Analyzing Student Enrollment at Clark State Community College

July 22nd 2013

Aimee Belanger-Haas, GISP
ayb5288@psu.edu

MGIS Capstone Project
The Pennsylvania State University

Faculty Advisors:
Dr. Stephen Matthews

Abstract

With less funding coming from government, colleges are being forced to internally generate additional revenues.   Increasing enrollment can play a significant role in this task especially when tuition dollars is the major source of funding. Recruitment activities have the potential to consume a large amount of time as well as resources so increasing the efficiency and effectiveness of recruiting is very beneficial. Before educational institutions refine the current recruiting strategies, care must be taken to understand the student population. This is even more important in smaller institutions without large endowments such as Clark State Community College in Southwest Ohio with a combined enrollment of just over 5,000 students. This presentation uses various exploratory statistical methods to analyze current student demographics with freely available data and utilizes this information to improve recruitment strategies as well as to identify other areas similar in composition. A secondary goal is to provide various socio-economic and institutional research maps that will help the school visualize their data.

Contents

Problem Description

Though rarely recognized outside of an educational environment, recruitment of students is a large component of the day to day activities in Higher Education Institutions. With less funding coming from State and Federal governments, colleges and universities are being forced to internally generate additional revenues and increased enrollment activities plays a major role in accomplishing this task. Recruitment activities consume a large amount of time and resources so increasing the efficiency and effectiveness of recruiting could save the institution money which could be redeployed to improve educational activities. This is even more important in smaller institutions without large endowments such as Clark State Community College (CSCC) in Southwest Ohio.

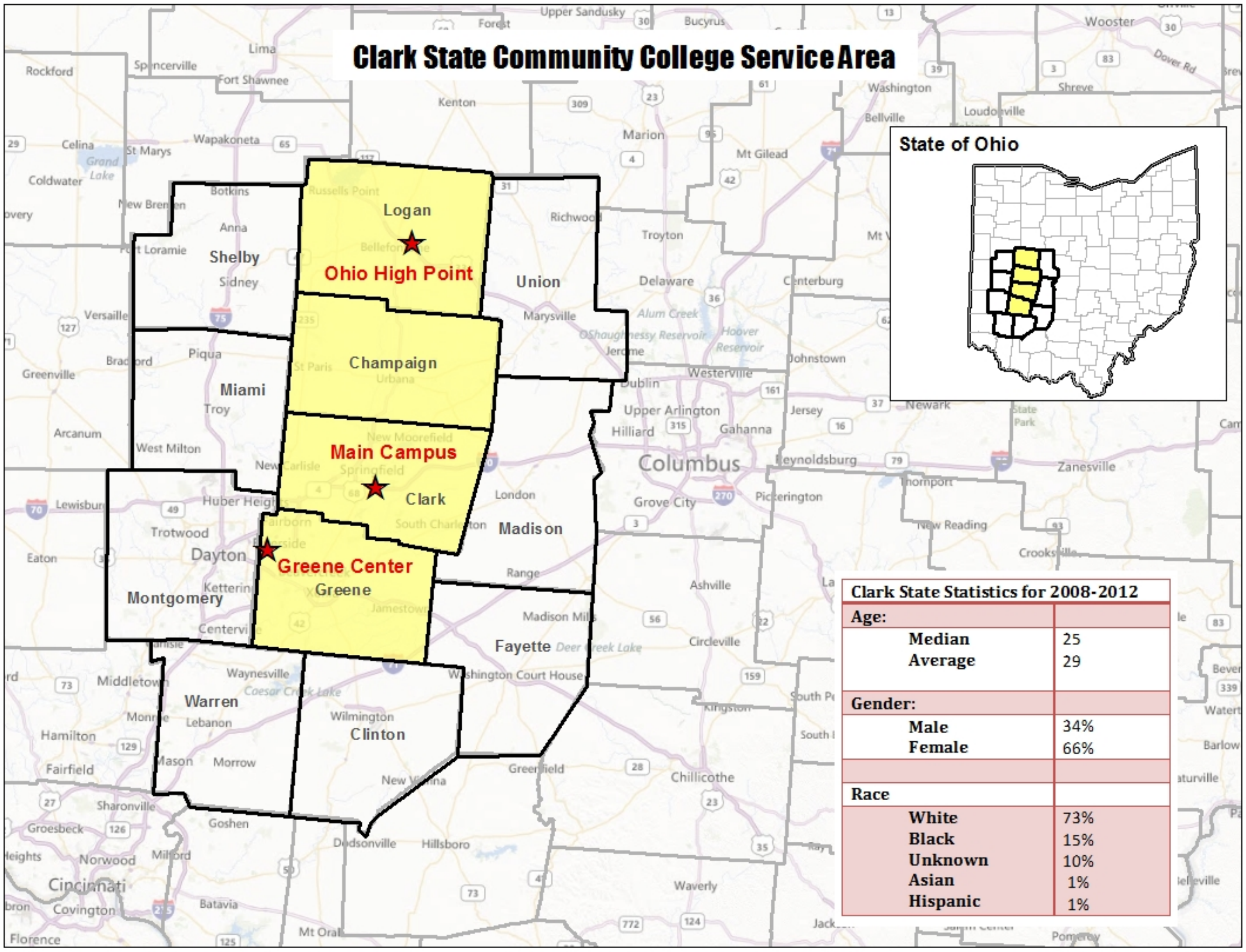
CSCC is a relatively small community college with a combined enrollment of just over 5,000 students spread across its 4 county service district (Logan, Champaign, Clark and Greene) (Figure 1). The majority of the students are located between the Main Campus located in Springfield, Clark County and the Greene Center located in Greene County (Figure 1). There is a third smaller campus (Ohio High Point) located in Logan County but it has relatively low enrollment and is not an isolated campus in this project. Clark State was first opened in 1962 and the Greene Center campus expansion occurred in 2007.The main Springfield campus is situated in close proximity to Dayton (20 minutes), Columbus (50 minutes), and Cincinnati (90 minutes). CSCC draws approximately 25% of its students outside of its service district and therefore the study area for this project falls within this 12 county area.

Before refining the current recruiting strategies, one must better understand the current student population at CSCC. In order to accomplish this task, this project uses statistical/geodemographic methods to analyze the current student demographic and uses this information to target similar areas in the surrounding area.

Each fall on the 14th day of the term, CSCC collects student demographic information to monitor how the demographic is changing over time. The following table (Table 1) summarizes this information for all campuses for the past 3 years. The drop in the 2012 enrollment data was expected as the entire state converted from quarters to semesters. Most institutions experienced increased enrollment in the years prior to the conversion and most also suffered a drop in the Fall of 2012.. This culminated in Spring 2012 with CSCC (as well as others) graduating the largest class on record.

Table 1: Clark State Community College Student Enrolment Profiles.

Year	Total Enrollment	Average Age	Male	Female	Full Time	Part Time
2012	4,977	28.2	33.8 %	66.2 %	41.1 %	58.9 %
2011	5,139	28.5	32.4 %	67.6 %	43.6 %	56.4 %
2010	4,993	28.4	34.0 %	66.0 %	45.9 %	54.1 %



Clark State Statistics for 2008-2012

Age:	
Median **Average**	25 29
Gender:	
Male **Female**	34% 66%
Race	
White **Black** **Unknown** **Asian** **Hispanic**	73% 15% 10% 1% 1%

Figure 1: Location of Clark State Service Area. The counties highlighted in yellow are part of the service district and the counties outlined in black are areas that CSCC pulls in from. The red stars represent the 3 different campus locations.

The retail sector has fully embraced the use of Geographic Information Systems (GIS) and geodemographics as a tool to help increase business and profits by better identifying potential customers. The same technology can be applied for Higher Educational institutions as there are common themes between the two types of institutions. Both have customers (students) that can be geographically identified by address and use this information to help uncover varying themes through their geodemographic profile.

In order to get a better idea of the area demographics, the general characteristics of the study area can be found below in Table 2 which gives a snapshot of the demographic profile. The data is based on American Community Survey (ACS) 5-year estimate data (2007-2011) from the Census Bureau.

Table 2: Basic demographic information for CSCC 4 county service area and the total for the 12 county study area. Information is based on ACS 5-year estimate.

	Service Area (4 counties)	Study Area (12 counties)
Percent of Population 20-29	13.4%	12.4%
Median Age	39.4	38.7
Median Income	$32,636	$33,952
Percent White	88.3%	84.5%
Percent with no College	59.86%	62.16%
Percent Unemployment	9.54%	9.99%

Research Question:

Based on five years of registration data, what areas is CSCC successful in recruiting students. Can socio-economic variables from the American Community Survey help identify the characteristics of these higher enrollment areas. What are the statistically significant variables that help predict enrollment. Can this relationship be successfully modeled via spatial regression techniques. Finally can we target areas that currently have lower enrollment but based on the model have the potential of producing more students.

Since the college has never had access to GIS mapping, a secondary objective to this project is to produce additional maps for the institution that characterize census demographic information.

Literature Review

Geodemographics is the study of people according to where they live and is loosely based on the assumption of “birds of a feather flock together”. This field provides demographers the capability to predict consumer behavior based on neighborhood classification. Charles Booth’s “Descriptive Map of London Poverty” was first published in 1889 and can be credited as being the first example of geodemographics (Troy, 2008).

The power of geodemographics has long been embraced in the business, retail, and housing community as is evidenced by any internet search on market or customer segmentation (Troy, 2008) (Krestle, 2004) and (Hanewicz, 2012)to name a few. Some Higher Education Institutions have also begun to apply geodemographics as an aid in student recruiting. However, one must remember that larger educational institutions have access to numerous resources such as experienced analyst, access to large datasets, monetary funding, and abundance of applicants that allow them to accomplish such analysis, but smaller community colleges usually do not have access to some or all of the listed resources.

This section reviews geodemographic literature that was previously done and was a starting point for this analysis. It is important to review existing literature so that lessons learned from those projects (e.g., methodologies, data sources, approaches that did not work, etc.) can be assimilated and applied to this project. The following synopses provide a review of previous projects as well as which aspects of them can be applied to this project.

Applications in Ohio and the US

The Ohio State University (TOSU) is the largest public Higher Education institution in Ohio. They introduced the use of targeted marketing to specific segments of the population in the early 1990's. This research was accomplished with the help of Doug Marble and several graduate students. These studies span several years (Marble, Applying GIS Technology to the Freshman Admissions Process, 1995) (Marble, A Model for the Use of GIS Technology in College and University Admissions Planning, 1997) (Mora, 2003) and details how TOSU utilized census demographic data (at the block level) combined with local information to identify students that are likely to enroll at TOSU.

Although these studies were done nearly 20 years ago, the papers still represents several ideas that are pertinent today and that can be applied to my analysis. TOSU uses several datasets including internal databases (which contain multiyear data on *prospective* students that were ultimately accepted) as well as external commercial datasets from Claritas (for market segmentation analysis), Ohio Department of Education (for High School enrollment statistics) and national testing services (ACT and SAT for test scores and addresses) (Mora, 2003). The model filters the census block level data through a demographic filter which in turn identifies subgroups that are likely to enroll. These results are then reprocessed with more traditional admission activities. These areas represent target areas (hot spots) for recruitment activities. TOSU then purchases names and addresses of students from the testing services to target for marketing (Mora, 2003). However, the typical CSCC student is not necessarily taking the SAT or ACT test (we have an open door policy that admits everyone regardless of grades or educational attainment). In addition, I have no budget for purchasing datasets.

At CSCC, perspective students almost always become students. In this paper, I am seeking to identify geographic areas that CSCC can target based on historical data of its students. I identify the characteristics of past and current students by investigating student records gained from Institutional Research. Mora (2003) also explains that the market segments for TOSU are based on census block groups. However, it is impossible to perform this study with the same unit of census geography. The US Census Bureau has updated the way it obtains socio-economic data. Previously it sent out the long form SF3 questionnaire to 1 in 6 households for a specific point in time. The new American Community Survey (ACS) collects data on 250,000 households monthly. It is a period estimate, which, at the tract level, is based on the five prior years. As an estimate, the US Census Bureau includes a margin of error (MOE) that reflects potential sampling problems and recommends caution when the MOE is greater than 10% (Census, 2013). The MOE can be very high for smaller units in less populated areas; therefore block group data is not reliable for this study, the Census Bureau also suggests caution when using block group data. Consequently, I have chosen to utilize census tract data as the more reliable estimates for the variables of primary interest. The OSU analysis also takes into account out of state students, and while CSCC does have out of state students (online), the percentage is low and this study does not address these students for that reason.

Bowling Green University in Northwest Ohio has built on Ohio State's success with similar studies (Zhou, 2005). Their study is quite different from my project as they were more interested in determining areas that provide more (or less) students than expected based on distance and potential population numbers (Zhou, 2005). I do not feel that this paper is very pertinent to my study. I was initially attracted to the article since it was accomplished in Ohio but other than location, not many ideas can be applied.   The community college student is quite different than a typical University student in that they tend to attend colleges that are close to where they currently live and do not move to a new location. Miami University in Southwest Ohio has also tried to encourage enrollment through specific marketing back in 1999 by investing $300,000 in different marketing techniques such as billboards and commercials in the larger cities in Ohio that possessed their own target demographic (Livinsgton, 2000). But again, CSCC does not have that kind of budget for any type of marketing and cannot attempt a similar exercise on that scale. Approximately 1 month prior to the beginning of a semester, CSCC do run commercials on local radio.

Stephen DesJardins of The University of Michigan Center for the Study of Higher Postsecondary Education has also published information on recruiting students. Their study focused on potential students that are on the fence about enrolling into college (admitted but not enrolled). The article is interesting and details the variables (students admitted in early January for the next year that are not recruited athletes and who have submitted SAT/ACT student profile questionnaires) needed for the model.   However, it is not very applicable to my study because CSCC students do not need to take the SAT/ACT and because we enroll students' right into the first week of school, no many students ever enroll months in advance.

Non-US Applications

The United Kingdom has also been interested in the combination of geodemographics as a tool for student recruitment (Batey, 1999; Read, 2005; Singleton, April 2012; Tonks, 1995). They seem to have embraced the combined use of geodemographics with GIS to gain a better understanding of their student population. In a recent article in the Journal of Geographic Systems, Singleton (2012), describes the use of a JAVA framework model to predict the flow of students from various market segments. The article indicates that the model is a good predictor but that a possible breakdown by major would help refine it. In another UK article, Read et all (2005) performed a case study of five different institutions in the hope of creating an "Admissions GIS" model. Their research was built on previous studies (Batey, 1999) (Tonks, 1995) and their paper was successful in identifying barriers that need to be addressed as well as issues of datasets accuracy. Of the barriers, only access to commercial geodemographic data is pertinent to my study, but the benefits are numerous and include identification of low participation and deprived areas, visualization of change and identification of targeted marketing areas. They also suggest comparison for different academic programs and changes over time. Although a geographically different area, analysis in the UK is fairly consistent to that of the US and of other markets.

The unique situation of Community Colleges

As previously hinted, one must understand the fundamental differences between 4 year institutions and 2 year community colleges. Most of the papers reviewed are targeting students graduating from high school (approximately 18 years old). However, the average age of a CSCC student ranges from 29-31 depending on the year (well out of high school). The article written by Lane (2003) helped me understand that the fundamentals of other studies from various 4 year institutions can serve as a starting point for my study, but that I need to take into account the "open door" policy that exists at CSCC. While Lane (2003) does not address recruitment strategies, he does explain the differences between the two types of institutions. Unlike the papers listed above, CSCC does not selectively admit students; all are accepted and put into an appropriate level to get them ready for college level courses.

Some research has been accomplished at a community college level. Crosta (2006) presents a paper that describes the steps undertaken by the Community College Research Center (CCRC) to determine the socioeconomic status (SES) of community college students in Washington State using Census block groups. I feel that this paper details many aspects that are relevant to my research and is used as a starting point in my analysis. The paper details the steps undertaken to link the student data to the census block groups to estimate the SES. This paper is different than others detailed in this report as they formulated their own demographic groups rather than just using those provided by commercial datasets. It even details which particular variables within the census data were utilized to compute the demographic groups as well as the steps necessary for accomplishing this task. The authors then examined the SES of the 15 demographic groups on both 1990 and 2000 Census datasets to ensure consistency. The paper ends with recommendations and examples that could help sway administrators of the value of such analysis. The results from the study can help asses such things as market penetration and to help gain a better understanding of the make-up of their own service areas. This paper is instrumental in my own analysis.

Missing information

Of all the articles listed in this paper, most are utilizing parts of datasets that I have envisioned for my study. Regardless of geographic location (US versus UK), researchers use area-based census variables such as sex and age breakdowns, education attainment, employment status and economic datasets to help categorize their market segments. Some of these census variables have already been incorporated into commercial datasets and projects (Marble, A Model for the Use of GIS Technology in College and University Admissions Planning, 2001) (Singleton, April 2012). .

It is also interesting that although I was successful at identifying the needed resources, surprisingly there is not an abundance of articles that detailed the specifics of the analysis but only give general information. As noted by DesJardins (2002) many higher education institutions do this type of research but are not disclosing their techniques as they regard them as proprietary and are not willing to reveal methods to prevent giving other institutions advantages at recruiting students that might be swayed by their tactics.

The articles listed in this section clearly show that this type of analysis is fairly common at larger four year institutions and while it is relatively less common at community colleges, there is no real barrier at attempting something similar at Clark State. The one issue that comes to mind is the lack of a budget with which to buy geodemographic segmentation information. Crosta et al (2006) does describe how to create one's own societal groups with Census data while this may be a solution it is a rather time consuming alternative.

Approach/methods:

The approach for this project consisted of identifying and acquiring the appropriate data sets, processing the data with the appropriate software, and then analyzing the processed data so that the research question could be answered. As a summary of the approach, the student address dataset was first requested, obtained, and geocoded. Next, Census data was identified, downloaded, and processed. Then, the data was examined with exploratory spatial data analysis (ESDA) and finally the results were analyzed to answer the various research questions. The following subsections provide in depth details on each stage of the approach. Table 3 lists the respective software programs were chosen for both ease of use and the ability to excel at certain tasks.

Table 3: List of Software utilized.

Software	Task
ArcGIS	Geocoding, locator and standard choropleth maps, and density mapping. Some spatial statistics tools such as linear regression was also accomplished Ordinary Least Squares (OLS).
Excel	Data cleaning and manipulation
SPSS	Individual data analysis (summary statistics) Tract level data: univariate, bivariate (correlations, difference of means), and OLS regression
GeoDA	Moran's I, LISA maps, OLS and spatial lag regression

Acquire Student Data

The first step was to acquire a dataset of student address information (Table 4). As previously mentioned, Clark State is too small to have its own IRB board and my case needed to be heard by another institution. This caused some additional issues since this analysis is part of a capstone project for Penn State University and the local IRB board questioned whether my petition should have been heard at Penn State. However I pushed to have it heard locally as I knew it would be easier to deal with questions. Also the project is a direct benefit to CSCC and the board relented.

The first task is deciding which type of petition needs to be filed. The US Department of Health and Human Services offers flow charts that should greatly simplify the choice (Health and Human Services). At first it was determined that this project would need a full review and I therefore completed the petition and extensive paperwork. However, once the petition was reviewed by the IRB clerk it was decided that this type of research is expedited and that I needed to update my petition and accompanying paperwork. However, once the petition was finally evaluated by the IRB they quickly determined that this research is exempt and required yet another change to the paperwork submitted with the petition.

As a side note, anyone who has a petition heard by IRB will need to take a short course offered via the Collaborative Institutional Training Initiative (CITI) for investigators prior to the petition being heard. This tutorial will take approximately 8-10 hours to complete and expires after 2 years.

Anytime research involves human subjects (students) research, you must first obtain consent from the participants. Because this study was retrospective, I could not acquire consent from thousands of previously enrolled students. As a public institution, CSCC is bound by all relevant state and federal statutes. The Family Educational Rights and Privacy Act (FERPA) prohibit institutions from releasing student information outside the "public directory" information (i.e. demographic information). In order to obtain consent and not break FERPA, I had to claim an exemption to FERPA by having the President from CSCC write an exemption letter stating that this research was for a legitimate educational interest/research study. However I was not aware that this needed to be done so it was not included with the petition. In addition the IRB also missed that the exception was not filed. It was not until I was about to take ownership of the data that this problem was discovered by CSCC. This caused a delay in receiving the data and a start to the project. Once the application was approved with the exemption, CSCC released the information. Additional information on the IRB process can be found in Appendix A.

Table 4: Information collected by CSCC that will be provided for the analysis.

Student Information
Address
Gender
Age
Ethnicity
Degree/Major

CSCC provided the requested information within an excel spreadsheet that had different worksheets for each registration year. I then created a geodatabase and imported the 5 different sheets. The student dataset was then geocoded with the ArcGIS online address locator service. Once geocoded the different feature classes were combined into one. In order to facilitate queries later on, I also created a few new fields with included an age group attribute that groups students into various age decades based on their age (Figure 2). This step was necessary to facilitate grouping students into each census tract. As previously seen in Table 1, Clark State students are in their twenties and predominantly female. From observing Figure 2 it is apparent that the majority of the students are in their twenties.

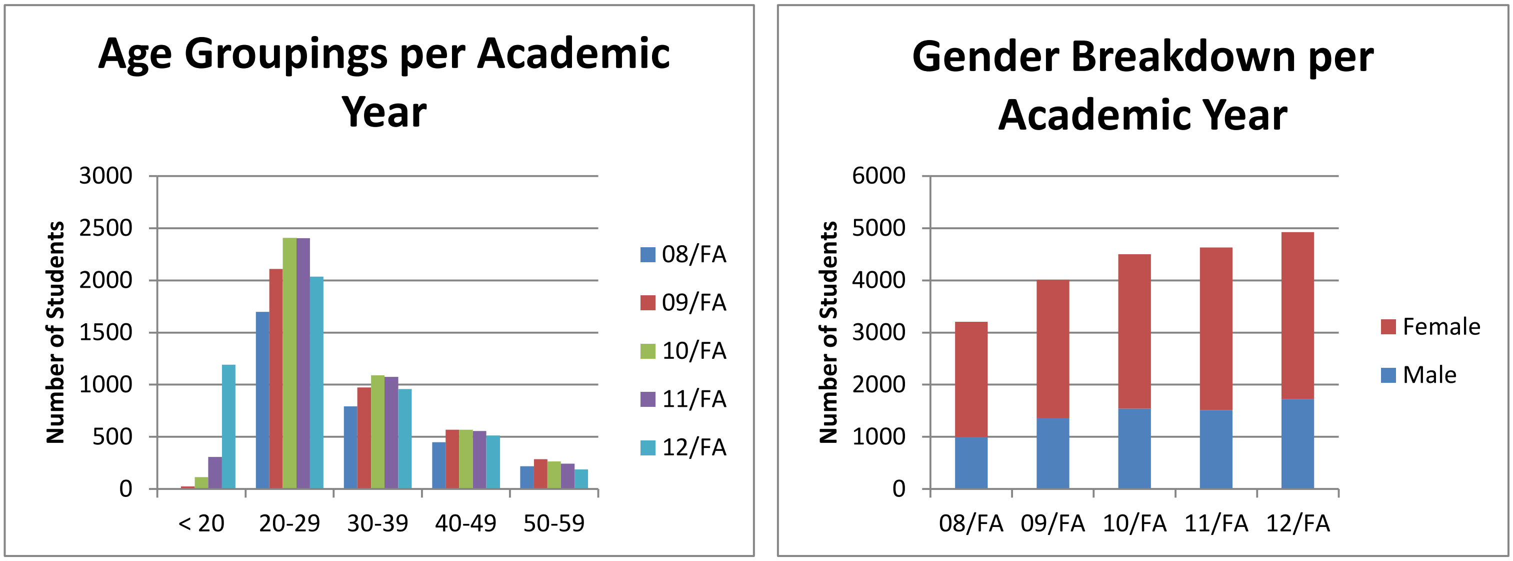


Figure 2: Age Groupings and Gender Breakdown for 2008-2012

The second field that was created was to generalize the majors. Although this paper does not address any research on the different majors, I did add a second field that generalizes the majors as I thought I would be looking at differences between majors. As an example, instead of having 20 different health related certificates and Associate Degrees, a field was added that grouped all of these together. This had the added benefit of changing the majors to a numeric field. I had originally wanted to do some analysis on the different majors but it quickly became apparent that this would not be possible as I first needed to understand the general student distribution. This information will prove to be quite useful for future studies that are discussed later in the report.

Download and Process Census Data

Census tract information was also procured. There are 355 different census tracts in the 12 county area of interest that surrounds the 3 different campuses.

Topologically Integrated Geographic Encoding and Referencing (TIGER) line shapefiles for each of the counties were downloaded from the Census.gov website and merged together into 1 feature class within a geodatabase.

Based on the literature review I was able to identify potentially important socio-economic tables to download from the Census/Factfinder website (http://factfinder2.census.gov) (Table 5). American Community Survey (ACS) 5-year estimate data (2007-2011) were utilized for demographic, social, economic and housing characteristics and Census 2010 (SF1) count data were utilized for basic counts, race and gender (Table 5). The downloaded files were examined in excel and some data processing helped clean up the information prior to the files being imported to the geodatabase. (Please refer to Appendix B for various tips and tricks that can help streamline the process of incorporating the data into your study.)

Once the tables were imported into the geodatabase they were later joined to TIGER Line feature class and ready for analysis.

Table 5: Variables utilized from American FactFinder.

Dataset	Census Table
Population Count	DP-1
Sex and Age breakdown Counts (Male and Female breakdown per ten year groupings)	DP-1
Race (White, Black, Hispanic, etc)	DP-1
Educational Attainment (HS, Associate, Bachelor, Graduate)	S1501
School enrollment (High School, Undergraduate, graduate)	S1401
Employment Status in past 12 months (in labor force, employed, unemployment rate)	S2301
Income (earnings and poverty (percentages)	DP03
Vehicle available (none, 1, 2, 3, etc)	DP04

However, once the data was downloaded and the tables joined to the geography files, I discovered that one census tract in Montgomery County did not contain any data (probably an area with no permanent population, which reduced my total count of tracts to 354.

Exploratory spatial data analysis (ESDA)

The next step was to map and examine the data, which can help you determine if there might be errors, view possible trends, and get a better understanding of the dataset you are working with. Exploratory spatial data analysis (ESDA) consists of a variety of techniques that can assist in discovering information about your dataset (Krivoruchko, 2011) p. 596). This was accomplished within Excel, SPSS, ArcGIS and GeoDa. Variables were evaluated and paired down based on correlation results and/or factor analysis.

Student Dataset

With the student dataset geocoded, analysis of the pattern of distribution can be addressed. Figure 3 graphically represent the residential location for each student who was enrolled at CSCC (2008-2012). This mapping activity has never been undertaken. There are a number of maps and reports included in this report that are not technically "high level" but are of great interest to the college administration since they have never had the opportunity to view their data within this medium.

Another meaningful mapping exercise is looking at the density of the student distribution. Mapping density is useful for looking at general patterns and allows you to visualize the concentration of features per area. This was accomplished with the Kernel Density tool in ArcGIS with a 1 mile radius for the neighborhood variable. These maps differ from point maps like the one in Figure 3 since they help reveal the concentration of student distribution.

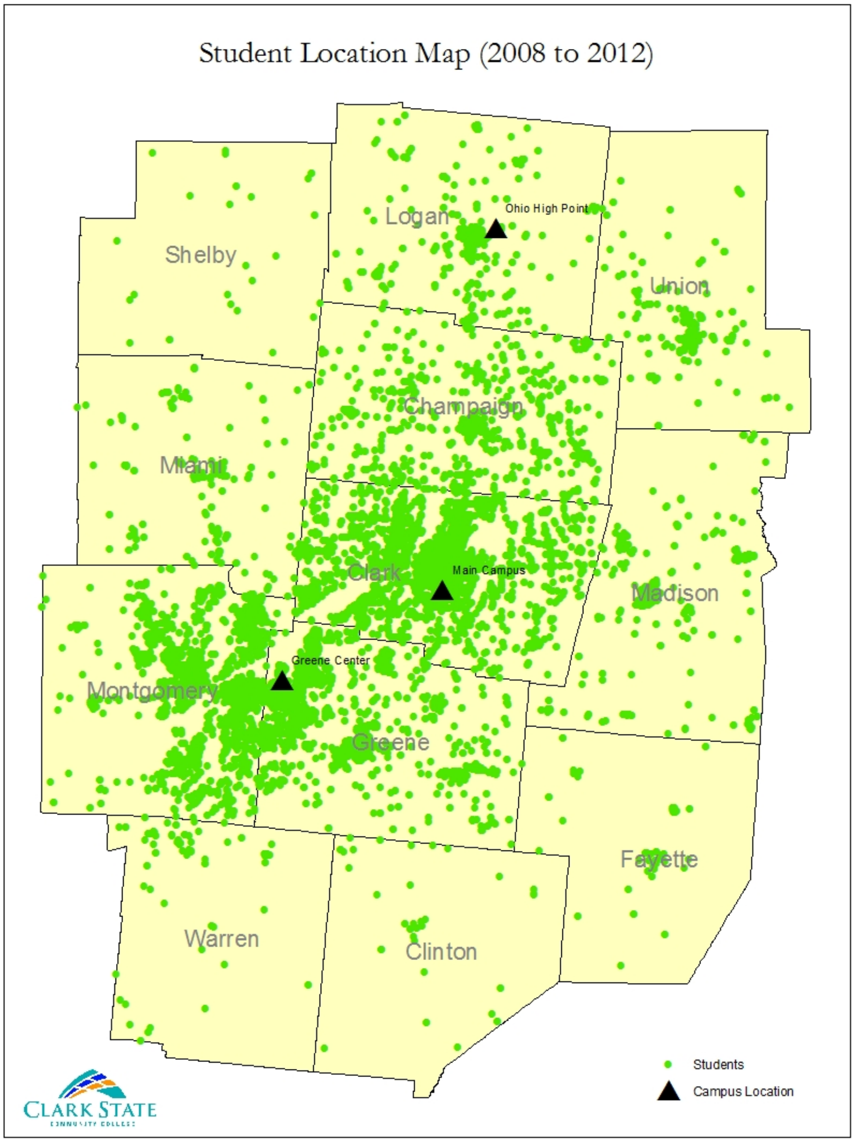


Figure 3: Student Location Map.

Figure 4 is an example of these maps, the one on the left is the density from 2008 and the map on the right is the density for 2012. The Fall 2012 Student Density Map clearly indicates that there has been an increase in student density in the southwest part of the study area. The Greene Center Campus was first opened in this area in spring 2007, the additional students centered around this campus in 2012 are simply a result of this campus expanding within the community. However, the highest concentration of students is still located around the Main Campus. Additional density maps can be found in Appendix D.

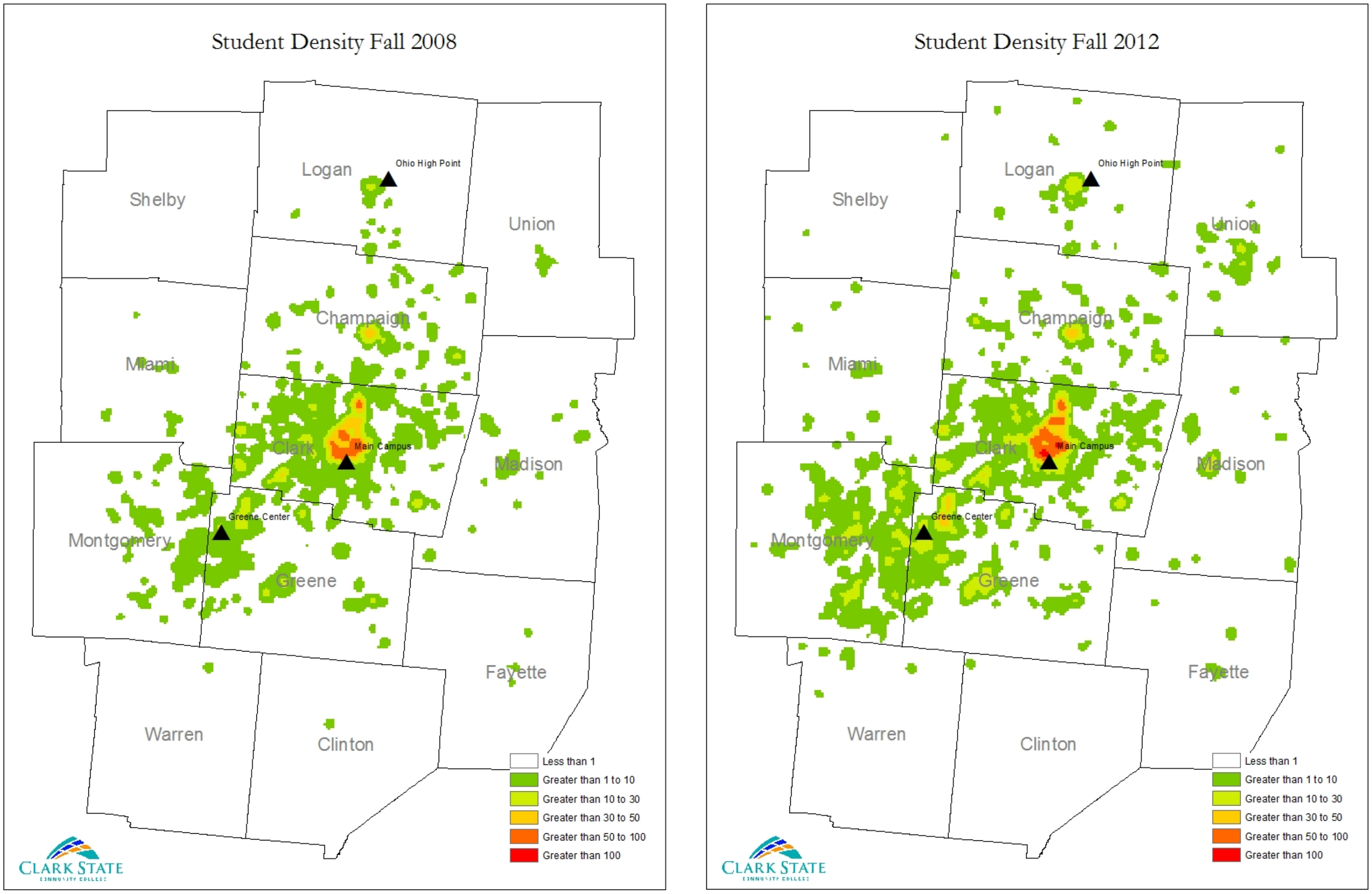


Figure 4: Student Density Maps.

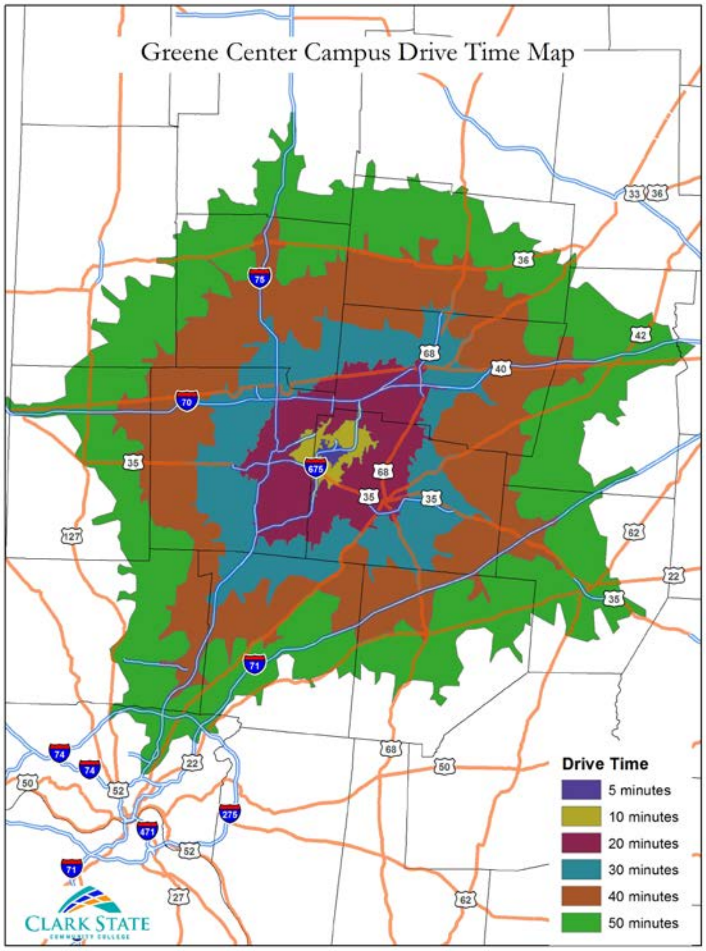


Figure 5: Drive Time Map for the Greene Center Campus.

Drive time (of 5, 10, 20, 30, 40 and 50 minutes) maps (Figure 5.) were created for each of the 3 different campuses with the help of Network Analyst in ArcGIS. These are useful to evaluate the accessibility of each campus. Further analysis (Figure 6) reveals that most students drive around 20 minutes to reach either the Main or Greene Center Campuses. Other possible variables that could be assessed include drive time evaluated by age breakdown and gender. We have already discovered that the school has a higher proportion of females but it would be interesting to know if they are willing to drive further than their male counterparts. The time boundaries were created with the help of the Network Analysis extension within ArcMap. As previously mentioned, a shortfall of the student dataset provided by CSCC is that the “home” campus of the student is not recorded and is consequently unavailable. This made analysis rather difficult since there was no way to assign each student to a college.

Additional drive time maps can be found in Appendix D.

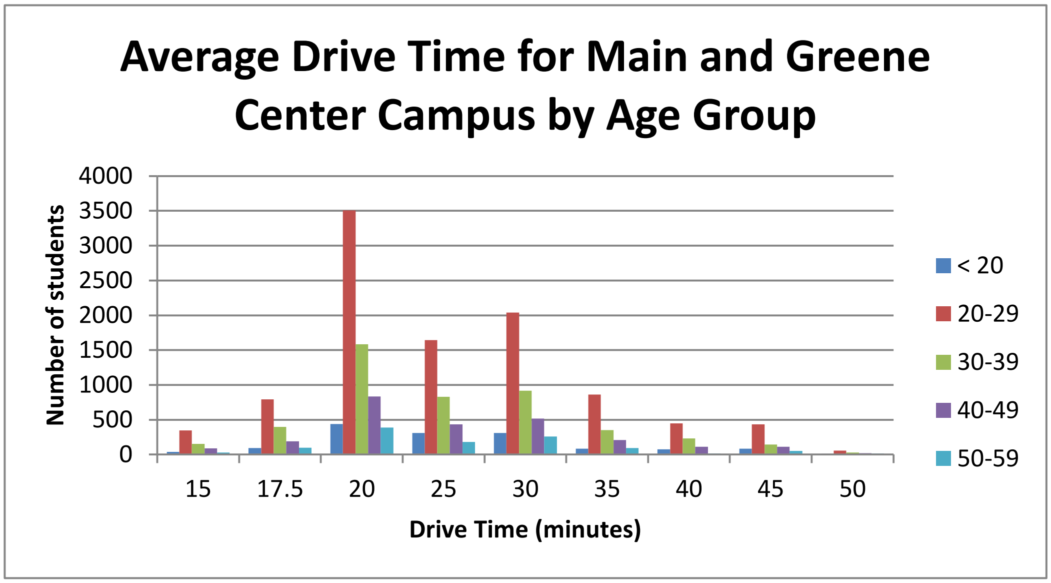


Figure 6: Average Drive Time for Main and Greene Center Campuses.

Census Data

The census data was examined in ArcGIS with the help of the geostatistical analyst tools, GeoDa with various Local indicators of spatial association (LISA) indicators, and SPSS for descriptive statistics (Table 6) and histograms (which be found in Appendix C). By conventional definition census tract data is considered to be regional data as they are counts (and estimates) for a known population within a polygon (Krivoruchko, 2011). The variables listed below were chosen based on the previous study by Crosta (2006) and others as previously discussed in the literature review. The variables chosen help describe the vast differences in the makeup of community college students in relation to four year institutions and highlight the differences between them.

Employment information is an important variable as some students return to school as a result of losing employment. Within my study period, Ohio was an industry driven state and was hit with some very high unemployment rates due to the collapse of the North American auto industry. Many people sought out new career fields and training in order to gain employment and community colleges were the logical choice for many of those.

Educational attainment is another important variable to analyze. As previously mentioned some students are recent high school graduates but many come to CSCC later in life (median age being 29) and may or may not have completed high school. Therefore if people have a high school diploma regardless of age, we are interested in them attending our school. CSCC accepts all students regardless of educational attainment. CSCC will get students ready for college (via college prep classes and other programs) and CSCC does not require a high school diploma or GED for admittance. Therefore I was interested in including people with some high school and less than 9th grade.

Additional variables that needed to be applied to this study are the vehicle availability. Of the 3 campuses only the main campus is on a regular bus route therefore students will need access to a car in order to mobilize themselves to school.

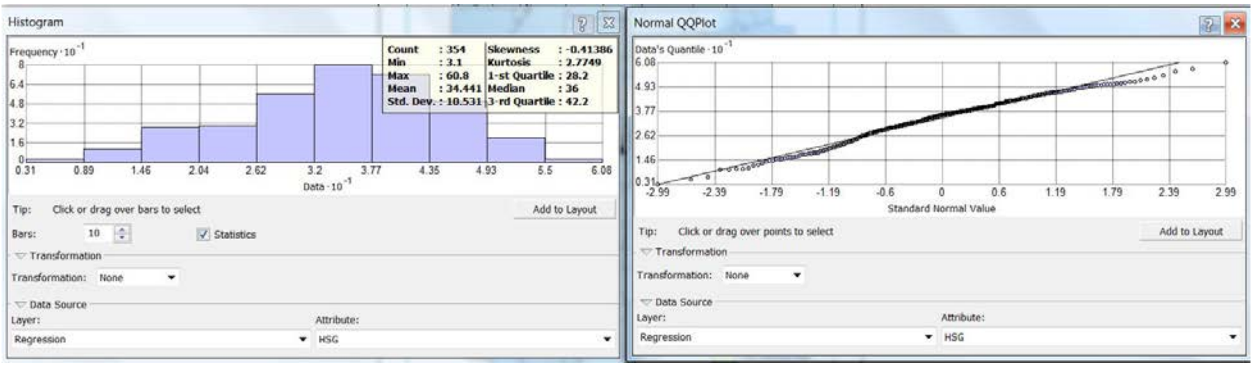
Even though the study area is also predominantly non-Hispanic white including information on race as well as factors such as poverty are important in a study such as this. In order to gain a basic understanding of the variables utilized in the study, one must start with descriptive statistics. Table 6 below provides simple summaries about the data. Together with simple graphics analysis found in Appendix C, they form the basis of virtually every quantitative analysis of data. Usually one will look at the distribution which in this case are the histograms in Appendix C, the central tendency which is the mean in Table 6 and the dispersion of the data which is the standard deviation which is the spread of the values around the mean (Social Research Methods, 2006).

Distance is another important variable for the analysis. Distance from each census tract centroid was established to both the Main and Greene Center Campuses. As previously mentioned, CSCC does not maintain information on the home campus of each student. To overcome this shortfall the average distance to either campus was calculated.

Table 6: Descriptive Statistics for Census ACS data

	N	Minimum	Maximum	Mean	Std. Deviation
Employment Rate	354	0.00	74.30	57.01	9.84
Unemployment Rate	354	0.00	39.00	10.00	6.51
Less than 9th grade	354	0.00	19.60	3.41	2.87
9 -12th grade (no diploma)	354	0.00	32.60	9.40	6.07
High School Graduate	354	3.10	60.80	34.44	10.53
Some College	354	8.70	40.60	22.50	5.69
Associate Degree	354	1.50	20.60	8.09	2.97
Bachelor Degree	354	0.00	38.50	13.59	8.35
Graduate or Professional Degree	354	0.00	39.60	8.57	7.22
Enrolled in School between the ages of 20-24	354	0.00	100.00	38.77	22.77
Enrolled in School between the ages of 25to34	354	0.00	71.30	15.47	10.65
Enrolled in School 35+	354	0.00	27.30	3.53	2.90
No Vehicle Available	354	0.00	41.80	7.75	8.18
One Vehicle Available	354	0.00	64.60	32.53	11.44
Two Vehicles Available	354	0.00	60.60	37.85	10.21
Three+ Vehicles available	354	0.00	100.00	21.58	11.33
Median Earnings	354	$8,996.00	$73,261.00	$33,952.47	$10,218.13
Population Over 18 below the Poverty Level	354	0.00	71.50	14.42	12.61
Population Over 18 in College	354	0.60	95.80	9.37	9.75
Percent White	354	2.15	99.43	81.86	24.48
Enrollment Percent	354	0.00	2.63	0.35	0.57
Average Distance (miles)	354	8.38	51.92	20.90	10.09
Closest Tract Distance	354	0.43	36.06	12.92	8.87

The ArcGIS tutorial for Geostatistical Analysis indicates that “if the mean and the median are approximately the same value, you have one piece of evidence that the data may be normally distributed” (ESRI, 2012). It contains graphs for exploratory spatial data analysis to help you understand your data distribution which will be important for choosing an appropriate regression technique (ESRI, 2012). As an example of the work accomplished on all variables listed above in Table 6, the histogram below in Figure 7 indicates that High School Graduate is normally distributed (bell shaped curve). Summary statistics are also included in the upper right corner of the histogram. The QQ (quantile-quantile) plot (on the right) is used to compare the distribution of the data to a standard normal distribution. The closer the points are to the straight (45-degree) line in the graph, the closer the sample data follows a normal distribution.



```
Count      : 354      Skewness      : -0.41386
Min        : 3.1      Kurtosis      : 2.7749
Max        : 60.8     1-st Quartile : 28.2
Mean       : 34.441   Median        : 36
Std. Dev.  : 10.531   3-rd Quartile : 42.2
```

Tip: Click or drag over bars to select

Bars: 10 ☑ Statistics

Transformation

Transformation: None

Data Source

Layer: Regression

Attribute: HSG

Tip: Click or drag over points to select

Transformation

Transformation: None

Data Source

Layer: Regression

Attribute: HSG

Figure 7: Distribution of High School Graduates (HSG)

The QQ plot above does support that the dataset (for HSG) is normally distributed. From looking at the histograms in Appendix C, you can quickly see that while some data in normally distributed not all of it is and this will lead to issues with regression analysis. A possible solution to this will be to transform the data.

In order to get a better understanding of the area, additional analysis was done in GeoDa (software that specializes in exploration of spatial data). GeoDa was used to examine the spatial distribution of the census variables of interest. Tobler first law of geography states “All places are related, but nearby places are more related

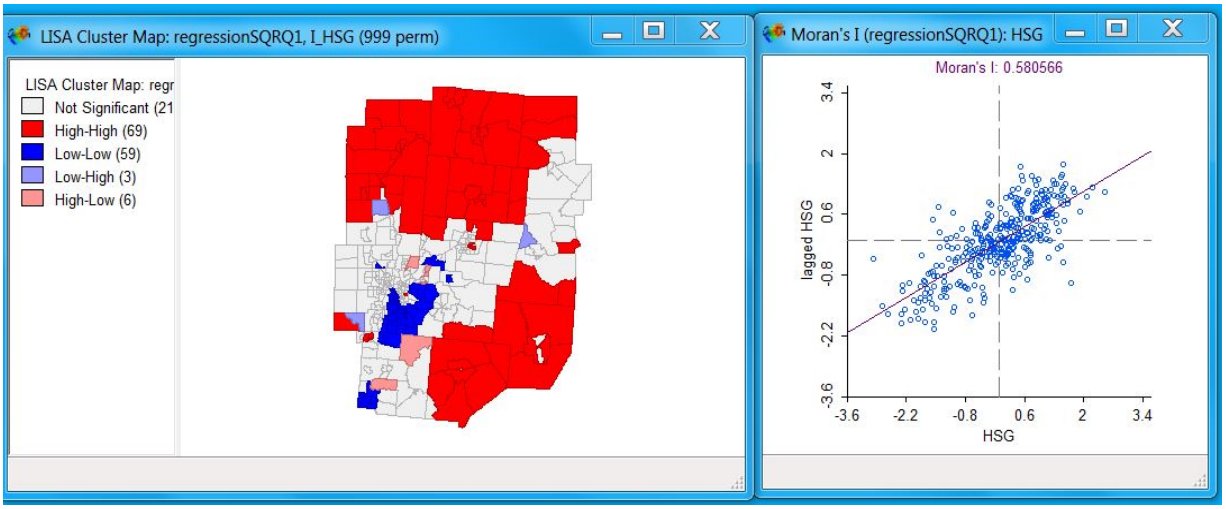


Figure 8: LISA Analysis for High School Graduates

than distant places" (O'Sullivan & Unwin, 2010 p.199). Spatial autocorrelation is the formal property that measures the degree to which near and distant things are related or dispersed. If the data sets tend to cluster together, then they are positively correlated, and if they are dispersed, they are negatively correlated (ESRI, n.d.). Local indicators of spatial association (LISA) help determine the local areas that significantly contribute to spatial autocorrelation (GeoDa Center, n.d.). In other words, LISA is used to analyze which features are the strongest contributors in the overall pattern. In order to compute these measures a spatial weights file is created to define which polygons are adjacent. For this analysis, I used a first order queen's weight based matrix. This type of matrix uses a shared border or vertex to define the neighbors (GeoDa Center, n.d.). From looking at Figure 8, the areas in blue have a low concentration of people with High School Graduates (HSG) whereas areas in red indicate higher concentrations of HSG. The higher percentage of HSG is clustered together in both the north and southeast corner of the study area. This leaves an area south of the City of Dayton with relatively low concentrations. Moran's I scatterplot can be used to measure how similar spatial data is to neighboring features and can be used to validate my subjective observations in Figure 8. The range of possible values extend from -1 (high negative_spatial autocorrelation) to 0 (no spatial autocorrelation random distribution) to +1 (high positive spatial autocorrelation) (Mitchell, 2009 p. 123-124). The slope of the regression line is Moran's I value. From observing Figure 8, we see that the Moran's I value is 0.58 and the slope of the regression line is positive indicating an overall positive autocorrelation that is statistically. The dark red and blue are locations strongly contribute to the overall positive spatial autocorrelation.

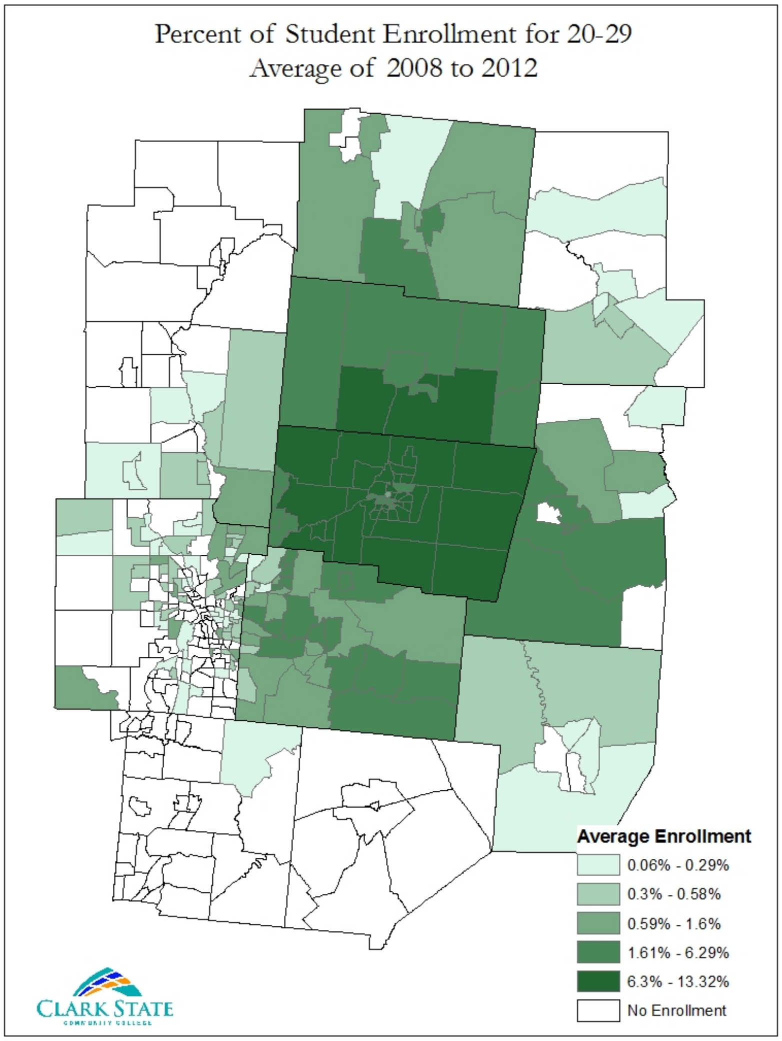


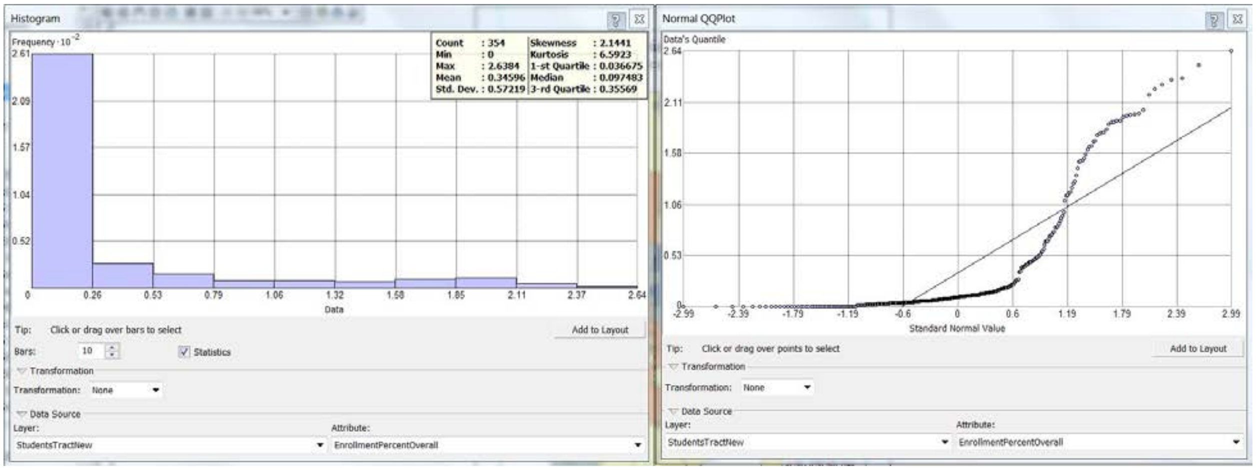
Figure 9: Average enrollment for 20-29 year olds from 2008 to 2012.

Combined dataset

The next logical step is to spatially join the student dataset to the census tracts and examine several relationships. In order to create tract-level summary statistics for each year of data, I used an excel plugin called DDXL. By utilizing this add-on, one is quickly able to do “contingency” tables, where you get a summary per tract for each category of interest (age group, gender and major description). Once the tables were created they were then imported and joined to the combined feature class for investigation in ArcGIS.

One of the first things of interest to visualize is the percent enrollment. This variable is easy to determine, it is the merely the count of students by census tract divided by the total population of the tract. This can be further refined by isolating various age groups (Figure 9).

Just as the census dataset was examined for distribution so was the enrollment percent variable. Figure 10 is the distribution of the enrollment data. This variable is skewed to the right (positive skew) and is not normally distributed. Another way of interpreting this dataset is to say that the majority of the census tracts contain low enrollment. Data should be normally distributed for linear regression; this distribution indicates that this variable may need to be transformed.



```
Count     : 354      Skewness       : 2.1441
Min       : 0        Kurtosis       : 6.5923
Max       : 2.6384   1-st Quartile  : 0.036675
Mean      : 0.34596  Median         : 0.097483
Std. Dev. : 0.57219  3-rd Quartile  : 0.35569
```

Tip: Click or drag over bars to select

Bars: 10 ☑ Statistics

Transformation

Transformation: None

Data Source

Layer: StudentsTractNew

Attribute: EnrollmentPercentOverall

Tip: Click or drag over points to select

Transformation

Transformation: None

Data Source

Layer: StudentsTractNew

Attribute: EnrollmentPercentOverall

Figure 10: Distribution of the Enrollment Percent dataset.

Analysis in SPSS

Correlation

Correlation analysis tests the strength of the relationship between two variables. This is important as I will need to reduce my variables for the upcoming regression analysis. Variables that are highly correlated will introduce multicollinearity in to my regression-based models. The correlation coefficient is a number between -1.0 (strong negative linear relationship) and +1.0 (strong positive linear relationship). The closer the value is to 0 indicates that there is a very weak relationship between the 2 variables.

In general, if two variables are correlated, it is possible to make a prediction, with better than chance accuracy, of the standing in one of the variables from knowledge of the standing in the other. The closer the relationship is between the two variables, the higher the correlation coefficient, and the better the prediction (Shaffer, 2010).

Table 7: Correlation Matrix. Cells in orange are significant to 95% and cells in red are significant to 99%.

/	Enrollment	Employment	9th to 12th (no diploma)	High School Graduate	Some College	Median Earnings	Closest Distance	White Percent	Average Distance
Enrollment	1								
Employment	-0.11	1							
9th to 12th (no diploma)	0.12	-0.59	1						
High School Graduate	0.15	-0.11	0.43	1					
Some College	-0.01	-0.06	-0.09	-0.19	1				
Median Earnings	-0.15	0.51	-0.70	-0.53	-0.21	1			
Closest Distance	-0.41	0.29	-0.05	0.22	-0.32	0.20	1		
White Percent	0.00	0.52	-0.34	0.14	-0.36	0.37	0.25	1	
Average Distance	-0.40	0.27	-0.03	0.32	-0.37	0.18	0.81	0.25	1

It appears that as the number of high school graduates (and those without high school degrees) increase so does enrollment. Conversely as enrollment increases distance decreases (the closer to the campus the more enrollment there is) as well as median earnings and employment. All of these associations are not surprising but we now have limited statistical evidence that relationships do exist. The more significant relationships are those of distance and number of high school graduates (red cells in Table 7). From looking at the results, you can see that none of the relationships are highly correlated with the exception of average and closest distance. In my models I do not include both measure of distance, I only use average distance.

Difference of Means/ANOVA and Tukey HSD

In order to gain insight into the combined dataset of census tracts and student count, the census track data was split into 5 different quintiles based on the percent of student enrollment within each tract.

That is to say that the data was rank ordered and split into 5 equally different categories so that each quintile represents an equal number of tracts. By doing this, I can now investigate how predictor (independent) variables varied by (dependent) enrolment quintiles. Group number 1 is the lowest enrollment percent and increases until you get to group 5 with the highest percent of enrollment.

A difference of means analysis (SPSS) was used to depict the variation in the means of a variable of interest by each quintile grouping of data. The results are presented below in Table 8. In most variables the means seem to increase (or decrease) within the quintiles but group 4 does not follow that trend off the grouping (meaning it sharply increases or decreases from the trend of the means preceding it). An example of this would be looking at median earnings, the lowest quintile (group with the lowest percent of enrollment) make the most with an amount of $36,558 and slowly decreases to $31,819 in quintile three. But the amount sharply increases to $36,344 in the fourth quintile and then drops down to the lowest amount in quintile 5 with an amount of $30,377 which is the group with the highest enrollment. The dataset was split into quartiles and into thirds to see if this anomaly could be eliminated but this anomaly persistent and analysis was done on quintiles.

Analysis of Variance (ANOVA) is used to analyze data further. One-way ANOVA was run in order to determine if differences among the 5 quintile groups existed based on enrollment percent. The results revealed statistically significant differences among the quintiles for several variables as listed inTable 8**.**. Any variable with 1 asterisk was found to be significant at the 95% level, variables with 2 asterisks were found to be significant at the 99% level.

If the variable was identified as statistically significant in ANOVA, there needs to be a post-hoc test (Tukey HSD) to determine *where the statistically significant differences exist.* This is an analysis that tests for the statistical significance differences between the individual groups of means. In other words, after ANOVA is conducted, one must determine which groups differ significantly. Post-hoc tests allow you to determine where significant differences lie. This post-hoc test revealed statistically significant differences between certain of the quintile grouping. These have been highlighted inTable 8, cells highlighted in purple show significance between quartiles at the 95% level and orange cells are significant at the 99% level. All other cells represented no other significant differences between the other groups.

Table 8: Difference of Means, ANOVA and Tukey HSD results.

	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
	Mean	Mean	Mean	Mean	Mean
Employment Rate	58.79	56.95	56.24	57.36	55.25
Unemployment Rate	8.89	11.19	11.19	9.27	10.34
Population Over 25 with Less than 9th grade	2.94	3.44	3.79	2.88	4.18
Population Over 25 with 9-12 (no diploma)	8.78	9.99	9.60	7.74	10.62
Population Over 25 that are High School Graduates	34.34	34.30	33.73	33.60	37.91
Population Over 25 with Some College**	20.46	24.40	24.89	22.33	22.31
Population Over 25 with an Associate's Degree	8.02	8.01	8.23	8.88	8.05
Population Over 25 with a Bachelor's Degree**	15.41	12.43	12.52	14.28	10.73
Population Over 25 with a Professional/Graduate Degree*	10.04	7.47	7.26	10.27	6.21
Population Over 18 Enrolled in College*	7.30	8.46	8.86	12.73	8.74
No Vehicle Available Rate	7.74	8.19	9.48	6.14	8.25
One Vehicle Available Rate**	29.22	35.20	36.61	31.49	34.67
Two Vehicles Available*	40.07	36.43	36.53	39.34	34.95
Three Or More Vehicles Available*	22.96	20.16	17.37	23.03	22.13
Median Earnings Population Over 25**	$36,558.22	$32,540.85	$31,819.39	$36,344.41	$30,377.84
Population 18-64 that are Below Poverty Level**	12.06	14.64	15.98	12.11	17.37
Average Tract Distance (miles)**	25.35	21.55	18.99	18.39	12.83
Percent White**	89.11	77.29	71.67	81.25	85.25
Percent Enrollment**	.0326	.0674	.1197	.3233	1.464
* Significant at 95%					
** Significant at 99%					
Difference between quartiles 95%					
Difference between quartiles at 99%					

Regression

The purpose of linear regression analysis is to find a relationship between a dependent variable (percent enrollment) and a set of explanatory (independent) variables. With other statistical analysis such as LISA, we are only able to answer the question of where something is happening (e.g., where is there a cluster of high school graduation rates, etc.). However, with regression analysis you are able to examine how potential predictors maybe related to the outcome of interest (enrollment), controlling for other predictors and control variables. Regression analysis allows you to model and examine associations; specifically it will be used to test my hypotheses about predictors of CSCC enrollment.

While I report the OLS models note that both models described below use the first order queen connectivity spatial weights file. The spatial weights file facilitates the calculation of spatial diagnostics to test assumptions of a non-spatial OLS model. In my work, as tracts are not independent observations (and do not satisfy the i.i.d assumptions) we might also anticipate the need for a spatial regression model.

OLS Spatial Regression

Ordinary Least Squares (OLS) is the most commonly used regression model for non-spatial data. It is the usual starting point for most models as it provides a *global* model of the processes that are being analyzed (ESRI, 2013). Model performance can be assessed using various statistical measures. One such statistic is the correlation coefficient squared $(R)^2$. This is a measure of the proportion of variation in the data that is explained by the model. The adjusted R^2 takes into account the number of variables that are specified. The larger the R^2value the more variability is explained by the model (MathWorks, n.d.). It can range from 0 (the independent variables are not related to the dependent variable) to 1 (the independent variables explain all variation in the dependent variable).

Although previous studies and ESDA helped identify variables of interest, much time was still spent trying to select relevant variables to correctly specify a model predicting enrollment. Original variables included all variables identified as significant during Difference of Means, ANOVA and Tukey. The variables were paired down in relation to their significance calculated during the regression. (It was surprising to see that some variables did not translate such as “Population over 18 enrolled in College”). The lack of data on which campus the student attends caused some problems, so in addition to running the model with closest tract centroid distance, the regression was also run for average distance (to Greene and Main campuses) to see if that improved the results. In the end, I decided upon using the average distance as it seemed to be the best fit. Also, traditionally community college students attend institutions that are closer to home so this is the logical choice.

The adjusted R-squared value for my model is 0.27, which is low (Table 9). This statistic essentially indicates that 27% of the enrollment is explained by my 5 variables. The F-statistic measures the ratio of the variation among the sample means to the variation within the samples.

Table 9: OLS Regression Analysis Results

Variable	T-Statistic	Significance
Constant	1.263	0.205
Average Distance	-11.07	0.000
High School Graduate	5.90	0.000
Some College	-2.46	0.014
Earnings	2.16	0.031
%White	-0.16	0.874
Adjusted R-Square	0.27	
F-Test	27.46	
Log Likelihood	-245.31	
AIC	502.83	
Multicollinearity condition number	28.01	
Moran's I	0.79	0.000
LM (lag)	654.60	0.000
LM (error)	591.09	0.000
Robust LM (lag)	64.19	0.000
Robust LM (error)	0.68	0.408

The individual regression coefficients tell you the direction (+ or-) and the strength of the relationship between X (independent variables) and Y (dependent variable). They report the change in x for a unit change in Y (enrollment). From Table 9, one can see that average distance, some college and % white are negatively affected by enrollment meaning that as enrollment increases, these variables decrease. On the other hand, as the percent of high school graduates and earnings increase, so does enrollment. The p-values (significance) of each variable tell if the results are statistically significant. T-statistics below -1.96 or above 1.96 indicate that they are influencing the model. Percent white does not fall within that tolerance but sometimes it is important to keep some variables that may not be significant for theoretical reasons regardless of whether or not the variable is significant.

The model also gives measures of comparability to compare between models. For all information criteria (AIC, or Schwarz criterion), the smaller they are the better the fit of your model is, these are aspatial diagnostics.

The next section deals with the regression diagnostics. A multicollinearity condition number over 30 indicates if 2 variables are highly correlated, which presents difficulties if continuing with the analysis. A condition number of 28 indicate that multicollinearity is not too serious an issue.

Another diagnostic is the Moran's I of the residuals, which should be random and not clustered. The standard deviational map for the residuals is useful to illustrate areas of over- and under-prediction, as well as the magnitude of the residuals. Residuals are the unexplained portion of the dependent variable. If the residuals appear to show little or no spatial pattern it supports the view that the fitted model provides a good representation of the observed spatial patterns. Figure 11 is a map of the residuals that clearly indicates residuals are clustered and that the OLS model is not a correctly specified model.

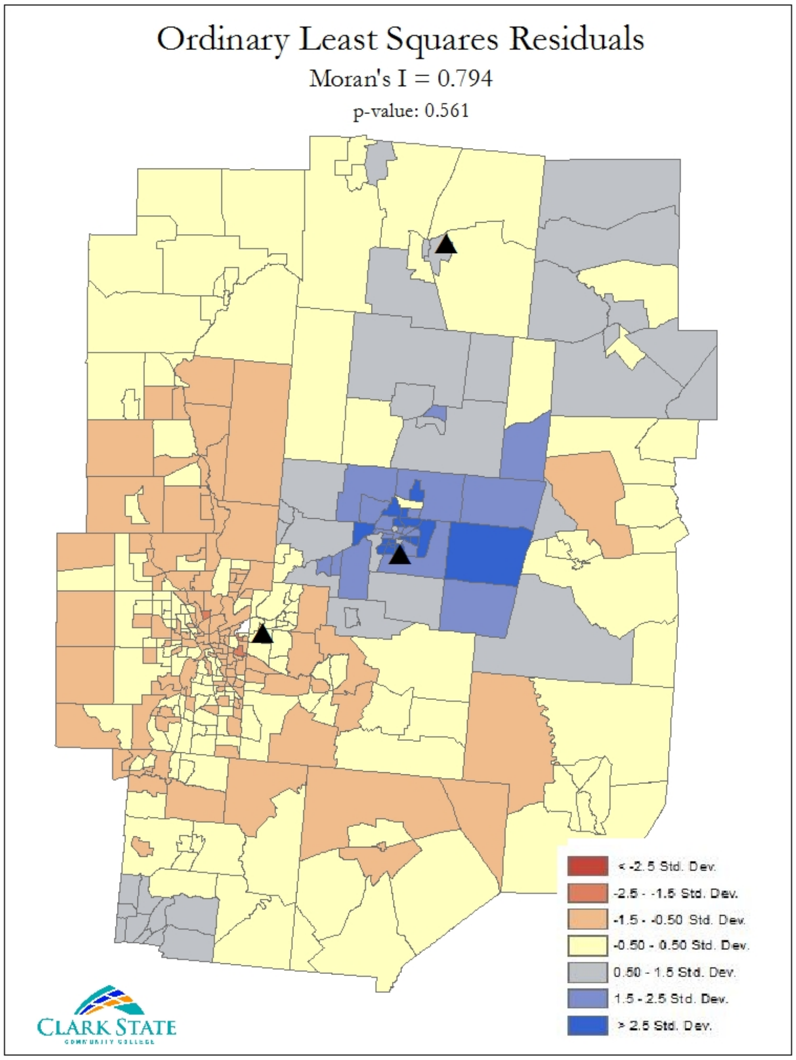


Figure 11: OLS Residuals

It is also worth mentioning that OLS models assume that the data is composed of "independent" observations. This is almost never the case with spatial data (such as census tracts). It is highly probable that some type of dependence or interaction between neighboring units is occurring such as is referenced in Tobler law of geography.

The Lagrange multiplier (LM) tests are calculated for the 'effectiveness' of spatial regression model along with their robust forms. If the LM lag and LM error are significant then one must compare the robust LM and pick the higher of the two robust scores (Anselin, 2005). From the results listed in Table 9, confirm that I should run a spatial lag model.

Spatial Lag Model

The standard OLS model assumes that the residuals are uncorrelated. Based on the diagnostic tests above I ran a spatial lag regression model. The spatially lagged model is effective for data in which "neighboring" values of the dependent variable affect one another which is also known as spatial dependence (Ward and Gleditsch, 2007).

Table 10: Spatial Lag Model Results

Variable	Z-Value	Significance
Constant	-1.43	0.154
Average Distance	-3.70	0.001
High School Graduate	3.292	0.001
Some College	0.642	0.521
Earnings	2.27	0.023
%White	-0.697	0.486
W_Enrollment	53.69	0.000
Adjusted R-Square	0.922	
Log Likelihood	98.39	
AIC	-182.79	

Comparing the measures of the overall fit for the model between Table 9 and Table 10, the model with the spatially lagged y indicates that spatially lagged model fits the data better. The spatial effects have been corrected for and the resultant coefficients are unbiased. We can observe that the effect of average distance coefficient has been greatly reduced (dropping from -11.06 in the OLS model to -3.70 in the spatially lagged model). We also observe an increase in magnitude for the effect of the earnings and percent white; though the latter is not significant. We see a reduction in magnitude for the High School Graduate variable but this remains highly significant. In addition, the "Some College" variable is no longer significant in the spatial regression model as it switched sign from a negative association to one that is positive. When we compare the spatial lag model with the OLS model using a likelihood ratio test, we see a significant improvement in fit of the spatial lag model over the OLS model (-245.03 for OLS to 98.39 for the spatial lag). In sum, the weighted spatial lag model shows some notable differences in coefficients compared with OLS model. The Akaike Information Criteria (AIC) values of the two models indicate a slightly better model fit for the spatial lag model (AIC = -182.79) than the OLS model (AIC = 502.83.05). However, it is important to note that the lag of enrollment is overwhelming the model as the z-value is extremely high and statistically significant. The model indicates that not all spatial dependence has been eliminated and that there may be some variables that have not been accounted for.

Transforming the Data

In my preliminary analysis I noted that my outcome variable of enrollment not normally distributed (Figure 10), I tried to transform the variables to improve the distribution with various transformations using log, exponential, sine, and square root of the values from my variables. The best resulting model came when I used square root but when I ran the same model in OLS and the spatial lag model diagnostic tests indicated that my model included severe multicollinearity and was not a good fit. I return to model specification issues later in my discussion.

Results and Conclusions

The main purpose of this project was to give the college a better understanding of student enrollment patterns based on the physical address of the students. This has already been reviewed in Figure 3, Figure 4 and Figure 7. One of the most interesting analyses was to look at the “market penetration” (count of students/divided by total amount of the same population count). We have previously discovered that the mean and median age for CSCC is in the twenties. Figure 9 depicts the average enrollment from our investigation time frame for people aged 20-29 normalized by the total count of 20-29 year olds (SF1 information).

For the overall college enrollment rate (Figure 12), one can see that the highest areas are located around the Main Campus in Springfield. This result is not surprising, the campus has been established for several years and there is no other competition in this area with the exception of Wittenberg University (a private 4 year school) that is expensive to attend. However, do not be misdirected by this information, the highest enrollment rate is still low in the highest areas with a maximum value of 2.64%.

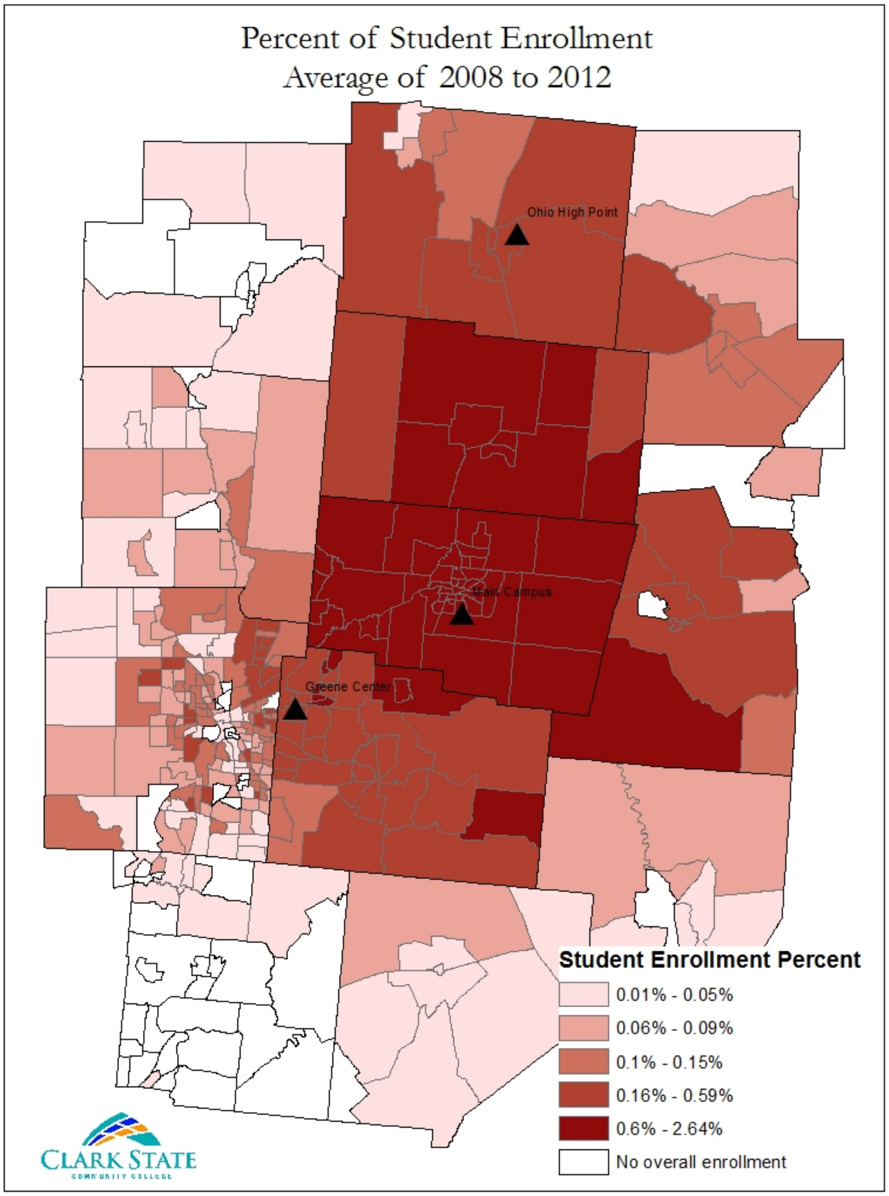


Figure 12: Overall enrollment average for 2008-2012

It should also be pointed out is that the white areas on the map are places where average enrolment could not be calculated. There may have been students enrolled in some isolated years but a clean average could not be calculated. These maps closely resemble the student density maps but provide another way of visualizing the data.

Another way of visualizing the data is examined in Figure 13. The areas in blue represent census tracts that have lost students and the areas in red represent census tracts that have seen an increase in student enrollment during the 5 year study period.

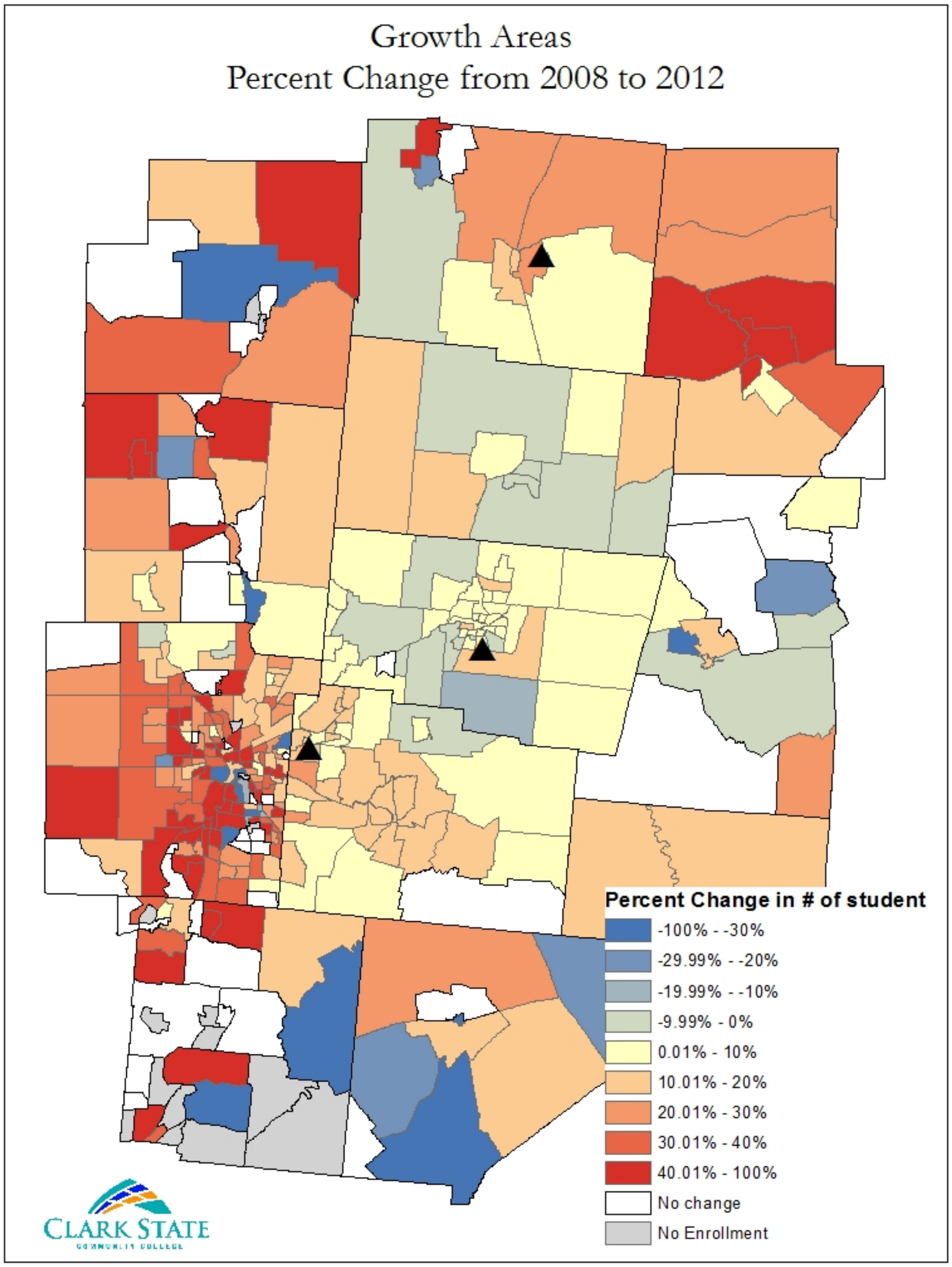


Figure 13: Growth Areas. This map depicts the areas that have increased (red) and decreased (blue) in enrollment between Fall 2008 to 2012

Staff and faculty were not surprised by to see that areas in Greene and Montgomery Counties have experienced the most growth (Figure 14). It is widely known that the numbers at the Greene Center have been increasing every semester to help offset the loss from the Springfield campus. One should remember that this campus opened in 2007 and is surrounded by numerous higher educational institutions. Even with those handicaps, enrollment is doing very well.

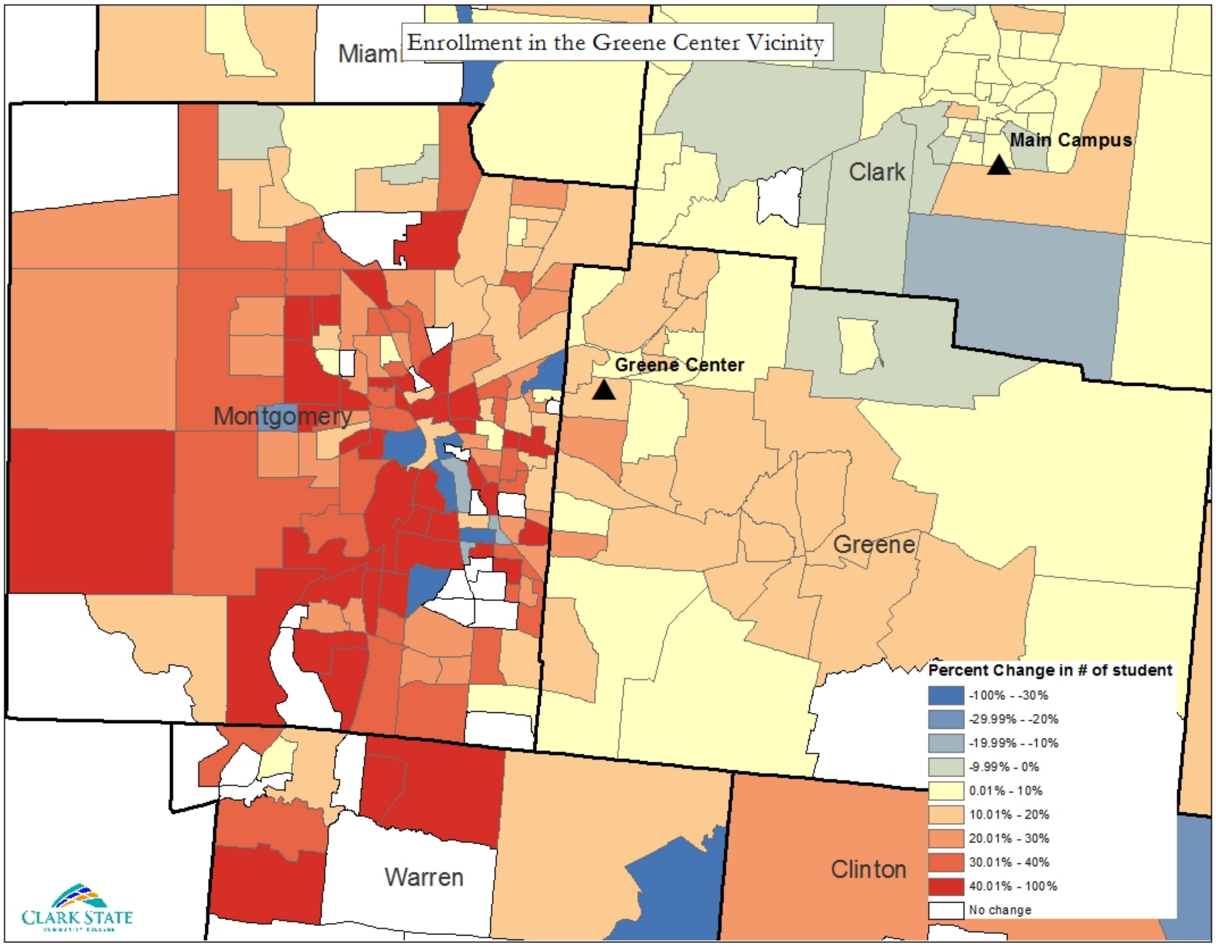


Figure 14: Percent Enrollment Change in the Greene Center Vicinity. This map depicts areas around the Greene Center Campus that have increased (red) and decreased (blue) in enrollment between Fall 2008 to 2012.

I also want to identify census tracts that could be targeted to increase enrollment. Even though I was unable to create a properly specified and successful regression model, I can still use the output to identify which variables statistically influence enrollment. From looking at the results from the various ESDA techniques, we can argue that “Average Distance”, “High School Graduates” and “Median Earnings” are statistically significant and appear to have a limited association with enrollment. When you combine this information with the mean values for the highest quintile (presented inTable 8) you can then do a simple ArcGIS query to find these areas. The resulting query selects census tracts that are not located in the highest quintile and that have values at or above the mean (from the highest quintile) for our 3 variables (Figure 15).

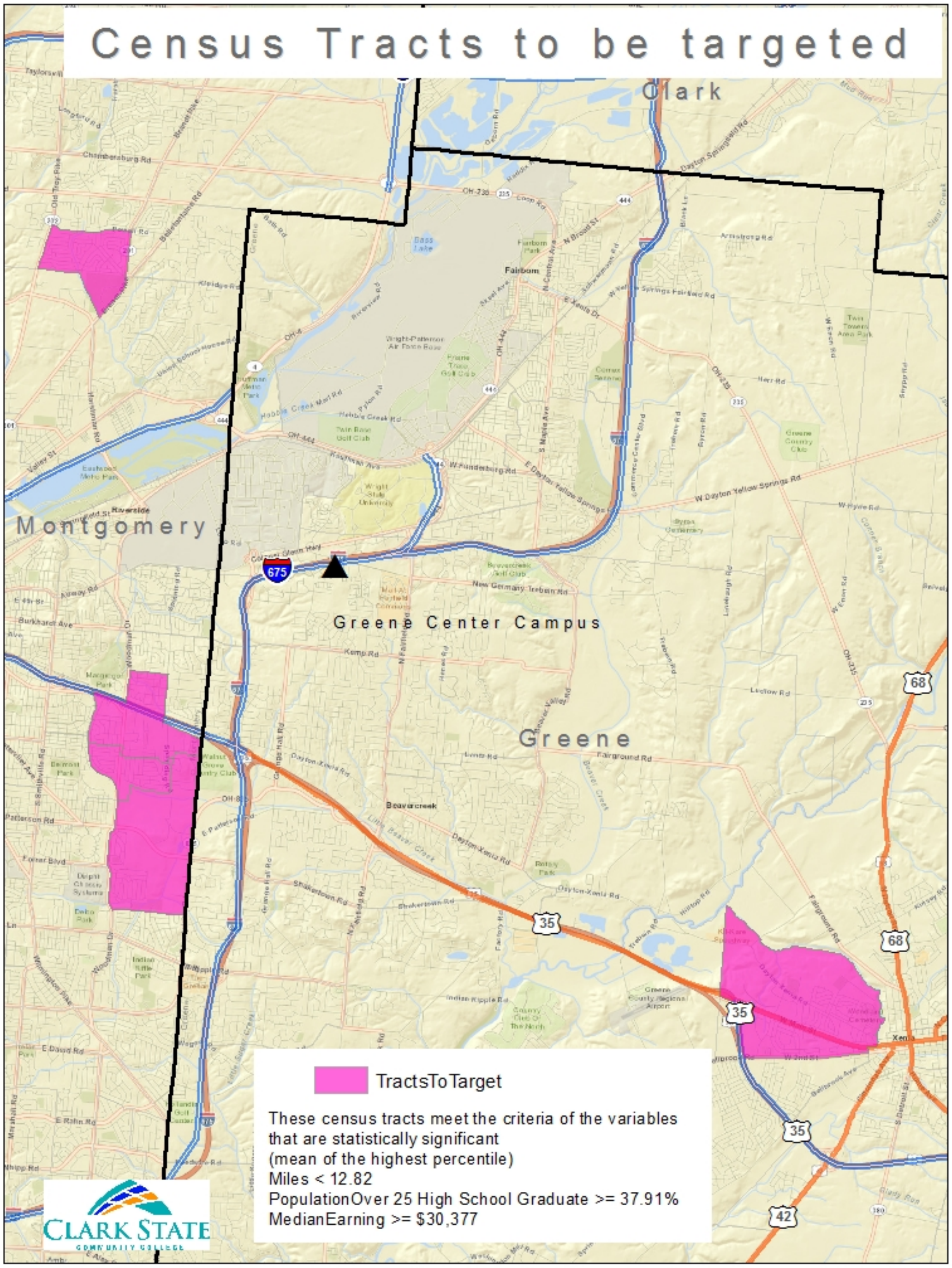

Census Tracts to be targeted
Clark
Montgomery
Greene Center Campus
Greene
Fairborn
Beavercreek
Xenia
675
35
68
42
TractsToTarget
These census tracts meet the criteria of the variables
that are statistically significant
(mean of the highest percentile)
Miles < 12.82
PopulationOver 25 High School Graduate >= 37.91%
MedianEarning >= $30,377
CLARK STATE
COMMUNITY COLLEGE

Figure 15: Census Tracts that CSCC should target.

Within these 5 census tracts there are 15,500 people of which there are 1,881 or 12% of the population that have no college degree.

As previously discussed in the ESDA for the census variables, there seems to be evidence of a clustering people who have graduated high school; based on the educational attainment variables (Figure 10). The higher percentage of high school graduates is clustered in both the north and southeast corner of the study area. This leaves an area south of the City of Dayton with relatively low concentrations. This clustering of people is important as it represents areas that CSCC could potentially recruit within. More specifically it may be of interest to possibly purchase some advertising via billboard or targeted mailing as there is a cluster of people that have not attained a higher level of education than high school.

Lessons Learned and Limitations

CSCC

I encountered several limitations as I worked on this project. The most pronounced issue is the fact that CSCC does not maintain information on the home campus for each student. Not knowing this information affected almost every facet on the study and I had to rely on an average distance between the Main Campus and the Greene Center Campuses.

CSCC indicated that they would not be able to give me access to all of the variables I was interested in. They would not be able to provide information on whether or not students had ever taken any remedial classes (CPE) nor was grade point average available for the study. This could have added additional insight into the project.

The marketing department was interested in determining the “market penetration” and this was easily accomplished. However, there was no internal marketing information that could be used to evaluate against the enrollment data so that CSCC could determine if previous marketing attempts were worthwhile.

IRB

I had been forewarned about needing to acquire IRB approval prior to getting the student dataset. But what I was not prepared for was the amount of time and effort required for garnering approval. Both institutions also originally missed the exemption under FERPA that was needed to get access to the student dataset. It was only at the last minute that this was discovered and delayed the data delivery until approval was once again obtained. I was under the impression that once approval was granted on the original petition back in April that I would quickly acquire the student dataset, but that was not the case due to the FERPA exception.

Methodologies

After looking at my data statistically and using ESDA, the variable of interest “percent enrollment” was not normally distributed (Figure 10). A normal distribution of a continuously measured variable is usually expected for traditional regression models (whether run in a statistical package or within ArcGIS). Other possible studies are described in the “Next Steps” section below.

Also, on the ecologic side of the analysis there are limitations that need to be addressed. First, ACS data are only estimates of the population and the margin of error (MOE) of smaller units can be problematic. Most of this was avoided by using generalized categories and not age specific variables. Secondly the assumption was made that census tracts was the relevant ecologic unit for my analysis. This is standard practice but does not preclude the possibility that MAUP issues may exist, and that other geographical units may be more relevant to the study of enrollment at CSCC.

Next Steps

There are several next steps that could be pursued. Probably the most practical for the college would be an analysis based on the rich CSCC enrollment data; specifically focusing on an examination of variations in enrolment by gender, ethnicity, and major. In the literature review there were several interesting studies based on cluster analysis that could be applied. Adnan et al. (2010) examined different clustering techniques for classifying geodemographic data near real time via the internet: k-means, clustering large applications (CARA) and genetic algorithms (GA) for various benchmarks. Although the clustering techniques were applied to the polygons and not to point data, this article did give some insight into the advantages and disadvantages of each clustering technique and could be applied to point data.

Another interesting avenue worth pursuing would be to obtain listings of students that have just graduated high school and enter the college needing CPE classes. This would help both CSCC and the local school districts identify areas of improvement. In a similar study, being able to identify areas that produce successful students by analyzing their grade point average could also help shed some light on the makeup of the school districts. A possible outcome could be to utilize high school students from the successful areas to assist and tutor students from the weaker areas.

In an ideal world additional analysis of student enrolment based on a student's "home" campus would be useful. However before that study can take place CSCC will need to maintain that information. If there is one recommendation I might make to CSCC it would be to add this field to their student data set.

Other studies could incorporate information on other sources of education in the study area (i.e., the competition). This would be especially interesting for the Greene Center campus as it is surrounded by several higher educational institutions. In addition, Greene County is the only county in the state that has the presence of 2 different state schools.

Additional regression-based approaches could also be considered. Traditional regression models are often (but not always) applied to continuously measured outcome variables. In my case, the outcome measure was not normally distributed – THOUGH IT WAS OK. I tried several different variable transformations but when running regression models on these specifications alternative methodological issues would arise (e.g., multicollinearity). Other specifications that I did not try include, converting the dependent variable to a categorical variable and estimating models using logistic regression type methods. Similarly, it was beyond the scope of this study, but in the future I might explore the use of modeling approaches that are designed for nonlinear distributions. A commonly used approach in statistics includes the use of Poisson regression methods. As my data are inherently spatial, the obvious twist is the need to examine spatial Poisson regression methods. Another avenue might also to explore the use of Bayesian statistical methods. These latter methods may be more appropriate but are beyond the capabilities of ArcGIS and require knowledge of methods and statistical softer packages such as SPSS, R or SAS.

Works Cited

Anselin, L. (2005, 03 06). *GeoDa Workbook*. Retrieved 07 01, 2013, from GeoDa: https://geodacenter.asu.edu/system/files/geodaworkbook.pdf

Batey, P. (1999). Participation in higher education: A geodemocratic perspective on the potential for further expansion in student numbers. *Journal of Geographical Systems*, 277-303.

Census. (2013, 07 02). *American Community Survey*. Retrieved 07 20, 2013, from Census.gov: http://www.census.gov/acs/www/

Clemson University. (2012). *Clemson.edu.* Retrieved 05 29, 2013, from Adminstration/Documents: http://www.clemson.edu/administration/ogc/documents/FERPA.pdf

de Smith, G. L. (n.d.). *Geospatial Analysis 4th Edition.*

ESRI . (2010). How GWR work. Redlands, CA.

ESRI. (2012). Geostatistical Analyst Tutorial. Redlands, CA.

ESRI. (2013, 06 28). ArcGIS 10 Help. Redlands, CA.

ESRI. (n.d.). *Exploratory Regression*. Retrieved 06 30, 2013, from ESRI.com: http://www.esri.com/news/arcuser/0111/files/exploratory.pdf

GeoDa Center. (n.d.). *Glossary of terms*. Retrieved 2012 йил 16-07 from GeoDa Center: https://geodacenter.asu.edu/node/390#lisa2

Gleditsch, W. a. (2007, 06 15). An Introduction to Spatial Regression.

Gleditsch, W. a. (2007, 06 15). An Introduction to Spatial Regression.

Hanewicz, D. C. (2012 йил 28-02). *Geographic Information Systems and the Political Process.* Retrieved 2012 йил 17-10 from wpsa.research.pdx: http://wpsa.research.pdx.edu/meet/2012/hanewicz.pdf

Health and Human Services. (n.d.). *HHS.gov*. Retrieved 02 01, 2013, from Human Subject Regulations Decision Charts: http://www.hhs.gov/ohrp/policy/checklists/decisioncharts.html#c3

Krestle, J. (2004). Geodemographic target clusters: A case study. *Monday Report on Retailers*, 2-4.

Krivoruchko, K. (2011). *Spatial Statistical Data Analysis for GIS Users.* Redlands: ESRI Press.

Livinsgton, A. (2000, 07 16). Colleges search for applicants, with glitz and geodemographics. *The Associated Press*. New York.

Marble, D. (1995 йил 07). *Applying GIS Technology to the Freshman Admissions Process*. Retrieved 2012 йил 01-10 from ESRI User Conference Proceedings: http://proceedings.esri.com/library/userconf/proc95/to200/p182.html

Marble, D. (1997 йил 07). *A Model for the Use of GIS Technology in College and University Admissions Planning*. Retrieved 2012 йил 01-10 from ESRI User Conference Proceedings: http://proceedings.esri.com/library/userconf/proc97/proc97/to250/pap218/p218.htm

Marble, D. (2001). A Model for the Use of GIS Technology in College and University Admissions Planning. *ESRI User Conference*, (p. 14). San Diego.

MathWorks. (n.d.). *Linear Regression output diagnostic*. Retrieved 06 23, 2013, from MathWorks: http://www.mathworks.com/help/stats/linear-regression-output-and-diagnostic-statistics.html#btkvxq6-10

Mora, V. (2003). Applications of GIS in Admissions and Targeting Recruiting Efforts. *New Directions for Institutional Research*, 15-21.

O'Sullivan, U. a. (2010). *Geographic Information Analysis.* Hoboken: Wiley and Sons.

Shaffer, R. (2010). *Statistics Primer.* McGraw Hill.

Singleton, A. (April 2012). Geodemographics and spatial interaction: an intergrated model for higher education. *Journal of Geographical Systems*, Volume 14:223-241.

Social Research Methods. (2006, 10 20). *Descriptive Statistics*. Retrieved 07 18, 2013, from Knowledge Base for Social Research Methods: http://www.socialresearchmethods.net/kb/statdesc.php

Tonks, D. G. (1995). Market sements for higher education: using geodemographics. *Marketing Intelligence and Planning*, 24-33.

Troy, A. (2008). Geodemographic Segmentation. In S. Sekkar, *Encyclopedia of GIS* (pp. 347-355). Springer.

Zhou, Y. (2005). *Modeling University Enrollments with ArcGIS*. Retrieved 10 16, 2012, from ESRI User Conference Proceedings: http://downloads2.esri.com/campus/uploads/library/pdfs/58255.pdf

Appendix A: Institutional Research

This type of research also requires that you file an exemption under FERPA. When invoking an exception for the use of educational records, the holder of the records must specifically cite the exception to the regulation in writing. The exceptions that may be used for educational research are:

- If the researcher is a school official with legitimate educational interest [34 CFR 99.31(a)(1); or
- If the researcher is conducting studies for or on behalf of the school [34 CFR99.31 (a)(6).

When planning to conduct research involving educational records, the FERPA exception letter should be submitted to the Institutional Review Board (IRB) along with the IRB application (Clemson University, 2012). Unfortunately for me, this was missed in my original application and was not identified until I was to take ownership of the student dataset. This can often happen in small institutions that do not do a lot of research. Please take this under advisement.

Appendix B: Data Tips and Tricks

There are several things that must be done with your data set prior to the beginning of analysis. While this may seem to be obvious to most people who do analysis on a daily basis it was not for me so I wanted to include this information within my report.

FactFinder data

- My first inclination was to download the data with descriptive variable names. I then took quite a bit of time of formatting the columns to remove characters that are not supported within ArcGIS such as - ; ,
- In Excel you can accelerate this process by using find and replace but I was unable to use this process to remove the * used by Census data (it would delete the entire page). In order to bring the column in as a numeric field (important for analysis) then these asterisks must be removed. I ended up doing a manual search and just deleting those cells by hand
- The linking field is different between the census tracts and the Factfinder data. In order to streamline the process it is much easier to add the new variable to TIGER information than to all of your ACS datasets.

ArcGIS data

- Join the table to the tracts (much easier since you already have matching variables)
- ArcGIS did seem to reach a maximum of columns to about 255 variables. Unless you need the entire table, first remove the extra fields prior to importing your data into the geodatabase.

Overall:

Since I wanted to accomplish data exploration in GeoDa, I then had to export out all the files from the geodatabase to a shapefile format. This proved to be problematic with my long variable names as they became truncated to 8 characters and the first 7 characters were almost all the same with the last one changing incrementally.

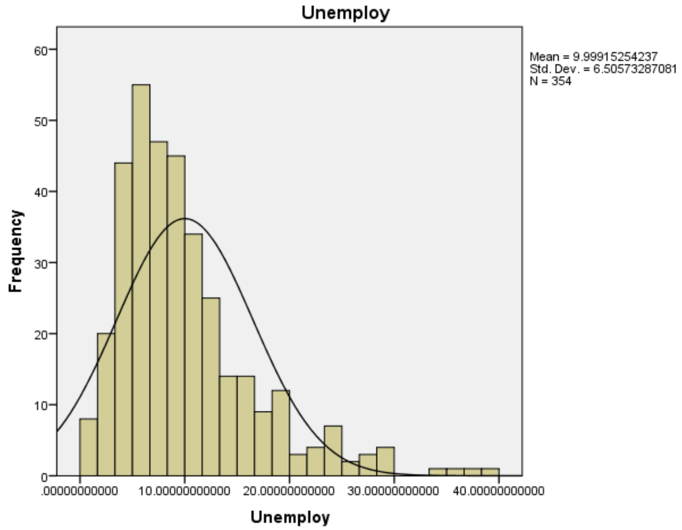
I did end up going back and redoing the data downloads but this time with the default FactFinder variable names. I also discovered that one should remove the NULL as this will cause problems later. Remember that if you have no population within the tract then that data will more than likely have a NULL field(s).

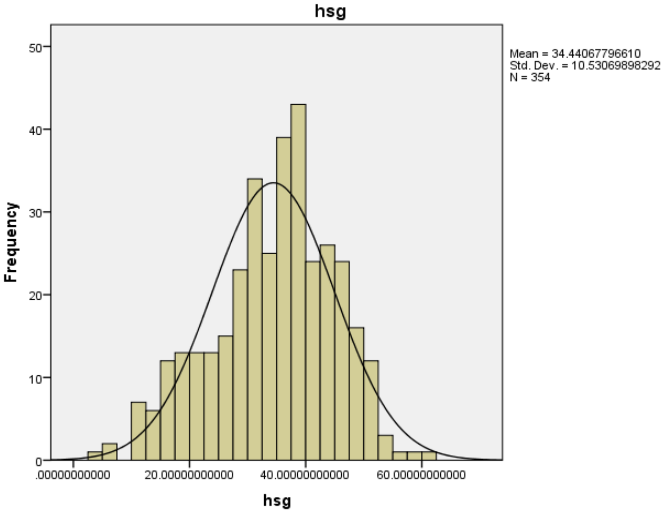
Since the NULL fields seemed to occur in different variables within the same table (School Enrollement), I thought I would create the shapefiles and remove the NULL that way. However, when converting the geodatabase feature class to a shapefile the NULL fields became 0. Therefore, this step must be done prior to converting to shapefiles.

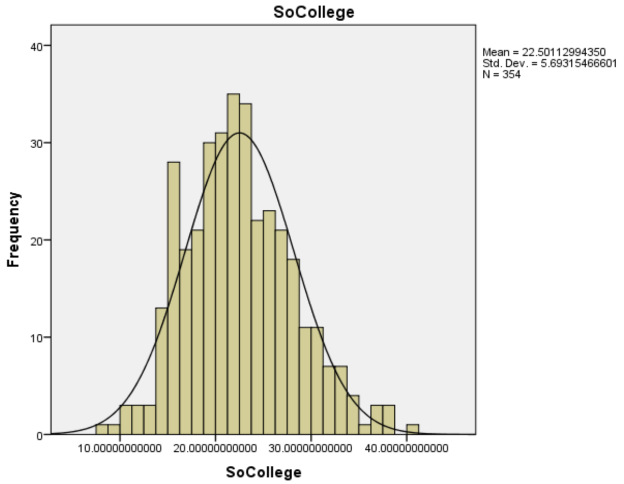
Excel

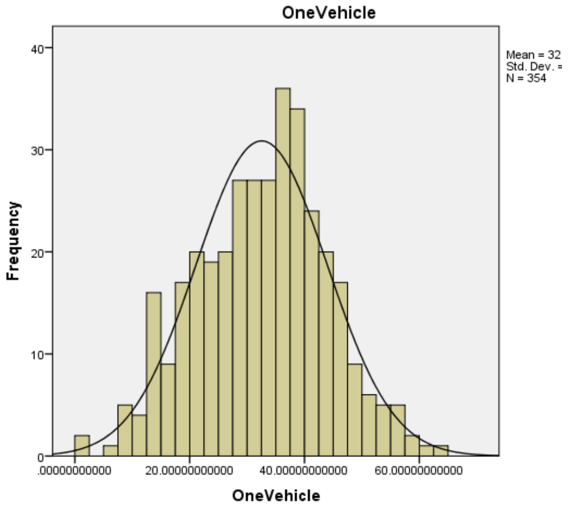
As previously mentioned I used excel to do summary statistics. It is rather straight forward but I did however hit a few snags that people should be aware of. Since the census tracts are all numerically identified, this causes problems. By taking the native number of the tracts the result come in as scientific notation and tremendously groups the data (going from 355 tracts to 17). Several different methods were tried; converting the column to text, to general all provided the same results as above. However the addition of a letter before the first number in the tract (for Ohio all tracts are 39) proved to be the easiest method via the concatenate function. I used the filter command to isolate each of my 5 years of registration data and preformed my contingency tables and copying the information back to a spreadsheet. The descriptive information from each table must be removed both at the beginning and end of the table. However the results have spaces between each line, I resolved that issue by performing a sort. Then I needed to remove the extra "Z" in my tract names which was accomplished with a find and replace and finally the column was converted to text. Also you must remember to not name a column in Excel starting with a number as ArcGIS does not like this format and will insert an underscore before each number thereby shortening the 8 characters down to 7. Next each worksheet is imported into a geodatabase which is then joined by the tractID. I discovered my next issue while doing data exploration in my age group feature classes; it appeared that the data did not change from year to year. I came to realize that when the filter was applied in Excel, the summary table still did the analysis for the entire table, hence all the results were the same and I did not catch it until everything had been imported. This can be easily avoided by doing the filter for each year and copying that information to a new workbook. This way all data is organized by year and each of the 3 worksheet contains the necessary information.

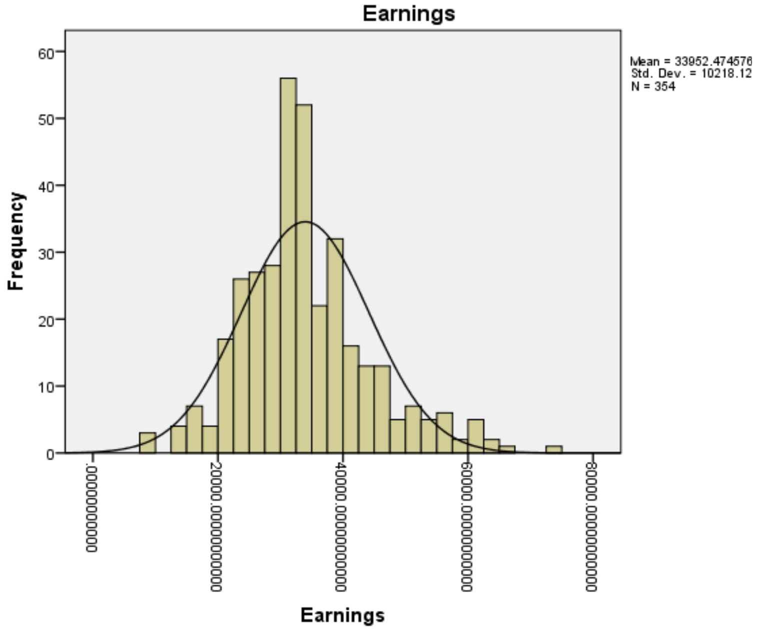
Appendix C: Selected Histograms from Census Tracts.

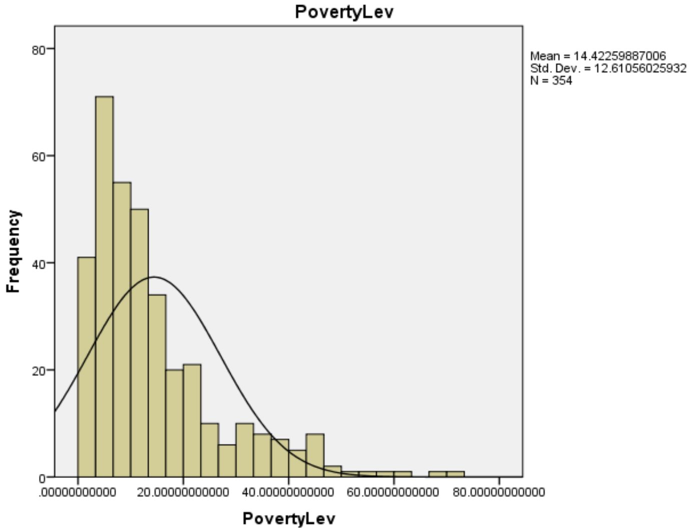




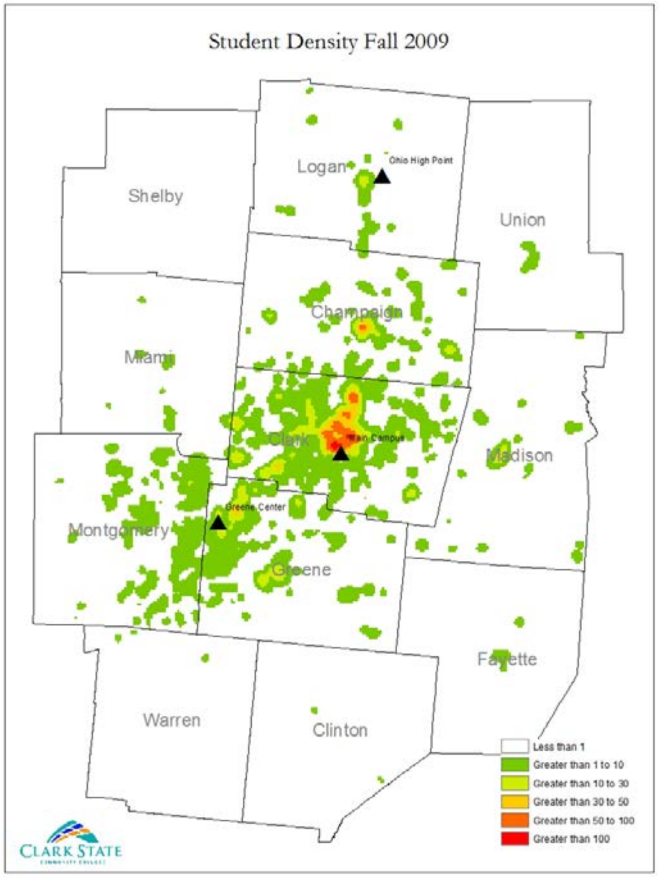


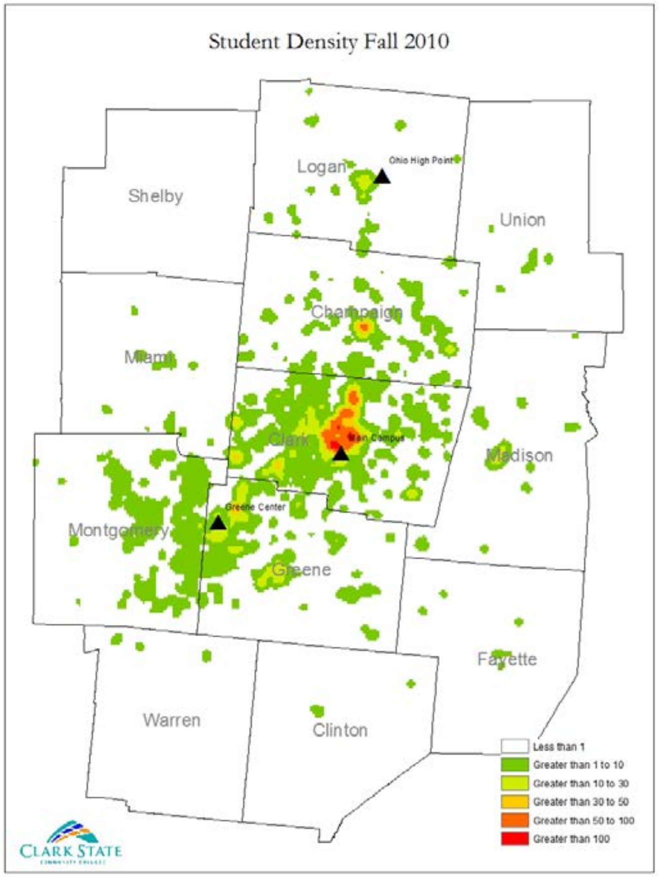


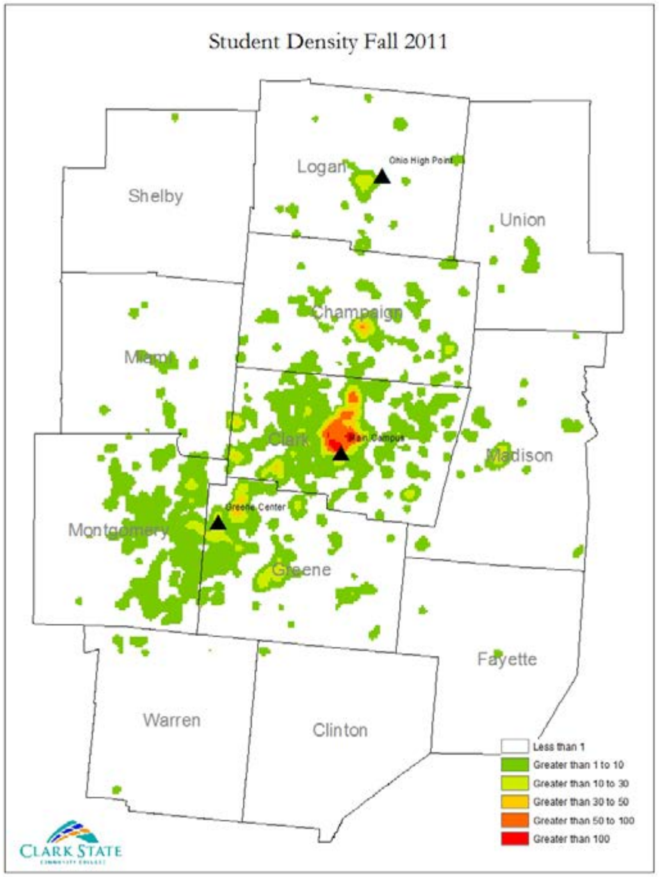


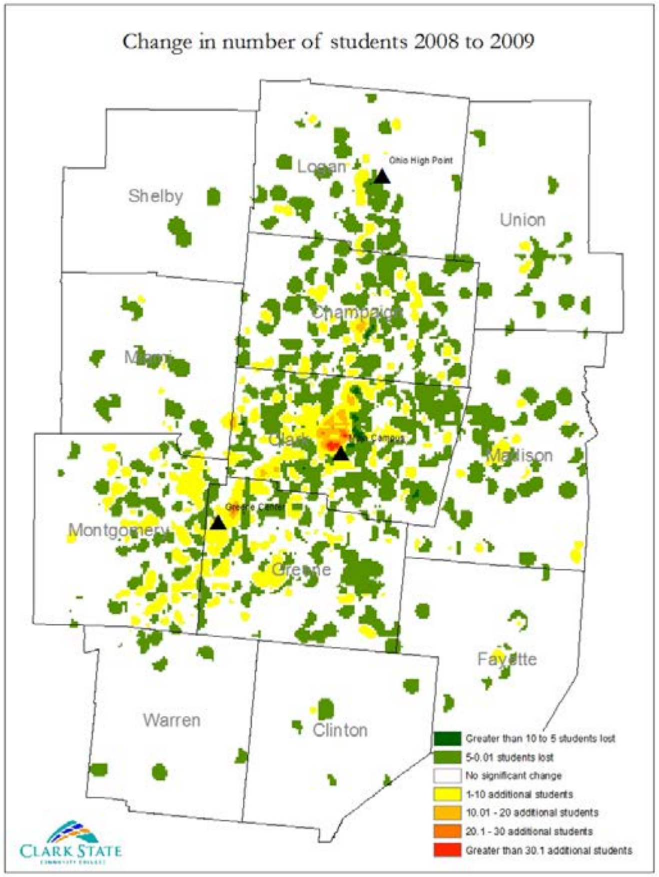


Appendix D: Additional Maps

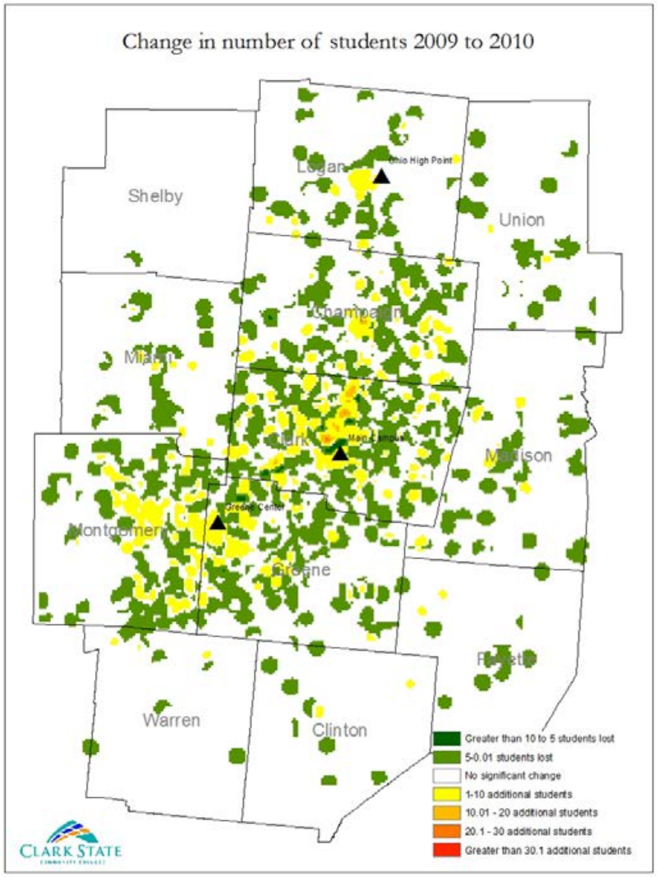

Student Density Fall 2009


Student Density Fall 2010

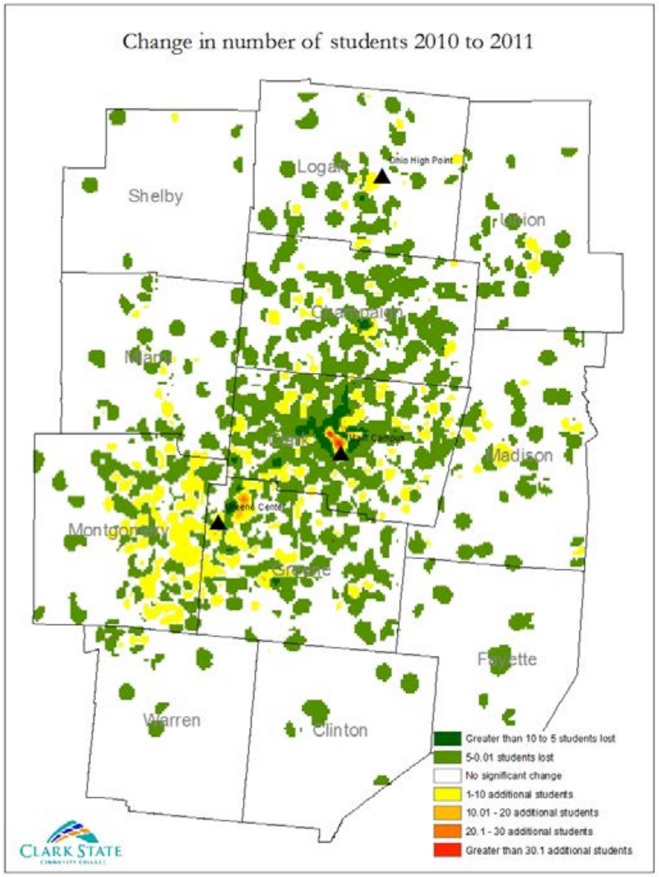

Student Density Fall 2011



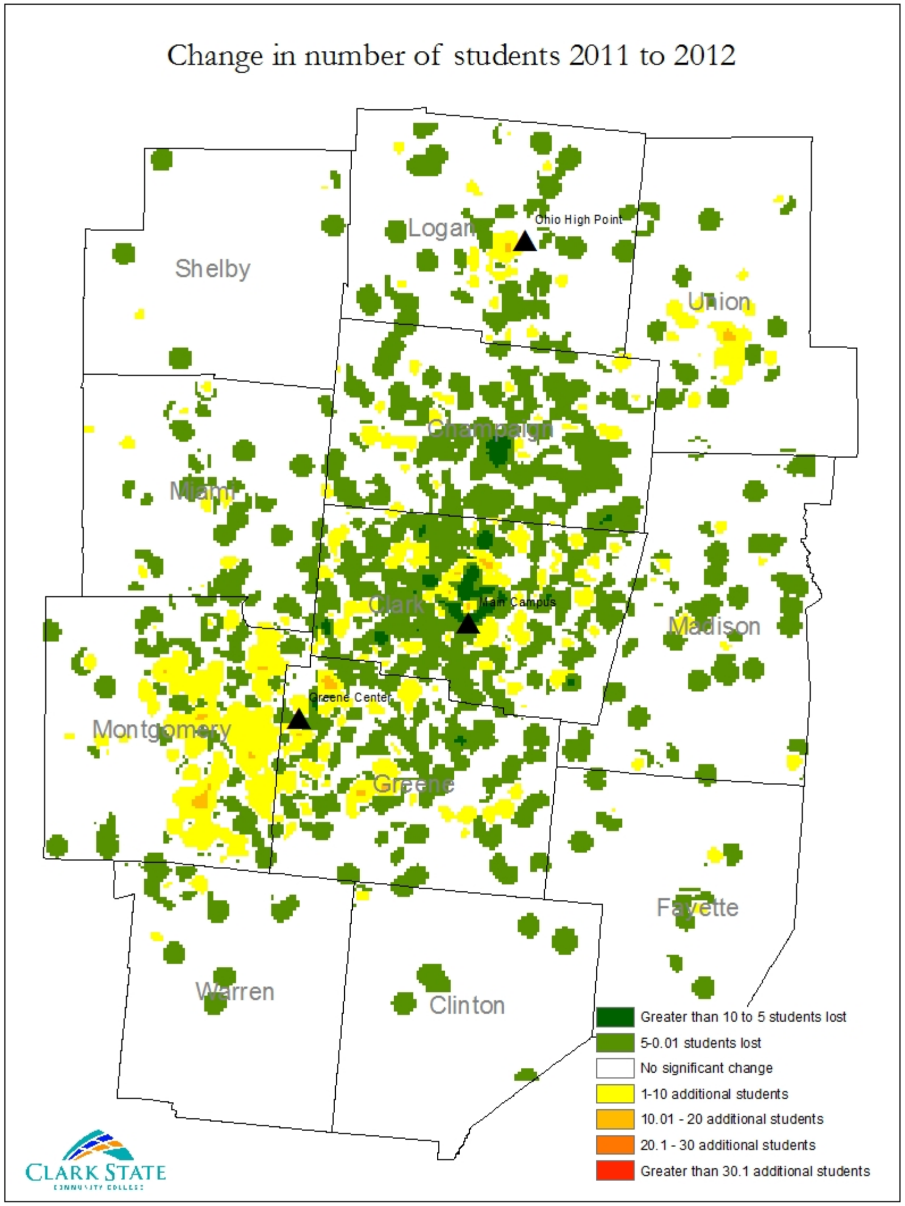
Change in number of students 2008 to 2009

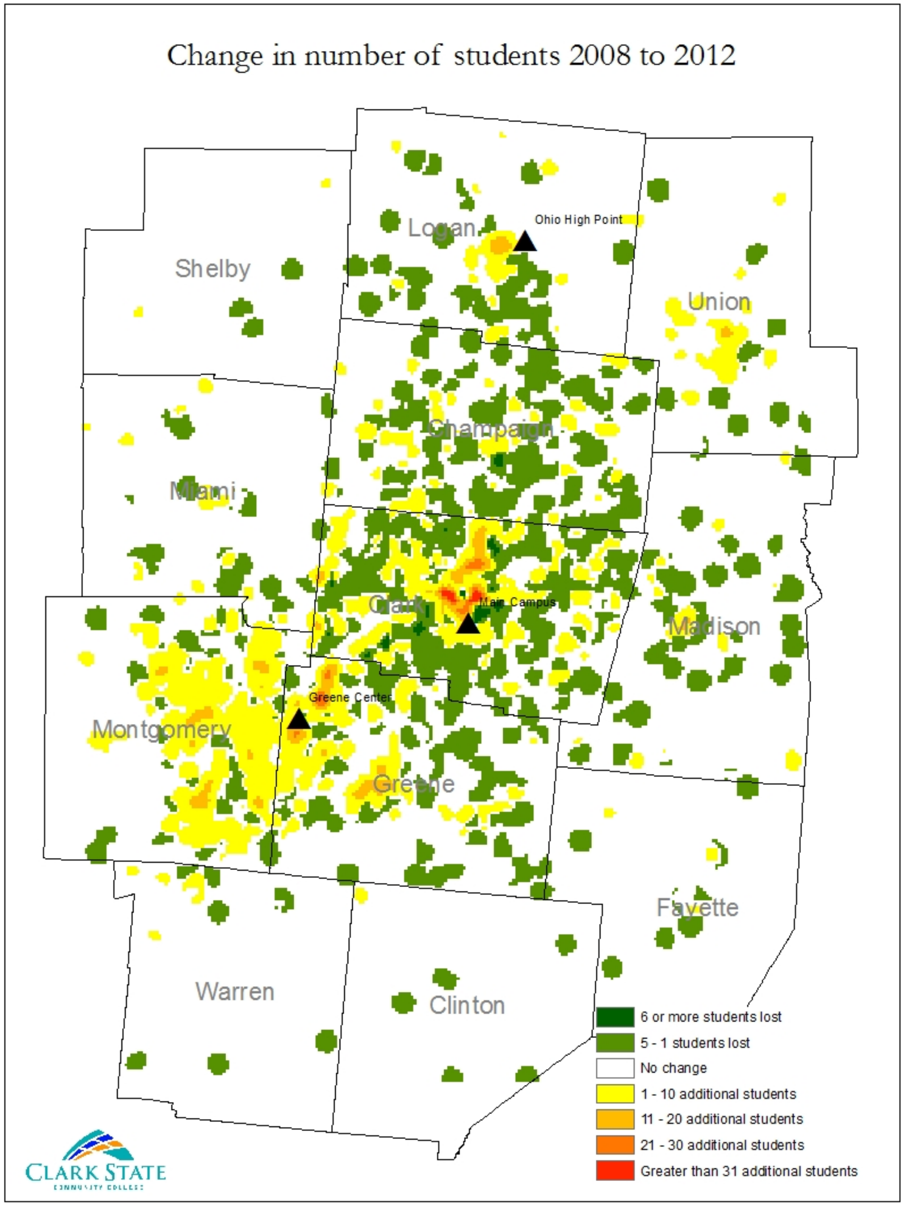


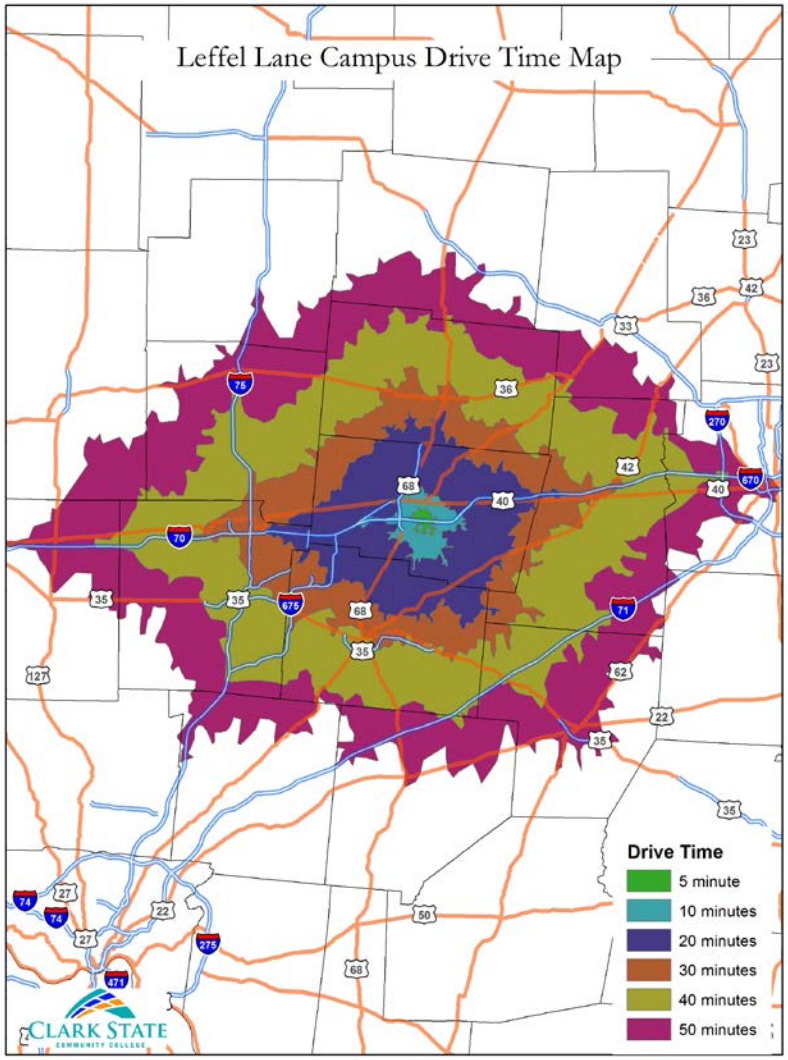
Change in number of students 2009 to 2010

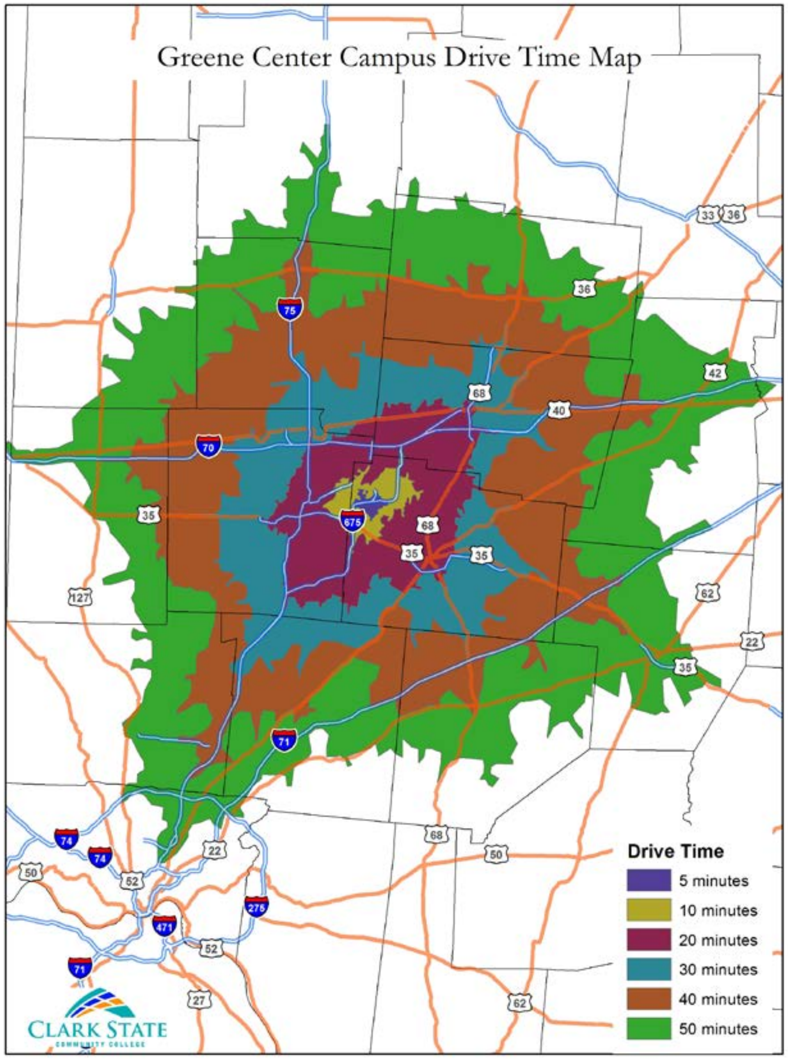


Change in number of students 2010 to 2011


Change in number of students 2011 to 2012


Change in number of students 2008 to 2012


Leffel Lane Campus Drive Time Map
Drive Time
5 minute
10 minutes
20 minutes
30 minutes
40 minutes
50 minutes
CLARK STATE
COMMUNITY COLLEGE


Greene Center Campus Drive Time Map
Drive Time
5 minutes
10 minutes
20 minutes
30 minutes
40 minutes
50 minutes
CLARK STATE
COMMUNITY COLLEGE