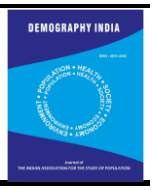
Demography India
A Journal of Indian Association of Study of Population
Journal Homepage: https://demographyindia.iasp.ac.in/



Forecasting Scheduled Tribe Student Enrollment in India Using the ARIMA Algorithm: A Data Mining Approach

Suman Pal[1] and Vinod Kumar[2*]

Keywords

AIC, Algorithm, ARIMA, Data Mining, Enrollment Forecasting, IHES, ST Enrollment Trend

Abstract

Background and purpose: Forecasting is crucial in strategic decision-making processes that facilitate effective and efficient management practices. In the context of higher education institutions (HEIs), mining and forecasting prove advantageous in their ongoing endeavor to analyze historical, current, and longitudinal data associations about specific scenarios within an educational framework. The primary objective of this study is to forecast the enrollment figures of students belonging to Scheduled Tribes (STs).

Methods: This study employed the widely recognized ARIMA (p, d, q) model to forecast the total STs' students' enrollments for the academic sessions from 2022-23 to 2030-31, utilizing the comprehensive enrollment time series data of Schedule Tribes (ST) students from the years 2012-13 to 2021-22 to the Indian Higher Education System (IHES). The SPSS, JMP, R studio, and MS Excel (V. 21) software are used for the analysis. The various combinations of p, d, and q parameters were evaluated, with the model yielding the lowest Akaike Information Criterion (AIC) value selected for the predictive analysis.

Key Results: The simulation outcomes indicated that the ARIMA (0,1,2) model emerged as the statistically appropriate model for forecasting enrollment trends in the IHES for ST community students. The forecast revealed a noticeable upward trend in enrollment figures for the subsequent academic years up to 2030-31, with a total of 113% increasement in enrollment of the STs in Indian Higher Education.

Implications: The findings of this study are anticipated to yield several advantages, including socio-educational policy perspectives, enhancement of the current admission and retention policies for STs' student, and the facilitation of central decision-making regarding the enrollment management framework of IHES. The forecasted enrollment figures may also serve as a foundational basis for advocating the construction of additional classroom facilities and faculty to accommodate the rising demand.

[1] ***Ex- Research Scholar,*** *School of Education, Mahatma Gandhi Antarrashtriya Hindi Vishwavidyalaya, Wardha, Maharashtra, India, PIN- 442001. E-mail Id: suman.mgahv777@gmail.com ORCID: https://orcid.org/0009-0003-6108-2337*
**Corresponding Author*
[2]***Assistant Professor****, Faculty of Education, Udai Pratap (Autonomous) College, Varanasi, India, PIN- 221002. E-mail Id: dr.vinodpal777@gmail.com ORCID: https://orcid.org/0000-0003-2383-0640*

Introduction

India cultivates a distinctive identity on the global stage through its multifaceted cultural paradigm, epitomized by the philosophy of 'Vasudhaivkutumbakam', which signifies the global family ethos. India is not merely characterized by its eclectic culture, but also possesses extensive geographical dimensions. The geographical narrative of India suggests its potential to emerge as a formidable nation within South Asia, assessed not only from a spiritual perspective but also from an economic standpoint. The nation, characterized by the world's biggest democratic governance, showcases a remarkable cultural heterogeneity and social cohesion that collectively manifest a unique cultural solidarity, recognized as a significant facet of its global identity. In India, one is unlikely to find a singular university or college where individuals from only one specific culture or community engage in education and instruction, thus reflecting the nation's inherent diversity. The Indic institutions within India also embody their unity in diversity. Indian society is divided into five social groups, which are the General Class, Economically Weaker Section (EWS), Other Backward Classes (OBC), Scheduled Castes (SC), and Scheduled Tribes (ST). The ST category ranks last in the Indian social classification based on socio-educational status. These social groups are delineated constitutionally into unreserved and reserved categories. EWS, OBC, SC, and ST belong to the reserved category, and the general class belongs to the unreserved category. The unreserved category encompasses all individuals irrespective of their class, that is, open to all categories. The reserved category permits access to educational institutions for specific individuals who possess particular identities, thereby accommodating their socio-cultural convictions. The IHES follows the Indian constitutional provisions at the time of admission in letter and spirit. While ensuring the application of all the provisions for reserved category students mentioned in the Indian Constitution, the IHES ensures maximum access to higher education to all categories of students.

Data mining (DM), alternatively referred to as knowledge discovery (KD), constitutes the systematic process of deriving implicit information or knowledge from extensive databases, a practice rooted in the discipline of statistics, which integrates mathematical methodologies and machine learning algorithms (A. J. P. Delima et al., 2019; Huda, n.d.; Suresh et al., 2016). The utilization of DM or KD is fundamentally contingent upon the specific research question that the investigator aims to address. Within the realm of Higher Educational Institutions (HEIs), where the application of DM or KD is notably prevalent, there exists a significant focus on analyzing student enrollment data to identify patterns and discern potential factors influencing a student's decision to enroll in a particular institution (P, 2020). Prediction, identified as one of the frequently employed data mining methodologies within the academic literature, is recognized as an effective strategy for HEI administrators to generate insights that can inform decision-making processes (Abideen et al., 2023; Haris et al., 2016). The role of prediction is pivotal in HEI administration as it facilitates the extraction of inferences pertinent to decision-making and policy formulation endeavors. Both administrative personnel and researchers possess the flexibility to implement

predictive analytics across a spectrum of challenges with varying degrees of complexity. When applied to student enrollment, prediction primarily enhances comprehension of enrollment trends, thereby contributing to an understanding of the substantial implications of enrollment on revenue-related outcomes. The insights derived from this analysis could significantly inform future strategic planning and resource allocation decisions (Abideen et al., 2023; Masinading et al., 2024; Ward, 2007). The advantages associated with knowledge extraction encompass the understanding of patterns, associations, temporal changes, critical structures, and the detection of anomalies.

Forecasting holds significant importance for both strategic and tactical decision-making, which is regarded as a fundamental component of efficient and effective management practices (Haris et al., 2016). The act of prediction is typically advantageous for analyzing the current, historical, and ongoing data relationships of specific circumstances. The appropriate prediction methodologies will yield a high degree of accuracy at a reduced cost, taking into account the volume of historical data, whereby the relevance of the information can be assessed for dependable decision-making; additionally, the temporal framework necessary for formulating the prediction is crucial (Gareth Shepherd et al., 2024; Waddell & Sohal, 1994). Data mining methodologies could be employed to enhance the prediction processes within higher education institutions (Ahmed, 2024; Delavari et al., 2008; Velasco et al., 2023; Wachowicz et al., 2024). The forecasting of student enrollment is advantageous for Higher Education Institutions (HEIs) as the insights derived can enhance the implementation of optimal decision-making frameworks essential for future strategic planning. Nonetheless, the process of identifying the appropriate predictive methodologies for a specific issue remains a challenge, as the selection of data mining algorithms is contingent upon the accessibility of the data pertinent to the HEI. Stationary ARIMA models are capable of modelling time series that exhibit uniform fluctuations around a fixed mean level, as well as those that are non-stationary (Box et al., 1994). In this paper, the well-known ARIMA (p,d,q) algorithm, which constitutes a form of time series analytical model, is employed due to the utilization of a univariate historical dataset about ST enrollees from India, covering the academic years from 2006-07 to 2021-22, which was sourced from the website of Ministry of Education (MoE), Government of India (GoI). Various ARIMA (p, d, q) models were evaluated, and the most suitable model for forecasting was identified and selected from these assessments. Enrollment data related to ST category students in Indian higher education for the academic years from 2022-23 to 2030-31 were forecasted.

The primary objective of this study was to develop an ARIMA (p, d, q) algorithm for Higher Education Enrollment Forecasting (HEEF) using data mining techniques, which will enable the maximization of accurate enrollment forecasts with minimal error. Similarly, another objective of this study was to project the anticipated enrollment figures of ST category students in India for the next 10 academic years, specifically covering the academic years from 2022-23 to 2030-31.

Literature Review

Enrollment forecasting is a crucial strategic and operational task for higher education institutions (HEIs), especially given the swiftly shifting demographic

trends, socio-economic factors, and national policy requirements like India's National Education Policy (NEP) 2020. Precise enrollment forecasts allow higher education institutions to distribute resources effectively, develop infrastructure, and formulate policies that align with objectives of access, equity, and quality. An expanding collection of international literature has examined different statistical and machine learning methods for predicting enrollment, including traditional time-series models and deep learning frameworks. Nonetheless, there is still a lack of context-relevant studies aimed at the Indian higher education landscape, which poses distinct challenges because of its diversity and intricacy. This analysis explores current literature in both international and regional contexts, recognizes methodological advancements, and points out research deficiencies that guide the present investigation. A fundamental method for enrollment forecasting is the application of traditional time-series models. Among these, the ARIMA (Autoregressive Integrated Moving Average) model has attracted considerable interest because of its reliability and ease of interpretation. Mao et al. (2024) detail that recent advancements in the collection and analysis of sequential educational data have placed time series analysis at a pivotal point in educational research, highlighting its essential role in fostering data-driven decision-making. The predictive studies regarding the enrollment rates of general education with vocational education are crucial for improving regional talent structure and the reorganization of industry. Rahman & Hasan (2017) suggest that several Autoregressive Integrated Moving Average (ARIMA) models were developed to forecast carbon dioxide emissions using time series data covering forty-four years from 1972 to 2015. It has been observed that within the various ARIMA models, ARIMA (0, 2, 1) stands out as the best-fitting model for forecasting carbon dioxide emissions in the setting of Bangladesh.

P et al. (2020) employed the well-known ARIMA (p,d,q) model to project overall student enrollment numbers for the academic years 2019–2020 to 2024–2025, using the extensive enrollment data from the university for the years 2011–2012 to 2018–2019. Different combinations of p, d, and q were evaluated, and the model with the lowest Akaike Information Criterion (AIC) value was chosen for forecasting. The simulation results showed that the ARIMA (0,2,1) model was statistically confirmed as the best model for predicting university enrollment. A methodology for time series modeling was utilized to forecast student enrollment numbers in basic public educational institutions throughout Ghana. A detailed dataset consisting of fifty-four data points was obtained from the enrollment records spanning the years 1961 to 2014, provided by the Ghana Ministry of Education. The results indicated that the ARIMA (0,2,2) model, identified using the Akaike Information Criterion (AIC), predicted an increase in enrollment over the next five years, followed by a steady decrease in later years (Baker et al., 2017). Dela Cruz et al. (2020) used the ARIMA model to predict student enrollment at Cebu Technological University in the Philippines. The research utilized ARIMA (0,2,1) as the ideal setup due to the minimum Akaike Information Criterion (AIC). The prediction indicated a rising enrollment trend from 2019 to 2025 and provided practical guidance for campus planning, such as infrastructure improvements and faculty recruitment. Similarly, Sabanal (2023) employed various time-series methods—

including linear, polynomial (sextic), exponential, and autoregressive approaches—to examine enrollment statistics from Saint John Berchmans High School in the Philippines. The research identified the sextic polynomial trend model as the most precise, achieving an R^2 value greater than 0.89, and highlighted the effect of independent variables like gender ratio, strand specialization (e.g., ABM vs. HUMSS), and year-level distribution on enrollment trends. Shen et al. (2024) presented a multiregional, probabilistic cohort-component model for the Australian Capital Territory. Their School Transition Estimation and Projection (STEP) model included multiple demographic and administrative elements like migration, grade advancement, and preschool enrollment. A significant advancement was the incorporation of Monte Carlo simulations to generate prediction intervals, thus measuring uncertainty in forecasts. The model was confirmed through reliable administrative data from 2009 to 2023 and showed robust in-sample performance. While this model targets primary and secondary education, its foundational logic provides a useful benchmark for Indian higher education institutions. Gnoh et al. (2024) performed a comparative study of Prophet from Facebook, LSTM networks, and Polynomial Regression. Analyzing enrollment data from Malaysian higher education institutions, the research demonstrated that LSTM surpassed the other models. They also created an interactive dashboard for visualizing predictions, which is highly relevant for Indian HEIs. LSTM networks effectively capture long-term relationships in sequential data, rendering them suitable for recognizing intricate, non-linear patterns related to student enrollment influenced by various policy and demographic elements. Numerous studies have investigated hybrid approaches by combining data mining techniques with statistical prediction. For example, support vector machines, decision trees, and neural networks have been utilized to identify trends in student demographics and academic success that affect enrollment. Dela Cruz et al. (2020) and related research indicate that incorporating data mining enhances forecasting accuracy.

Although there is a strong body of global literature, the use of enrollment forecasting models in India is still needed. Many studies fail to break down social categories, concentrate on individual campuses, and do not correspond with The NEP 2020 aims for a GER of 50% by 2035. The literature review indicates that no significant research has applied ARIMA, LSTM, or Prophet models to Indian HEI enrollment data in my knowledge. Lack of an equity perspective, the majority of studies do not break down forecasts by caste, gender, or marginalized groups, no current forecasting model matches NEP 2020 or GER objectives, and no decision support with dashboard capabilities. Indian higher education institutions lack resources for visual forecasting assistance. Insufficient emphasis on forecasting uncertainty, the absence of prediction intervals, and a lack of forecasting ARIMA modeling are noted in Indian higher education enrollments. International literature emphasizes the importance of forecasting in managing student movements and strategic planning. Although global models provide technical capabilities, India requires solutions tailored to its context that are inclusive and aligned with policy. This research tackles these deficiencies by utilizing ARIMA models on data from

Indian HEIs, breaking down forecasts, and conforming to NEP 2020's goals.

Methods and Materials

The data illustrated in Figure 1 reveals a non-stationary pattern characterized by an upward trend. To select an appropriate stochastic model, we adhered to the four phases of identification, estimation, diagnostic verification, and forecasting, as suggested by Box and Jenkins (1976). The formulation of ARIMA is meticulously aligned with the Box–Jenkins methodology for time series analysis, encompassing the processes of identification, estimation, diagnostic verification, and forecasting as outlined in the Box–Jenkins framework. The diagnostic verification is employed to ascertain whether the time series data conform to the conditions of stationarity and goodness of fit (Shim et al., 1994; Fuller, 1976). The initial phase in the application of the ARIMA methodology involves an assessment of stationarity. "Stationarity" denotes that the series maintains a relatively stable level over time. In the absence of these stationarity conditions being fulfilled, numerous calculations pertinent to the process become infeasible. During the identification phase, we delineate the response series and determine potential ARIMA models applicable to it. Should a graphical representation of the data suggest non-stationarity, the "differencing" technique ought to be implemented on the series. Differencing serves as an effective method for converting a non-stationary series into a stationary one. This transformation is executed by subtracting the observation from the current period from that of the preceding one. If this transformation is conducted only once on a series, we refer to the data as having been "first differenced." This procedure effectively mitigates the trend, provided that the series is expanding at a relatively uniform rate. In the estimation and diagnostic verification phase, we employ the estimate statement to delineate the ARIMA model to be fitted to the variable and to estimate the parameters of that model. The estimate statement simultaneously generates diagnostic statistics that assist in evaluating the model's adequacy. Significance tests for parameter estimates reveal whether certain terms within the model may be extraneous. In the forecasting phase, we utilize the forecast statement of the ST students' enrollment and to construct confidence intervals for these projections derived from the ARIMA (0,1,2) model established by the preceding estimate statement. The materials (datasets) and various components of the methods used in this study are presented below simultaneously.

Time series Dataset

The Dataset used in this study is the historical data of the overall enrollment from 2012-13 to 2021-22 of ST category students in the Indian higher education system. The time series datasets were obtained from Table 14 of the report of the All India Survey on Higher Education (AISHE) from each academic session from 2012-13 to 2021-22, which is available on the official website (https://aishe.gov.in/aishe-final-report/) of the Department of Higher Education, Ministry of Education, Govt of India. The dataset was authenticated and reliable because it was carefully compiled from reputable government source.

Data Analysis procedure

Secondary time-series data was utilized in this study to determine the number of STs' students enrolled in Indian

higher education for the academic years 2022-23 to 2030-31. The analysis process was carried out with the help of three software tools—SPSS, Microsoft Excel, and R. Each software with specific analytical purposes were used to analyze the data. To verify that the collected dataset met the requirements for successful time-series modeling, the stationarity of the time-series dataset was assessed using unit root testing in the preliminary stage using SPSS (v. 23). Following that, Microsoft Excel was used to create graphical presentation of enrollment trends, which allowed for a clear visualization of enrollment trends with yearly time frames. After that, we used JMP and R studio software for advanced time-series analysis, which involved model optimization, making Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) correlograms, calculation and presentations of ACF and PACF residuals, and future enrollment forecasting of STs. We used multiple software tools approach for combine statistical validation, clear visual presentation, and strong forecasting methods in this study.

ARIMA Algorithm

The Box and Jenkins (BJ) Methodology is technically known as the ARIMA equation modeling. The ARIMA Methodology neither constructs a single nor a simultaneous equations model, but analyses the probabilistic or stochastic properties of time series data. In this model, four steps are Identification, Estimation, Diagnostic checking, and Forecasting. The ARIMA(p,d,q) model is used in time series forecasting. The p variable denotes the autoregressive order (Number of autoregressive terms), d represents differenced t times (Number of times the series is highest to be differenced before it becomes stationary), and q represents the moving average order. The ARIMA(p,d,q) model is shown in equation (1) below:

$$\emptyset(B)\,(w_t - \mu) = \theta(B)a_t \quad (1)$$

where t is represented as the time index and backshift operator for symbol B, the autoregressive parameter is assigned as $\emptyset(B)$, $\theta(B)$ for MA, w_t for d value in the ARIMA(p,d,q) model, and a_t for white noise (A. J. Delima, 2019; P, 2020).

Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF)

The basis of the ARIMA p and q assignment is determined through its ACF and PACF plots. The equation is stressed as:

$$P_k = \sum^{r}_{t=k+1}(Y_t - \bar{Y})(Y_{t-k} - \bar{Y}) / \sum^{r}_{t=1}(Y_t - \bar{Y})^2 \quad (2)$$

where $\boldsymbol{P_k}$ denotes the ACF coefficient in lag k, the observed period is expressed as t, while observations in period t are denoted by y_t. The $\bar{Y}$ denotes the mean, and the observation in t-k is expressed as Y_{t-k} (A. J. Delima, 2019; P, 2020). The autoregressive (p) order is represented in the Partial Autocorrelation Function (PACF) plot, while the Autocorrelation Function (ACF) plot denotes the moving average (q) of the model. (P, 2020).

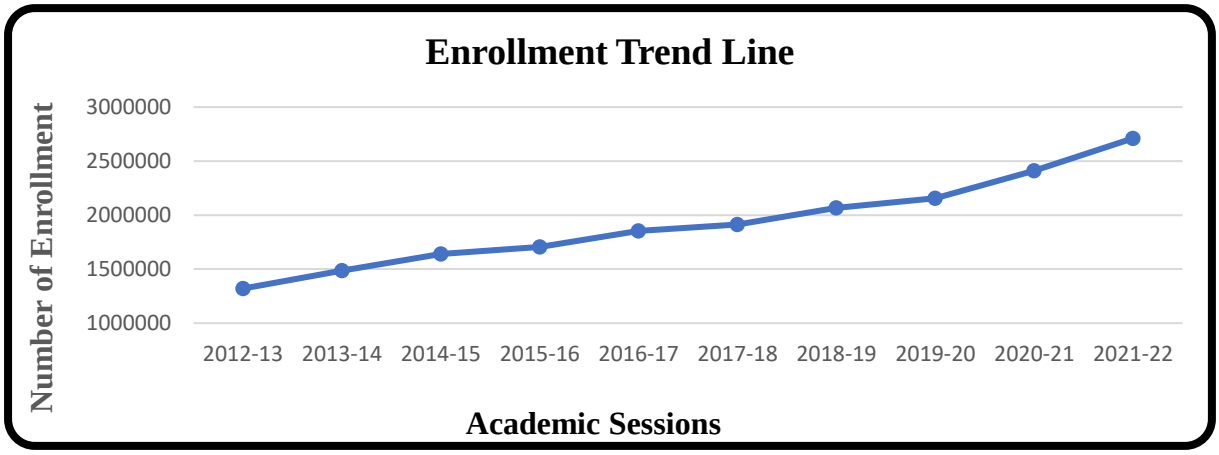
Akaike Information Criterion (AIC)

For decision-making of the appropriate model to be used to forecast, one criterion is the Akaike Information Criterion (AIC) (Akaike, 1974). Other various methods exist, like Schwarz's information criterion (SIC) (Schwarz, 1978). In this paper, the author used the widely accepted AIC. The decision is made based on the AIC's lowest value parameter for appropriate forecasting. The AIC is expressed as an equation (3) below:

$$AIC = 1n\,[\{(\sum^{t}_{i=1} \hat{\varepsilon}_i^2)/(T\text{-}n)\} + 2n\,/\,T] \quad (3)$$

Where $\hat{\varepsilon}_i^2$ Denotes squared residual estimates, T denotes observation size within samples, and N denotes the estimated parameters (Molebatsi & Raboloko, 2016; P, 2020).

Results



Source: Created by Author

Figure 1 Enrollment Trends in the ST Category Over Time

Figure 1 shows an upward trend in the ST community enrollment time series plot. This is the most powerful evidence of the increasing number of ST enrollees over time, which means the data series is non-stationary.

Stationarity Testing

A stationary process is characterized by a mean and variance that remain constant over time, devoid of any upward or downward trends. To facilitate the estimation of an ARIMA model, the series must be stationary. So, we employed the Augmented Dickey-Fuller (ADF) test to ascertain the presence of stationarity within the dataset of ST students' enrollment in India, approached from a quantitative perspective. For the ADF test, the hypotheses are formulated as follows:

H_0: The ST students' enrollment data series is not stationary.

H_1: The ST students' enrollment data series is stationary.

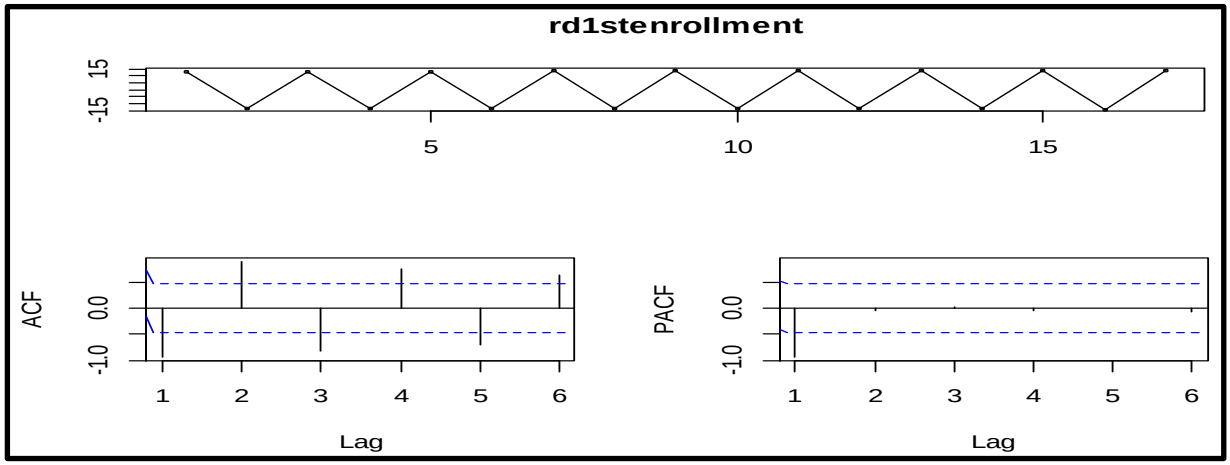
In conducting the stationarity test, a significance level of 95% is utilized to test the null hypothesis. A p-value that is less than 0.05 signifies a rejection of H_0, thereby indicating the acceptance of H_1, which implies that the series is stationary. Otherwise, if the p-value exceeds 0.05, we are unable to reject H_0, indicating that the data series is not stationary.

Table 1 Augmented Dickey-Fuller (ADF) Test Report

Original Dataset		Level		1st Difference	
DF	p-value	DF	p-value	DF	p-value
0.48353	0.99	-2.1941	0.4984	-5.4549	0.01

Source: Created by Author

From the perusal of Table 2, it is clear that the DF is 0.48353 and the p-value is 0.99 of the original datasets. The p-value of the original datasets is greater than 0.05 (0.99 > 0.05). In this way, we are unable to reject $H_{0,}$ so we retain $_{H0,}$ which means the original data series is not stationary. So, first of all, we make the data series stationary by providing various values of d and perform the ADF test with the help of R Studio. The p-value of the log return file is 0.99 (see Table 2), which is greater than 0.05 (0.99 > 0.05). That means the log return file has no constant variance. The p-value of 1st difference file is 0.01 (see Table 2), which is less than 0.05 (0.01< 0.05). That means the 1st differenced file has no unit root (data series is stationary). In this way, the data series is converted into a stationary time series after 1st difference. So, it is clear that the order of d is 1. The trend line and ACF & PACF correlogram plots of the 1st difference data file (rd1stenrollment) are presented in Figure 2 below:



Source: Created by Author with the help of R Studio

Figure 2 Trend line and ACF & PACF Correlogram

Figures 2 show the ACF and PACF correlogram plots for MA (q) and AR (p) values for lag 1 to 6 simultaneously. The trend line (see Figure 2) and ADF test report (see Table 2) show enough evidence to determine that the 1st difference file has a constant variance, which means this file is stationary. From the perusal of Figure 2, it is clear that the ACF plot showed a mixture of exponential decay patterns throughout the lags, implying that the required ARIMA model is an ARIMA (P,1,0), and the PACF plot shows that the 1st lag is significant. This means the AR (p) value is also 0. So, the best ARIMA model is obtained by comparing various ARIMA models based on their AIC value. Therefore, we manipulate the value of d in the ARIMA model. The various ARIMA (p,d,q) models are presented in Table 3 with the AIC value below:

Table 2 ARIMA (0,1,0) and ARIMA (0,2,0) based on AIC

ARIMA (0,1,0)	AIC	ARIMA (0,2,0)	AIC
(0,1,0)	153.30	(0,2,0)	164.64
(0,1,1)	136.87	(0,2,1)	149.59
(0,1,2)	122.35	(0,2,2)	136.16

Source: Created by Author

From the perusal of Table 2, it is clear that when the value of d is increased from 1 to 2, the AIC value of the ARIMA (0,1,2) is found to be lowest. In this way, the ARIMA (0,1,2) model can be selected as the best-fit model for enrollment forecasting in the ST category in India, which is presented in bold letters in Table 3.

Model Construction

The author developed the ARIMA (0,1,2) model with the help of R Studio. The Coefficients of MA with the standard error of the proposed best model are presented in Table 4 below:

Table 3 Coefficients of Proposed Models

Best Fit Model		AR1	AR2	MA1	MA2
ARIMA (0,1,2)	AR/MA	NA	NA	-1.9022	0.9999
	Std. Error	NA	NA	0.3155	0.2192

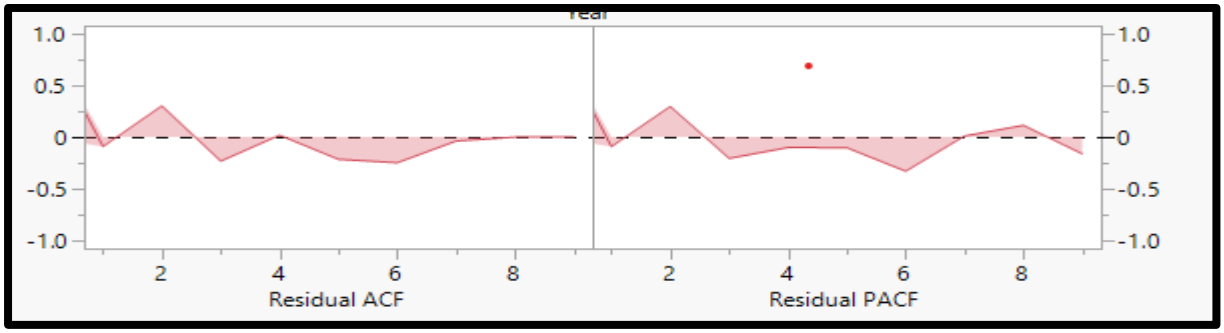
Source: Created by Author

With the help of Table 4, the best-fit ARIMA (0,1,2) model can be expressed as follows:

$$X_t = -1.9022W_{t-1} + 0.9999W_{t-2} \quad (1)$$

Model Diagnostic check:

For the diagnostic checking of the created model, the author used the ACF plot of residuals. The residuals ACF plot is presented in Figure 3 below:



Source: JMP software output of residuals ACF plot

Figure 3 Residuals ACF Plot

In Figure 3, the time plot of the ST category enrollment forecast residuals (errors) shows that the variance of the forecast residuals is constant over time. This means the time series shows that the forecast residuals are normally distributed and the mean of the residuals is close to zero. Therefore, it is determined that the forecast residuals are distributed normally with zero mean and the variance of the residuals is constant. The ACF plot of residuals shows that there are non-autocorrelations in the forecast residuals. Thus, the ARIMA (0,1,2) seems to provide an adequate predicted model for the ST category enrollment in the Indian Higher Education System.

Enrollment Forecasting from 2022-23 to 2030-31 sessions

After the diagnosis of the ARIMA (0,1,2) model, the author forecasts enrollment in the ST Cotegary up to the academic sessions 2030-31 with the help of JPM software. The output of the enrollment forecast for the ST category can be seen in Table 5 and the graphical presentation in Figure 4.

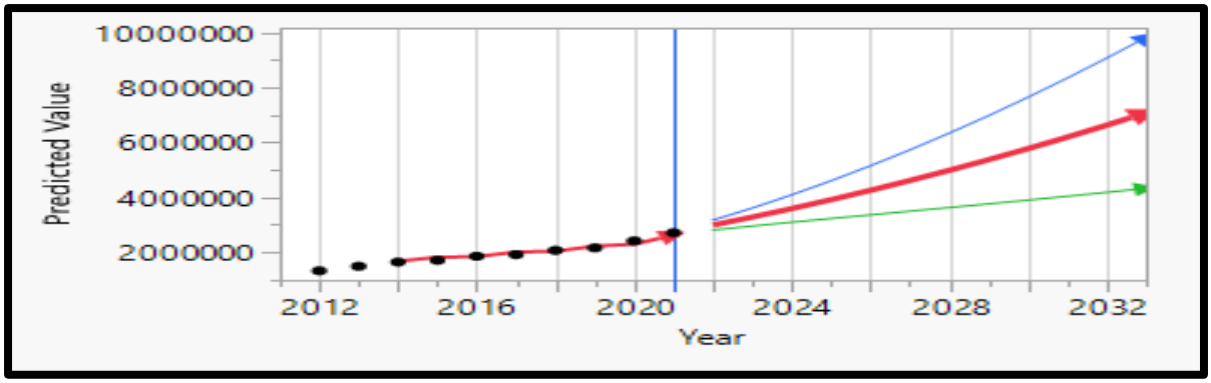


Figure 4 Forecasted Enrollment from 2022-23 to 2030-31

Table 4 Details of Forecasted Enrollment from 2022-23 to 2030-31

Academic Sessions	UCL	LCL	Std. Error	Predicted Enrollment	Yearly Increases (in %)
2022-23	3167040.17	2808740.58	91404.63	2987890	-
2023-24	3610922.02	2950828.24	168394.36	3280875	9.81
2024-25	4089366.27	3089898.25	254971.02	3589632	9.41
2025-26	4602441.67	3225881.86	351169.67	3914162	9.04
2026-27	5148969.99	3359957.34	456389.17	4254464	8.69
2027-28	5727764.98	3493310.86	570024.28	4610538	8.37
2028-29	6337788.04	3626981.08	691545.09	4982385	8.07
2029-30	6978149.60	3761857.56	820497.74	5370004	7.78
2030-31	7648086.04	3898703.92	956492.61	5773395	7.51
Total increase in enrollment (from 2022-23 to 2030-31) 113%					

Source: Created by Author

Table 4 indicates that the forecasted enrollment of the ST category in Indian higher education institutions will reach 5,773,395 by the academic session 2030–31, representing an approximate 113% increase compared to the 2022–23 enrollment. The table also presents the projected year-on-year percentage increases in enrollment from 2022–23 to 2030–31, which are 9.81%, 9.41%, 9.04%, 8.69%, 8.37%, 8.07%, 7.78%, and 7.51%, respectively.

Discussions

This study had explored the main goals, to create an effective ARIMA (p,d,q) model to forecast enrollment trends for Scheduled Tribe (ST) students in the Indian Higher Education System (IHES) and to project enrollment numbers for this group over the next ten years, covering academic sessions from 2022–23 to 2030–31. These goals were aimed at helping to support policy development and educational planning for marginalized communities in India. By using a time series modeling framework, we applied the ARIMA algorithm to analyze

time series enrollment data from 2012–13 to 2021–22. The initial analysis showed a non-stationary upward trend, indicating that participation among ST students in higher education had been rising over the past decade. To ensure the model's reliability, the stationarity of the data was tested using the Augmented Dickey-Fuller (ADF) test. It was found that the original series was non-stationary, requiring first-order differencing to achieve stationarity, which is necessary for ARIMA modeling. After first-order differencing, both the trend line and ACF/PACF plots confirmed that the transformed data were suitable for building the model. Various ARIMA models were evaluated using the Akaike Information Criterion (AIC), with the model having the lowest AIC value considered the best fit. Through this AIC-based evaluation, the ARIMA (0,1,2) model was identified as the most optimal and appropriate forecasting model for the ST students' enrollment data. The forecasts from the ARIMA (0,1,2) model showed a steady increase in ST students' enrollment in the IHES, with estimates rising from 2,710,678 in 2021–22 to 5,773,395 by 2030–31. This indicates a total increase of approximately 113% showcasing significant improvement in access to higher education for ST communities. A diagnostic check on the model residuals further supported the model's strength. The residuals showed constant variance, a near-zero mean, and no significant autocorrelation, confirming the model's readiness for prediction. In summary, the results of this study add valuable insights to the current literature on enrollment forecasting in higher education. The use of the ARIMA (0,1,2) model offers a solid and data-backed way to forecast enrollment trends. These findings are crucial for educational administrators and policymakers looking to improve access, support fairness, and allocate resources effectively in higher education for the most marginalised groups, such as ST communities in India.

Conclusion

This study used time series forecasting with the ARIMA (p,d,q) model to analyze and predict enrollment trends of Scheduled Tribe (ST) students in the Indian Higher Education System (IHES). It utilized enrollment data from 2012 to 2022, systematically applying the Box–Jenkins method to identify, evaluate, and select the most suitable ARIMA model for forecasting. After performing stationarity tests and model evaluation using the Akaike Information Criterion (AIC), the ARIMA (0,1,2) model was determined to be the best fit for the dataset. This model effectively captured key patterns in the data and satisfied all necessary statistical criteria, confirmed through residual diagnostic checks. The forecast from this model indicates a steady upward trend in ST student enrollment, projecting an increase from 2.71 million in 2021 to approximately 5.77 million by the 2030–31 academic year, representing nearly a total of 113% rise. This expected increase in enrollment has significant implications for policy development, resource allocation in higher education institutions, and the promotion of inclusive teaching strategies. The findings highlight the effectiveness of data-driven methods like ARIMA in educational forecasting and planning. By providing clear and statistically grounded predictions, this study offers valuable insights to support informed decision-making and enhance access, equity, and growth in Indian higher education, especially for marginalized ST communities. Future research could extend this methodology to other social groups or

incorporate combined data mining models to further improve forecast accuracy and inform targeted educational interventions.

Acknowledgment

The authors extend their sincere appreciation to the refereed and anonymous reviewers of *Demography India* for their time, rigorous evaluation, thoughtful and constructive suggestions on this paper. Their detailed thoughtful comments and critical insights have significantly enhanced the scholarly quality, clarity, and overall contribution of this study; and also helped us refine the interpretation of our findings, making the study more relevant for both academic and policy audiences.

Declarations of AI Uses

In accordance with COPE guidelines, the authors declare that generative AI tools (ChatGPT, Grammarly, and SciSpace) were used exclusively for grammar checking, paraphrasing, and sentence restructuring. No AI tools were used for the generation of research ideas, study design, data collection, statistical analysis, or interpretation. The authors take full responsibility for the integrity and accuracy of all content.

Authorship Declaration

All authors confirm that they have substantially contributed to the conception, design, analysis, and/or interpretation of the study, participated in drafting or revising the manuscript, and approved the final version for submission. There are no conflicts of authorship, and all authors agree on the order of authorship as presented in the manuscript.

Funding Details

The authors declare that no financial support was received for the research, authorship, and/or publication of this paper.

Data Availability Statement

The data used in this study are publicly available from official government website (https://aishe.gov.in/aishe-final-report/) of the Department of Higher Education, Ministry of Education, Govt of India. The authors made no changes to the original data except for standard formatting required for analysis.

References

Abideen, Z. ul, Mazhar, T., Razzaq, A., Haq, I., Ullah, I., Alasmary, H., & Mohamed, H. G. (2023). Analysis of Enrollment Criteria in Secondary Schools Using Machine Learning and Data Mining Approach. *Electronics, 12(3), Article 3*. https://doi.org/10.3390/electronics12030694

Ahmed, E. (2024). Student Performance Prediction Using Machine Learning Algorithms. *Applied Computational Intelligence and Soft Computing, 2024(1),* 4067721. https://doi.org/10.1155/2024/4067721

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19(6),* 716–723. IEEE Transactions on Automatic Control. https://doi.org/10.1109/TAC.1974.1100705

Baker, J., Swanson, D. A., Tayman, J., & Tedrow, L. M. (2017). Forecasting School Enrollment Size and Composition. In J. Baker, D. A. Swanson, J. Tayman, & L. M. Tedrow (Eds.), Cohort Change Ratios and their Applications (pp. 107–118). Springer International Publishing. https://doi.org/10.1007/978-3-319-53745-0_7

Box, G.E.P., & Jenkins, G.M. (1976). Time Series Analysis: Forecasting and Control (rev. ed.). San Francisco: Holden-Day.

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. https://doi.org/10.5539/ies.v15n6p109

Dela Cruz, A. P., Basallo, M. L. B., Bere, B. A., Aguilar, J. B., Calvo, C. K. P., Arroyo, J. C. T., & Delima, A. J. P. (2020). Higher education institution (HEI) enrollment forecasting using a data mining technique. *International Journal of Advanced Trends in Computer Science and Engineering, 9(2),* 2060–2064. https://doi.org/10.30534/ijatcse/2020/179922020

Delavari, N., Phon-Amnuaisuk, S., & Beikzadeh, M. R. (2008). Data Mining Application in Higher Learning Institutions. *Informatics in Education - International Journal, 7, No. 1,* 31–54.

Delima, A. J. (2019). Application of Time Series Analysis in Projecting the Philippines' Electric Consumption. *International Journal of Machine Learning and Computing, 9.* https://doi.org/10.18178/ijmlc.2019.9.5.860

Delima, A. J. P., Sison, A. M., & Medina, R. P. (2019). Variable Reduction-based Prediction through Modified Genetic Algorithm. *International Journal of Advanced Computer Science and Applications (IJACSA), 10(5),* Article 5. https://doi.org/10.14569/IJACSA.2019.0100544

Fuller, W. A. (1976). Introduction to Statistical Time Series, NY: John Wiley & Sons, Inc., 3–6.

Gareth Shepherd, N., Lou, B., & Maynard Rudd, J. (2024). Going with the gut: Exploring top management team intuition in strategic decision-making. *Journal of Business Research, 181,* 114740. https://doi.org/10.1016/j.jbusres.2024.114740

Gnoh, H. Q., Keoy, K. H., Iqbal, J., Anjum, S. S., Yeo, S. F., Lim, A. F., Lim, W. L., & Chaw, L. Y. (2024). Enhancing business sustainability through technology-enabled AI: Forecasting student data and comparing prediction models for higher education institutions. *paperASIA, 40(2b),* 48–52. https://doi.org/10.59953/paperasia.v40i2b.86

Haris, N. A., Abdullah, M., Hasim, N., & Rahman, F. A. (2016). A Study on Students' Enrollment Prediction using Data Mining. Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication, 1–5. https://doi.org/10.1145/2857546.2857592

Huda, M. (n.d.). DECISION SUPPORT SYSTEM OF SCHOLARSHIP GRANTEE SELECTION USING DATA MINING. Retrieved 2 November 2024, from https://www.academia.edu/37101550/DECISION_SUPPORT_SYSTEM_OF_SCHOLARSHIP_GRANTEE_SELECTION_USING_DATA_MINING

Mao, S., Zhang, C., Song, Y., Wang, J., Zeng, X.-J., Xu, Z., & Wen, Q. (2024). Time Series Analysis for Education: Methods, Applications, and Future Directions (arXiv:2408.13960; Version 2). arXiv. http://arxiv.org/abs/2408.13960

Masinading, G. M., Saturnino E. Dalagan, J., & Manib, S. K. R. (2024). Forecasting the semestral enrollment of dorsu curricular programs. *Advances and Applications in Statistics, 91(12),* 1579–1592. https://doi.org/10.17654/0972361724080

Molebatsi, K., & Raboloko, M. (2016). Time Series Modelling of Inflation in Botswana Using Monthly Consumer Price Indices. *International Journal of Economics and Finance, 8,* 15. https://doi.org/10.5539/ijef.v8n3p15

P, A. (2020). Higher Education Institution (HEI) Enrollment Forecasting Using Data Mining Technique. I*nternational Journal of Advanced Trends in Computer Science and Engineering, 9(2),* 2060–2064. https://doi.org/10.30534/ijatcse/2020/179922020

Rahman, A., & Hasan, M. M. (2017). Modeling and Forecasting of Carbon Dioxide Emissions in Bangladesh Using Autoregressive Integrated Moving Average (ARIMA) Models. *Open Journal of Statistics, 7(4),* Article 4. https://doi.org/10.4236/ojs.2017.74038

Sabanal, R. M. (2023). Senior high school student enrollment forecasting model: An application of time series analysis. I*nternational Journal of Research Publication and Reviews, 4(9),* 1102–1113. https://doi.org/10.55248/gengpi.4.923.52729

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics, 6(2),* 461–464. https://doi.org/10.1214/aos/1176344136

Shen, T., Raymer, J., & Hendy, C. (2024). Forecasting school enrollments in the Australian Capital Territory. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. Advance online publication. https://doi.org/10.1093/jrsssa/qnae094

Shim, J. K., Siegel, J., and Liew, C. J. (1994). Strategic Business Forecasting, Probus Publishing Company, Chicago, England, pp. 152–243.

Suresh, C., Reddy, T., & Sweta, N. (2016). A Hybrid Approach for Detecting Suspicious Accounts in Money Laundering Using Data Mining Techniques. *International Journal of Information Technology and Computer Science, 8,* 37–43. https://doi.org/10.5815/ijitcs.2016.05.04

Velasco, C. L. R., Villena, E. G., Ballester, J. B., Prados, F. Á. D., Alvarado, E. S., & Álvarez, J. C. (2023). Forecasting of Post-Graduate Students' Late Dropout Based on the Optimal Probability Threshold Adjustment Technique for Imbalanced Data. *International Journal of Emerging Technologies in Learning (iJET), 18(04),* Article 04. https://doi.org/10.3991/ijet.v18i04.34825

Wachowicz, T., Roszkowska, E., & Filipowicz-Chomko, M. (2024). Decision-makers' behavioral characteristics and multiple criteria decision aiding. Impact of decision-making style and experience on methods' use, evaluation, and recommendation. *Operations Research and Decisions, 34(3).* https://doi.org/10.37190/ord240315

Waddell, D., & Sohal, A. S. (1994). Forecasting: The Key to Managerial Decision Making. *Management Decision, 32(1),* 41–49. https://doi.org/10.1108/00251749410050697

Ward, J. (2007). Forecasting Enrollment to Achieve Institutional Goals. 82.