Proceedings of the 2022 Winter Simulation Conference
B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C.G. Corlu, L.H. Lee, E.P. Chew, T. Roeder, and P. Lendermann, eds.

AN APPROACH TO POPULATION SYNTHESIS OF ENGINEERING STUDENTS FOR UNDERSTANDING DROPOUT RISK

Danika Dorris
Julie Ivy
Julie Swann

North Carolina State University
915 Partners Way
Raleigh, NC 27695, USA

ABSTRACT

Dropping out of STEM remains a critical issue today, and it would be useful for universities to have reliable predictive models to detect students' dropout risks. Generating a synthetic population of the true population could be useful for simulating the system and testing scenarios. We outline an approach for creating a synthetic population of students in STEM and build a microsimulation which simulates students' risk behaviors over time. This process has identified several areas that must be addressed before the synthetic population represents the true population in a simulation.

1 INTRODUCTION

Sixty percent of college students dropout within their first two years of school (Chen et al. 2018), which highlights the importance of early identification of students at risk of dropping out of college and intervention to reduce this risk (Dewantoro and Ardisa 2020; Ortiz-Lozano et al. 2020; Villano et al. 2018; Golding and Donaldson 2006). Studies have identified a number of potential factors related to departure from STEM programs. Some studies have shown that some students drop out of STEM programs due to curriculum difficulty (Marra et al. 2012; Seymour and Hewitt 1997), and many others have found that students primarily dropout due to a lack of sense of belonging and institutional support (Watkins and Mazur 2013; Burke 2019). Furthermore, not only is the need for institutional support critical, but the timing of the support can influence its effectiveness (Ishitani and Desjardins 2002). If at-risk students can be identified early, then interventions could be targeted to reduce departures in ways that are cost-effective.

Machine learning has been used to predict student dropout using a variety of information such as academic, financial, and admissions data at different points throughout the first semester (Dewantoro and Ardisa 2020; Fernandes et al. 2019; Naseem et al. 2019). One study demonstrated a dynamic decision model in which a student dropping out is influenced by their beliefs about future academic performance (Stinebrickner and Stinebrickner 2014). Another study demonstrated a process for predicting the dropout of distance learners while also considering the perceived reliability of model results by experts (Freitas and Salgado 2020).

Population synthesis has been used in a wide range of social science studies from analyzing the spread of infectious diseases among people in romantic relationships (Scholz et al. 2016) to addressing workplace location assignments (Fournier et al. 2021) Although many studies have generated synthetic social populations (Krauland et al. 2020; Wu et al. 2018; Wu et al. 2022), there is no clear consensus on the best way to synthesize a population that most accurately represents the actual population, which is likely due to the nuances of the attributes of the real population and the specific outcomes of interest (Chapuis and Taillandier 2019). There have been several methods for creating a synthetic population that

can address the correlations among attributes in a multivariate distribution, such as the Gibbs sampler and Iterative Proportional Fitting (IPF). However, these common methods become too computationally intensive as the number of features increases (Farooq et al. 2013). Bayesian Networks can be used to represent underlying relationships between the large number of variables without requiring as much computational effort (Sun and Erath 2015). Bayesian networks have been used to create synthetic representations of large populations such as Jakarta, Indonesia (Ilahi and Axhausen 2019) and social influence on travel behaviors (Zhang et al. 2019).

With a synthetic population that is representative of students over time, we are able to observe how dropout risk fluctuates over time and understand how specific university interventions may impact students. In this paper, we construct a simulation built upon a synthetic population, with the overarching goal of evaluating potential interventions. We generate the synthetic population to represent the actual population with more than 100 different attributes accumulated over time. We match individual people from an historical test cohort with a representative synthetic agent and simulate dropout risk over several semesters. We assess the approach in several ways to determine appropriateness.

2 DATA

We used historical data from 5,348 undergraduate engineering first-year students across four undergraduate engineering cohorts at a large public institution between 2013 and 2016 to train our prediction models. The 2017 undergraduate engineering cohort consisting of 1,428 students, was used to test both the prediction models and the simulation. The university defines a cohort as first-year and transfer students who were admitted in either the summer or fall terms and enrolled in the fall term. Students whose first term is in the spring are not included in the cohort. Note that we only included first-year students in our analyses. The university defines "dropout" as a student who leaves the university and does not return within six years of their first term. In our analysis, we focus on students whose dropout occurs by the beginning of the third academic year, or within two years from their start at the university.

The university records data on both the census day and the last day of the semester. Census day occurs nine class days after the beginning of the semester and is the last day for students to change their course schedule and receive tuition refunds. Table 1 provides a sample of the information included in each type of model and when each variable is known within a given a semester. For example, we know how many credits a student is enrolled in for the semester on census day, but we do not know the grade for those credits until the end of the term. Note that the before start column only refers to the model built before the beginning of their college career. There is only one instance of this model.

Our data primarily consists of nine types of information: academic performance, community engagement, course load, demographics, financial aid, housing, admissions, residency and relatives' education level. Academic performance consists of GPA, AP credits earned, and SAT Score, among other indicators for failing or dropping a course. Community engagement refers to a student's involvement in various student organizations that promote social connection. Course load describes the number of credits a student is enrolled in during a semester. Demographics collected include race, ethnicity, age, and gender. Financial aid information includes the types of financial aid received (e.g. Pell grant, private scholarship, federal loan). Housing variables describe the area of campus a student lives in or indicates if the student lives off campus. Admissions information includes factors related to a student's registration and academic program (e.g. student athlete, major). Residency refers to the location of the student's permanent address. In our data, relatives' education level refers to the number of relatives who have received a college degree.

3 METHODS

3.1 Generating Agent Attributes

A Bayesian network was constructed at each point in time to randomly assign values to newly revealed attributes. In this way, we are able to capture dependencies inherent in a multivariate distribution of a

Table 1: Abbreviated list of factors included in each type of model. The "Census" and "End of Term" columns show which types of information were known at those points in time for a given term.

Variable	Type	Before Start	Census	End of Term
Dropped out within two years	Response	x	x	x
AP Credits	Academic	x	x	x
SAT Composite Score	Academic	x	x	x
Student Athlete	-	x	x	x
Engineering First Year Major	Program	x	x	x
Federal Loan Recipient	Financial Aid	x	x	x
Private Scholarship Recipient	Financial Aid	x	x	x
First Generation Student	Education Level	x	x	x
Gender	-	x	x	x
In-State Residency	Residency	x	x	x
On-Campus Housing near Dining Hall	Housing	x	x	x
Race and Ethnicity (White, Black or African American, Hispanic/Latino, Asian, American Indian/Alaska Native, Not Specified, Native Hawaiian or Other Pacific Islander)	Race/Ethnicity	x	x	x
Off-Campus Housing	Housing	x	x	x
On-Campus Housing near Student Center	Housing	x	x	x
Civil Engineering Major	Program	x	x	x
Change in Degree Program	Program			x
Did Not Enroll in Term	Registration	x	x	x
Engineering Living & Learning Village Member	Community Engagement	x	x	x
Women in Science and Engineering Member	Community Engagement		x	x
Suspended	Registration			x
Graded Credits Enrolled in at Census	Course Load		x	x
Part-time at Census	Course Load		x	x
Dropped at least 1 course	Course Load			x
International Student	Residency	x	x	x
Failed at least 1 course	Academic			x
Permanent Address in Southern US	Residency	x	x	x
Term GPA > 0.0, < 1.0	Academic			x
Term GPA $>= 1.0$, < 2.0	Academic			x
Both Relatives have College Degrees	Education Level	x	x	x
Term GPA $>= 3.0$, < 4.0	Academic			x
Term GPA 4.0	Academic			x

large set of attributes. Numeric variables were normalized by the min-max method and discretized before constructing the Bayesian network. The structure of the network was determined using the tabu search method, which is a heuristic method that identifies the best network structure according to some objective without terminating at local optima that are not globally optimal (Glover and Laguna 1998). In our case, we chose the network which minimized the Bayesian Information Criterion (BIC). Once the network structure was determined, we learned the conditional distributions from fitting the test data and minimizing the BIC. Sets of attributes known before the start of the first semester were randomly generated using this network. In subsequent time steps, the additional information that is revealed is inferred based on the known attributes.

Specifically, the distribution of a new attribute conditional upon the set of known attributes is determined by using the "bayes-lw" method in the R package 'bnlearn' , which averages numerous likelihood weighting simulations using all available information (Scutari 2010).

3.2 Dynamic Logistic Regression

We built multiple step-wise logistic regression models using these data at seven points in time. For each of the seven types of prediction, we created 100 fully balanced training sets by pairing all of the dropouts with an equal-sized bootstrap sample of non-dropouts. Training sets were balanced to enhance prediction of relatively low proportion of drop-out response. Figure 1 displays each type of prediction built and which information was used in each prediction. The first model ("Before Start") was built using information known before the start of the first fall semester. Then, we built a "Fall 1 Census" model using information known on census day of the first fall semester, in addition to the average risk score across the 100 regression models built before the start of the semester. Similarly, we built the subsequent models using the information known at that time in addition to the average predicted dropout probability across all models built at the previous time step. Our models span the first three terms of students' academic careers excluding the summer term between the first spring term and the second fall term. Numerical data was normalized according to the min-max method, in which the minimum value of a given numeric variable x was subtracted from each instance of x and divided by the range of x.

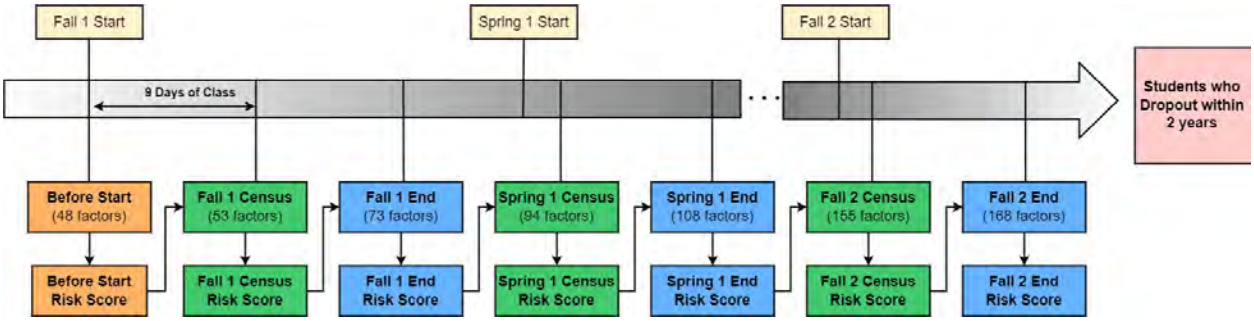


Figure 1: Predictions built and the information known for each prediction. There are always nine days of class between the start of a semester and its census day. The average risk score across the 100 models from a given time step was used as a predictor in models built during the subsequent time step.

3.3 Matching Agents to Actual Student Outcomes

In order to establish a ground truth to evaluate our simulation performance, we matched agents with students from the test cohort and used the outcome of the matched student as the "true" outcome of the agent. We first randomly generated 3,000 agents from the Before Start Bayesian Network. Then, we calculated the distance between a student and an agent in the test population based on the 48 attributes known before the start of the first semester. Since we have a mix of numerical and binary variables, we used the Heterogeneous Euclidean-Overlap Metric (HEOM) distance metric as discussed in (Wilson and Martinez 1997). As shown by Equations 1 through 3 (reproduced from (Wilson and Martinez 1997)) we define a distance function for numerical variables based on the normalized range of the attribute and an overlap function for categorical variables. Because we did not have any missing values in our dataset, the distance function to handle unknown or missing values is omitted.

$$d_a(x,y) = \begin{cases} overlap(x,y), & \text{if } a \text{ is nominal, else} \\ rn\_diff_a(x,y) \end{cases} \tag{1}$$

$$overlap(x,y) = \begin{cases} 0, & \text{if } x = y \\ 1, & \text{otherwise} \end{cases} \tag{2}$$

$$rn\_diff_a(x,y) = \frac{|x-y|}{max_a - min_a} \tag{3}$$

We followed an iterative process for matching students to agents. For a given student and agent, if the HEOM distance was less than or equal to 1.5, then the student was matched with the agent. Otherwise, the HEOM distance would be calculated between that student and other agents until the student matched with an agent. If all distances between a student and available agents are greater than 1.5, then the student remains unmatched and the next student is considered for matching. Once an agent is matched with a student, that agent is no longer available for matching with other students. Once all possible matches have been made within a distance of 1.5, we increased the distance threshold for an acceptable match to 2 and repeated the matching process for the unmatched students and the available agents. We continued increasing the distance threshold and repeating the matching process until all students were matched with an agent.

3.4 Simulation Process Flow

Our simulation models the risk behavior of each agent over time as more information is revealed, spanning the first year and a half of a student's academic career. Profiles for each agent were built prior to running the simulation and consist of two components: student attributes and initial risk score. As shown in Figure 2, the Bayesian Network built before the start of the first semester generated an agent with a set of attributes. Based on these attributes, an initial risk score is then calculated using a randomly selected "Before Start" prediction model.

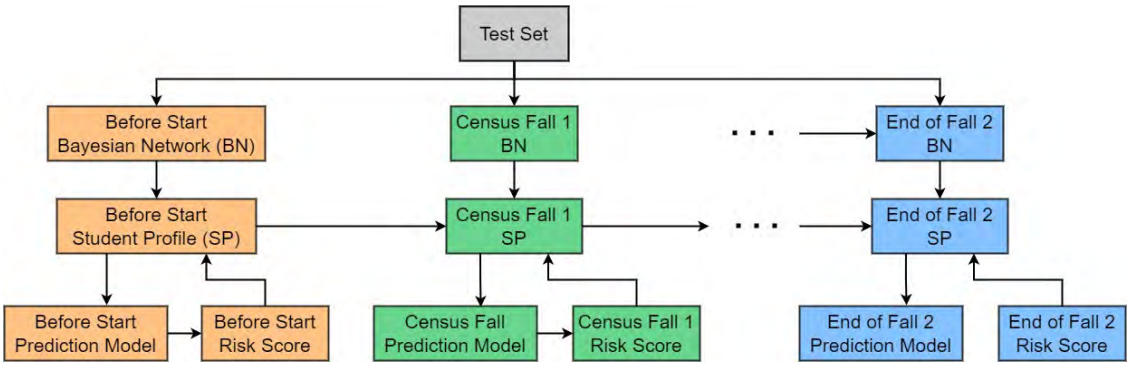


Figure 2: Process flow of simulation.

The agent attributes and risk score carry over to the next time step. At the next time step ("Census Fall 1"), a new Bayesian network is built using all information from the test set known at census of the first fall semester. This new Bayesian network is used to randomly assign values for the new attributes that are revealed based on the values of the known attributes. Once the new attributes are determined, a new risk score is calculated from a randomly sample Census Fall 1 prediction model. This process is repeated at each time step to update the agent's profile as new information is revealed. Note that each of the prediction models may use a different subset of attributes, so the attributes that are used to calculate the risk score in one time step may not be used to calculate the risk score in another time step.

4 RESULTS

Table 2 presents several characteristics of the training and testing data. Each racial group was considered separately in the analysis. However, American Indian/Alaska Native, Native Hawaiian/Other Pacific Islander, and Not Specified have been aggregated in the table for reporting purposes. Among the four cohorts included in the training set, 440 students or roughly 8.22 percent, dropped out within two years. Similarly, 8.19 percent, or 117 students in the test population dropped out within two years.

Table 2: Summary of characteristics describing the training, testing, and synthetic populations. The training population is the set of students used to build the prediction models.

Variable	Training (n=5,348)	Testing (n=1,428)	Synthetic (n=1,428)
Gender			
Male	75.22%	72.27%	75.28%
Female	24.78%	27.73%	24.72%
Race			
White	80.60%	76.75%	76.26%
Black or African American	4.52%	4.97%	4.41%
Asian	12.86%	15.83%	16.46%
American Indian/Alaska Native, Native Hawaiian/Other Pacific Islander, or Not specified	5.49%	6.09%	5.04%
Two or More Races	3.77%	3.99%	2.73%
Ethnicity			
Hispanic/Latino	4.31%	5.32%	3.99%
Residency			
In-State	81.11%	82.21%	83.89%
Student Status			
First generation student	13.74%	13.24%	13.94%
Dropped out within two years	8.22%	8.19%	8.19%

4.1 Bayesian Network

Figure 3 shows the Bayesian Network that was built based on information known before the start of the first semester. The colors in the figure indicate the type of information the attribute is related to. Note that there are no attributes related to course load since this model is built before those attributes are known. The types of information tend to be grouped together, however note that we are seeing direct relationships between some housing attributes and demographics or admissions. This may be influenced by certain housing restrictions imposed by the university. In Figure 4, we see that the means for all of the attributes of the synthetic population fall within the 95 percent confidence interval of the attribute means for the test population. The characteristics of the synthetic population are mostly consistent with both the training and testing sets. The number of dropouts in the synthetic population reflect the number of matched outcomes that were dropouts which should be identical to the number of dropouts in the test population.

4.2 Dynamic Logistic Regression Models

The average coefficients of significant predictors that appear in 100 models built before the start of the first semester and at the end of the the second fall semester are shown in Figure 5. For brevity, we present only significant predictors that appeared in at least 10 of the models at a given point in time. Note that many of the attributes known in the first semester are still influential at the end of the second fall semester.

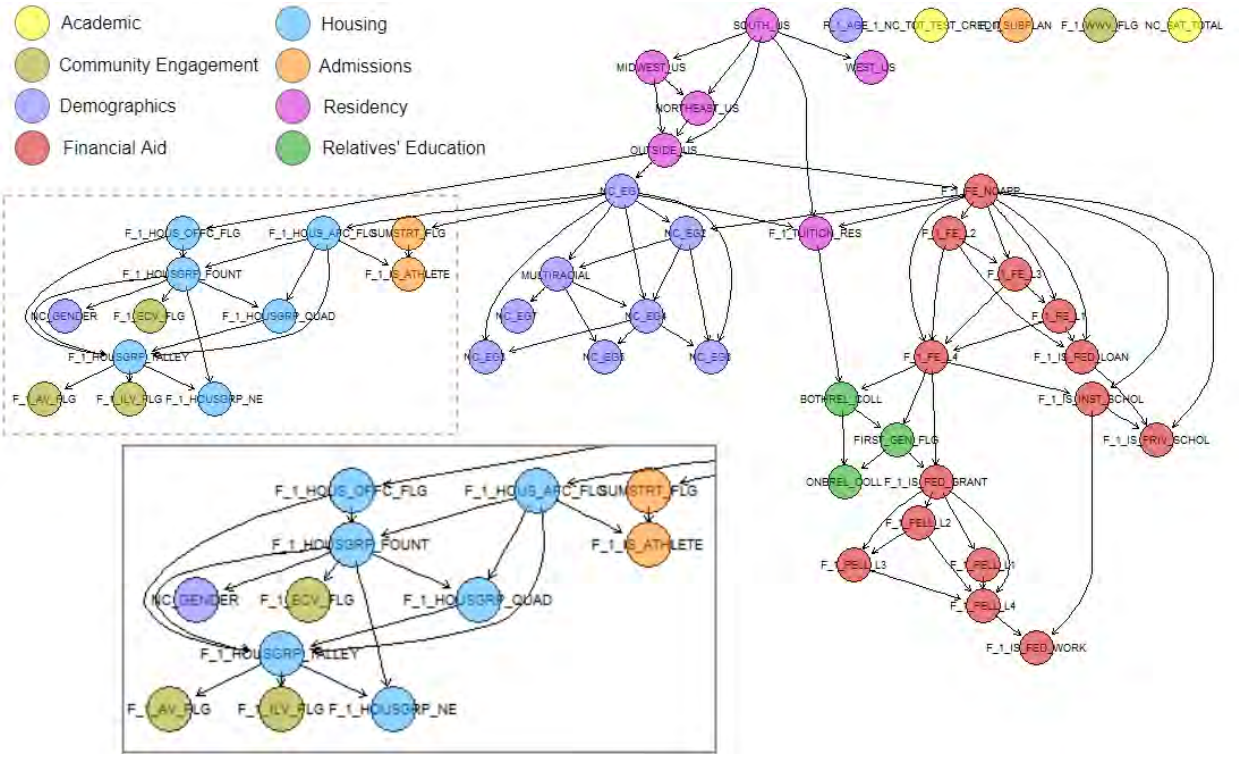


Figure 3: Bayesian Network built before the start of the first semester.

4.3 Matching Agents to Test Cohort

Figure 6 shows the distribution of distances between an agent and its matched student. Ninety-two percent of the matches were accepted in the first round of matching under the 1.5 distance threshold, and 98 percent of the matches fell below the distance threshold of 2.

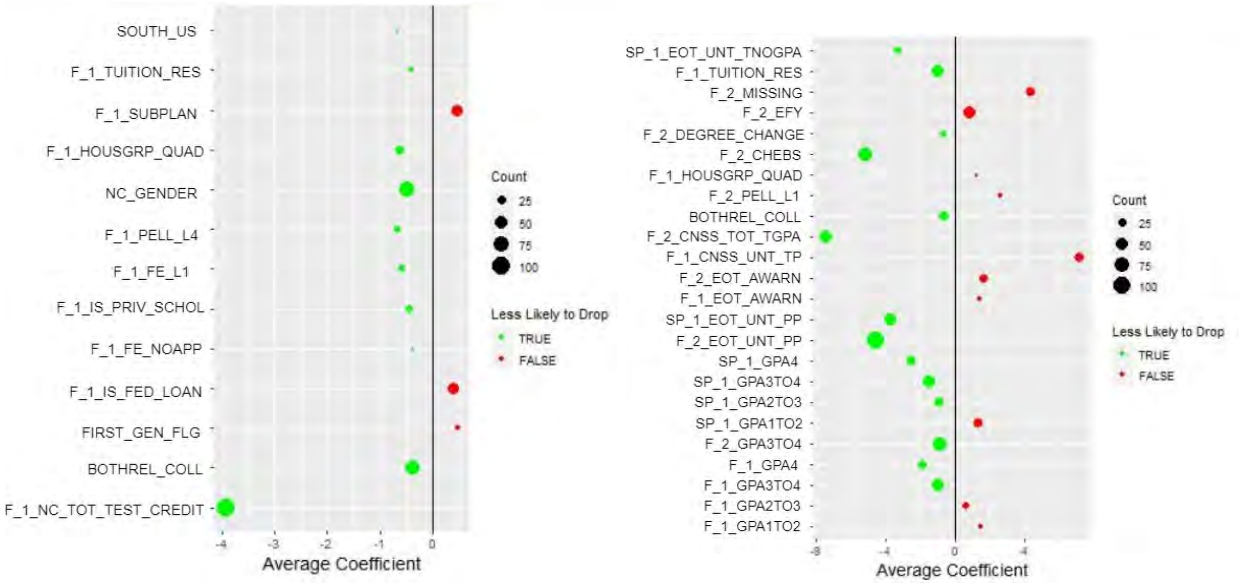
4.4 Simulation

To compare the risk behaviors of the test set and the synthetic population, we present several performance measures for predicting students in the test set and with those of the synthetic agents. Figure 7 compares the average area under the ROC curve (AUC), average sensitivity, and average specificity across all 100 prediction models built at each point in time for the test set and synthetic populations. Recall that the ground truth for the agents is the outcome of the matched student in the test set.

5 DISCUSSION

The Bayesian Network was fairly effective for characterizing the conditional relationships for the agents after the before start seeding. The test and synthetic populations are not significantly different for all but one of the attributes. Because the HEOM distance metric uses a normalized range for numerical variables, the possible distances between a given numerical attribute for any two observations typically fall between 0 and 1 (Wilson and Martinez 1997). Thus, squaring the overall HEOM distance could roughly indicate the number of variables that are different between the two observations. This interpretation can be meaningful especially in our case where we only have three numerical variables. Thus, most of the distances are either 0 or 1. Using this interpretation of the HEOM distance, we can see that for most matches, the students and their corresponding agents differ by less than two attributes. In terms of predictive performance, we see that the Census Fall 1 (time 1) and the Census Spring 1 (time 5) models correctly identified a higher

Figure 4: Ninety-five percent confidence intervals around the mean of binary attributes for the synthetic population and the test set.



(a) Before Start (b) End of Fall 2

Figure 5: Average coefficient estimates for significant predictors that appeared in at least 10 training iterations for models built before the start of the first semester (a) and at the end of the second fall semester (b). The size of the dot is proportional to the number of models the corresponding factor appeared significant in at a specific point in time. Green dots represent a negative coefficient, or a factor that reduced the likelihood of dropout. The red dots represent an average coefficient above zero, or a factor that increases the likelihood of dropout.

proportion of the dropouts in synthetic population than in the test set. Before Start (time 0) and End of Fall 1 (time 2) models correctly identify a higher proportion of non-dropouts in the synthetic population than in the test population. The variability in the metrics suggests there are some sources of randomness that have not been addressed.

From our analysis, we have identified potential improvements to model for the synthetic population to represent a real population. Since matches are based only on information known at the beginning, the advantage of using these attributes may decrease as the later prediction models are able to incorporate more recent information. The high variability in the sensitivity and specificity may be caused by the variability across the 100 models built at a given point in time. Specifically, some attributes may be used to determine one agent's risk while another agent's risk may be determined by a different set of attributes. When constructing the Bayesian Networks, the numerical variables are represented as categorical variables after discretization, which may cause differences in the distribution of the values that are generated for the synthetic agents. Moreover, the discretization of numerical variables, namely AP credits which largely influence dropout risk, may be affecting how its relationship to other variables is considered when constructing the network.

The method for matching actual students to synthetic agents can also influence the predictive performance metrics of the simulation. Matching students with agents who are not alike may result in establishing a non-representative ground truth for the agents and a poor comparison to the simulated outcomes regardless of the performance of the prediction models embedded in the simulation. Therefore, other matching methods should be explored to improve the performance metrics. For instance, it may be worthwhile to prioritize the distance between a subset of attributes that have shown to be more influential of dropout risk. In our current matching process, we do not discriminate between attributes when calculating the distance, rather

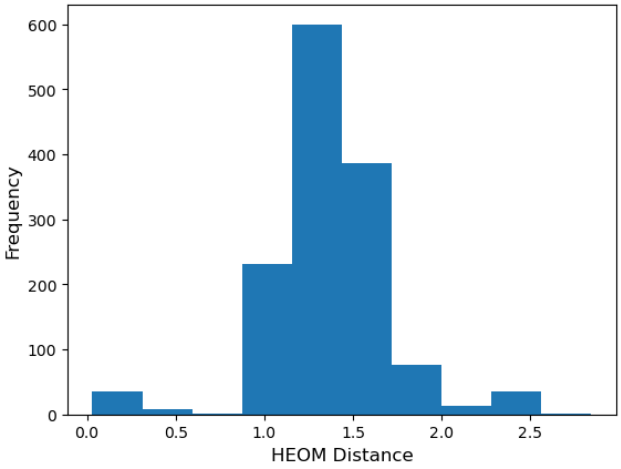
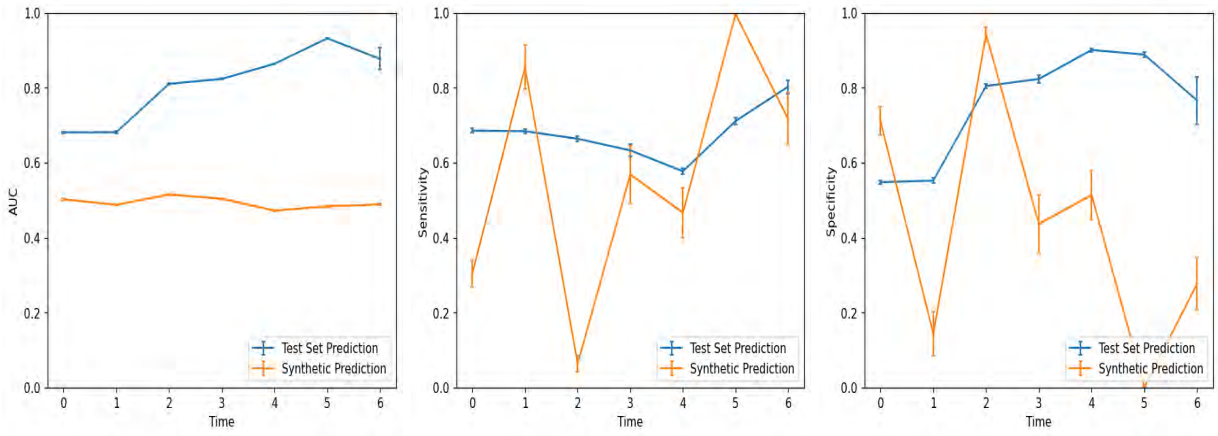


Figure 6: Histogram of HEOM distances between matched student-agent pairs.



(a) Average AUC (b) Average Sensitivity (c) Average Specificity

Figure 7: Comparison of performance metrics between the test set and the synthetic population predictions.

we include all available attributes when computing the HEOM distance. However, a student and an agent may have attributes in common that are not significant predictors and differ on attributes that consistently appear significant across prediction models.

5.1 Limitations and Future Work

There are some limitations to this study, we present a single replication of the simulation creating a single synthetic population to illustrate the proposed methods. The influential factors in the historical data may not be the same in the future, especially during substantial disruptions such as the pandemic. Our models are also based on a population from one university, although multiple years are represented. It is also possible that there are some simulated agents with a combination of attributes that do not appear in the actual population.

Our initial efforts are focused on generating a synthetic population that is representative. For future work, additional randomness may be incorporated into the simulation. While matching on risk score holds potential, there could be advantages to matching on both risk and a small set of variables. This process has identified areas for future research to enhance the synthetic population. More replications of the simulation

would allow us to compare performance across multiple synthetic populations. Generating additional synthetic populations would also help us identify the sources of variability, however the computation time for one replication of the simulation process is approximately 10 hours, which makes it difficult to conduct many replications. Parallel processing should be considered when running any additional replications given the lengthy computation time.

6 CONCLUSION

Dropping out of higher education remains a critical issue, and early dropout limits the opportunity for universities to intervene. Additionally, the vast amount of sensitive data collected by universities suggest a need for a representative synthetic population in order to better understand dropout risk behaviors and to realize the effect certain interventions can have on dropout risk. While we explored an approach that incorporates multiple matching methods and prediction models, our process highlighted many areas that must be considered when creating a representative synthetic population.

ACKNOWLEDGMENTS

This research has been supported by National Science Foundation Grant: Matriculation and Well-Being Under Emergent Events (MWEE) NSF RAPID collaborative (NSF Award Number: 2040072). The opinions expressed in the paper represent those of the authors and not necessarily those of the National Science Foundation.

REFERENCES

Burke, A. 2019. "Student Retention Models in Higher Education: A Literature Review". *College and University* 94(2):12–21.

Chapuis, K., and P. Taillandier. 2019. "A Brief Review of Synthetic Population Generation Practices in Agent-Based Social Simulation". In *Proceedings of the 2019 Social Simulation Conference*. Mainz, Germany.

Chen, Y., A. Johri, and H. Rangwala. 2018. "Running out of STEM: A Comparative Study Across STEM Majors of College Students At-Risk of Dropping Out Early". In *Proceedings of the 2018 International Conference on Learning Analytics and Knowledge*, 10. Sydney: ACM.

Dewantoro, G., and N. Ardisa. 2020. "A Decision Support System for Undergraduate Students Admissions Using Educational Data Mining". In *Proceedings of the 2020 International Conference on Information Technology, Computer, and Electrical Engineering*, 105–109: Institute of Electrical and Electronics Engineers Inc.

Farooq, B., M. Bierlaire, R. Hurtubia, and G. Flötteröd. 2013. "Simulation-Based Population Synthesis". *Transportation Research Part B: Methodological* 58:243–263.

Fernandes, E., M. Holanda, M. Victorino, V. Borges, R. Carvalho, and G. V. Erven. 2019. "Educational Data Mining: Predictive Analysis of Academic Performance of Public School Students in the Capital of Brazil". *Journal of Business Research* 94:335–343.

Fournier, N., E. Christofa, A. P. Akkinepally, and C. L. Azevedo. 2021. "Integrated Population Synthesis and Workplace Assignment Using an Efficient Optimization-Based Person-Household Matching Method". *Transportation* 48(2):1061–1087.

Freitas, R., and L. Salgado. 2020. "Educators in the Loop: Using Scenario Simulation as a Tool to Understand and Investigate Predictive Models of Student Dropout Risk in Distance Learning". In *Proceedings of the 2020 International Conference of Human-Computer Interaction*, Volume 12217, 255–272. Cham: Springer.

Glover, F., and M. Laguna. 1998. "Tabu Search". In *Handbook of Combinatorial Optimization*, 2093–2229. Springer.

Golding, P., and O. Donaldson. 2006. "Predicting Academic Performance". In *Proceedings of Frontiers in Education 36th Annual Conference*, 21–26.

Ilahi, A., and K. W. Axhausen. 2019. "Integrating Bayesian Network and Generalized Raking for Population Synthesis in Greater Jakarta". *Regional Studies, Regional Science* 6(1):623–636.

Ishitani, T. T., and S. L. Desjardins. 2002. "A Longitudinal Investigation of Dropout from College in the United States". *Journal of College Student Retention: Research, Theory & Practice* 4(2):173–201.

Krauland, M. G., R. J. Frankeny, J. Lewis, L. A. Brink, E. G. Hulsey, M. S. Roberts, and K. A. Hacker. 2020. "Development of a Synthetic Population Model for Assessing Excess Risk for Cardiovascular Disease Death". *JAMA Network Open* 3(9):e2015047–e2015047.

Marra, R. M., K. A. Rodgers, D. Shen, and B. Bogue. 2012. "Leaving Engineering: A Multi-Year Single Institution Study". *Journal of Engineering Education* 101(1):6–27.

Naseem, M., K. Chaudhary, B. Sharma, and A. G. Lal. 2019. "Using Ensemble Decision Tree Model to Predict Student Dropout in Computing Science". In *Proceedings of the 2019 IEEE Asia-Pacific Conference on Computer Science and Data Engineering*, 1–8: Institute of Electrical and Electronics Engineers Inc.

Ortiz-Lozano, J. M., A. Rua-Vieites, P. Bilbao-Calabuig, and M. Casadesús-Fa. 2020. "University Student Retention: Best Time and Data to Identify Undergraduate Students at Risk of Dropout". *Innovations in Education and Teaching International* 57(1):74–85.

Scholz, S., B. Surmann, W. Greiner, S. Elkenkamp, and M. Batram. 2016. "Creating Populations with Partnerships for Large-Scale Agent-Based Models-a Comparison of Methods". In *Proceedings of the 2016 SummerSim-SCSC*, 339–344. Montreal, Quebec, Canada.

Scutari, M. 2010. "Learning Bayesian Networks with the bnlearn R Package". *Journal of Statistical Software* 35(3):1–22.

Seymour, E., and N. Hewitt. 1997. *Talking About Leaving*. Boulder, CO: Westview Press.

Stinebrickner, R., and T. Stinebrickner. 2014. "Academic Performance and College Dropout: Using Longitudinal Expectations Data to Estimate a Learning Model". *Journal of Labor Economics* 32(3):601–644.

Sun, L., and A. Erath. 2015. "A Bayesian Network Approach for Population Synthesis". *Transportation Research Part C: Emerging Technologies* 61:49–62.

Villano, R., S. Harrison, G. Lynch, and G. Chen. 2018. "Linking Early Alert Systems and Student Retention: A Survival Analysis Approach". *Higher Education* 76(5):903–920.

Watkins, B. J., and E. Mazur. 2013. "Retaining Students in Science, Technology, Engineering, and Mathematics (STEM) Majors". *Journal of College Science Teaching* 42(5):36–41.

Wilson, R., and T. Martinez. 1997. "Improved Heterogeneous Distance Functions". *Journal of Artificial Intelligence Research* 6:1–34.

Wu, G., A. Heppenstall, P. Meier, R. Purshouse, and N. Lomax. 2022. "A Synthetic Population Dataset for Estimating Small Area Health and Socio-Economic Outcomes in Great Britain". *Scientific Data* 9(1):1–11.

Wu, H., Y. Ning, P. Chakraborty, V. Tech, J. Vreeken, N. Ramakrishnan, N. Tatti, . Y. Ning, and N. Ramakrishnan. 2018. "Generating Realistic Synthetic Population Datasets". *ACM Transactions on Knowledge Discovery from Data* 12(4):1–22.

Zhang, D., J. Cao, S. Feygin, D. Tang, Z. J. Shen, and A. Pozdnoukhov. 2019. "Connected Population Synthesis for Transportation Simulation". *Transportation Research Part C: Emerging Technologies* 103:1–16.

AUTHOR BIOGRAPHIES

DANIKA DORRIS is a Ph.D. student in the Edward P. Fitts Department of Industrial and Systems Engineering at North Carolina State University. She received her B.S. in Industrial and Systems Engineering at the University of Tennessee and her Master's in Operations Research at North Carolina State University. Her email address is dmdorri2@ncsu.edu.

JULIE IVY is a professor in the Edward P. Fitts Department of Industrial and Systems Engineering and Fitts Faculty Fellow in Health Systems Engineering. She previously spent several years on the faculty of the Stephen M. Ross School of Business at the University of Michigan. She received her B.S. and Ph.D. in Industrial and Operations Engineering at the University of Michigan. She also received her M.S. in Industrial and Systems Engineering at Georgia Tech. She is an active member of the Institute of Operations Research and Management Science (INFORMS), Dr. Ivy served as the 2007 Chair (President) of the INFORMS Health Applications Society and the 2012 – 13 President for the INFORMS Minority Issues Forum. Her email address is jsivy@ncsu.edu.

JULIE SWANN is the department head and A. Doug Allison Distinguished Professor of the Fitts Department of Industrial and Systems Engineering. She is an affiliate faculty in the Joint Department of Biomedical Engineering at both NC State and the University of North Carolina at Chapel Hill. Before joining NC State, Swann was the Harold R. and Mary Anne Nash Professor in the Stewart School of Industrial and Systems Engineering at the Georgia Institute of Technology. There she co-founded and co-directed the Center for Health and Humanitarian Systems (CHHS), one of the first interdisciplinary research centers on the Georgia Tech campus. Starting with her work with CHHS, Swann has conducted research, outreach and education to improve how health and humanitarian systems operate worldwide. Her email address is jlswann@ncsu.edu