INTRODUCTION

Amid skyrocketing educational costs and stagnant recruitment budgets, enrollment professionals are under increasing pressure to attract and retain a high quality and diverse student population. This is no easy task. “In the past five years, the average cost of in-state tuition and fees at public colleges has jumped 35% —after adjustment for inflation” (Block 2007). Recruitment costs have also increased; as of 2005 “the average college admission office expended $714 per admitted student,” (Hawkins 2006) a 9.68% increase over 2003. According to the NACAC Admissions Trend Survey, 2005; enrollment office budgets have remained stagnant with 58% of admissions offices reporting either no adjustment or a decrease in recruiting budgets. With stabilizing enrollment budgets and increasing recruitment costs, traditional recruitment methods could be improved. GIScience technology as well as Bayesian statistical modeling offer enrollment professionals the ability to analyze the unique influential enrollment factors among prospective students, thereby allowing recruiters to identify those students with similar factors.

Higher education recruiters have relied heavily on prior enrollment data to dictate future recruitment activities, through strategic planning and tactical execution. The use of prior enrollment data in an enrollment funnel, which measures the ratio of students between each stage of the enrollment process, allows for the establishment of future recruiting goals. The use of an enrollment funnel, which is shown in Figure 1, is an example of using prior enrollment data to strategically plan for the future. The enrollment process includes the following four stages: prospects, applicants, acceptances, and finally enrolled students.
Another method of using prior enrollment data involves the identification of influential enrollment factors which are common among successfully enrolled students. Further analysis of influential enrollment factors allows for the identification of undiscovered insights and trends. “These trends and patterns form the basis of predictive models that enable analysts to produce new observations from existing data” (Luan 2002). The use of predictive modeling is an example of tactical execution using prior enrollment data. The existing methods used to develop enrollment funnels and predictive models rely on the use previous enrollment data to develop an understanding of currently enrolled students. While the existing methods used to develop enrollment funnel and predictive models vary in complexity, they provide the necessary foundation for future enrollment activities. The efficiencies gained by leveraging existing enrollment data through Bayesian modeling and GIScience, provides additional insights into successfully enrolled students. These efficiencies strengthen the enrollment model, which provides the necessary intelligence to create and executable strategic plan. A well developed strategic plan identifies the key components while simultaneously outlining the tactical methods needed to achieve success.
BACKGROUND

Colleges and universities use enrollment funnels and predictive models to exceed the established enrollment goal, yet real world case studies of institutional practices are often difficult to find for several reasons. Consistent with Paul Read’s (2005) analysis regarding the limited acceptance of GIScience within higher education, in his paper titled ‘The potential and barriers to the use of geographical information systems for marketing applications in higher educational institutions,’ institutions generally underutilize their ability to:

- Create and maintain geodemographic and socioeconomic databases
- Coordinate the knowledge, skills, and abilities of individual academic and administrative professionals
- Interpret enrollment trends using emerging or unfamiliar technologies

DesJardins (DesJardins 2002) identifies that institutions utilize various analytical approaches to develop a competitive advantage when recruiting students. The analysts who initiate this type of research often collaborate with executive management and provide a summary of the results, while disclosing relatively few details about the data and methods employed. This is not to say that this type of technology is not currently in use within higher educational enrollment. A variety of companies have created various products and services to predict student enrollment. The College Board has created a product called ‘Descriptor Plus’, while RuffaloCody has created ‘Enhanced Search Strategies’, and Noel-Levitz has created ‘ForecastPlus’. While this research is not a critique of existing predictive modeling systems, it is important to note that very little documentation exists. Due to the limited documented studies it is necessary to investigate other disciplines to find data and methods that may offer additional insights into higher education enrollment processes; these include Suitability Models, Spatial Interaction Models, and P-Median Modeling.
Suitability models are one method which uses a raster or grid to calculate the attractiveness of geographical areas based on various influential factors (Berry 2007). “Like any other models, suitability models are a generalized statement, or abstraction, of the important considerations in a real-world situation” (Berry 1999). Grid sizes can vary depending upon the spatial scale at which the analysis takes place, from a few inches to many miles. Analysis methods may also vary from conducting a local analysis, which is each individual grid cell, to a neighborhood type of analysis, which is using one grid cell to represent the phenomena of the surrounding grid cells. These models are calculated by reclassifying the influential factors and then calculating a score within the analysis area, with higher scores representing more attractive areas. Different industries refer to these types of models in different ways, they include using habitat models to associate environmental factors to plant and animal habitats in which species thrive or land planning models to identify geographic areas which will maximize growth while controlling development and enhancing the livability of communities.

While raster based predictive modeling offers many advantages, such as availability of hardware and software to process large amounts of data; and the ability to understand, measure, and calculate spatial adjacency. Disadvantages of raster based predictive models also exist. Raster based predictive models can vary depending on spatial resolution, with differing results between smaller and larger grid cells; and difficulty converting vector data to raster format, which could lead to data integrity problems. Adequate spatial scale needs to be considered when using raster based models, such as a local or neighborhood analysis. It is important to understand that local analysis provides detailed information on individual cells, while not taking into account the surrounding cells. Neighborhood analysis allows for a centralized cell to represent the attributes of surrounding cells. In either event, when using local or neighborhood
analysis attributes that fall outside the analysis area is not considered. Still, raster based predictive models can work well provided influential factors are measurable over a continuous geographic area, and can be reclassified in a raster format.

Spatial interaction modeling offers the ability to model the movement of goods and services between origins and destinations (Fotheringham 1989) (Hayes 1984). Many multinational retailers use this type of modeling to identify new store locations and determine trade areas. In an effort to provide goods and services to their target customers, companies utilize the latest technology to identify suitable locations for new development. Merging geographic data and customer data, such as customer zip code and point of sale data, allows retailers to spatially view where their best customers are located, or identify possible voids in the markets. The first model of this type was developed by William Reilly in 1931 and is known as the Law of Retail Gravitation. He theorized that consumer choice is proportional to distance between the origin and destination, and the size of the destination. “Reilly’s model has two limitations: an exponentially increasing distance-decay parameter, which overemphasize travel distance, and a two shopping center specification, which limits store location analysis to two locations” (Eppli 1996). Another example of retail spatial modeling comes from David Huff, who developed the Huff model in 1964 to predict consumer shopping behavior. The Huff model calculates the attractiveness of shopping destinations by taking into consideration distance, attractiveness of the center, and other constants which account for unpredictable shopping behaviors (Huff 2003). These unpredictable shopping behaviors, which include merchandise discounts, one-time sales, or extreme weather, are difficult to calculate and are often estimated. The consequence of modeling unpredictable shopping patterns without acceptable statistical methodologies will lead to inappropriate assumptions, thus rendering the results ineffective.
Several published studies on higher education recruiting have used spatial interaction models ((Herries n.d.) (Marble 1995)) such as Reilly’s Law of Retail Gravitation (Martin 2001) and The Huff Model (Zhou 2005) to model student enrollment. While these models were originally designed to determine the proposed trade areas or estimate the sales volume of potential retail sites, their use within higher education recruiting is limited. Reilly’s Law of Retail Gravitation and The Huff Model are based on the assumption that customers will travel the shortest distance to meet their consumption habits. These models will allow for the identification of the recruiting area of the school, thus identifying the type of students within the recruiting area, but not necessarily students interested in attending the study institution. Consider a prospective student living in Indianapolis, IN, looking to attend an in-state division I school to study business. Both Reilly’s law and The Huff Model will identify that IUPUI is the best fit, while in actuality Indiana University in Bloomington, Purdue University in Lafayette, and Ball State in Muncie maybe a better fit to this student’s needs.

A third type of spatial interaction model is based on the P-median models. The P-median problem arises out of operations management and its “objective is to locate \(p\) supply points in order to minimize the total distance of the demand points to their nearest supply point.” (Jackson 2007) The P-median problem is a popular facility location model used to locate distribution centers, factories, or new retail sites. While this type of spatial interaction model has been effective in solving location based problems, it may not be the most appropriate to use in modeling student enrollment of an existing institution. If an institution is looking for a new location to open a satellite campus, or cooperative extension, the P-median model would be model to evaluate.
Spatial Interaction models, in particular Reilly’s Law, The Huff Model, and the P-median model, offer insight into the recruiting areas of colleges and universities, they are most effective in the analysis of new sites or locations and modeling the sales impact a new location will have on existing store sales. While these types of spatial interaction models work well within trade area analysis, the adaptation of these methods within university enrollment is limited.

Another approach, which will be used in this analysis, is multivariate statistical modeling, using multiple regression. Multivariate statistical modeling addresses the problem of identifying influential variables and their respective weights to predict enrollment outcomes. Traditionally, multivariate regression did not have a proper way to incorporate spatially and temporally auto-correlated data. With the development of hierarchical models using random space and time components multivariate statistical methodology allow us to model the complexity of recruitment and enrollment prediction. Bayesian Hierarchical Models are one set of models that can incorporate data with significant space and time lags/autocorrelation effects.
DATA AND METHODS

The use of Bayesian Hierarchical modeling to analyze the spatial distribution of higher education enrollment in relation to influential enrollment variables enables the statistical analysis that combines the currently available data and technology. Bayesian modeling is based on the statistical methods developed by Thomas Bayes, in which the prior distribution and the likelihood of an event’s occurrence are used to determine the posterior distribution or Bayesian inference. The Besag, York, and Mollie (BYM) (Lawson 2003) model is a Bayesian statistical model that calculates the posterior distribution of an event’s occurrence while taking into consideration fixed effects, random effects of correlated spatial heterogeneity and uncorrelated spatial heterogeneity, prior enrollment history, and the likelihood of enrollment. Uncorrelated spatial heterogeneity are random effects which may or may not exist within the influential enrollment factors, while correlated spatial heterogeneity is also referred to as spatial autocorrelation. “Spatial autocorrelation refers to the correlation of a variable with itself through space” (Burt 1996). Patterns that exhibit highly negative spatial autocorrelation appear to have a checkerboard pattern, whereas highly positive spatial autocorrelation patterns are two distinct regions, which share a boundary. The use of Bayesian Statistics to model higher education enrollment offers a unique perspective on enrollment trends because of its ability to calculate the posterior distribution through the use of the prior distribution, likelihood, fixed effects and random effects.

Modeling student enrollment with the BYM model allows for area-count analysis, which aggregates individual student records into defined spatial regions, and offers many advantages. These advantages include modeling student enrollment contiguously, maintaining student privacy, and analyzing influential enrollment variables. The 1412,
block groups within the state of Indiana offer a contiguous spatial surface in which individual student records can be aggregated. The office of admissions provided the 2,159 enrolled student records for first time full time instate freshmen for the fall of 2005. These records were geocoded using the website Batchgeocoder.com, which uses the Yahoo! Geocoding API to create a latitude and longitude based upon the student's application address. These individual student records were aggregated into their respective block groups using a spatial-join feature, which is available in most desktop mapping applications. Synergos Technologies, a provider of socioeconomic and demographic data, provided the necessary influential enrollment variables at the block group level for the state of Indiana; these records were joined to the block group geography file using an attribute join, which is also available in most desktop mapping applications. School district data was also collected from the Indiana Department of Education data extractor. Area count analysis enables student enrollment to be modeled contiguously while maintaining student privacy and analyzing influential enrollment variables simultaneously. The following is the formula for the BYM Model that will be used in this analysis, show in Figure 2:
Figure 2. Besag, York, and Mollie Model

\[
\psi_i = \frac{y_i}{e_i} \\
\log(\psi_i) = \log \left( \frac{y_i}{e_i} \right) \\
\log(\psi_i) = (\log y_i - \log e_i) \\
\log y_i - \log e_i = \alpha + \sum \beta x_i + U_i + V_i \\
\log y_i = \log e_i + \alpha + \sum \beta x_i + U_i + V_i
\]

Where:
- \( Y_i \sim \text{Poisson}(e_i, \psi_i) \)
- \( \alpha = \text{Constant} \)
- \( \beta = \text{Influential Enrollment variables} \)
- \( U_i = \text{Correlated Heterogeneity} \sim \text{Spatial Autocorrelation} \)
- \( V_i = \text{Uncorrelated Heterogeneity} \)

The BYM Model allows for the inclusion of influential enrollment variables to be modeled along with the relative risk, uncorrelated heterogeneity, and correlated heterogeneity (Lawson 2003). The influential enrollment variables used in this analysis can be classified as demographic, educational, or geographical. The demographic variables used include total population between 18 and 20 years old, households earning more than $100,000 per year, population currently enrolled in college/university, African American population, population who have earned at least a masters degree. Geographic variables include distance from each block group centroid to study institution and population density. The educational variables which were collected at the school district level include average SAT Score, graduation rate, and 10th grade ISTEP score. Relative risk is a measure of expected enrollment, calculated as the accepted population divided by the total population. The uncorrelated heterogeneity is the observed enrollment, or prior distribution within each census block group. The correlated
heterogeneity uses an adjacency matrix to account for the spatial distribution of block groups.

Modeling student enrollment using Bayesian statistics offers the ability to create different models based on different influential variables. In this research three models will be tested, one using only demographic variables, one using geographic and academic variables and the third utilizing demographic, geographic, and academic variables. These three models were selected because of the source and availability of the data. The first model containing the demographic data contains data that was donated by the data provider, but typically would need to be purchased. The second model using geographic and academic variables illustrates a model that can be created using data which can be obtained free of cost from a variety of different sources. The third and final model is a combination of the two data sets, combining the demographic, geographic, and academic variables.

The first model will only include demographic data, shown in Figure 3, in this model the influential enrollment variables will include; population between 18 and 20 years old, households earning more than $100,000 per year, population currently enrolled in college/university, African American population, and population earned at least a masters degree.

Figure 3. Besag, York, and Mollie Enrollment Model 1

\[
\log Y_i = \alpha + \beta_1 + \beta_2 + \beta_3 + \beta_4 + \beta_5 + U_i + V_i
\]

Where: 
\( Y_i \sim \text{Poisson} \left( e^{\log Y_i} \right) \)
\( \alpha = \text{Constant} \)
\( \beta_1 = \text{Population aged 18-20} \)
\( \beta_2 = \text{Households earning greater than $100,000 per year} \)
\( \beta_3 = \text{Population currently enrolled in College/University} \)
\( \beta_4 = \text{African American Population} \)
\( \beta_5 = \text{Population education greater than masters degree} \)
\( U_i = \text{Correlated Heterogeneity~ Spatial Autocorrelation} \)
\( V_i = \text{Uncorrelated Heterogeneity} \)
The second model will include the geographic and academic variables, shown in Figure 4, in this model the influential enrollment variables will include; the school district average SAT score, average school district SAT score, average school district 10th grade ISTEP score, Euclidean distance from study institution to block group centroid, and block group population density.

**Figure 4. Besag, York, and Mollie Enrollment Model 2**

\[ \log \theta_i = \alpha + \beta_1 + \beta_2 + \beta_3 + \beta_4 + U_i + V_i \]

Where: \( Y_i \sim \text{Poisson}(\theta_i) \)
- \( \alpha \) = Constant
- \( \beta_1 \) = Average SAT Score
- \( \beta_2 \) = Graduation Rate
- \( \beta_3 \) = Average 10th grade ISTEP Score
- \( \beta_4 \) = Distance from IUPUI to block group
- \( \beta_5 \) = Population density
- \( U_i \) = Correlated Heterogeneity ~ Spatial Autocorrelation
- \( V_i \) = Uncorrelated Heterogeneity

The third model will include the same demographic variables as the first model and the same geographic and academic variables as the second model, shown in Figure 5, in addition to average SAT score, graduation rate, average 10th grade ISTEP score, distance from IUPUI to block group, and population density.

**Figure 5. Besag, York, and Mollie Enrollment Model 3**

\[ \log \theta_i = \alpha + \beta_1 + \beta_2 + \beta_3 + \beta_4 + \beta_5 + \beta_6 + \beta_7 + \beta_8 + \beta_9 + \beta_{10} + U_i + V_i \]

Where: \( Y_i \sim \text{Poisson}(\theta_i) \)
- \( \alpha \) = Constant
- \( \beta_1 \) = Population aged 18-20
- \( \beta_2 \) = Households earning greater than $100,000 per year
- \( \beta_3 \) = Population currently enrolled in College/University
- \( \beta_4 \) = African American Population
- \( \beta_5 \) = Population education greater than masters degree
- \( \beta_6 \) = Average SAT Score
- \( \beta_7 \) = Graduation Rate
- \( \beta_8 \) = Average 10th grade ISTEP Score
- \( \beta_9 \) = Distance from IUPUI to block group
- \( \beta_{10} \) = Population density
- \( U_i \) = Correlated Heterogeneity ~ Spatial Autocorrelation
- \( V_i \) = Uncorrelated Heterogeneity
This research provides an opportunity to assist higher education recruiters with the ability to utilize the prior enrollment data to dictate future recruitment activities, through strategic planning and tactical execution. The three models were selected to illustrate the ability to use not only data which is collected from individual students, but data that is collected from the State Department of Education, and purchased variables from demographic data providers. The ability to combine traditional enrollment management techniques and advanced statistical modeling has the potential to identify students most likely to enroll, while providing the right message to the right student.
RESULTS AND CONCLUSIONS

The conclusions which can be drawn from this analysis demonstrate the ability to combine traditional enrollment theory and the data and technology that currently exists. These models represent a way to identify prospective students, based on demographic, academic, and geographic variables, which have the highest likelihood to enroll. By deploying these models early in the enrollment cycle, enrollment professionals are given the opportunity to spend their time with students who have the highest propensity to enroll. The skills which are necessary to create strategic and tactical enrollment plans vary from basic descriptive statistics to more complex regression techniques.

Enrollment professionals are in a unique situation, in which they have access to highly educated professors, as well as access to companies which specialize in enrollment modeling. This analysis was completed using available resources as well as utilizing knowledge found within the study institution.

Three models were created and tested. In summary, the first model uses only demographic variables, while the second model uses geographic and academic variables, and the third model represents the most complete model using demographic, geographic, and academic variables. The WinBUGS (The BUGS Project Welcome 2010) software was used to calculate the theta, or relative risk, of these models. Results are significant if the results are less than or greater than 1.96. The results for the first model, shown in Table 1, show that the number of households earning greater than $100,000 per year, total African American population, and correlated and uncorrelated heterogeneity are significant, while total age 18-20, college enrollment, and education higher than a masters were not significant.
Table 1. Results of Besag, York, and Mollie Enrollment Model 1

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>sd</th>
<th>2.5%</th>
<th>median</th>
<th>97.5%</th>
<th>significant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-0.50570</td>
<td>0.04954</td>
<td>-0.60330</td>
<td>-0.50560</td>
<td>-0.4092</td>
<td>-10.2079</td>
</tr>
<tr>
<td>Age 18-20</td>
<td>-0.00128</td>
<td>0.00098</td>
<td>-0.00322</td>
<td>-0.00125</td>
<td>0.00062</td>
<td>-1.30795</td>
</tr>
<tr>
<td># HH Inc &gt; 100K</td>
<td>0.00236</td>
<td>0.00041</td>
<td>0.00154</td>
<td>0.00236</td>
<td>0.00322</td>
<td>5.72573</td>
</tr>
<tr>
<td>College Enrollment</td>
<td>-0.00054</td>
<td>0.00037</td>
<td>-0.00128</td>
<td>-0.00053</td>
<td>0.00017</td>
<td>-1.46833</td>
</tr>
<tr>
<td>Black Population</td>
<td>0.00065</td>
<td>0.00008</td>
<td>0.00049</td>
<td>0.00065</td>
<td>0.00081</td>
<td>7.85939</td>
</tr>
<tr>
<td>Earned masters</td>
<td>-0.00076</td>
<td>0.00069</td>
<td>-0.00217</td>
<td>-0.00076</td>
<td>0.00059</td>
<td>-1.10906</td>
</tr>
<tr>
<td>Correlated Heterogeneity</td>
<td>0.00127</td>
<td>0.00003</td>
<td>0.00120</td>
<td>0.00127</td>
<td>0.00133</td>
<td>36.88047</td>
</tr>
<tr>
<td>Uncorrelated Heterogeneity</td>
<td>1.14700</td>
<td>0.03239</td>
<td>1.08600</td>
<td>1.14600</td>
<td>1.21400</td>
<td>35.41216</td>
</tr>
</tbody>
</table>

The first model, which uses demographic and spatial data, is highly spatial, as shown by the high correlated heterogeneity. The constant is the average of all variables not taken into account in the model. The income variable, number of Households with income greater than $100,000, is significant in this model because it represents the areas with greater influence and affluence, as higher incomes generally are associated with higher educated households. The African American Population is also significant in this model. The geographic area that comprises the primary trade area for the study institution contains a concentration of densely populated, minority friendly, inner city neighborhoods. The relationship between the observed and predicted values in the first model, which has an $R^2$ of .9317, and shown in Figure 6, indicates that the predicted values closely follow the observed values.
Shown in Figure 7 is the observed enrollment, while Figure 8 shows the predicted enrollment. When comparing the maps there is noticeable similarities around Indianapolis, Merrillville, and Fort Wayne, while the more rural areas throughout the state have few predicted cases of enrollment.
The results for the second Model, shown in Table 2, shows that average SAT score, distance from IUPUI, high school graduation rate, and correlated heterogeneity are significant factors in predicting enrollment. Additionally, the model found that population density, Average 10th grade ISTEP score, and uncorrelated heterogeneity are not significant factors.

Table 2. Results of Besag, York, and Mollie Enrollment Model 2

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>sd</th>
<th>2.5%</th>
<th>median</th>
<th>97.5%</th>
<th>Significant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.9449</td>
<td>0.38480</td>
<td>0.26690</td>
<td>0.90830</td>
<td>1.67600</td>
<td>2.45556</td>
</tr>
<tr>
<td>Population Density</td>
<td>0.00047</td>
<td>0.00026</td>
<td>-0.0002</td>
<td>0.00046</td>
<td>0.00102</td>
<td>1.78144</td>
</tr>
<tr>
<td>Average SAT Score</td>
<td>-0.00141</td>
<td>0.00040</td>
<td>-0.0022</td>
<td>-0.00144</td>
<td>-0.00066</td>
<td>-3.56151</td>
</tr>
<tr>
<td>Distance From institution</td>
<td>-0.01968</td>
<td>0.00093</td>
<td>-0.02103</td>
<td>-0.01992</td>
<td>-0.01768</td>
<td>-21.20690</td>
</tr>
<tr>
<td>Graduation Rate</td>
<td>0.02050</td>
<td>0.00326</td>
<td>0.01358</td>
<td>0.02062</td>
<td>0.02674</td>
<td>6.29800</td>
</tr>
<tr>
<td>ISTEP 10th Grade Passing %</td>
<td>-0.00081</td>
<td>0.00206</td>
<td>-0.00476</td>
<td>-0.00081</td>
<td>0.00334</td>
<td>-0.39549</td>
</tr>
<tr>
<td>Correlated Heterogeneity</td>
<td>0.00127</td>
<td>0.00003</td>
<td>0.00120</td>
<td>0.00126</td>
<td>0.00133</td>
<td>37.42604</td>
</tr>
<tr>
<td>Uncorrelated Heterogeneity</td>
<td>0.22180</td>
<td>0.33040</td>
<td>0.01081</td>
<td>0.01718</td>
<td>0.80650</td>
<td>0.67131</td>
</tr>
</tbody>
</table>
The average SAT score variable is significant, particularly in higher education enrollment. There is a double acceptance in higher education, the student must accept an institution and the institution must accept the student. One aspect of the enrollment process is SAT score, which the institution requires a minimum in order to be considered for admission. Distance is a significant factor, this could be contributed to the fact that the study institution is an urban campus, and a member of a network of schools spaced throughout the region. For similar reasons the distance variable contributes to the correlated heterogeneity variable. Graduation Rate is a significant factor, similar in nature of SAT score, schools and school districts with higher graduation rates have more students who qualify for admission. The relationship between the observed and predicted values in the second model, which has an $R^2$ of .5678, and shown in Figure 9, indicates that the predicted values and observed values contain considerable differences.

Figure 9. Scatter Plot of Besag, York, and Mollie Model 2

![Model 2](image)

Shown in Figure 10 is the observed enrollment, while Figure 11 shows the predicted enrollment. When comparing the maps there is noticeable clustering in the predicted
values around and radiating outward from Indianapolis, while the further away the fewer predicted cases of enrollment.

The results for the third model, shown in Table 3, shows that population density, African American population, average SAT score, distance from IUPUI, high school graduation rate, and correlated heterogeneity are significant factors in predicting enrollment. Additionally, the model found that population aged 18-20, number of households earning greater than $100,000 per year, college enrollment, population who have earned at least a masters degree, average 10th grade ISTEP score, and uncorrelated heterogeneity are not significant factors.
Table 3. Results of Besag, York, and Mollie Enrollment Model 3

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>sd</th>
<th>2.50%</th>
<th>median</th>
<th>97.50%</th>
<th>Significant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>1.241</td>
<td>0.4436</td>
<td>0.4176</td>
<td>1.276</td>
<td>2.108</td>
<td>2.79757</td>
</tr>
<tr>
<td>Age 18-20</td>
<td>-0.00067</td>
<td>0.00064</td>
<td>0.00225</td>
<td>-0.0006</td>
<td>0.00046</td>
<td>-1.04009</td>
</tr>
<tr>
<td>Population Density</td>
<td>0.00062</td>
<td>0.00029</td>
<td>0.00008</td>
<td>0.0006</td>
<td>0.00121</td>
<td>2.12513</td>
</tr>
<tr>
<td># HH Inc &gt; 100K</td>
<td>0.00074</td>
<td>0.00051</td>
<td>0.00002</td>
<td>0.00068</td>
<td>0.00177</td>
<td>1.45558</td>
</tr>
<tr>
<td>College Enrollment</td>
<td>-0.00031</td>
<td>0.00026</td>
<td>0.00089</td>
<td>-0.00029</td>
<td>0.00014</td>
<td>-1.21721</td>
</tr>
<tr>
<td>Black Population</td>
<td>0.00021</td>
<td>0.0001</td>
<td>0.00006</td>
<td>0.00019</td>
<td>0.0004</td>
<td>2.04573</td>
</tr>
<tr>
<td>Earned masters</td>
<td>-0.00025</td>
<td>0.00048</td>
<td>0.00146</td>
<td>-0.00017</td>
<td>0.00051</td>
<td>-0.52074</td>
</tr>
<tr>
<td>Average SAT Score</td>
<td>-0.0012</td>
<td>0.00054</td>
<td>0.00245</td>
<td>-0.00112</td>
<td>0.00029</td>
<td>-2.21934</td>
</tr>
<tr>
<td>Distance From institution</td>
<td>-0.01829</td>
<td>0.00107</td>
<td>0.02008</td>
<td>-0.01826</td>
<td>0.01648</td>
<td>-17.1576</td>
</tr>
<tr>
<td>Graduation Rate</td>
<td>0.01931</td>
<td>0.00284</td>
<td>0.01361</td>
<td>0.01946</td>
<td>0.02441</td>
<td>6.7993</td>
</tr>
<tr>
<td>ISTEP 10th Grade Passing %</td>
<td>-0.00141</td>
<td>0.00222</td>
<td>0.00552</td>
<td>-0.00153</td>
<td>0.00329</td>
<td>-0.63264</td>
</tr>
<tr>
<td>Correlated Heterogenity</td>
<td>0.00127</td>
<td>0.00003</td>
<td>0.0012</td>
<td>0.00127</td>
<td>0.00133</td>
<td>36.93878</td>
</tr>
<tr>
<td>Uncorrelated Heterogeneity</td>
<td>0.424</td>
<td>0.3732</td>
<td>0.00927</td>
<td>0.7245</td>
<td>0.8218</td>
<td>1.13612</td>
</tr>
</tbody>
</table>

The population density variable is a measure of how concentrated the population is within a given geographic area. It is significant in this model because the denser a neighborhood is the better the opportunity that potential students exist. The African American population is also significant in this model. The geographic area that comprises the primary trade area for the study institution contains a concentration of densely population, minority friendly, inner city neighborhoods. The average SAT score variable is significant, particularly in higher education enrollment. One aspect of the enrollment process is SAT score, which the institution requires a minimum in order to be considered for admission. Distance is a significant factor, this could be contributed to the fact that the study institution is an urban campus, and a member of a network of schools spaced throughout the region. For similar reasons the distance variable contributes to the correlated heterogeneity variable. Graduation rate is a significant
factor, similar in nature of SAT score, schools and school districts with higher graduation rates have more students who qualify for admission.

The relationship between the observed and predicted values in the third model, which has an $R^2$ of .7, and show in Figure 12, indicates that the predicted values and observed values contain some differences, with a few outliers, most noticeably areas close to the study institution.

Figure 12. Scatter Plot of Besag, York, and Mollie Model 3

![Model 3](image)

Shown in Figure 13 is the observed enrollment, while Figure 14 shows the predicted enrollment. When comparing the maps there is noticeable clustering in the predicted values around and radiating outward from Indianapolis, while the further away the fewer predicted cases of enrollment.
Comparing the significant variables across the three models provides additional insight into these variables. Table 4 shows how all three models compare against one another.

Table 4. Summary of Besag, York, and Mollie Enrollment Models 1, 2, & 3

<table>
<thead>
<tr>
<th>Influential Variables</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-0.50570</td>
<td>0.94490</td>
<td>1.24100</td>
</tr>
<tr>
<td>Age 18-20</td>
<td>-0.00128</td>
<td>-0.00067</td>
<td></td>
</tr>
<tr>
<td># HH Inc &gt; 100K</td>
<td>0.00236</td>
<td>0.00074</td>
<td></td>
</tr>
<tr>
<td>College Enrollment</td>
<td>-0.00054</td>
<td>-0.00031</td>
<td></td>
</tr>
<tr>
<td>Black Population</td>
<td>0.00065</td>
<td>0.00021</td>
<td></td>
</tr>
<tr>
<td>Earned Masters</td>
<td>-0.00076</td>
<td></td>
<td>-0.00025</td>
</tr>
<tr>
<td>Population Density</td>
<td></td>
<td>0.00047</td>
<td>0.00062</td>
</tr>
<tr>
<td>Average SAT Score</td>
<td>-0.00141</td>
<td>-0.00120</td>
<td></td>
</tr>
<tr>
<td>Distance From Institution</td>
<td>-0.01968</td>
<td>-0.01829</td>
<td></td>
</tr>
<tr>
<td>Graduation Rate</td>
<td>0.02050</td>
<td>0.01931</td>
<td></td>
</tr>
<tr>
<td>ISTEP 10th Grade Passing %</td>
<td>-0.00081</td>
<td>-0.00141</td>
<td></td>
</tr>
<tr>
<td>Correlated Heterogeneity</td>
<td>0.00127</td>
<td>0.00127</td>
<td>0.00127</td>
</tr>
<tr>
<td>Uncorrelated Heterogeneity</td>
<td>1.14700</td>
<td>0.22180</td>
<td>0.42400</td>
</tr>
</tbody>
</table>
While all three models point to the possibility of identifying prospective students, some variables are more predictive than others. The first model is the most predictive when comparing the $R^2$. It is also the model that uses data that was generated by a professional data supplier. The second model uses data that is widely available, usually free of charge from internal databases and state department of education websites. This model is the least predictive. The third model combines the two data sets into one model. When comparing the results of the three models, it is necessary to compare the individual variables. The age variable of total population aged 18-20 and the total population enrolled in higher education, which is the typical demographic of college students was not significant in either model. Households with an average household income greater than $100,000, was significant in the demographic model, but not in the combined third model, this could be caused by the introduction of the density variables. While the density variable is not significant in second model, it was significant in the third model. Household income and population density in the study area are divergent, with the more dense the geographical area, the lower the income, while the more affluent areas are less dense. Total African American population is significant in both the first model and the third model; this can be attributed to the minority friendly areas that are in the immediate areas surrounding the study institution. Using the variables of earned masters, was to identify areas in which the population was well educated, this variables was not significant in either the first model or the combined third model, this can be attributed to urban and local pull of the campus. The variables of average SAT score and graduation rate are significant in both the second model and third model. These are due to the factors of double acceptance, where the study institution has a minimum standard, and areas with low academic achievements do not typically have educational ambitions. The distance from the study institution to the center of the census block group is significant in both the second and third models; this can be associated with the
urban and local draw. The percentage of 10th graders passing the ISTEP score was not significant in either the second or third models, the ISTEP test is a statewide test that all students are required to take. The difference between the average SAT score and the ISTEP variable is that the SAT score is a voluntary test taken by students interested in continuing their studies, while the ISTEP test is mandatory, and includes students who do not necessarily interested in academic achievements. The uncorrelated heterogeneity was significant in the first model and not in either the second or third models; this difference could be attributed to the data which was provided. When comparing the point level data of individual students, to the continuous surface that the demographic data was provided, large sections of rural areas and some sections of urban areas did not have any enrolled students. Correlated heterogeneity was significant in all three models; this demonstrates the existing spatial enrollment patterns at the study institution.

The maps show that although all three models identify with the area around central Indiana correctly, their correspondence with enrollment at the state periphery is less so. Since the adjacency matrix is based on county neighborhoods, an Euclidean distance based neighborhood mapping, those counties tied with ‘shorter’ non-Euclidean or network based distances are not accounted for in the model. There are no known theoretical solutions to identifying neighbors that are non-Euclidean and hence can be applied to the hierarchical Bayesian concept or any other multivariate statistical concept. However, the Euclidean model is appropriate in representing spatial lags in a multivariate environment.

While there will always be prospective students who “break” the model, this analysis provides a glimpse into the delicate balance between the art and science that challenge
enrollment professionals on a daily basis. This analysis represents the science, while the art challenges admissions professionals to energize, understand, and commit prospective students into making favorable enrollment choices to the study institution. Understanding where students come from is one way to modify the traditional enrollment funnel. The modification of the traditional enrollment funnel from wide and flat, into one that is narrow and skinny, which may resemble a cylinder, offers enrollment professionals the ability to target the students most likely to enroll, as illustrated in Figure 15.

Figure 15. Enrollment Funnel & Enrollment Cylinder

The purpose of creating an enrollment model is to better understand the existing students, so enrollment professionals can identify prospective student earlier in the enrollment cycle. The identification of these students earlier in the enrollment cycle allows for these students to receive marketing materials based upon their individual interests and propensity to enroll. This research identifies one way enrollment professionals can extend their existing data and research capabilities to identify potentially successful students, with limited resources, doing more with less.
 enrollment data to estimate future recruiting classes.

GIScience-Geographic Information Science, is the study and interpretation of geographic data, cartographic principles, and database management to better understand fundamental spatial concepts and analysis.

Spatial Interaction Models-Statistical models which take into account spatial adjacency in addition to traditional statistics.

P-Median Models-These models are primarily found in operations management to determine optimal locations between distribution centers and retail outlets.

Bayesian Hierarchical Models-Multi-level modeling technique which uses methods developed by Thomas Bayes. These types of models calculate posterior distribution of an event's occurrence while using prior distribution, correlated and uncorrelated spatial heterogeneity, and the likelihood of enrollment.

Besag, York, and Mollie Model-(BYM Model) A Bayesian Hierarchical Model which calculates the posterior distribution of an event's occurrence that takes correlated and uncorrelated heterogeneity, prior distribution, and the likelihood of an event's occurrence into account.

Geocode-The process of creating a Latitude and Longitude point based on an address.
Correlated Heterogeneity—Also referred to as Spatial Auto-Correlation, a measurement of
the same variable through space.

Uncorrelated Heterogeneity—Is unstructured error, error which is not dependent on
adjacency.
WORKS CITED


Eppli, Mark J. and James D. Shilling. "How Critical is a Good Location to a Regional Shopping Center." *Journal of Shopping Center Research*, 1996: 97-111.


