Predicting Undergraduate Student Success using Geographically Weighted Logistic Regression

James Roddy[1], Samantha Robinson[1]

[1]University of Arkansas, Department of Mathematical Sciences, 1 University of Arkansas, Fayetteville, AR 7270

Abstract
Due to interest in undergraduate student success, including individual success in early coursework and overall timely college graduation, attempts to identify 'at-risk' students using both demographics and pre-college variables (e.g., SAT scores) are numerous. Interventions and programs on college campuses utilize these 'at-risk' identifications in an effort to increase student success. Despite the interest in accurate identification of 'at-risk' students, few studies investigate the impact that pre-college location (i.e., hometown geographic location) has on student success.

Geographically Weighted Logistic Regression (GWLR), a local spatial model, is used to explore the impact location has on student success of 2012-2014 freshman (i.e., first year) cohorts at one midsized, public university. Mappings of the results are presented, which can assist policy makers at institutions of higher education identify and intervene with 'at-risk' students. The predictive accuracy of these GWLR models, with and without covariates, is then compared to logistic regression models that do not take into consideration location, highlighting the benefits of spatial models in education research.

Key Words: Geographically Weighted Logistic Regression (GWLR), Local Spatial Modeling, Undergraduate Education, Student Success, Education Policy

1. Introduction and Motivation

Though college is deemed an important factor for future career success, only 60% of undergraduate students in the United States (US) will graduate within six years at the university where they begin their degree. According to the National Center for Education Statistics (NCES), this six-year graduation rate is as low as 30% at less selective public institutions such as those that have open admissions policies. Given the obvious and pressing need to improve timely graduation and increase the overall six-year graduation rate, numerous studies have attempted to identify 'at-risk' students using both demographics and pre-college variables (e.g., SAT scores).

For example, Woods et al. (2018) applied Logistic Regression to predict student success in individual college courses such as introductory English and mathematics based upon high school course preparedness. A significant link between high school coursework and success in early college courses was discovered (Woods et al., 2018). Numerous studies

have related pre-college variables to student success in college (both in terms of individual course success and overall student persistence).

Consequently, many speculate that the failure to achieve a degree in a timely manner is attributable to a lack of academic preparation. However, Tinto (1993) emphasizes that student persistence, student retention, and, consequently, timely graduation are related to more than the often studied pre-college variables of high school grade point average (GPA) and scores on college entrance examinations (e.g., ACT/SAT scores, etc.). Success in early college courses, regardless of pre-college *academic* preparation, can subsequently increase student commitment to an institution, student persistence, and timely graduation. Conversely, negative experiences *during* college weaken the commitment of students to an institution and increase the likelihood of voluntary withdrawal, which accounts for approximately 75-85% of all departures from institutions of higher education (Tinto, 1993).

While a student can have many negative experiences during college, one of the most negative experiences that a student can have during college is an individual course failure. Consistent achievement in individual courses eventually results in a degree and timely completion of that degree. On the other hand, individual course failure (including withdrawing from an individual course) increases the cost of college for a student, increases debt, increases the risk of dropout, and definitively delays graduation (Boldt et al., 2017; Valentine et al., 2011). This is especially true in the first semester of college when the earnings benefits gained from higher education are negligible and the student connection to the institution is often minimal (Valentine et al., 2011).

Early success in college then (regardless of and distinct from pre-college variables that might not capture individual specific effects such as study habits and effort and might also have systemic biases) might be highly predictive of both retention and, ultimately, timely graduation. Thus, it is of great interest to understand how performance in first semester courses impacts timely graduation. At the same time, pre-college factors related to timely graduation cannot be completely ignored. While high school GPA and scores on college entrance examinations might not fully account for individual specific effects, some consideration of community background (i.e., pre-college geographic location) when predicting timely graduation is of interest.

Despite the interest in increasing the rate of timely graduation among college students and despite the interest in accurate identification of ‘at-risk’ students for the purposes of intervention, few studies investigate the impact that pre-college *location* (i.e., hometown geographic location) has on student success.

2. Purpose of the Study

The purpose of this study is to compare the performance of logistic regression and Geographically Weighted Logistic Regression (GWLR) in predicting the timely graduation of college students. The predictive accuracy of these models is compared to determine if consideration of pre-college geographic location is beneficial when predicting timely graduation of undergraduate students. Mappings of the results are presented, which can assist policy makers at institutions of higher education identify and intervene early with ‘at-risk’ students in order to increase timely student graduation.

3. Method

3.1 Sample and Data

The sample consisted of 1045 students that both entered the university as new (i.e., non-transfer) freshmen in Fall 2012-2014 and took five specific introductory courses. The introductory coursework of interest included Principles of Biology, English Composition, College Algebra, General Psychology, and General Sociology (i.e., BIOL 1543, ENGL 1013, MATH 1203/1204, PSYC 2003, and SOCI 2013). Only students that took all of the classes were included in the sample during analysis. Overall sample characteristics were similar to those of the university as a whole, with the exception of sex, as females appeared to be overrepresented in the sample. Table A1 (in the appendix) provides additional demographic information for the sample.

The data consisted of time to graduation (recorded in months) as well as course letter grades across five introductory courses for each student in the sample.

The response variable in the current study was an indicator for whether or not the time to graduation for a student was less than 48 months (i.e., less than 4 calendar years) or not. The predictor variables are indicator variables related to individual course success in the five introductory courses described above. Course grades were recorded for five introductory freshman courses (i.e., Biology, English, Algebra, Psychology, and Sociology), which would be representative of a 15 hour first semester course load. Indicator variables for mastery of each class were created such that an 'A' or 'B' indicated mastery in the course and all other grades represented non-mastery.

3.2 Analytical Methods

In order to identify 'at-risk' students and to explore the impact location has on timely graduation at one university, two regression techniques were implemented for comparison: (1) logistic regression and (2) Geographically Weighted Logistic Regression (GWLR). Both regression techniques were implemented with no covariates (i.e., intercept only models) and with indicator variables for individual course success (as described above) as covariates.

3.2.1 Logistic Regression

Logistic regression is a very common generalized linear model (GLM) that uses a logit link function to model the log odds of success as a linear combination of the *m* covariates in the following form:

$$logit(p) = \frac{\log(p)}{\log(1-p)} = \beta_0 + \beta_1 x_1 + \ldots + \beta_m x_m$$

Logistic regression does not account for spatial autocorrelation and necessarily assumes that the relationships between explanatory and response variables are stationary across a spatial region. However, the assumption of spatial stationarity is often violated in practice due, potentially, to the existence of relationships that intrinsically differ across space (Fotheringham et al., 2002). Matthews and Yang (2012) note that violations of spatial stationarity (i.e., spatial nonstationarity) arise in several application areas including but not limited to health care, infectious disease, real estate, poverty, religion, traffic, crime, and voting.

3.2.2 Geographically Weighted Logistic Regression

Local spatial models, unlike logistic regression, are location dependent and allow spatial processes to exhibit nonstationarity across space. Such local spatial models have been likened to 'spatial microscopes' that can uncover spatial patterns that are hidden and/or obscured at more aggregate levels (Fotheringham et al., 2002). Moreover, local spatial models, with location dependent parameter estimates and direct links to GIS, are mappable providing a visualization of any previously unobserved spatial patterns. This visualization can facilitate interpretation that is not possible with global models (Matthews & Yang, 2012). While only one local spatial modeling technique (i.e., GWLR) is described in the current manuscript, a thorough overview of these modeling techniques may be found in Lloyd (as cited in Matthews & Yang, 2012).

Geographically Weighted Regression (GWR) models are one of the most often used methods for modeling data that exhibits spatial nonstationarity (Finley, 2011). These local spatial models attempt to address spatial nonstationarity directly through the local calibration of estimated model parameters (Gollini et al., 2015). These models assume that relationships vary spatially, such that near places are more similar than distant ones. Separate regression models are calibrated at each of several pre-specified regression points. Observations within a certain distance of each regression point are included in local model calibrations with the weights of these individual observations specified according to a weighting kernel scheme (Fotheringham et al., 2002).

Geographically Weighted Logistic Regression (GWLR) utilizes a logit link function to model the log odds of success as a linear combination of the *m* covariates specific to location $\boldsymbol{u}$ in the following form:

$$logit(p) = \frac{\log(p)}{\log(1-p)} = \beta_{0i}(u) + \beta_{1i}(u)x_{1i} + \ldots + \beta_{mi}(u)x_{mi}$$

where $\boldsymbol{u}$ is a vector of coordinates (u_i, v_i) for each regression point i. Estimations of β_k are based upon weights conditioned on the location $\boldsymbol{u}$.

For each model calibration at the pre-specified regression points, a weighting matrix $\boldsymbol{W}(u_i, v_i)$ is computed. The weights for the *j*th observation during the model calibration at regression point location i, w_{ij}, are specified using a weighting kernel scheme, which depends upon a fixed (or adaptive) bandwidth h that can be used to smooth the surface of parameter estimates.

In the current work, equidistant regression points were specified by utilizing a uniform grid to cover the study region. Gaussian weights with an adaptive bandwidth were used and have the form:

$$w_{ij} = exp\left\{-\frac{1}{2}\left(\frac{d_{ij}}{h}\right)^2\right\}$$

where d_{ij} is the distance from location i to j.

Gaussian weights, which decay continuously as a function of the increasing distance between i and j, were selected for the current study as was an adaptive bandwidth. Due to

the clustering and irregularity of the observed data points (i.e., individual student hometown locations), there was potential for large standard errors, failure in parameter estimation at particular regression points, and/or an undersmoothed surface of parameter estimates. Thus, as is recommended in such circumstances, an adaptive bandwidth was used (Fotheringham et al., 2002).

3.2.3 Model Comparison

Model selection was performed, utilizing AICc, for both regression techniques (i.e., logistic regression and GWLR). The best model for each regression technique, according to AICc, as well as the null models (i.e., intercept only models) were fit and the results are reported in the next section.

4. Results

The primary goal of the present study was to compare the performance of logistic regression and GWLR in predicting the timely graduation of college students and to determine if pre-college geographic location is beneficial when doing so.

The full models (including all course success indicator variables) were best for both logistic regression and GWLR according to AICc. These models, as well as the intercept only models, were implemented using the sample data. A summary of the model accuracy for logistic regression and GWLR models (both intercept only and the full models) is provided in Table 1 below.

Table 1.
Summary of Model Accuracy

Model	Correct	False Negative	False Positive
Logistic (intercept only)	50.81%	49.19%	0%
GWLR (intercept only)	50.81%	49.19%	0%
Logistic with Classes	60.86%	30.72%	8.42%
GWLR with Classes	64.31%	18.56%	17.13%

The intercept only GWLR model appeared to have the exact same results as the intercept only logistic regression model which, presumably, was due to the adaptive bandwidth selected. The adaptive bandwidth was quite large and, consequently, the results of the local spatial model were expected to approach that of the global model as can be seen in Table 1. However, the full models (i.e., those including all course success indicator variables) did give different results, as can be seen in Table 1.

The intercept-only models (both logistic regression and GWLR) predict that no one will graduate in a timely fashion. This leads to correct outcomes and false negatives. However, these intercept only models appear to be no better than simply flipping a coin to make predictions, as they are correct in only 50% of cases. The Logistic and GWLR

models that include class success as indicator variables provide a more complete picture. Moreover, the results of these full models highlight the beneficial aspects of accounting for spatial nonstationarity when predicting timely graduation. While both approaches have apparently similar performance, the GWLR approach outperforms the logistic regression approach slightly in terms of overall predictive accuracy (see Table 1).

A further analysis of the two approaches, examining the parameter estimates of each model in Tables 2-3, demonstrate the apparent similarity of these models.

Table 2.
Logistic Regression – Parameter Estimates

Variable	Estimate	Significance
Intercept	-1.7181	< 0.0001
Algebra	0.5342	0.0004
Biology	0.3664	0.0100
English	0.5516	0.0016
Psychology	0.5122	0.0003
Sociology	0.5668	0.0005

Table 3.
Geographically Weighted Logistic Regression – Parameter Estimates

Variable	Minimum	Lower Quartile	Median	Upper Quartile	Maximum
Intercept	-1.8738	-1.8283	-1.7929	-1.7651	-1.7422
Algebra	0.5211	0.5546	0.5706	0.5964	0.6464
Biology	0.3050	0.3181	0.3282	0.3357	0.3876
English	0.5419	0.5592	0.5734	0.5962	0.6323
Psychology	0.5284	0.5976	0.6134	0.6354	0.6561
Sociology	0.5465	0.5664	0.5742	0.5824	0.5990

Despite this apparent similarity of the model parameter estimates, the strength of the GWLR approach comes from the ability to visualize the results. The direct link to GIS inherent in GW models is demonstrated by mapping the parameter estimates (see Figure 1). With the visualization afforded by this GWLR method, it is possible to identify locations of interest where early (i.e., first semester) class performance plays a greater role in determining timely graduation.

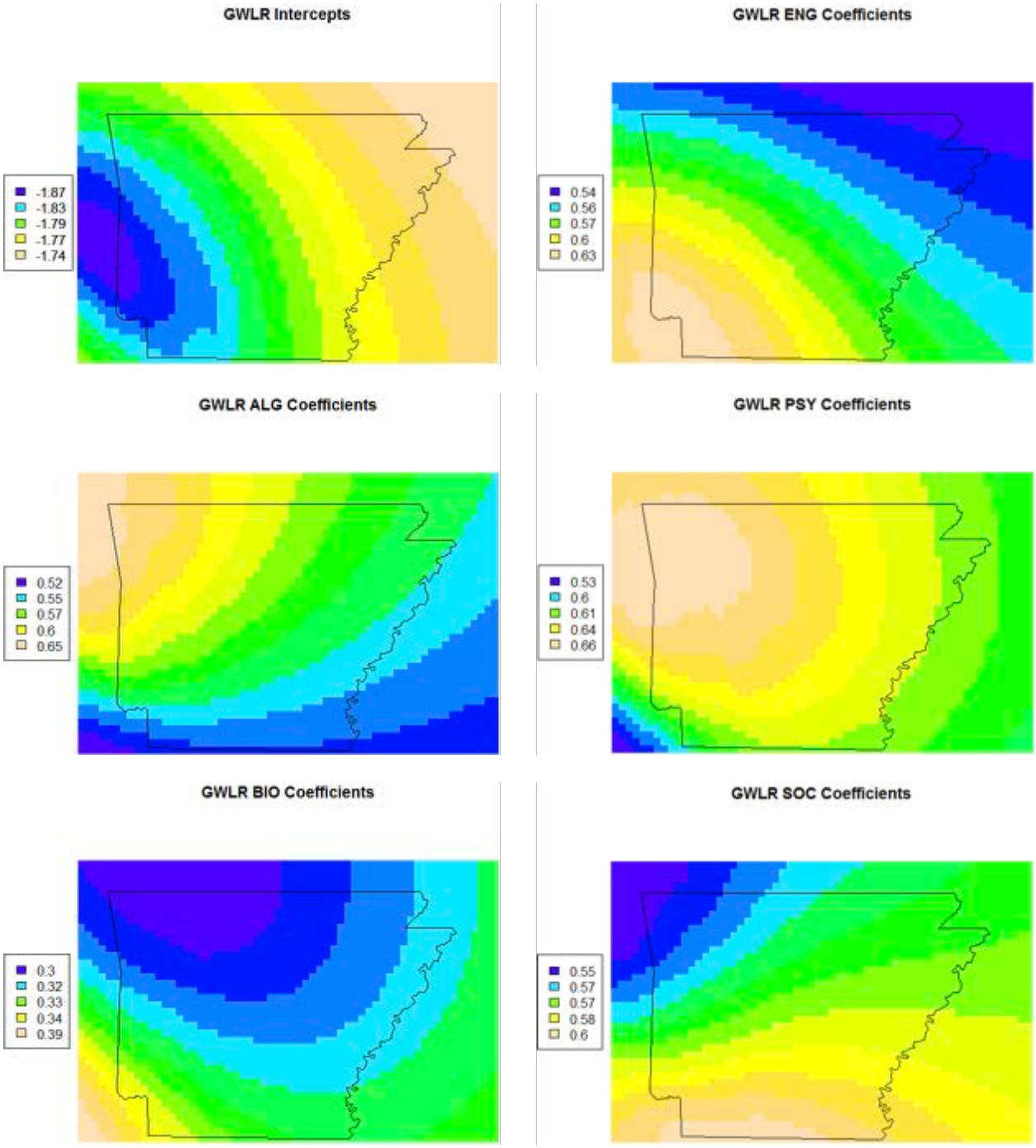

Figure 1.
Geographically Weighted Logistic Regression – Mapped Parameter Estimates

From the mappings provided in Figure 1, it is clear that the relationship between class performance and timely graduation in the northwest portion of the state is different than that of the rest of the state in many of these introductory courses. This implies that those students that come from the northwest portion of Arkansas (nearer to the university) have

a higher chance of timely graduation when they perform well in Algebra and Psychology compared to students that come from other parts of the state, but have a lower chance of timely graduation with strong performance in Biology and Sociology. Moreover, timely graduation for those in the southwest portion of the state seems to be more associated with English (i.e., English Composition) than any other introductory subject.

5. Discussion

5.1 Implications

The goal of the current study was to predict timely graduation (a form of student success) making use of only success in introductory college coursework and pre-college geographic location as predictor variables. Two regression techniques (i.e., logistic regression and GWLR) were utilized in the study and results demonstrated that there was benefit to incorporating spatial structure when making predictions.

Although results show no improvement when adding a spatial component to the intercept only models, spatial models generally outperformed their non-spatial counterparts when course success indicator variables were added. These spatial models had higher overall rates of correct classification and lower false negative rates. It also seemed apparent that Northwest Arkansas is an especially distinct region, with the relationships between timely graduation and course success differing greatly from other parts of the state with the exception of the relationship between timely graduation and English Composition. The intercept and coefficient estimates for Biology and Sociology are lower, while the Algebra and Psychology coefficient estimates are higher. Overall, spatial models not only outperform their non-spatial counterparts but the ability to visualize the results can assist policy makers at institutions of higher education when identifying and intervening with ‘at-risk’ students.

While previous studies have utilized introductory coursework and pre-college variables such as high school GPA when making predictions about various measures of undergraduate student success, few studies have examined how these relationships function across space (i.e., few have utilized geographic location). This study indicates that spatial nonstationarity might exist when modeling relationships between introductory coursework and timely graduation. This result is of interest to university administrators and various other stakeholders, as it provides further information to these individuals. This additional information might help increase the effectiveness of various intervention measures for ‘at-risk’ students and might facilitate the earlier graduation of all students.

5.2 Limitations and Future Work

There were many limitations in this work, including the fact that females were overrepresented in the sample. Despite the fact that the university is approximately 55% female, over 71% of the sample in the current study was female. This is an obvious limitation and one that should be addressed in future work. Perhaps the five introductory courses chosen for the study should be evaluated and altered to capture more undergraduate students such as those in business, engineering, and related STEM fields.

Moreover, while local models allow for an analysis of the sensitivity and/or stability of spatial model parameter estimates by allowing model refitting over a wide range of data aggregation levels (controlled by the bandwidth), the current study utilized only one adaptive bandwidth. Given the clustering and irregularity of the observed data points, an

adaptive bandwidth was recommended. However, future work could examine the impact that differing bandwidths has on the results.

While this study had numerous limitations and there is much room for future work, this preliminary study does indicate that spatial nonstationarity exists between success in introductory coursework and timely graduation. Researchers, university administrators, and all relevant stakeholders should take note of pre-college location when implementing any sort of intervention for the 'at-risk' student population, as this might decrease overall student time to graduation.

References

Boldt, D. J., Kassis, M. M., & Smith, W. J. (2017). Factors impacting the likelihood of withdrawal in core business classes. *Journal of College Student Retention, 18*(4).

Finley, A. O. (2011). Comparing spatially-varying coefficients models for analysis of ecological data with non-stationary and anisotropic residual dependence. *Methods in Ecology and Evolution, 2*.

Fotheringham, A. S., Brunsdon, C. & Charlton, M. E. (2002). *Geographically weighted regression: The analysis of spatially varying relationships*. New York, NY: Wiley.

Gollini, I., Lu, B., Charlton, M., Brunsdon, C., & Harris, P. (2015). GWmodel: An R package for exploring spatial heterogeneity using geographically weighted models. *Journal of Statistical Software, 63*(17).

Lu, B., Harris, P., Charton, M., Brunsdon, C., Nakaya, T., Gollini, I. (2019). Package 'Gwmodel'. CRAN Project Online. Retrieved from https://cran.r-project.org/web/packages/GWmodel/GWmodel.pdf

Matthews, S. A. & Yang, T. C. (2012). Mapping the results of local statistics: Using geographically weighted regression. *Demographic Research, 26*(6).

R Development Core Team. (2017). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Online http://www.rproject.org

Saefuddin, A., Andi Setiabudi, N., Fitrianto, A. (2012). On Comparison Between Logistic Regression and Geographically Weighted Logistic Regression: with Application to Indonesian Poverty Data. *World Applied Sciences Journal, 19*.

Tinto, V. (1993). *Leaving College: Rethinking the causes and cures of student attrition* (2nd Ed.). Chicago: University of Chicago Press.

U.S. Department of Education, National Center for Education Statistics. (2019). The Condition of Education 2019 (NCES 2019-144), Undergraduate Retention and Graduation Rates.

Valentine, J. C., Hirschy, A. S., Bremer, C. D., Novillo, W., Castellano, M. & Banister, A. (2011). Keeping at-risk students in school: A systematic review of college retention programs. *Educational Evaluation and Policy Analysis, 33*(2).

Woods, C., Park, T., Hu, S., Jones, T. B. (2018). How High School Coursework Predicts Introductory College-Level Course Success. *Community College Review, 46*(2).

Appendix

Table A1.
Demographic Features Representative of the Sample Participants

Demographic	N	%
Sex		
Male	300	28.71
Female	745	71.29
Race/Ethnicity		
Caucasian	773	73.97
Hispanic	105	10.05
African or African American	86	8.23
Asian, Pacific Islander or Asian/American	26	2.49
Other	55	5.26
Home State		
Arkansas	625	59.81
Texas	238	22.78
Missouri	58	5.55
Other	124	11.87
Admit Term		
Fall 2012	344	32.92
Fall 2013	345	33.01
Fall 2014	356	34.07
Graduation Status		
Enrolled	145	13.88
Withdrawn	335	32.06
Graduated	565	54.07