WHITE PAPER SERIES

Using Lagged Enrollment to Predict High School Graduates

Decision Support and Institutional Research

Austin Peay State University
CLARKSVILLE TENNESSEE

About the Authors

Dr. Andrew L. Luna, is Executive Director of Decision Support and Institutional Research. He has served over 35 years in higher education, with almost 25 of those years in institutional research. He has published research studies on many topics including salary studies, assessment, market research, and quality improvement. Dr. Luna received his Ph.D. and M.A. degrees in higher education administration and his M.A. and B.A. degrees in journalism, all from the University of Alabama.

Dr. Ryan King is a Research Analyst in the office of Decision Support and Institutional Research. Prior to coming to APSU, Ryan received a B.A. in psychology from Montclair State University and an M.A. and Ph.D. in Cognitive Psychology from New York University.

Table of Contents

Executive Summary

During the summer of 2023, Decision Support and Institutional Research (DSIR) decided to conduct a study on enrollment trends of selected K-12 grade levels in order to predict the number of high school graduates for a given year.

This research compared a number of modeling strategies for projecting the number of Tennessee high school graduates. Models based on population tended to produce poor forecasts as assessed via cross-validation. However, models based on Lagged K-12 grade counts proved to be robust to the influence of educational policy.

For the state of Tennessee, the models used in this study tended to have less Mean Average Percent Error than both the NCES and WICHE models, while providing for reasonable forecasts into the mid 2030s. Furthermore, the projections generated within this study seemed to align with the WICHE projections, although the projections in this study tended to be conservatively lower than WICHE.

Of note, all the studies relating to the projected number of high school graduates in Tennessee seem to indicate a steady or flat trend rather than the decreases projected elsewhere. The strongest model in this study was based upon the lagged kindergarten count 13 years in the past, indicating that high school graduate numbers will continue on a steady state as the population increases and more young children attend school.

Because the model in this study relies on past enrollments of kindergarten students to make projections, it is clear that strategic changes in state-wide economic policy could significantly induce sustained or stronger growth. Contrarily, policy changes that hinder economic growth could be detrimental to the number of students receiving high school diplomas in the future.

Introduction

Identifying trends of population growth and decline across Tennessee counties can help institutions of higher education plan for the future. In particular, reliable projections of the expected number of high school graduates can help institutions adapt to changing market conditions and strategically plan recruitment and marketing efforts.

Specifically, understanding how many high school students are in the pipeline for entrance into higher education is vital to the planning, economic development, policymaking, and infrastructure development for all colleges and universities (WICHE, 2020). Accurate projections will enable enrollment management personnel to buttress limited financial resources to areas where they are most needed.

Enrollment projections have been widely used throughout the education environment, particularly within public elementary, secondary, and higher education. Reliable and valid enrollment forecasts give educational administrators, policymakers, and government officials a vital "crystal ball" look into impending trends, shifts, or anomalies so that they can better plan for the future. In most cases, effective enrollment forecasting becomes an important financial tool that helps determine where resources are dwindling or where additional revenue should be increased to sustain infrastructure, larger classrooms, and more instructors.

For almost 50 years, the US Department of Education has provided projections for key education statistics including enrollment, graduates, instructors, and expenditures for both public and private K-12 and higher education (Hussar & Bailey, 2020). Not only does the report look at national data within these areas, it also gives state-level predictions.

Other entities like the Western Interstate Commission for Higher Education (WICHE) and the Southern Regional Education Board (SRBE) have created similar models with comparable, if not more conservative, outcomes. These projection data will become increasingly important to higher education administrators as they try to navigate the choppy waters of college enrollment uncertainty.

"Enrollment projections have been widely used throughout the education environment, particularly within public elementary, secondary, and higher education"

While ARIMA and ratio models have been the staple for enrollment and graduation projections, other models that utilize independent variables to predict a dependent variable are also used. Models that utilize predictor variables in projecting enrollment (whether lagged or not) may be stronger in that they rely on other data relationships rather than just historical enrollment. Specifically, this study uses lagged historic data of K-12 enrollment within specific grade levels and historic state population numbers to project high school graduates within the state of Tennessee for the next 13 years. Not only are the models used reliable, they clearly demonstrate the relationships K-12 enrollments to future high school graduates.

"Some of these models rely only on the trend of high school graduates over time, while others analyze predictor variables to determine association with and, therefore, predictability of high school graduates."

Types of projection models

Various models are used to project high school graduates by using procedures within to explain change in graduate numbers. Each type of model uses specific methodology to detect and predict change of high school graduate numbers. Therefore, while these models use different methodologies, their effects can all be strong and reliable. Some of these models rely only on the trend of high school graduates over time, while others analyze independent variables to determine association with and, therefore, predictability of high school graduates. It is up to the researcher to understand each model, to recognize their strengths and weaknesses, and to choose the best model for the data at hand. What follows is a brief discussion of some of the more prominent models used in projecting the number of high school graduates.

Ratio Method – One of the simplest projection models to use is the ratio method. According to Wing (1974), It is the ratio of high school graduates of the current year to high school graduates of the previous year, multiplied by current high school graduates. This method is expressed in the following formula:

$$Forcasted\ High\ School\ Graduates = Actual\ HS\ Grads, Current\ Year \frac{Actual\ HS\ Grads, Current\ Year}{Actual\ HS\ Grads, Previous\ Year}$$

While this method is quick, easy, and can be effective, it is also highly dependent on the previous year's numbers which could be skewed (Pettibone & Bushan, 1990). As will be seen later, this same issue applies to Markov Chains.

Cohort Survival Method – This method takes the total number of students enrolled from a given cohort (i.e. high school freshmen) and computes the percentage of those students who actually graduated during the current year. According to Lyell & Toole (1974), this percentage is then applied to the number of the next cohort class in order to project how many students will graduate in the future. For instance, if it is found that the number of current year graduates is 81% of the cohort four years ago, the projected number of graduates will be .81 multiplied by the current number of high school freshmen.

Markov Model – According to Gandy et. al., (2019), the Markov model predicts the probabilities of future occurrences based on probabilities of current known values. The Markov Method can have various states of graduating and not graduating. For instance, this model could differentiate between college prep and non-college prep graduates. While this model has an intuitive nature of determining student flow characteristics, the model can only give accurate projections a few years out.

Time Series Analysis – This method uses data points collected sequentially through equally spaced periods of time. According to Brinkman & McIntyre (1997), time series forecasting assumes that the future depends on the present while the present depends on the past. There are various forecasting methods using time series data, such as moving averages, exponential smoothing, autoregression, ARIMA, neural network models, etc.

Choosing the best method is predicated upon the type of data, data patterns, as well as the level of forecasting accuracy. For instance, exponential smoothing is a time series method that gives weights to previous data to predict future data. There are three types of exponential smoothing methods. Single exponential is used on data that have a stable fluctuating pattern. Double exponential is used on data that have a trending pattern. Triple smoothing is used on data that have both trend and seasonal patterns.

Another popular time series method is the Autoregressive Integrated Moving Average (ARIMA). This method is a statistical analysis that predicts future values based on past values by using lagged values of the time series itself and moving averages as predictors. While this method is good for short-term forecasting, it is not useful for more than a few years out. Furthermore, this model is poor at predicting turning points within the data.

Regression Analysis – This method of forecasting shows the relationship between two or more variables. The analysis yields a predicted value of the dependent (criterion) variable and one or more independent (predictor) variables. In essence, the model will determine which possible predictors have a relationship with, in the case of this study, the number of students to graduate high school. Once a relationship has been determined, the strongest predictors can be used to develop predictions. A simple regression model can be expressed by the following expression:

$$Y_i = f(X_i\beta) + e_i$$

Whereby Y_i is the dependent variable, f is the function, X_i is the independent variable, β represents the unknown parameters, and e is the error term.

The simplest form of the model is the linear regression/curve fitting method. It is referred to by Webster (1971) as the "law of growth principle" method. This method uses the number of high school graduates as the dependent variable and year as the independent variable. While this method is quick and simple, it suffers from a lack of explanatory value because only year is used. Furthermore, the model misses other important values that could strengthen the prediction.

Another type of simple regression is to use a variable other than year to predict high school graduates. The independent variable could be current in that one could use the total number of seniors enrolled to predict how may will graduate. However, the independent variable could also be lagged whereby the total number of 9th graders four years back could be used to predict current high school graduates. It should be noted that, in this study, lagged variables are used.

Multiple regression is another type of method within this category. This method utilizes multiple independent measures to collectively predict high school graduates. These variables could include population, economic, and education statistics. The difficulty with this method is, while it is desirable to have each independent variable have a relationship with the dependent variable, it becomes problematic when one independent variable has a relationship with another independent variable.

"However, the independent variable could also be lagged whereby the total number of 9th graders four years back could be used to predict current high school graduates."

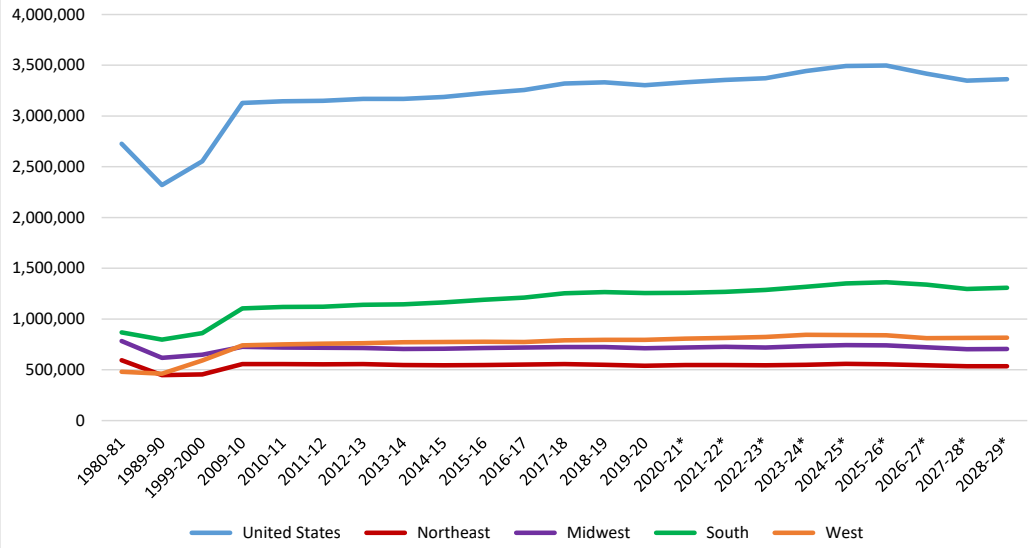
Current Predictions of High School Graduates

The National Center for Education Statistics (NCES) has been projecting national, regional, and state-wide high school graduates for the past 28 years with good success. Depending on the range projected, NCES calculates the number of high school graduates as a percentage of 12th grade enrollment based on historical data. Exponential smoothing is then used to project the percentage. In essence, NCES uses a combined survival analysis method with a time series method.

To arrive at their projections, the DOE uses both single and double exponential smoothing techniques which is a type of autoregressive integrated moving average (ARIMA) model. In general, exponential smoothing places more weight on recent observations than on earlier ones. The weights for observations decrease exponentially as the data move further into the past. Single smoothing is used for making forecasts based in a time series that has no trend or seasonality. Double exponential smoothing allows for more accurate projections of trend and seasonality of data. According to the federal data, the overall number of high school graduates will be flat to decreasing depending on the geographic area.

The most recent NCES report projects high school graduates up to 2029. According to **Figure 1,** the number of high school graduates increased nationally by 14% between 2003-2004 and 2012-2013. The number of high school graduates is projected to be 7% higher in 2028-2029 than in 2012-2013. Regionally, NCES projects that the South and Midwest will have a higher number of graduates in 2028-2029 than in 2012-2013. These projections are supported by other posi-

Figure 1: NCES HS Graduate Trends and Projections



tive factors in these regions including economic and population growth.

For the past ten years, the Western Interstate Commission for Higher Education (WICHE) has published high school graduate projections disaggregated by public and private schools, and race and ethnicity. Their most current report projects the number of graduates out to 2037. To arrive at their projections, WICHE uses the cohort survival ratio which is an observation from the count data sources, of the progression of the number of students from birth to first grade, through each grade, and eventually from the 12th grade to high school graduate.

WICHE uses these calculated ratios to project the number of enrollments and graduates in the years to come. They use a five-year smoothed average ratio for making the projections to place relatively greater weight on the most recent year's data without masking or eliminating any trends that would be evident by taking a longer view. Each cohort survival ratio is calculated as:

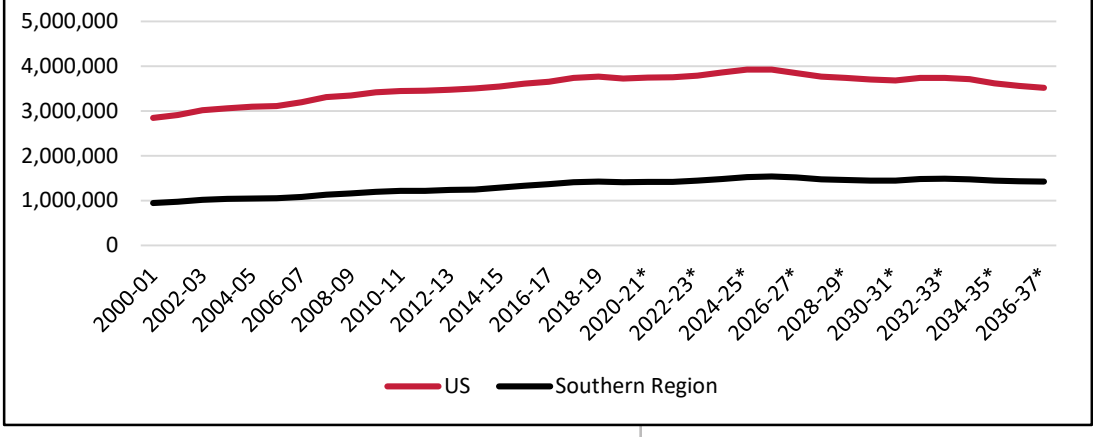
$$Y_{pt} w Y_{p(t-1)} + (1 - w) \frac{\sum_{=2}^{5} Y_{p(i-1)}}{4},$$

,

Where Y_{pt} is the cohort survival ratio at a given progression point p in year t, and w is the smoothing weight (equal to 0.4 in the first year and .15 for each of the four prior years.)

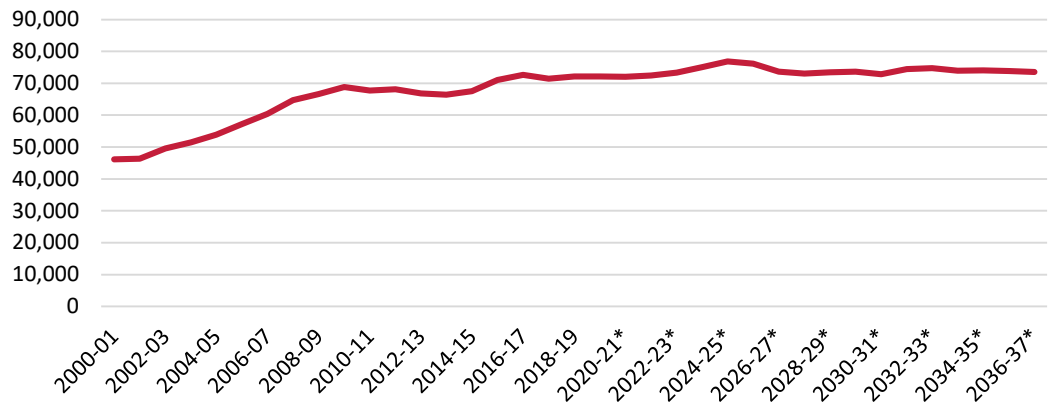
According to the data in Figure 2, the number of high school graduates is expected to peak nationwide by 2025 and begin a downward trend to 2037, equaling the number of graduates from 2014-2015. Within the southern region, however, the 2025 peak is not as apparent, but the decay to 2037 is very small and the trend is almost flat.

Figure 2: WICHE US and Southern HS Graduation Trends and Projections



Within Tennessee, as shown in **Figure 3**, WICHE still predicts a 2025 peak however, the future trend is flat with the 2037 projection equaling the number of high school graduates from 2020-2021. While these projections go out further than the NCES projections, it is clear that Tennessee should be able to weather a projected downward trend in high school graduates nationwide.

Figure 3: WICHE Tennessee HS Graduation Trends and Projections



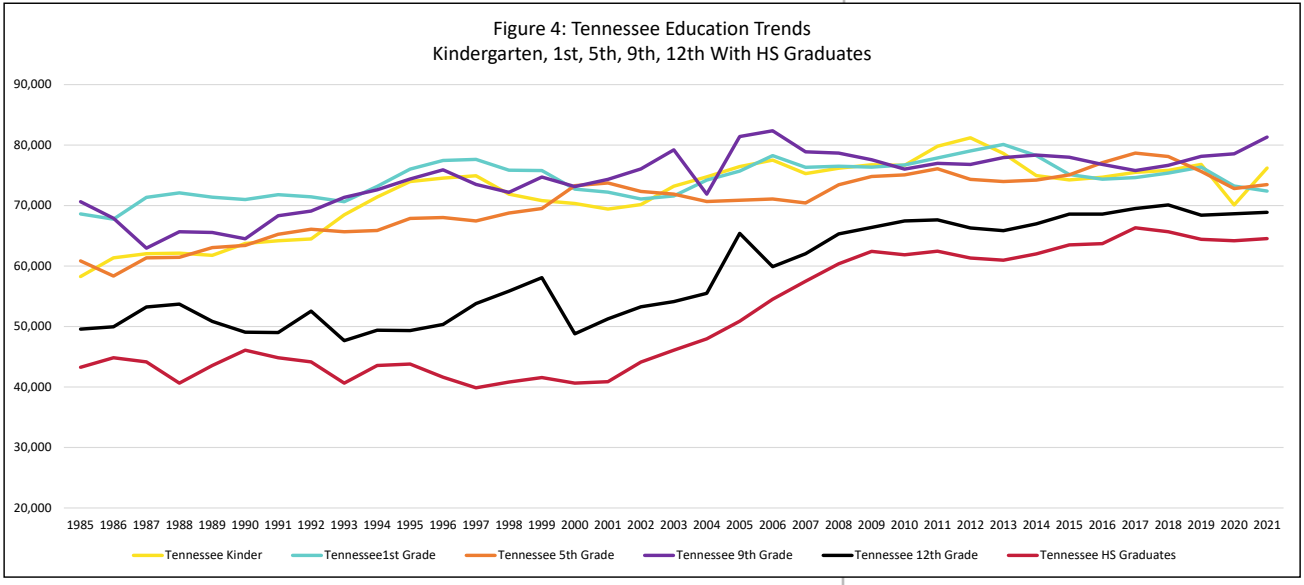
Methodology

This study uses population projections and lagged grade counts from the Tennessee census to project the number of HS graduates up to 13 years in the future. We demonstrate that projections based on lagged grade counts are more robust to changes in education policy than projections based on population alone. Furthermore, this study, which uses various prediction models, will compare the projected outcomes from the strongest models to the WICHE study that uses cohort survival analysis.

The first step in creating the models was to collect historical Tennessee enrollment data. These data included kindergarten, 1st, 5th, 9th, and 12th grade enrollments from 1985 to the present. These data were obtained through the Digest of Education Statistics from the Department of Education. Due to the two-year lag of projection made by the NCES, the latest data that could be obtained was for 2021.

The results of the data collection are shown in **Figure 4**. While the data show a moderate increase in enrollments for kindergarten as well as 1st through 9th grades, there is a steep increase in the number of 12th graders and, subsequently, high school graduates starting in 2003. This increase in the number of 12th graders and high school graduates coincides with and is plausibly related to the No Child Left Behind Act signed by President George W Bush in 2002.

Figure 4: Tennessee Education Trends
Kindergarten, 1st, 5th, 9th, 12th With HS Graduates



Further evidence that this sharp increase is due to educational policy rather than what would be expected based on state population growth alone comes from inspecting the following plots: **Figure 5** shows Tennessee population growth from 1985 to 2021 (left) and also shows Tennessee population plotted against high school graduates (right). It is clear that Tennessee population has been steadily increasing at a relatively constant rate from 1985 to the present. High school graduation numbers, however, show approximately three qualitatively different states during this same time period. Despite the steady increase in population, the number of HS graduates appears to be flat or declining between 1985 to 2000, followed by a sharp, constant increase (stronger than expected via population increase alone) from 2001 to about 2010, and culminating in a flat or slightly positive trend from 2011 to the present.

Figure 5: Tennessee Population, 1985 to 2021

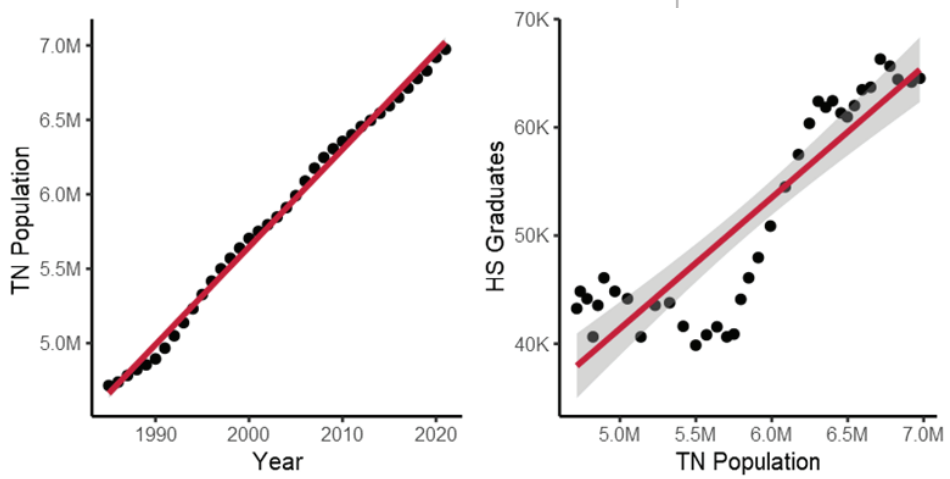
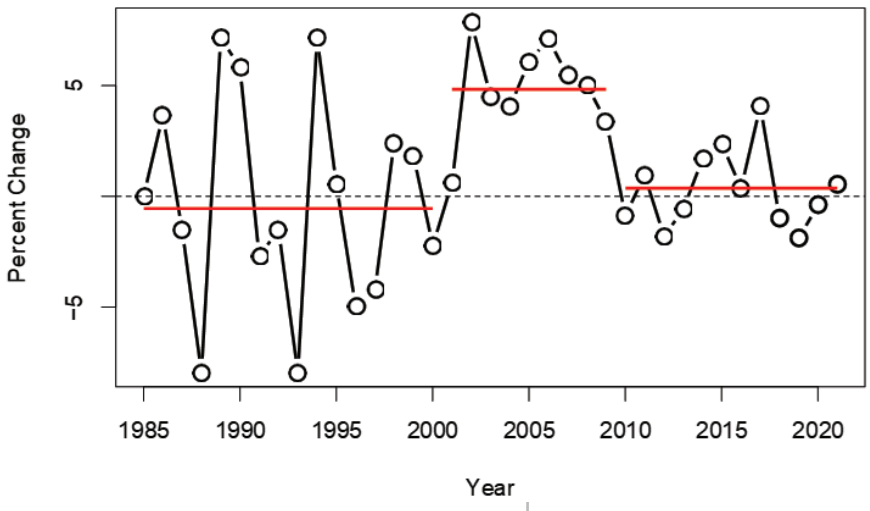


Figure 6: Tennessee Population Change



Characteristics of the three different phases can be seen by plotting year-to-year percent change in high school graduate numbers over time (**Figure 6**). The red line segments indicate the average percent change during each of the three phases. The first phase (mean = -0.54) looks like a white noise process, with values randomly

fluctuating around 0 or slightly below. This first phase appears to be a steady-state, but is more likely flat as a function of two opposing processes: a steady increase in population on the one hand and a decline in retention rate during this time period on the other. Then, there is a change around 2001 which produces a steady increase in HS Graduation numbers (mean = 4.83), which eventually culminates in a new, higher steady state around 2010 until the present (mean = 0.37). This second steady state looks similar to the first in that the percent change year-to-year fluctuates around 0, but with substantially less variability.

Thus, it is clear that the relationship between Tennessee population and Tennessee high school graduates is indirect. Given the sensitivity of high school graduate numbers to changes in education policy and other potential factors, models predicting the number of high school graduates based on population alone are unlikely to produce reliable projections. There is an obvious need to incorporate measures that are more robust to changes in education policy. To this end, the study presented here uses lagged K-12 public school grade counts in addition to population projections. In particular, we use counts of the number of kindergarteners with a 13-year lag, 1st graders with a 12-year lag, 5th graders with an 8-year lag, and 9th graders with a 4-year lag. In the ideal scenario, a model based on kindergarten counts can produce projections for high school Graduates 13 years into the future. Comparisons between the projections made by the most ambitious model (i.e. kindergarten 13-year lag) and models using other grade counts is important for assessing the reliability and consistency of the projections being made. Additionally, fitting models based on different lagged grade counts and performing model averaging across each model's projection where those projections overlap can improve the precision of these estimates.

Unlike many previous approaches to projecting Tennessee high school graduates which adopt a particular modeling framework, here we fit a number of model variants and directly compare their performance using cross-validation. The set of models considered here is shown in **Table 1**. As a baseline measure, we include an ARIMA model. **Table 1** shows a simple AR(1) model (i.e. Autoregression of order one) as an example ARIMA model, but, in general, ARIMA refers to a class of models that can have any number of autoregression or moving average components. While ARIMA models have been widely adopted for enrollment projections, autoregressive models alone typically produce poor long-term forecasts. Here, we compare the following additional

models: three simple regression models (population-only, lagged grade count-only, and population + lagged grade count) and three regression models with ARIMA errors which is an extension of ARIMA models to include covariates in the prediction of the dependent variable. These models can capture any residual autocorrelation in the time series after regressing on the covariates of interest.

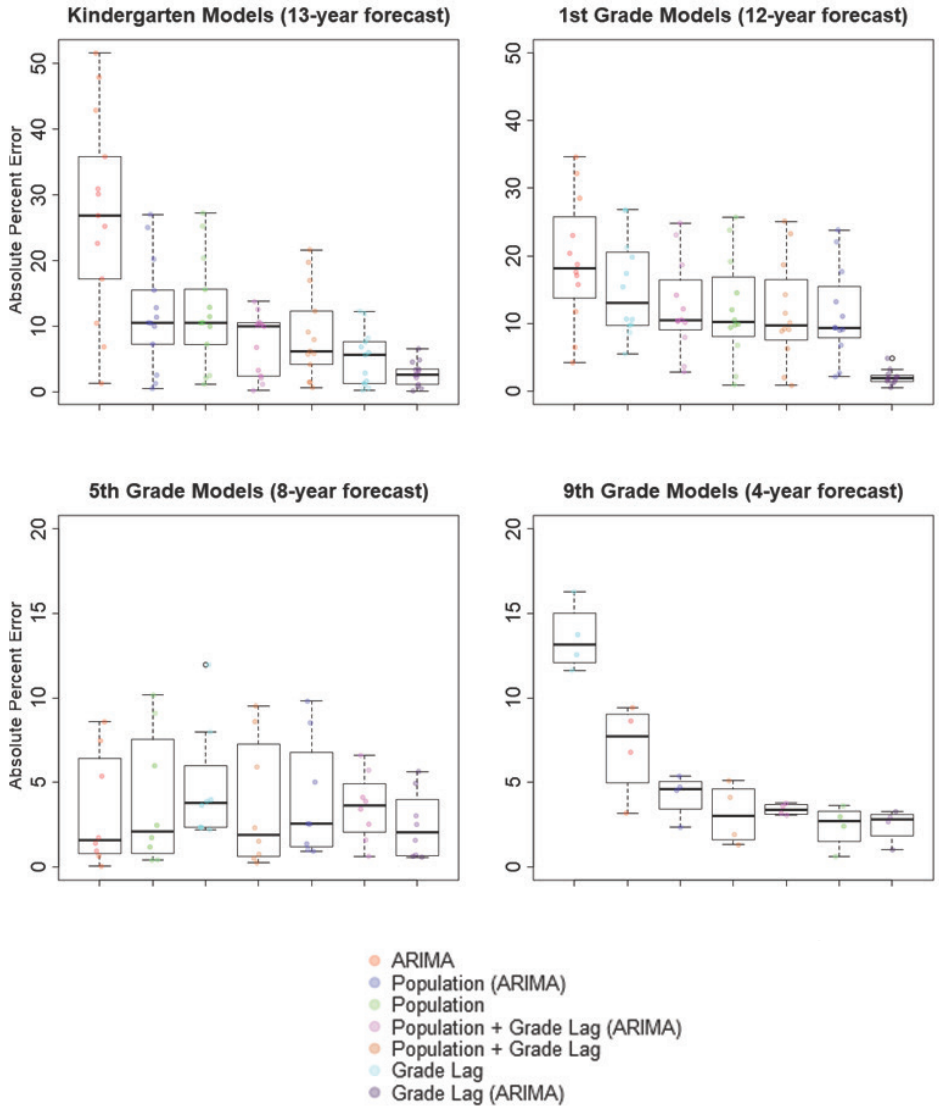
Table 1: Model types and example equation

ARIMA	$HS\ Graduates_t = \beta_0 + \beta_1 HS\ Graduates_{t-1} + \varepsilon_t$
Population Regression	$HS\ Graduates_t = \beta_0 + \beta_1 Population_t + \varepsilon_t$
Lagged Grade Regression	$HS\ Graduates_t = \beta_0 + \beta_1 Lagged\ Grade_t + \varepsilon_t$
Population + Lagged Grade Regression	$HS\ Graduates_t = \beta_0 + \beta_1 Population_t + \beta_2 Lagged\ Grade_t + \varepsilon_t$
Population Regression + ARIMA error	$HS\ Graduates_t = \beta_0 + \beta_1 Population_t + \eta_t$, where $\eta_{t=} \eta_{t-1} + \varepsilon_t$
Lagged Grade Regression + ARIMA error	$HS\ Graduates_t = \beta_0 + \beta_1 Lagged\ Grade_t + \eta_t$, where $\eta_{t=} \eta_{t-1} + \varepsilon_t$
Population + Lagged Grade Regression + ARIMA error	$HS\ Graduates_t = \beta_0 + \beta_1 Population_t + \beta_2 Lagged\ Grade_t + \eta_t$, where $\eta_{t=} \eta_{t-1} + \varepsilon_t$

Results

In order to compare the performance of each model, we employed the following cross validation strategy: The HS graduate data was partitioned into a training set and a test set where the size of the test set was equal to the number of lags in the grade count. This was done to ensure that the validation procedures were as close as possible to the model's intended use case. For example, the intended use case of the Kindergarten model was to produce reliable 13-year forecasts. Therefore, the test set (i.e. forecast horizon) was comprised of the most recent

Figure 7: Mean Absolute Percentage Error Between Actual and Predicted Number of Graduates



13-years of HS graduate counts (i.e. 2009-2021) and the training set was comprised of the remaining years (i.e. 1998-2008). Each of the models described above was fit on the training set and n-step ahead forecasts were produced. The mean absolute percentage error (MAPE) between the model's projected number of HS graduates and the actual number of HS graduates was calculated. **Figure 7** shows how MAPE varied across model types for each grade. In every case, the best performing model was the Lagged Grade Regression with ARIMA errors.

Table 2 shows the MAPE values for each model. Highlighted in bold is the best performing model for each group. The Lagged Grade regression model with ARIMA errors consistently performed the best on the cross-validation test sets for each grade. In all cases, the MAPE value was lower than 3%. Moreover, the model in this study seems to out-perform both the NCES and WICHE studies in relation to overall MAPE. For instance, the ten-year error for the NCES model is 5.1 (Hussar & Bailey, 2020) while the error for the WICHE model is 6.2 (WICHE, 2020). The model in this study that projects 13 years into the future has an error of 2.8.

TABLE 2: Comparison of MAPE By Model

Grade	Model	MAPE
Kindergarten (13-year forecast)		
	ARIMA	26.9
	Population	12.1
	Grade Lag	5.1
	Population + Grade Lag	8.8
	Population (ARIMA)	11.9
	Grade Lag (ARIMA)	**2.8**
	Population + Grade Lag (ARIMA)	7.3
1st Grade (12-year forecast)		
	ARIMA	19.2
	Population	12
	Grade Lag	15.2
	Population + Grade Lag	11.6
	Population (ARIMA)	11.3
	Grade Lag (ARIMA)	**2**
	Population + Grade Lag (ARIMA)	12.4

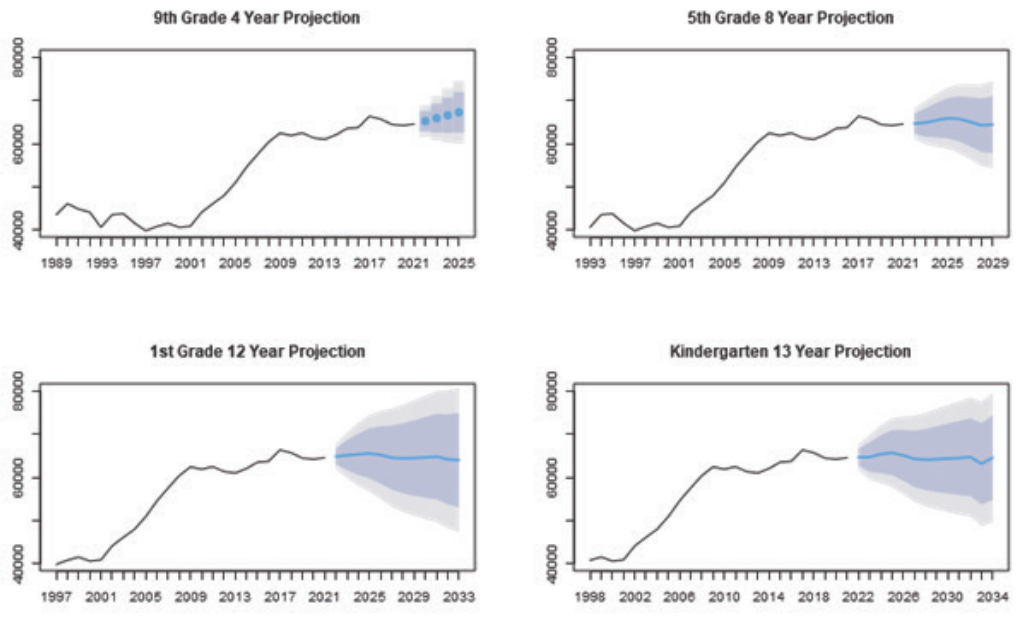
Cont.

5th Grade (8-year forecast)		
	ARIMA	3.3
	Population	3.9
	Grade Lag	4.8
	Population + Grade Lag	3.7
	Population (ARIMA)	4
	Grade Lag (ARIMA)	**2.4**
	Population + Grade Lag (ARIMA)	3.6
9th Grade (4-year forecast)		
	ARIMA	7
	Population	2.6
	Grade Lag	13.5
	Population + Grade Lag	3.1
	Population (ARIMA)	4.2
	Grade Lag (ARIMA)	**2.4**
	Population + Grade Lag (ARIMA)	3.4

Model Projections

Based on the model validation procedure discussed in the previous section, the Lagged Grade Regression with ARIMA errors model was used to generate projections for high school Graduates after 2021 in each case (**Figure 8**). The projections of each model are fairly consistent with one another, with a modest increase in HS Graduates projected from 2022 until around 2025, at which point the projection dips and flattens out. Thus, the

Figure 8: High School Graduate Projections By Model



evidence presented here suggests that the number of Tennessee high school graduates is expected to remain relatively constant in the near to long term. **Table 3** shows the 13-year forecasts of number of high school graduates from the Kindergarten model as it compares to the WICHE and NCES projections.

As can be seen in the table, the model in this study contains projections that are close to and align with the WICHE model, albeit more conservative in its projections.

Table 3: Model Projections

	2022	2023	2024	2025	2026	2027	2028	2029	2030	2031	2032	2033	2034	2035	2036
Kindergarten Model	64651	64628	65350	65667	65076	64243	64079	64176	64363	64437	64662	63134	64528	-	-
WICHE	65200	66940	68480	67830	65300	64640	64700	65250	64530	65900	66140	65430	65560	65310	65080
NCES	65250	67010	68260	67850	65050	64230	64670	-	-	-	-	-	-	-	-

Conclusion

This research compared a number of modeling strategies for projecting the number of TN HS graduates. Models based on population tended to produce poor forecasts as assessed via cross-validation. This is likely due to changes in educational policy influencing K-12 enrollment, retention, and graduation rates in a way that obfuscates the relationship between population growth and the number of TN students graduating from high school. Models based on Lagged K-12 grade counts, on the other hand, proved to be robust to the influence of educational policy.

For the state of Tennessee, the models used in this study tended to have less error than both the NCES and WICHE models, while providing reasonable forecasts into the mid 2030s. Furthermore, the projections generated within this study seemed to align with the WICHE projections, although the projections in this study tended to be conservatively lower than WICHE.

More importantly, all the studies relating to the projected number of high school graduates in Tennessee seem to indicate a steady or flat trend rather than the decreases projected elsewhere. The strongest model in this study was based upon the lagged kindergarten count 13 years in the past, indicating that high school graduate numbers will continue on a steady state as the population increases and more young children attend school.

Because the model in this study relies on past enrollments of kindergarten students to make projections, it is clear that strategic changes in state-wide economic policy could significantly induce sustained or stronger growth. Contrarily, policy changes that hinder economic growth could be detrimental to the number of students receiving high school diplomas in the future.

While it becomes increasingly important to accurately forecast population growth, school enrollments, and high school graduations in order to adequately affect the planning, economic development, policymaking, and infrastructure development for all colleges and universities, projection models like the ones in this study should help make these tasks less challenging.

References

Gandy, R., Crosby, L., Luna, A., Kasper, D, & Kendrick, S. (2019). Enrollment projection using Markov Chains: detecting leaky pipes and the bulge in the boa. AIR Professional File, 147.

Hussar, W.J., & Bailey, T.M. (2020). Projections of education statistics to 2028 (NCES 2020-0024). US Department of Education, Washington, DC: National Center for Education Statistics.

Lyell, E. H. & Toole, P. (1974). Student flow modeling and enrollment forecasting. Planning for Higher Education, (3) 6, p 2-6.

Pettibone, T.J. & Bushan L.S. (1990) District enrollment projections: A comparison of three methods. Paper presented at the Annual Meeting of the Mid-South Educational Research Association: New Orleans, LA.

Western Interstate Commission for Higher Education and the College Board. (2020). Knocking at the college door: projections of high school graduates by state and race/ethnicity, 2020. Boulder, CO: Wester Interstate Commission for Higher Education.

Wing, P. (1974). Higher educational enrollment forecasting: A manual for state level agencies. Boulder, Colo.: National Center for Higher Education Management Systems.