correction experiment yields a negative result: error increases, and bias reverses sign; therefore, the model must address bias through forecast combination or asymmetric use, not by subtracting a constant.

Assuming 51 and 55 programs, the model forecasts 6,754 students for 2026 and 7,064 for 2027, with 95% prediction intervals of [4,684; 8,825] and [4,837; 9,291], respectively. When translated to planning units, the common error range amounts to 52 equivalent full-time faculty members in a student-faculty ratio of

25:1. In light of the fact that the forecast points lie beyond the observed data range and the prediction intervals are broad, we advise adopting the results as ranges rather than target points, restricting the forecast period to one or two years, and choosing the approach based on its purpose as outlined in §6.2.

References

[1] Sinuany-Stern, Z. (2021). Forecasting methods in higher education: An overview. In Z. Sinuany-Stern (Ed.), *Handbook of Operations Research and Management Science in Higher Education* (Vol. 309, pp. 131–157). Springer, Cham. DOI: 10.1007/978-3-030-74051-1_5

[2] On, V. V., To, B. T., & Tham, N. T. H. (2026). Forecasting university enrolment amid crisis and admission restructuring: A phase-aware regression model for a 15-year series (2011–2025). *International Journal of Artificial Intelligence and Machine Learning*, 6(9s), 1513–1528. DOI: 10.51483/IJAIML.6.9s.2026.1513-1528

[3] Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4), 679–688. DOI: 10.1016/j.ijforecast.2006.03.001

[4] Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3–4), 591–611. DOI: 10.1093/biomet/52.3-4.591

[5] Breusch, T. S., & Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, 47(5), 1287–1294. DOI: 10.2307/1911963

[6] MacKinnon, J. G., & White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29(3), 305–325. DOI: 10.1016/0304-4076(85)90158-7

[7] Durbin, J., & Watson, G. S. (1950). Testing for serial correlation in least squares regression. I. *Biometrika*, 37(3–4), 409–428. DOI: 10.1093/biomet/37.3-4.409

[8] Ljung, G. M., & Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65(2), 297–303. DOI: 10.1093/biomet/65.2.297

[9] Breusch, T. S. (1978). Testing for autocorrelation in dynamic linear models. *Australian Economic Papers*, 17(31), 334–355. DOI: 10.1111/j.1467-8454.1978.tb00635.x

[10] Godfrey, L. G. (1978). Testing against general autoregressive and moving average error models when the regressors include lagged dependent variables. *Econometrica*, 46(6), 1293–1301. DOI: 10.2307/1913829

[11] Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, 19(1), 15–18. DOI: 10.1080/00401706.1977.10489493

[12] Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25(3), 289–310. DOI: 10.1214/10-STS330

[13] Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting*, 16(4), 437–450. DOI: 10.1016/S0169-2070(00)00065-0

[14] Bergmeir, C., Hyndman, R. J., & Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, 120, 70–83. DOI: 10.1016/j.csda.2017.11.003

[15] Chatfield, C. (1993). Calculating interval forecasts. *Journal of Business & Economic Statistics*, 11(2), 121–135. DOI: 10.1080/07350015.1993.10509938

[16] Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and Practice* (3rd ed.). OTexts, Melbourne, Australia. https://otexts.com/fpp3/

[17] Hurvich, C. M., & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76(2), 297–307. DOI: 10.1093/biomet/76.2.297

[18] Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, 27(3), 1–22. DOI: 10.18637/jss.v027.i03

[19] Gardner, E. S., Jr., & McKenzie, E. (1985). Forecasting trends in time series. *Management Science*, 31(10), 1237–1246. DOI: 10.1287/mnsc.31.10.1237

[20] Assimakopoulos, V., & Nikolopoulos, K. (2000). The theta model: A decomposition approach to forecasting. *International Journal of Forecasting*, 16(4), 521–530. DOI: 10.1016/S0169-2070(00)00066-2

[21] Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55–67. DOI: 10.1080/00401706.1970.10488634

[22] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. DOI: 10.1023/A:1010933404324

[23] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189–1232. DOI: 10.1214 /aos/1013203451

[24] Tapio, R., & Tarepe, D. (2025). Comparative analysis of Random Forest and hybrid ARIMA–Random Forest models for student enrollment forecasting in higher education. *Journal of Advances in Mathematics and Computer Science*, 40(3), 124–136. DOI: 10.9734/jamcs/2025/v40i31982

[25] Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 253–263. DOI: 10.1080/07350015.1995.10524599
[26] Harvey, D., Leybourne, S., & Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, 13(2), 281–291. DOI: 10.1016/S0169-2070(96)00719-4
[27] Timmermann, A. (2006). Forecast combinations. In G. Elliott, C. W. J. Granger, & A. Timmermann (Eds.), *Handbook of Economic Forecasting* (Vol. 1, pp. 135–196). Elsevier. DOI: 10.1016/S1574-0706(05)01004-9
[28] Dwiansyah, A., Fahrurrozi, I., Fakhrurrifqi, M., Farooq, U., & Alfian, G. (2026). Forecasting graduate student enrollment in a university using regression analysis. *Bulletin of Electrical Engineering and Informatics*, 15(2), 1347–1355. DOI: 10.11591/eei.v15i2.9713