Journal of the Royal Statistical Society Series A: Statistics in Society, 2024, **00**, 1–18
https://doi.org/10.1093/jrsssa/qnae094

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series A
Statistics in Society
A

Original Article

Forecasting school enrollments in the Australian Capital Territory

Tianyu Shen[1], James Raymer[1] and Caroline Hendy[2]

[1]School of Demography, Australian National University, Canberrra, Australia
[2]School of Literature, Languages and Linguistics, Australian National University, Canberra, Australia

Address for correspondence: James Raymer, School of Demography, Research School of Social Sciences, Australian National University, 146 Ellery Cres, Acton, ACT 2601, Australia. Email: james.raymer@anu.edu.au

Abstract

School enrollment forecasts are vital for effective planning. This study introduces a probabilistic multiregional population projection model, which accounts for different components, including preschool entries, migration, grade progression, and graduations. Using different distributions with a cohort component projection and Monte Carlo simulation, this paper forecasts student enrollments for each school and academic year level in the Australian Capital Territory based on the annual record-level administrative data. In the in-sample validation tests, the model's overall performance is robust, and the probabilistic design offers reliable prediction intervals reflecting the variation in observed values. The paper ends with a discussion on the importance of prediction intervals for informing school planning.

Keywords: administrative data, forecasting, Monte Carlo simulation, multiregional population projection, school enrollment

1. Introduction

School jurisdictions use school enrollment projections for budget development and staff planning, building construction and utilization, school resource and equipment allocation, transportation decisions, and school priority enrollment area assignments (Schellenberg & Stephens, 1987). With such a wide range of applications, large inaccuracies in enrollment forecasting can have significant political and economic consequences. One way to reduce this risk for school jurisdictions is by including prediction intervals in enrollment projections. In this paper, we present a method for building stochastic prediction intervals for a multiregional sources of growth school projection model (Raymer et al., 2017), and discuss its advantages and drawbacks in a case study using data from the Australian Capital Territory (ACT) Education Directorate.

There is little academic work on prediction intervals for school enrollment projections. In the 1970s, Denham (1971), Braden (1972), and Rives (1977) put forward various approaches for calculating school enrollment confidence and prediction intervals. Both Denham and Braden's works were later criticized by Simpson (1987) for relying on arbitrarily estimated probabilities. For example, Denham assumed that 'high' and 'low' estimates decided upon by a forecaster represent the bounds of a 98% confidence interval (Denham, 1971, p. 37). In response to these issues, Simpson argued that uncertainty in school enrollment forecasts is best given not by confidence or prediction intervals, but by an examination of past forecasting errors and experimenting with a range of input assumptions, such as migration rates (Simpson, 1987, pp. 73–74).

Received: April 30, 2024. **Revised:** September 4, 2024. **Accepted:** September 5, 2024

More recently, Grip and Grip (2020) produced school enrollment prediction intervals for three New Jersey school districts based on 20 years of enrollment data. They compared the utility of confidence intervals based on past error rates with that of stochastic prediction intervals using Monte Carlo simulations. They determined that the empirically based confidence intervals were too wide for school planners to use meaningfully beyond 1–2 years. Similarly, the prediction intervals yielded by the stochastic methods, in which the change in school survival ratios were randomly selected from observed change in the past, were too large to be of value to planning authorities for any but the largest school districts, defined as those with 25,000 or more students.

Schellenberg and Stephens (1987) observed that 'enrollment projection seems to be one of those things that everyone does but very few people talk about' (p. 4). Indeed, despite the lack of peer-reviewed work on the subject, forecasters do appear to be providing confidence or prediction intervals to school jurisdictions. For example, in forecasts for the San Francisco Unified School District, Lapkoff and Gobalet used a Monte Carlo procedure randomly selecting kindergarten-birth ratios and grade-progression ratios from 20 years of observations to create the confidence intervals. As a further example, Andersen et al. (2014) for the Lexington Public Schools in Massachusetts, created confidence limits for their total enrollment forecasts by using the standard deviation of the errors for previous years as a standard error, and applying the usual confidence interval formulas for a 90% confidence interval for sample means. However, this method relies on the assumption that their forecast errors are additive and follow a normal distribution—assumptions which are unlikely to hold—and the authors warn that their intervals are not rigorous (Andersen et al., 2014, p. 27).

Other school enrollments modellers have declined to create confidence intervals in their work. Echoing the arguments put forward by Simpson (1987), Rynerson and Wei (2022) in their projections for the Portland Public School system, write:

> 'Due to the nature of forecasting, there is no way to estimate a confidence interval as one might for data collected from a survey. The best way to measure potential forecast errors is to compare actual enrollments with previous forecasts that were conducted using similar data and methodologies.' (p. 55)

Yet, not providing prediction intervals remains risky for forecasters and the school jurisdictions they serve. As Simpson (1989) writes, from the perspective of local governments, both over- and under-predicting school enrollments come with substantial cost. Large inaccuracies in both directions contribute to noticeable funding and resource inequality and an eventual loss of public confidence in the planning authority.

Forecasters and planning authorities thus stand to benefit from the development of sound methodology for creating prediction intervals for school enrollments. In this paper, we present such a methodology embedded within a multiregional cohort component projection model for school enrollments that was designed and implemented for the ACT Education Directorate in Canberra, Australia (Raymer et al., 2017; Xiang et al., 2023). Our key contribution is the inclusion of prediction intervals, which provide a range of likely outcomes rather than just a single estimate for each component of school enrollment change. The components of change include entries to the school system, transfers between schools, and exits from the school system. To demonstrate the effectiveness and robustness of the prediction intervals of the methodology, we use annual school census data provided by the ACT Education Directorate to predict school enrollments from 2017 to 2023. By removing 1–7 years of data at the end of the time series, we use in-sample testing to compare and assess the models.

2. Background

2.1 Demographic and school enrollment in the Australian Capital Territory

The Australian Capital Territory (ACT) encompasses Canberra—Australia's national capital—and adjacent areas. According to the 2021 national census, the Australian Capital Territory (ACT) had a population of 454,000 residents—a notable increase of 27.1% from the 357,222 residents recorded in the 2011 census (Australian Bureau of Statistics, 2022). About 50% of the population growth was due to net international migration, with the remaining growth due to natural increase (births minus deaths). Internal migration contributed a small amount towards the

ACT's population growth during this period. Reflecting its status as the seat of government for Australia's Federal Government and Parliament, and as a centre for several higher education institutions, the Australian Capital Territory (ACT) also experiences a dynamic and transient population, characterized by significant annual flows of both domestic and international migrants.

As of 2023 planning system reforms, the ACT now comprises nine districts: Belconnen, Gungahlin, Inner North and City, Inner South, East Canberra, Woden Valley, Weston Creek, Molonglo Valley, and Tuggeranong,[1] in which Molonglo Valley is the newest district, formally established in 2010. Of the original districts, Belconnen and Gungahlin have experienced the most rapid growth over the past 5 years, due to the construction of new suburbs and housing developments.

School education in the ACT is overseen at the territory level by the ACT Education Directorate. The educational framework in ACT public schools includes four stages: 'Preschool', nonmandatory schooling for children aged four, with further universal provision for children aged three from 2024 onwards; 'Primary School', commencing with kindergarten at age five and continuing through Years 1 to 6; 'High School', comprising Years 7–10; and 'College', comprising Years 11 and 12. While nongovernment schools organize their school administrations differently for all other purposes,[2] these four stages are used in the collection of student statistical data and in the projection model for student enrollments due to the movement between sectors at these junctures. Full-time schooling is compulsory from age 6 until the completion of Year 10, after which young people must engage in full-time education, training, or employment until they complete Year 12 or turn 17. There were 81,514 students enrolled in Preschool through Year 12 in the ACT in 2023—an increase of 7% over the 76,411 students enrolled in 2018.[3]

The ACT's school education system is divided into three sectors: public, Catholic, and independent. The public sector, managed by the ACT Education Directorate, offers free education and is the largest sector, with 61% of total student enrollments. Students residing within a public school's 'priority enrollment area' are guaranteed admission at their local school. Schools in the Catholic sector are managed by the Catholic Education Office, and account for 19% of total enrollments. The independent sector consists of a mixture of private schools and comprises the remaining 20% of enrollments. In addition to direct management of public schools, the ACT Education Directorate oversees the registration and regulation of Catholic and independent schools, as well as home education, ensuring that the educational needs of the territory's growing population are met. Catholic and independent schools have their own organizational structures and are allowed to charge tuition and be selective in the students they enroll.

The Education Directorate collect and update student records through the Enrolment Census of ACT Schools in February every year.[4] Students are assigned a unique identification number, which students retain for as long as they remain within the ACT education system, allowing for the analysis of year-to-year transitions of students between and within schools. In general, the administrative data are considered to be highly accurate as the Education Directorate has the responsibility to ensure all children in the ACT have access to schools and education.

2.2 School enrollment forecast models

School enrollment forecasting has been a long-standing responsibility for education departments in cities or district councils throughout the world (Simpson, 1988). Johnstone (1974), Shaw (1984) and Simpson (1987) provide reviews on different school enrollment models. In this paper, we extend these reviews to the measure of uncertainty in different models.

2.2.1 Ratio model

The ratio model assumes school enrollments are proportional to the population of students. It is a simple approach that is frequently used among school planners because it only requires the

[1] Map of districts can be found on page 13 in this document: https://www.legislation.act.gov.au/DownloadFile/ni/2023-540/copy/164770/PDF/2023-540.PDF.

[2] Non-government schools are organised in a variety of models ranging from preschool or kindergarten to year 6, preschool to year 10, preschool to year 12, years 4–12, or years 7–12.

[3] Exclusions were made in some special academic levels and/or special schools. For this reason, the total enrollment in the model does not correspond exactly with the numbers in the Education Census.

[4] The summary of Enrolment Census of ACT Schools in each year is published in this website: https://www.education.act.gov.au/about-us/policies-and-publications/publications_a-z/census.

aggregate number of student enrollments and the number of student-aged persons in the population. However, the drawbacks of this approach are also evident. Because of its simplicity, it is unable to account for changes in housing and/or school developments. This type of model is most useful for forecasting total school enrollments in large districts with a relatively stable trend. In our review of the literature, we were unable to locate any research that included uncertainty for this type of model.

2.2.2 Cohort survival model

The cohort survival model is another commonly used model for school planners (see e.g. Edmonston, 2000; Rushton et al., 1995). It is slightly more sophisticated than the ratio model by accounting for student progression through academic levels. Braden (1972) computed the confidence interval by assuming a certain probability distribution for the survival ratio. However, it was unclear how well the confidence intervals covered the true values. Grip and Grip (2020) proposed two methods for including projection uncertainty: empirical confidence interval based on in-sample forecasts and Monte Carlo simulations drawing from historical changes in the survival ratios. They reported that the confidence intervals from both approaches tended to be too wide for practical use. It is worth noting that almost all the enrollment forecasts produced with the cohort survival model only include point estimates. An exception is Lapkoff and Gobalet, who accounted for uncertainty associated with new housing developments and population change. They used a Monte Carlo procedure with 5,000 simulations, randomly selecting kindergarten-to-birth ratios and grade-progression ratios from 20 years of observations to create the 67% and 90% confidence intervals. Denham (1971) developed a probabilistic model based on a multivariable model with retention and drop-out transition probabilities. For each transition probability, Beta distributions are specified for high, low and most likely assumptions.

2.2.3 Regression-based model

Regression-based models are primarily used to extrapolate historical student enrollment trends. For example, Rives (1977) uses regression to predict student enrollments in the next year by academic level for inputs into the cohort survival model. Standard errors of the parameters are used to construct the prediction intervals. Another common type of regression-based forecast is the time-series model. Yang et al. (2020) discusses and compares a few different time-series models to forecast school enrollments in Taiwan. Confidence intervals were not included.

2.2.4 Prediction interval and evaluation

From previous studies, there are mainly three ways to estimate the uncertainty or prediction intervals: regression, empirical method and Monte Carlo simulation. A regression-based prediction interval is a slightly different framework than the other two as it is embedded in the estimation process (Rives, 1977). In Monte Carlo simulation, the prediction intervals are derived from the standard errors of the estimated regression parameters. Empirical prediction intervals, on the other hand, are based on forecast errors produced from in-sample tests (Lee & Scholtes, 2014). Monte Carlo simulations are randomly drawn from a set of numbers and assumed distributions. Denham (1971) applies probability distribution with limits, while Grip and Grip (2020) compute a set of discrete numbers from changes observed over time.

There are various methods and indicators to evaluate forecast performance (Swanson & Tayman, 2012). The most straightforward error indicators are the differences, absolute difference or percentage difference between the observed and the forecasted values. These indicators can be examined at the unit level or summed across units. Mean Algebraic Percentage Error (MALPE) and Mean Absolute Percentage Error (MAPE) are useful metrics for providing an overall assessment of model performance. MALPE is generally used as a measure of bias to determine whether the forecasts tend to over- or under-predict school enrollments. MAPE, on the other hand, is often used to assess the overall accuracy of the model. Hussar and Bailey (2020) and Yang et al. (2020), for example, employed MAPE to assess the accuracy of their forecasting models.

3. Method

3.1 Data

The underlying data used to produce school enrollment forecasts were obtained from the Australian Capital Territory's Education Directorate and Chief Minister, Treasury and Economic Development Directorate (CMTEDD). The Education Directorate provided annual school enrollment census information on all students enrolled in ACT schools collected in February each year from 2009 to 2023.[5] The census data is an administrative source that captures deidentified student record-level data and includes information on schools, academic level, and home suburb. For students with multiple home addresses, a primary home address is assigned in the dataset. The unit-level records are not available to the public, but the Education Directorate publishes student enrollment by school and academic level for each year online[4]. School enrollments in the ACT increased from 64 thousand in 2009 to 82 thousand in 2023 across 132 schools, including 4 academic categories: preschools, primary schools (Kindergarten, Year 1–Year 6), high schools (Year 7–Year 10), and colleges (Year 11–Year 12). There are three main sectors: public schools with 50 thousand students, independent schools with 16 thousand students, and Catholic schools with 16 thousand students. In 2023, the number of enrollments separated by school and academic category ranged from about 30 students (at an independent school, years 7–10) to around 1,400 students (at a Catholic school, years 7–10).[6]

Another source of information used for the school enrollment forecasting is the estimated resident population of persons aged 4 years old in suburbs across ACT. Both the observed and projected data come from CMTEDD. CMTEDD publishes single-age population projections by suburb for the years 2022–2026, which is based on the historical population by age, sex and suburb from 2011 to 2021 (ACT Chief Minister, Treasury & Economic Development Directorate, 2023). Since the population projection only includes point estimates, we assume uncertainty in the age four population follows a normal distribution with a standard deviation of 10%.

3.2 Model

The focus of the model development is to add the capability to forecast prediction intervals in the School Transition Estimation and Projection (STEP) model. The STEP model, described in Guan et al. (2022) and Xiang et al. (2023), is a multiregional cohort component projection model that simultaneously predicts enrollment change across academic levels and sectors for all schools within a system. The model incorporates multiple sources of enrollment change, including preschool entries, transfers between schools, migration into and out of the system, and graduations. In this paper, we extend the STEP model methodology to incorporate probabilistic forecasts for each component of change that are then integrated together to provide probabilistic forecasts for each school by academic level. The inclusion of prediction intervals acknowledges that point estimates are unlikely to be true and could potentially mislead education planners and policy makers.

Prediction intervals are incorporated within the STEP model through the following three sub-models:

(i) a sub-model for preschool entries;
(ii) a sub-model for in-migration of students enrolling in Kindergarten to Year 12 in each school;
(iii) a sub-model for transition probabilities between schools and academic levels, including out-migration and Year 12 graduation.

Sub-models (i) and (ii) capture students entering the ACT school system and sub-model (iii) captures students transitioning within or leaving the school system. Each sub-model is assumed to

[5] Data quality has improved since 2011, with a higher linkage rate across schools and year levels, compared with lower linkage rates in 2009 and 2010.

[6] As per section 2.1, for the purposes of the Enrolment Census of ACT Schools and school enrollment projection work, independent and Catholic school cohorts are split in the same way as public schools' academic categories, due to the movement between sectors at these junctures. However, it is important to note that non-government schools are not usually configured in this way in practice.

have specific probability distributions. Monte Carlo simulations are then used to combine the effects from these distributions for use in the multiregional cohort component projection model for every school and academic level. In addition, each source of school enrollment change contains uncertainty that can be adjusted independently.

3.2.1 Projecting preschool entries

Preschool entries are students from suburb *i* enrolled in a specific ACT preschool in year *t*. Students who repeat preschool the following year are not counted as entries. Preschool entries are modelled separately for students living in the ACT and those living outside the ACT (mainly students living in New South Wales regions surrounding the ACT). The majority of students come from the ACT.

3.2.1.1 For preschool entries living in ACT suburbs. To estimate the preschool entries from each suburb in the ACT, ratios are created using the Education Directorate's school census data and CMTEDD's population estimates and projections of persons aged 4 years old. For each suburb and each year from 2009 to the most recent year of available data (henceforth 'base year'), the ratio of preschool entries from a suburb to the number of 4-year-olds living in each suburb is given by the following relationship:

$$Q^t_{1,i} = \frac{P^t_i}{N^t_{4,i}}. \tag{1}$$

P^t_i represents preschool entries in ACT schools in year *t* whose home suburb are *i*; and $N^t_{4,i}$ represents population at age 4 years in suburb *i*. If most of the residents at age 4 years attend preschool, then $Q^t_{1,i}$ should centre around one. If the population estimates are consistently higher or lower than the preschool entries, the modal $Q^t_{1,i}$ ratio may be above one or between zero and one, respectively.

Similar to the cohort survivor ratio model (Grip & Grip, 2020) and STEP model (Xiang et al., 2023), we use average $\bar{Q}_{1,i}$ values for each suburb in the projection. However, instead of employing a fixed average, we use an algorithm to select the average with the least variation between three and 7 years based on the standard deviation values, working backwards in time from the most recent year. This approach allows us to detect and emphasize recent changes to the overall trend. Further, for each $\bar{Q}_{1,i}$, we assume a gamma distribution. To fit a gamma distribution, one needs two parameters, and these parameters can be estimated through method of moment, including ordinary least squares (OLS), or maximum-likelihood estimation. As both produce very similar parameters, we opted for the simplest and most intuitive, which is the method of moment estimator. The mean of the distribution matches the average $\bar{Q}_{1,i}$ calculated above.

Preschool is the lowest and entry level to the school system, so it is likely to be more fluctuating than the higher grades. We compute the maximum standard deviation ($\sigma^{\max}_{Q1,i}$) among the last 3–7 years for the distribution. The $\hat{Q}_{1,i}$ ratio for suburb *i* can be described by a gamma distribution using the shape parameter *k* and scale parameter θ with mean $\bar{Q}_{1,i}$ and standard deviation $\sigma^{\max}_{Q1,i}$,

$$\hat{Q}_{1,i} \sim Gamma\left(k = \frac{\bar{Q}^2_{1,i}}{(\sigma^{\max}_{Q1,i})^2}, \theta = \frac{(\sigma^{\max}_{q1,i})^2}{\bar{Q}_{1,i}}\right).$$

Note that the distribution does not vary over time in the projection. Adjustments to account for increasing variability over time is discussed below in Section 3.2.4.

The next step in the preschool sub-model is to distribute preschool students from each suburb to schools. They are distributed using proportions, i.e.

$$Q^t_{2,i,j} = \frac{P^t_{i,j}}{P^t_i}, \tag{2}$$

where $Q^t_{2,i,j}$ is the proportion of preschool students in suburb *i* who enrolled in school *j* in year *t*, and $P^t_{i,j}$ is the number of preschool entries from suburb *i* who enrolled in school *j* in year *t*. As with above, a range of averages over time are used to calculate, $\bar{Q}_{2,i,j}$. As most students are enrolled in schools located near their residential suburb, many zeros are present in the $Q_{2,i,j}$ proportions when students occasionally enrol in a school far away from their home. For situations with both zero and nonzero $Q^t_{2,i.j}$ values in the last 5 years, the 5-year average is used. For situations where all values are above zero, the average is based on the minimum standard deviation over time ($\sigma_{Q2,\,i,j}$), calculated for observations from the most recent 3 years to the most recent 7 years. The use of the most recent 3-year averages is applied to account for the sudden shifts in patterns.

Since $Q^t_{2,i.j}$ values should be between zero and one, we describe them using Beta distribution, analogous to Denham (1971). Beta distribution, $\hat{Q}_{2,i,j}$, has mean $\bar{Q}_{2,i,j}$ and standard deviation $\sigma^{\max}_{Q2,i,j}$ based on the maximum standard deviation from the last 3 to 7 years,

$$\hat{Q}_{2,i,j} \sim Beta\left(\alpha = \left(\frac{1-\bar{Q}_{2,i,j}}{(\sigma^{\max}_{Q2,i,j})^2} - \frac{1}{\bar{Q}_{2,i,j}}\right) \cdot \bar{Q}_{2,i,j}{}^2,\ \beta = \alpha \cdot \left(\frac{1}{\bar{Q}_{2,i,j}} - 1\right)\right).$$

Because different pairs of i and j are modelled independently, the sum of $\hat{Q}_{2,i,j}$ across schools from any suburb i needs to be rescaled to 100%, denoted as $\tilde{Q}_{2,i,j}$.

Randomly drawing from distributions $\hat{Q}_{1,i}$ and $\hat{Q}_{2,i,j}$, we can project forward the projected preschool entries, $\hat{P}^t_{i,j}$, for school j from suburb i, with projected 4-year old population estimates drawn from the normal distribution ($\hat{N}^t_{4,i}$) in each year,

$$\hat{P}^t_{i,j} = \hat{N}^t_{4,i} \cdot \hat{Q}_{1,i} \cdot \tilde{Q}_{2,i,j}. \tag{3}$$

However, instead of the algebraic relationship presented in Eq. (3), we project $\hat{P}^t_{i,j}$ by applying a multinomial distribution with the average number of preschoolers in suburb i, $\hat{P}^t_i = \hat{N}^t_{4,i} \cdot \hat{Q}_{1,i}$ with probability $\tilde{Q}_{2,i,j}$. The result has the same mean as Eq. (3) but with a probabilistic distribution included.

3.2.1.2 For preschool entries living outside the ACT. For preschool entries whose home addresses are located outside the ACT, the numbers are typically very small. For the projection inputs, a negative binomial distribution is fitted with 5-year average values, $\bar{P}_{e,j}$. Negative binomial distributions are used because they allow for possible overdispersion in the counts. The negative binomial distribution for $\hat{P}_{e,j}$ with mean μ and dispersion θ is specified as

$$\hat{P}_{e,j} \sim NB\left(\mu = \bar{P}_{e,j},\ \theta = \frac{\bar{P}_{e,j}{}^2}{(\sigma_{Pe,j})^2 - \bar{P}_{e,j}}\right).$$

Note that θ must be positive. When the variance $(\sigma_{Pe,j})^2$ is lower or equal to the mean, θ is set to infinity and it is equivalent to a Poisson distribution. Finally, preschool entries to each school, $\hat{P}^t_j$, are obtained by adding the randomly drawn preschool entries from ACT suburbs, $\hat{P}^t_{i,j}$ with those from outside the ACT, $\hat{P}^t_{e,j}$.

3.2.2 Projecting in-migration of students

In the STEP model, student in-migrants are those who were not in the ACT school system in year t but enrolled in an ACT school in year $t+1$ in academic levels Kindergarten to Year 12 (K-12). Similar to preschool students, in-migration is modelled separately for students moving to the ACT from those moving to areas outside (but near) the ACT. The within-ACT and outside-ACT in-migration numbers are added together to obtain the final projected number of student in-migrants for each school and each academic level.

3.2.2.1 K-12 student migration to the ACT. First, the number of K-12 student migrating to each suburb in the ACT is projected. Second, these numbers are distributed to schools and academic levels. To do this, we apply a 5-year moving average of the number of students migrating to each suburb i, denoted as $\vec{I}_i^t$. To include stochasticity, we fit a negative binomial distribution with mean equal to $\vec{I}_i^t$ and standard deviation $\sigma_{I,i}$ equal to the preceding 5 years of observed data,

$$\hat{I}_i^t \sim NB\left(\mu = \vec{I}_i^t, \theta = \frac{\vec{I}_i^{t2}}{(\sigma_{I,i})^2 - \vec{I}_i^t}\right).$$

To distribute this number to schools and academic levels, we compute the proportion of all in-migrants to suburb i who would enrol in academic level k of school j in year t,

$$S_{i,j,k}^t = \frac{I_{i,j,k}^t}{I_i^t}, \tag{4}$$

where I_i^t is the number of in-migrants from SA2 suburb i in year t; and $I_{i,j,n}^t$ is the number of in-migrants from SA2 suburb i who enrolled in academic level n of school j in year t. Since migration is not a frequent event, $S_{i,j,k}^t$ contains many zero values and fluctuations across years. A 5-year average, denoted as $\bar{S}_{i,j,k}$, helps to smooth out these values for projection, and the standard deviation of this proportion in these 5 years is denoted as $\sigma_{S,i,j,k}$. For uncertainty, a Beta distribution is used,

$$\hat{S}_{i,j,k} \sim Beta\left(\alpha = \left(\frac{1 - \bar{S}_{i,j,k}}{(\sigma_{S,i,j,k})^2} - \frac{1}{\bar{S}_{i,j,k}}\right) \cdot \bar{S}_{i,j,k}{}^2, \beta = \alpha \cdot \left(\frac{1}{\bar{S}_{i,j,k}} - 1\right)\right).$$

Since each $\hat{S}_{i,j,k}$ are drawn from the distribution independently, $\hat{S}_{i,j,k}^t$ for each suburb i are rescaled to a total of 1.0 to ensure all projected student in-migrants $\hat{I}_i^t$ are distributed across ACT schools, denoted as $\tilde{S}_{i,j,k}$. To obtain the projected in-migration for each school j and academic level k of students living in suburb i ($\hat{I}_{i,j,k}^t$), we apply a multinomial distribution with the size of the projected in-migrant numbers in year t for suburb i and the randomly drawn proportions, $\tilde{S}_{i,j,k}$. The multinomial distribution incorporates stochasticity to the forecast outcome while ensuring that the means of the results follow the relationship in Eq. 5,

$$\hat{I}_{i,j,k}^t = \hat{I}_i^t \cdot \tilde{S}_{i,j,k}. \tag{5}$$

3.2.2.2 For students living outside the ACT. For in-migration of students living outside the ACT, we simply use the moving average of the proceeding 5 years by school and academic level. These numbers are usually very small and sporadic.

3.2.3 Transition probability

Transition probabilities are calculated for students in the school system from time $t - 1$ to time t. The probabilities capture students progressing to higher academic levels, transferring to another school, as well as leaving the ACT school system through out-migration or Year 12 graduation. Transition probabilities, $G_{o,d}^{t,t+1}$, are specified for all schools and academic levels from time $t - 1$ to any school or academic level

$$G_{o,d}^{t-1,t} = \frac{M_{o,d}^{t-1,t}}{M_o^{t-1}}, \tag{6}$$

where $M_{o,d}^{t-1,t}$ represents the number of students transitioning from status o to status d between time $t - 1$ and time t; and M_o^{t-1} represents the numbers of students enrolled in status o at time $t - 1$. Note that status d also includes out-migration or graduation.

Similar to the approaches used for preschool entries and in-migration, we apply weighted averages of $G_{o,d}^{t,t+1}$ for projection. The minimum standard deviation ($\sigma_{G,o,d}^{\min}$) from the preceding 3 to 7 years (intervals) is used to define the range for the average calculations. There are two important school-to-school transitions: Year 6 primary school to Year 7 high school and Year 10 high school to Year 11 college.[7] To account for the extra unpredictability and uncertainty of these two transitions, the maximum standard deviation ($\sigma_{G,o,d}^{\max}$) from the last 3 to 7 years is calculated if status d is in Year 7 or Year 11. Since each transition probability should be centred between 0 and 1, we assume $\hat{G}_{o,d}$ follows a Beta distribution with the mean and standard deviation mentioned above,

$$\hat{G}_{o,d} \sim Beta\left(\alpha = \left(\frac{1-\bar{G}_{o,d}}{(\sigma_{G,o,d})^2} - \frac{1}{\bar{G}_{o,d}}\right) \cdot \bar{G}_{o,d}{}^2, \beta = \alpha \cdot \left(\frac{1}{\bar{G}_{o,d}} - 1\right)\right),$$

where $\sigma_{G,o,d}$ is dependent on the academic level in status d. The randomly drawn $\hat{G}_{o,d}$ is rescaled to sum to 1.0 across all d, $\sum_d \hat{G}_{o,d} = 1$, which produces the adjusted probability, $\tilde{G}_{o,d}$.

3.2.4 *Multiregional cohort transitional model*

The results from each of the sub-models described above are combined together in a multiregional cohort component projection framework, which can be expressed in a matrix form as (Guan et al., 2022; Xiang et al., 2023):

$$\mathbf{E}^{t+1} = \mathbf{G}^{t,\,t+1}\mathbf{E}^{t} + \mathbf{P}^{t+1} + \mathbf{I}^{t+1}, \tag{7}$$

where $\mathbf{E}^t$ is a vector of student enrollment by school and academic level at time t. $\mathbf{P}^t$ is a vector for preschool enrollments at time t, $\hat{P}_j^t$. $\mathbf{I}^t$ is a vector for in-migrants into each school and academic level at time t, $\hat{I}_{j,k}^t$. $\mathbf{G}^{t,\,t+1}$ is a matrix for transition probabilities between schools and academic levels from time t to time $t+1$, considering graduates and out-migration. Instead of using the matrix multiplication in Eq. (7), $\mathbf{G}^{t,\,t+1}\mathbf{E}^t$, we apply a multinomial distribution for enrollments in each school and academic level in year t and the randomly drawn probability, $\tilde{G}_{o,d}$ to account for stochasticity in the system.

As with other forecasts, the accuracy of the model is expected to decrease over time due to increased uncertainty in the future. As a result, the prediction intervals should span out over time. However, as demonstrated in the equations, the ratios and probabilities indicated in the method section are essentially time invariant, so the bound only grows as the youngest preschool cohort with full uncertainty progresses through time. However, preschool uncertainty is also predicted to be stable over time. Therefore, we make modifications to the standard deviations depending on the original data to correct for the underrepresentation of uncertainty in our medium-term forecast, akin to the empirical prediction interval approach (Lee & Scholtes, 2014). Every year, all standard deviations (except for the population projection) are increased by 8% without changing the mean. The uncertainty of the results prior to the adjustments can be found in See online supplementary material, Table A1. Without adjustments, the coverage of the observed enrollments by the prediction intervals falls gradually over time. As the decrease in coverage is more prominent among high schools and colleges, we inflated the standard deviation of the transition probability by 50% for students transitioning from Year 6 (last year of primary school) and Year 10 (last year of high school) to Year 7 (first year of high school) and Year 11 (first year of college), respectively. The code for the model is available on the online repository: https://github.com/tyaSHEN/ACTedu.

4. Results

To assess the coverage of our prediction intervals (PI), we produce forecasts of the more recent observed data, assuming this information is missing, and then compare the results—a process known as in-sample forecasting. Given that the data spans 2009–2023, we generate in-sample forecasts using 2016–2022 as base years. Although the prediction interval is the primary interest

[7] Kindergarten is also a different academic category from Preschool. However, since it usually has strong tie with the Preschool, students are likely to stay in the same school as Preschool, and it is not considered as an entry grade.

Table 1. Percentage difference between observed and forecasted (mean) ACT total enrollment by base year

Forecast year	Base year (%)						
	2016	2017	2018	2019	2020	2021	2022
1	−0.7	−0.7	0.0	0.0	0.1	1.0	1.6
2	−1.4	−0.8	−0.1	0.0	1.1	2.6	
3	−1.6	−1.2	−0.3	0.7	2.5		
4	−2.0	−1.7	0.4	2.0			
5	−2.6	−1.1	1.6				

Note. Forecast from the base year 2016 relies on 6 years of data. *Source*: Authors' calculation.

of this research, we first briefly examine the overall model performance before evaluating the coverage of the observed values by the prediction interval in the in-sample forecasts.

4.1 Model performance

In Table 1, we compare the percentage difference between the observed and our forecasted mean values by the base year (2016, 2017, ..., 2022) and year of forecast (1 year forward, 2 years forward, ..., 5 years forward). For example, the first cell (−0.7%) represents that the forecasted value in 2017 from the base year of 2016 is 0.7% lower than the total enrollment observed in 2017 of ACT Schools Census. All the forecasted values are within 3% of the actual values. The difference typically increases over the forecasted year unless the sign of the difference has flipped over the years. This measure illustrates the overall bias of the results where the over-forecasted values cancel out the under-forecasted values.

Similar to Xiang et al. (2023), we compute the mean absolute percent error (MAPE) to measure the accuracy of the forecasting model. It is expressed as,

$$MAPE_t = \frac{1}{n}\sum_{i}^{n}\left|\frac{Est_t^i - Obs_t^i}{Obs_t^i}\right| * 100\%, \tag{8}$$

where Est_t^i represents the forecasted mean value for school i at year t and Obs_t^i is the corresponding observed value. Simpson (1989) suggests that MAPE consists of two proportions, over-forecast and under-forecast, and it would be beneficial to examine them separately for planning purposes. These two proportions should add up to MAPE. In Table 2, we present MAPE and the two proportions in parentheses (under-forecast/over-forecast) for the 1–5-year forecasts using 2016, 2017, ..., and 2022 as base years. There are about 130 distinct schools (N in the base year) in the ACT with small increases over the years. MAPE increases over time from 4.6% on average in the first year to 12.8% in the fifth year. On average, the under-forecast proportions are roughly the same as the over-forecast proportions after 5 years: 6.7% and 6.1%, respectively. MALPE indicators by base year are presented in See online supplementary material, Table A2.

In summary, our model exhibits a reasonable level of robustness and introduces several improvements to the model proposed by Xiang et al. (2023). Firstly, the model leverages a more relevant population to calculate preschool enrollment. Secondly, it employs weighted averages for the ratios. Lastly, the model is constructed within a probabilistic framework, facilitating the derivation of prediction intervals. While the probabilistic framework might not directly enhance model performance, it adds insights into the forecast's reliability and uncertainty, aspects that are further evaluated in subsequent sections.

4.2 Prediction interval

Using the Monte Carlo simulation, we obtain 2,000 forecasts of enrollment by school and academic level (or unit level). Among these 2,000 iterations, we calculate the 80% prediction interval (PI)

Table 2. Mean absolute percent error (MAPE) by base year

Forecast year	Base year (%)							
	2016	2017	2018	2019	2020	2021	2022	Average
N	128	128	128	129	129	130	131	
1	4.8	5.3	4.5	4.9	3.8	4.0	5.1	4.6
	(2.7/2.1)	(3.0/2.3)	(2.3/2.2)	(2.2/2.7)	(2.0/1.8)	(1.7/2.4)	(1.6/3.5)	(2.2/2.4)
2	8.1	7.9	7.7	6.3	6.0	7.9		7.3
	(4.6/3.5)	(4.4/3.5)	(3.7/4.0)	(3.2/3.1)	(2.7/3.3)	(2.7/5.2)		(3.5/3.8)
3	10.7	10.5	8.7	8.0	9.2			9.4
	(6.1/4.6)	(5.6/4.9)	(4.6/4.1)	(3.8/4.2)	(3.4/5.8)			(4.7/4.7)
4	13.1	11.2	10.2	10.9				11.4
	(7.1/6.1)	(6.5/4.7)	(5.0/5.1)	(4.4/6.5)				(5.7/5.6)
5	13.6	12.1	12.8					12.8
	(7.8/5.8)	(6.8/5.4)	(5.4/7.3)					(6.7/6.1)

Note. Forecast from the base year 2016 relies on 6 years of data. *N* is for the base year. Proportions under-forecast and over-forecast are in parentheses (under-forecast/over-forecast). *Source*: Authors' calculation.

and 95% PI by the 10 (lower) and 90 (upper) percentiles, and 2.5 and 97.5 percentiles, respectively. To compare the bound of the PI, we also compute the margin of error percentage (MEP) for 80% PI and 95% PI. This measure (cf. Grip & Grip, 2020; Rives, 1977) describes the halved differences between the upper and the lower percentiles of the PI relative to the median,

$$\text{MEP} = \frac{\text{upper percentile} - \text{lower percentile}}{2 * \text{median}}.$$

This measure captures the variation of the forecast. A 1% of MEP represents that this PI is 1% above or below the median. The larger the MEP, the higher the uncertainty in the forecast.

To produce PIs for higher levels such as schools or regions, numbers of enrollment are aggregated to the required level in each of the 2,000 iterations and calculated from the iterations of the aggregated numbers. For example, at the ACT level, we sum all the enrollments in each of the 2,000 iterations and then compute the median, 80% PI and 95% PI from these 2,000 iterations of total enrollment. The resulting uncertainty or MEP is smaller than at lower levels of analysis because it aggregates more unit-level forecasts before calculating the PIs. In the following sections, we present and describe the PIs at different levels.

4.2.1 Unit level

The prediction interval is designed on the unit-level forecast (i.e. by school and academic level). Thus, it is important to evaluate whether the unit-level prediction interval can adequately account for the uncertainty in the observed values. For instance, 80% of the observed values should be covered by the 80% PI. Table 3 displays the 80% and 95% PI coverage for the 1–5-year forecasts using 2016, 2017, …, and 2022 as base years. There are about 1,000 combinations (*N* in the base year) of school and academic level in the ACT. The coverage slowly declines throughout the course of the forecast year but is comparatively stable for the same forecast year. The 5-year forecasts show that, on average, the 80% PI covers 82% of the observed values and the 95% PI covers 95%. The PIs are remarkably close to the designed coverages. See online supplementary material, Tables A3 and A4 disaggregate all unit-level observations by sector (public, CEO, and independent) and by academic categories (preschool, primary school, high school, and college). Although the information of these higher levels is not included in the model, PIs are robust and align with the design coverages across different groups over forecast years.

Table 3. Coverage of the prediction interval (PI) by base year

PI	Forecast year	Base year (%)							
		2016	2017	2018	2019	2020	2021	2022	Average
	N	971	972	975	983	985	995	1,006	
80%	1	84	83	85	84	90	88	84	85
	2	81	81	83	86	87	82		83
	3	77	79	83	84	84			81
	4	76	78	83	82				80
	5	75	79	81					78
	Average	78	80	83	84	87	85	84	82
95%	1	96	95	97	96	98	97	96	96
	2	93	94	96	96	96	94		95
	3	92	93	95	96	95			94
	4	92	93	96	93				93
	5	92	94	94					93
	Average	93	94	96	95	96	96	96	95

Note. Forecast from the base year 2016 relies on 6 years of data. *N* is for the base year. *Source*: Authors' calculation.

Table 4. Average margin of error percentage (MEP) of the aggregated prediction intervals (PI) at school level of by base year (public schools only)

PI	Forecast year	Base year (%)							
		2016	2017	2018	2019	2020	2021	2022	Average
	N	82	82	82	83	83	84	85	
80%	1	6.2	6.0	6.0	5.8	5.9	5.8	5.7	5.9
	2	8.0	7.8	7.7	7.6	7.6	7.5		7.7
	3	9.2	8.9	8.8	8.7	8.6			8.8
	4	9.9	9.8	9.7	9.5				9.7
	5	10.8	10.5	10.4					10.6
95%	1	9.6	9.2	9.3	9.0	9.1	8.9	8.8	9.1
	2	12.2	12.0	11.8	11.7	11.7	11.5		11.8
	3	14.1	13.6	13.7	13.3	13.2			13.6
	4	15.3	15.0	14.9	14.5				14.9
	5	16.6	16.1	16.0					16.2

Note. Forecast from the base year 2016 relies on 6 years of data. *N* is for the base year. *Source*: Authors' calculation.

4.2.2 School level

For the school level, we aggregate the PI to the level of each distinct school. Table 4 shows the average of the MEP of all the public schools across 1–5-year forecasts based on 2016, 2017, …, and 2022 base years. The row denoted as '*N*' provides the number of public schools involved for calculating the average. The MEPs are rather consistent in the same forecast year and increasing over forecast years. Notably, the 95% PI indicates a higher degree of uncertainty (as reflected in the MEP) compared to the 80% PI. On average, the 80% PI deviates by 5.9% above or below the median in the initial year of forecast and by 10.6% in the fifth year. Given the varying sizes of schools, ranging from approximately 300 to 1,200 students, a 10% deviation could correspond to an uncertainty of 30 or 100 students. While larger schools are generally more adept at accommodating fluctuations, we should anticipate comparatively smaller MEP values for a larger school.

Figure 1 presents examples of individual schools of different academic categories in the forecast with the base year of 2018. Since the model performs similarly across base years, we pick 2018 as an example as it is the last year with 5 years of forecast. In this figure, we show the observed and forecasted values with PIs for every academic level in this school and the total school enrollment of all these academic levels. The MEPs for 2019 (the first forecast year), 2021 (the third), and 2023

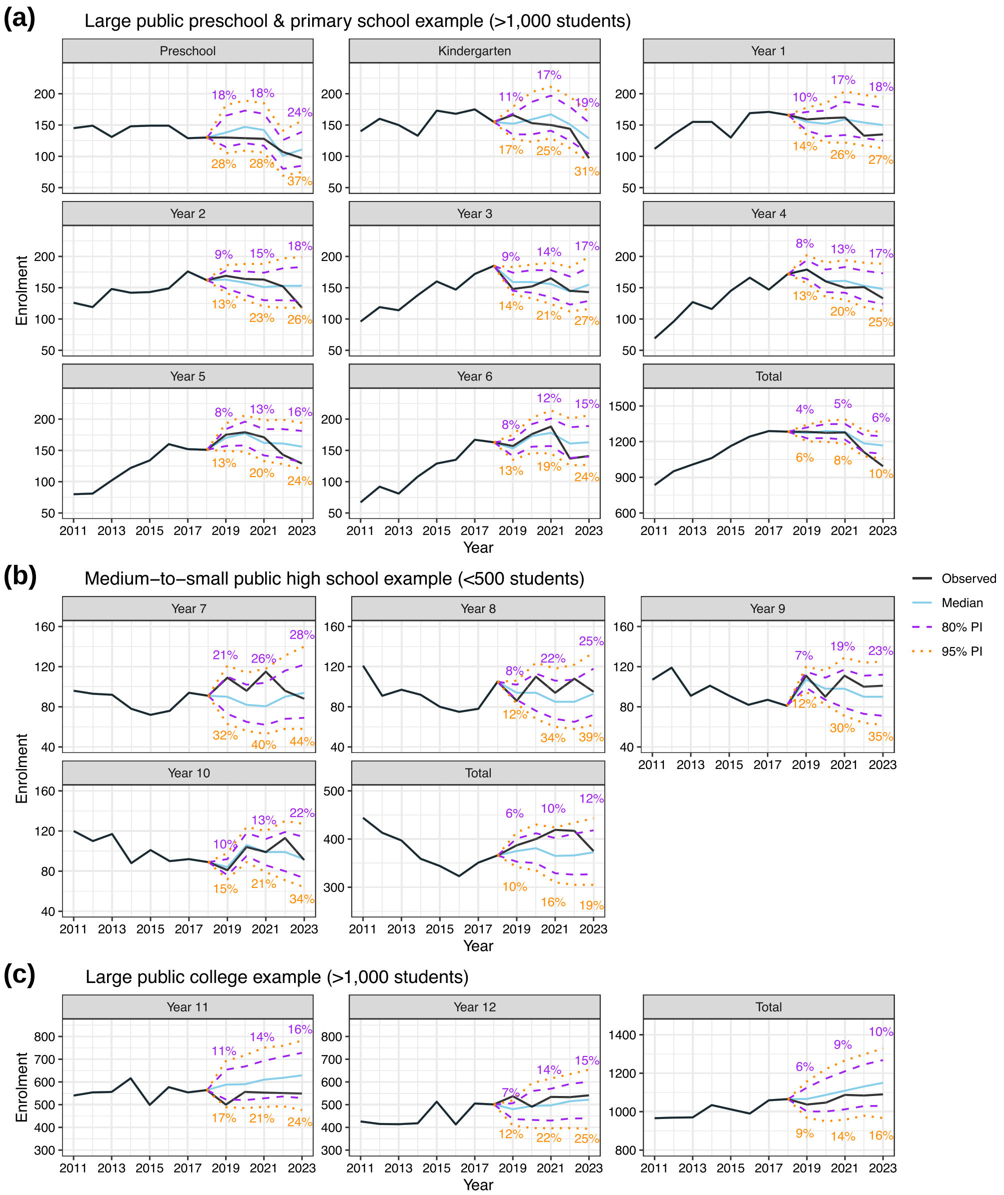

Figure 1. Large preschool & primary school (a), Medium-to-small high school (b), and large college (c) examples of public school forecasts with base year of 2018. *Source*: ACT School Enrolment Census and authors' calculation.

(the fifth) are illustrated as percentages in the figure. The numbers above (in purple) are for 80% PI and the ones below (in orange) are for 95% PI.

Figure 1a demonstrates a relatively large preschool and primary school accommodating more than 1,000 students at all levels combined. Figure 1b portrays a medium-to-small high school with less than 500 students and Figure 1c is a large college with about 1,000 students. The entry levels (Preschool, Year 7, and Year 11) in different panels have the highest uncertainty, while the other levels are predicted to be more stable. This is consistent with observed variations and our model assumption. See online supplementary material, Table A5 in the Appendix illustrates the MEP of the aggregated PI at the academic level of public schools, averaged across all base years. The MEPs for the entry levels are slightly higher than the average of the other levels after aggregating all the enrollment in the same level.

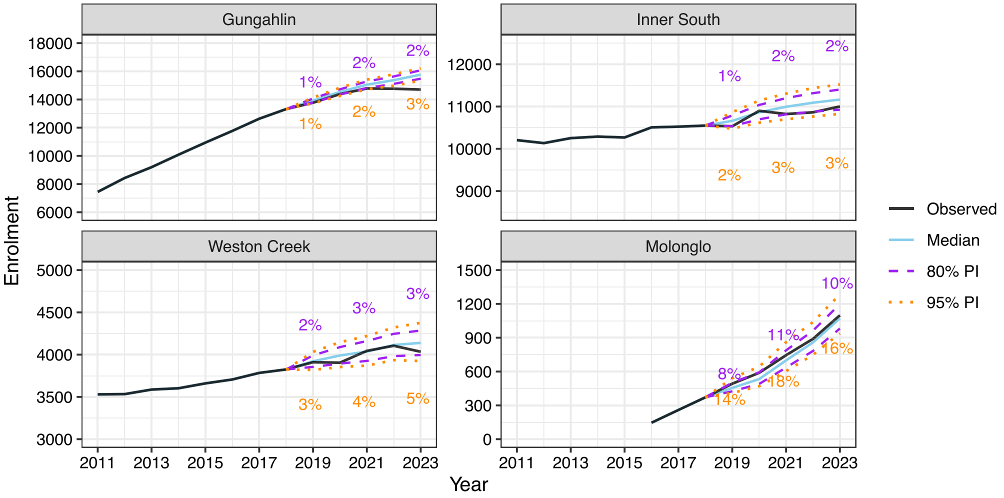

Figure 2. Enrollment by district (selected) and aggregated prediction intervals (PI) at district level, base year 2018. *Source*: ACT School Enrolment Census and authors' calculation.

The PI of individual academic levels in these schools may appear to be wide, but it captures the actual fluctuation in the unit level. For instance, the enrollment of Year 6 in Panel A has shown a historical upward trend, with observed values surpassing the forecasted before 2021. However, it drops below the 10th percentile in 2022 and 2023. The underlying reasons for this decline remain unobservable in the data and hence cannot be controlled in the model. Without the prediction intervals, the median (or the point estimate) would be inaccurate. Conversely, due to the presence of prediction intervals, this deviation is somewhat expected and possible, considering the inherent fluctuations in the data.

Compared to the relatively high uncertainty in each academic level of these schools (in Figure 1), the MEP in the total enrollment for each school is much smaller. Furthermore, the larger the school, the smaller the MEP across the forecast years as we expected. For the large primary school, the deviation between the 80% PI and the 95% PI is about 6% and 10%, respectively, after a 5-year forecast period. In contrast, these values reach 12% and 19% for the smaller high school. However, beyond the difference in the size of the school, the observed variation is also much higher in the high school and college examples compared with the primary school.

4.2.3 *Region and ACT level*

Figure 2 presents the PI and MEP at the regional level. As mentioned in the background, there are nine districts where most people live in the ACT. We selected four of them with slightly different traits. They can be broadly characterized as large (Gungahlin), medium (Inner South), small (Weston Creek), and new (Molonglo Valley) districts. Most of the forecasted values in these regions are within our PI and the MEP is very small, aside from Molonglo Valley. The variation is slightly higher for Molonglo Valley because school enrollments in this region are still relatively small and fewer observed values are available. The observed values in Gungahlin are slightly lower than the forecasted lower bound after 2022. Gungahlin is a fast-growing region that attracts many young migrant families, from elsewhere in Australia and from other countries. The reason for the gap between observed and predicted school enrollments in 2022 and 2023 is mainly due to lower preschool entries than expected. Lower than expected enrollment of preschool students occurred across the ACT in these years (as shown in Figure 3). Preschool entries are particularly challenging because no administrative data exists within the school system that could be used to forecast entries. For the STEP model, preschool entries are based on population estimates and projections of 4-year olds provided by CMTEDD.

4.2.4 *Sources of growth*

Figure 4 presents the results by source of growth components, illustrating the changes in the overall ACT enrollment. Forecasts for preschool entries and Year 12 graduations mostly fall within the

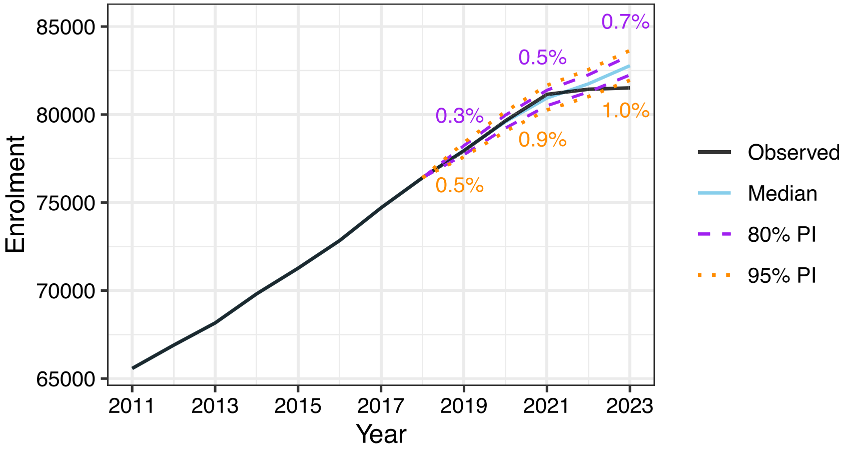

Figure 3. Enrollment and aggregated prediction intervals (PI) for ACT, base year 2018. *Source*: ACT School Enrolment Census and authors' calculation.

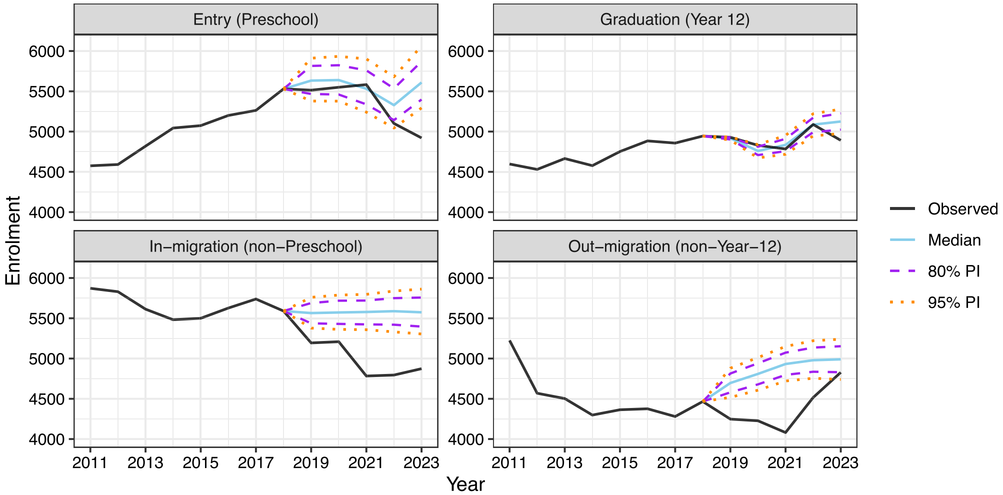

Figure 4. Sources of growth of ACT total enrollment change, base year 2018. *Source*: ACT School Enrolment Census and authors' calculation.

prediction interval. However, the model failed to capture the changes observed for the migration components. In-migration encompasses all academic levels except preschool entries, while the out-migration component incorporates all levels except Year 12. It is worth noting that except for the in-migration component, our model operates on the modelling of ratios or proportions, resulting in the counts in Figure 4.

We assume levels of in-migration will remain flat for the projection period. This represents a compromise between the downward trends indicated by the data and beliefs by experts that in-migration would increase.[8] Additionally, the sharp decrease after 2020 is likely resulting from the border closures during the COVID-19 Pandemic. As for the out-migration component, it is a more intricate issue since the count is predicted by proportions. The proportions of out-migration are slightly decreasing in the ACT; our model assumes fixed proportions. An increase in the total number of students, accompanied by a higher level of in-migration, resulted in a predicted increase in out-migration.

[8] The Education Directorate also undertook custom modelling to allow for a higher migration scenario consistent with CMTEDD official population projections.

Furthermore, discrepancies in the migration components may be attributed to the gradual improvements in the data linkage of student records. The Education Directorate has revealed that their administrative data team has invested in enhancing student linkages over time. Migration for each school and academic level involves sparse numbers that complicate the projection of future numbers.

5. Discussion

The projection model described in this paper provides probabilistic forecasts of school enrollments, taking into account cohort transitions, preschool and migration entries, and transfers between schools and out of the school system. A stochastic approach offers additional insight for school planners by providing measures of forecast uncertainty. This is important for planning and providing the short to medium term support (budget allocations, staffing levels, physical/built infrastructure, material school resources and equipment, operational programs such as 'Meals in Schools' pilot, etc.) required for students and teachers in all schools under the jurisdiction of the school planning entity. The model also includes the underlying components of school enrollment change at each academic level. By providing prediction intervals for both the enrollment forecasts and the sources of enrollment change, education departments can better anticipate the needs at school and district levels and plan longer term interventions such as the construction of new or expanded schools, and the creation or adjustment of priority enrollment area boundaries.

Grip and Grip (2020) found that the results from their probabilistic cohort survival ratio (CSR) model presented potential challenges for producing practical uncertainty measures that districts with fewer than 25,000 students could use. The advantage of our model is that it accounts for the variance in each of the sources of the enrollment change while the Grip and Grip (2020) CSR model treated them as unobserved variance in each grade. The probabilistic multiregional cohort component model highlights its effectiveness in producing practical uncertainty in schools as small as 1,000 students with around 5% above or below the median estimates after 5 years of forecast. Our predicted uncertainty would be even smaller if aggregating at a higher level. Both Rives (1977), and Grip and Grip (2020) suggest that 5% uncertainty is the cutoff of the practical interval for planning purposes when the student number is above 1,000. Yet, we also estimate by school and academic levels, a much smaller unit, that represent 50–200 students. In these cases, we would suggest that it is feasible for the education department to account for 10%–15% of uncertainty in the planning.

Although the outcome of the model demonstrates its capability in reliably forecasting school enrollments, there are some limitations and need for further research. Our model uses record-level administrative data as inputs for student enrollment projections. In other school districts, the level of detail available may vary or there may be obstacles regarding access and consistency over time. If detailed data are not available on the sources of enrollment change, then users will have to use simpler models that do not capture transitions between schools and instead use overall change ratios, as discussed in Section 2.2. However, with many school districts now having reliable and detailed administrative data, we believe the STEP model framework with information on sources of enrollment change and corresponding predictive intervals will greatly enhance school planning. Through extensive testing, we found the STEP model only requires approximately 8 years of data to produce reliable forecasts for at least 5 years.

There are several promising directions for future research worth exploring. Firstly, our model adopts commonly used probability distributions, such as the beta and sigma distributions. Alternative distributions could be explored, such as the two-sided power distribution and truncated normal distribution (Kotz & van Dorp, 2004; van Dorp & Kotz, 2002). These distributions offer greater flexibility, especially in scenarios involving small or high ratios. Secondly, when estimating the parameter for the distribution, maximum-likelihood estimation could be employed. In our preliminary tests, the estimated parameters yielded negligible differences across various methods. The decision to employ the method of moment estimator resonates with the intuitive understanding of mean and standard deviation in the observed data, facilitating a more straightforward implementation for school planners. Thirdly, integrating additional factors such as priority enrollment areas, proximity to schools, and the attractiveness of schools could significantly enhance the model. Including these variables would provide valuable insights, particularly when students transition to higher educational stages (e.g. moving from primary to secondary school). This approach would better inform the model about the dynamics affecting student enrollment and school choice. Lastly, it would be beneficial to evaluate and compare the prediction intervals resulting from various model assumptions by applying them to the same dataset over several years. This would aid in assessing model performance and robustness under consistent data conditions.

6. Conclusion

In the Australian Capital Territory (ACT), the probabilistic multiregional cohort component model described in this paper is now being used to forecast school enrollments based on anonymized annual record-level administrative data. The model simultaneously forecasts all students in each ACT school and grade, accounting for preschool entries, in-migration, transfers between schools, grade progression, out-migration, and Year 12 graduations. The model has demonstrated robust performance in the 5-year in-sample forecasts. Given the primary focus of this paper on the uncertainty surrounding school enrollment, we compare the prediction intervals against the observed data within the in-sample forecasts. Our assessment reveals that 82% of the observed enrollments by school and academic level fall within the 80% prediction interval, while 95% are within the 95% prediction interval over the 5-year period. The method produces reasonable and practical uncertainty, particularly for larger or well-established schools and at higher aggregate levels. In other words, the prediction intervals effectively encapsulate the actual fluctuations while maintaining a narrow enough bound for the education department to prepare for potential future scenarios. This model provides school planners capacity for detailed analyses and scenario building for specific sources of change. Finally, the model framework is adaptable and can be reproduced in other school districts, provided that the local administration has access to record-level data.

Acknowledgements

Many thanks to Elizabeth Howell, Sean Das, Hieu Nguyen, and Aizanel Patenia for providing valuable suggestions and helpful comments to construct the model and improve this paper.

Conflicts of interest: The authors declare no conflict of interest.

Funding

This research is made possible by funding from the ACT Education Directorate.

Data availability

The administrative records containing individual information cannot be published due to privacy concerns. However, aggregated enrollment numbers by school and academic level for each year are publicly available at ACT Education Census (https://www.education.act.gov.au/about-us/policies-and-publications/publications_a-z/census). The code for our model, along with the model's predictions, is accessible on GitHub: https://github.com/tyaSHEN/ACTedu and can be compared with the Education Census published by the ACT government.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series A*.

References

ACT Chief Minister, Treasury and Economic Development Directorate. (2023). *ACT population projections 2022 to 2060*. https://www.treasury.act.gov.au/__data/assets/pdf_file/0007/2181985/ACT-Goverment-population-projections-2022-2060.pdf

Andersen, M., Cole, R., Dunn, T., Krupka, D., & Pato, J. (2014). *Forecasts for the Lexington Public Schools: FY2015-FY2020 report of the enrolment working group, report prepared for district*. Lexington School Committee. https://www.lexingtonma.org/wp-content/uploads/2019/04/13Jan15SCAgendaPacket.pdf

Australian Bureau of Statistics (ABS). (2022). *Snapshot of Australian Capital Territory: High level summary data for Australian Capital Territory in 2021*. https://www.abs.gov.au/articles/snapshot-act-2021#:~:text=Housing-,Population,the%20Census%20counted%20137%2C000%20people.

Braden, B. (1972). *Enrollment forecasting handbook introducing confidence limit computations for a cohort-survival technique*. New England School Development Council. https://eric.ed.gov/?id=ED066781

Denham, C. H. (1971). *Probabilistic school enrollment predictions using Monte Carlo computer simulation. Final report*. Boston College. https://eric.ed.gov/? id=ED062729

Edmonston, B. (2000). *The path ahead: Future enrollments in Portland Public Schools, 2000–2010*. Portland State University. http://archives.pdx.edu/ds/psu/8969

Grip, R. S., & Grip, M. L. (2020). Using multiple methods to provide prediction bands of K-12 enrollment projections. *Population Research and Policy Review*, *39*(1), 1–22. https://doi.org/10.1007/s11113-019-09533-2

Guan, Q., Raymer, J., & Gray, E. (2022). Identifying different sources of school enrolment change in the Australian Capital Territory. *Australian Population Studies*, *6*(1), 37–40. https://doi.org/10.37970/aps.v6i1.99

Hussar, W. J., & Bailey, T. M. (2020). *Projections of education statistics to 2028* (47th ed.). National Center for Education Statistics. https://nces.ed.gov/pubs2020/2020024.pdf

Johnstone, J. N. (1974). Mathematical models developed for use in educational planning: A review. *Review of Educational Research*, *44*(2), 177–201. https://doi.org/10.2307/1170163

Kotz, S., & Van Dorp, J. R. (2004). *Beyond Beta: Other continuous families of distributions with bounded support and applications*. World Scientific Publishing Co. Pte. Ltd. https://doi.org/10.1142/5720#t=toc

Lee, Y. S., & Scholtes, S. (2014). Empirical prediction intervals revisited. *International Journal of Forecasting*, *30*(2), 217–234. https://doi.org/10.1016/j.ijforecast.2013.07.018

Raymer, J., Biddle, N., & Guan, Q. (2017). A multiregional sources of growth model for school enrolment projections. *Australian Population Studies*, *1*(1), 26–40. https://doi.org/10.37970/aps.v1i1.10

Rives, N. W. (1977). Forecasting public school enrollment. *Socio-Economic Planning Sciences*, *11*(6), 313–318. https://doi.org/10.1016/0038-0121(77)90017-9

Rushton, G., Armstrong, M. P., & Lolonis, P. (1995). Small area student enrollment projections based on a modifiable spatial filter. *Socio-Economic Planning Sciences*, *29*(3), 169–185. https://doi.org/10.1016/0038-0121(95)00007-9

Rynerson, C., & Wei, C. (2022). *Portland Public Schools enrollment forecast 2022–23 to 2036–37, based on October 2021 enrollments*. Portland State University. https://archives.pdx.edu/ds/psu/38215

Schellenberg, S. J., & Stephens, C. E. (1987). *Enrollment projection: Variations on a theme*. Annual Meeting of the American Educational Research Association, Washington DC, April 20–24, 1987.

Shaw, R. C. (1984). Enrollment forecasting: What methods work best? *NASSP Bulletin*, *68*(468), 52–58. https://doi.org/10.1177/019263658406846810

Simpson, S. (1987). School roll forecasting methods: A review. *Research Papers in Education*, *2*(1), 63–77. https://doi.org/10.1080/0267152870020105

Simpson, S. (1988). The use of school roll forecasts in LEA administration: The allocation of resources to schools. *Cambridge Journal of Education*, *18*(1), 89–98. https://doi.org/10.1080/0305764880180107

Simpson, S. (1989). School roll forecasts: Their uses, their accuracy and educational reform. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *152*(3), 287–304. https://doi.org/10.2307/2983127

Swanson, D. A., & Tayman, J. (2012). *Subnational population estimates* (1st ed.). Springer Netherlands. https://doi.org/10.1007/978-90-481-8954-0

Van Dorp, J. R., & Kotz, S. (2002). The standard two-sided power distribution and its properties: With applications in financial engineering. *The American Statistician*, *56*(2), 90–99. https://doi.org/10.1198/000313002317572745

Xiang, L., Raymer, J., & Gray, E. (2023). The school transition estimation and projection (STEP) model: A flexible framework for analysing and projecting school enrolments. *Population, Space and Place*, *29*(5), e2681. https://doi.org/10.1002/psp.2681

Yang, S., Chen, H.-C., Chen, W.-C., & Yang, C.-H. (2020). Student enrollment and teacher statistics forecasting based on time-series analysis. *Computational Intelligence and Neuroscience*, *2020*, 1246920. https://doi.org/10.1155/2020/1246920