Modeling Predictors of Cultural Amenities in South Minneapolis, Minnesota USA Using Ordinary Least Squares Exploratory Regression

Christopher Ross
Department of Resource Analysis, Saint Mary's University of Minnesota, Minneapolis, MN 55404

Keywords: Demographics, Gentrification, Geographic Information System, Ordinary Least Squares, Exploratory Regression, American Community Survey

Abstract

Understanding factors that influence urban organization has been valuable for many decades. Until the 21st century, the lack of specific data restricted analysis to broad and often conceptual methods. With the availability of demographic data from the American Community Survey (ACS), address locations and business types from Esri Community Analyst, and data derivation tools such as GIS, this study was able to narrow the scope and identify specific variables that serve as predictors of cultural amenities. Using Ordinary Least Squares (OLS) and Exploratory Regression, a combination of cultural amenities were tested for significance against a candidate list of predicting demographic, zoning, and land cover values. Explaining 68% of the cultural amenities in South Minneapolis, the best predicting model includes total population, average household income, commercial parcels, percentage of black race, and percentage of bachelor's degree holders. These findings support conclusions of previous literature that gentrified demographics can be found where there are more cultural amenities. This study goes further, however, to identify cultural amenities new to this type of study and their significant relationships to the demographics of the study area.

Introduction

The evolving composition of urban and rural populations has been a topic of value since the first census was completed. As census collection methods became more detailed, further patterns emerged and more useful knowledge were derived for city and business planning. Each decade has different factors that contribute to population increase and decrease. Beale (1975) discussed how the post-World War II industrialization caused a "rural exodus" that led to population increase in metropolitan areas.

By the 1970's the trend was showing signs of reversal as economic and social patterns accommodated life outside of the metropolitan areas (Frey, 1987). An increasingly diversified economy created urban demand once again in the 1980s (Frey, 1993).

While both metro and non-metro areas experienced population growth in the early 1990's, the non-metro areas saw a higher percentage of growth than they did in the 1980's (Johnson and Beale, 1994). In 2003 the Office of Management created a new category within the Core-Based Statistical Area to describe urban centers with populations between 10,000 and 50,000 as "Micropolitan" (Vias, 2012). The new classification led to a new way of viewing population distribution rather than

Ross, Christopher. 2015. Modeling Predictors of Cultural Amenities in South Minneapolis, Minnesota USA Using Ordinary Least Squares Exploratory Regression. Volume 18, Papers in Resource Analysis. 10pp. Saint Mary's University of Minnesota University Central Services Press. Winona, MN. Retrieved (date) from http://gis.smumn.edu.

the oversimplified metro versus non-metro distinction (Vias, 2012).

By comparison to the research of the 20th century, population trends of the most recent decade are primarily focused on issues of gentrification and the effects of specific social or economic influences. A useful example is the study of the gentrifying effects of public art in Hollywood, California (Reynolds, 2012). The incorporation of GIS into these studies is growing; however, most analyses choose to focus on a specific set of variables rather than identifying a larger combination of influences.

This study was focused on determining predictors of cultural amenities in the southern half of Minneapolis, MN. The use of GIS is essential due to the inherently spatial nature of urban development (Elwood and Leitner, 2003). Esri's ArcGIS software tools were used for data processing and analysis. The methods and findings of this study are valuable for city planning of infrastructure, business, and residential neighborhoods.

Methods

Study Area

The area of study is defined as the 72 census tracts, based on the 2010 decennial census, that are within the borders of Minneapolis, Minnesota USA and located south of Interstate 394 to Washington Ave N, and from E Hennepin Ave to the Mississippi River (Figure 1). Note that while the chain of lakes does not have any demographic data, their area is included in the tract they fall within. Initial interest in the study area came from demographic change observed by the author over the last twenty years. A perceived increase in younger populations moving to the area with more social and recreational amenities led to curiosity about the relationships assessed in the study.

The specific study area was selected because of the ease of cultural amenity identification due to areas of strong cultural influence as well as easily defined study borders.

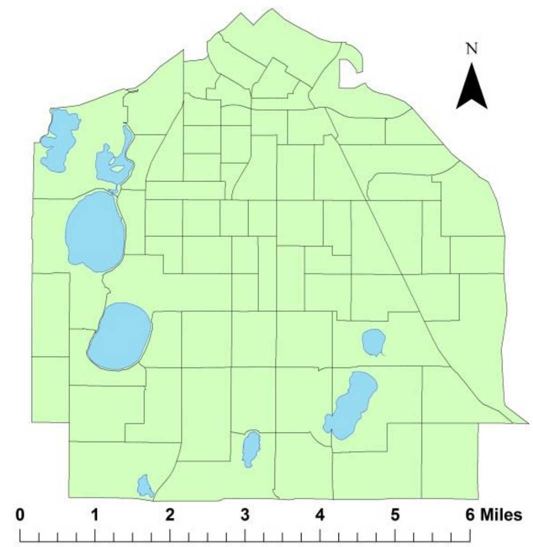

Figure 1. South Minneapolis census tracts and lakes included in the OLS/Exploratory Regression study.

Dependent Variable Data Selection

Conducting a study to identify predictors of cultural amenities raises the question of what qualifies as a cultural amenity. Reynolds (2012) described public art focused on cinema in Hollywood's 40 year gentrification project. Identifying the culture of the study area requires a general awareness of the area, such as cinema in Hollywood, along with identifiable trends.

One of the first cultural amenities identified for this study were brew pubs. Twenty-seven out of the 41 brew pubs identified in the study area were established after 2007 indicating an increased cultural propensity towards brew pubs.

Nice Ride is a system where anyone can rent a bicycle from any stall, and then return it to any stall location in the metropolitan area. These locations were included in the study to represent the Twin Cities boasted status as having the best urban cycling in the country. Alternatively, major roads were also included to support the importance of road vehicle transportation.

Farmers markets and community gardens were included to represent the promotion of the consumption of locally grown and organic foods. Spas and fitness centers were included as a representation of amenities to promote healthy lifestyles. Vias (2012) emphasized the importance of recreation as a common desirable amenity, which further supports the inclusion of the Nice Ride, fitness and spa locations, and parks. Retirement homes were included in the study because retirement age groups also have strong correlations to the availability of recreation (Beale, 1975).

Colleges were included as an amenity because there has been a consistent trend since World War II of populations shifting in proximity to colleges (Frey, 1993). Colleges are also known for hosting culturally themed events such as plays and athletic events.

Locational data was acquired from various sources including Hennepin County, Esri Community Analyst, Minnesota Geospatial Commons, and the Nice Ride website (Table 1).

Table 1. List of selected amenities, total of each in the study area, and data source used in OLS exploratory regression. *Major roads along tract boundaries were counted once for each bordering tract.

Amenity	# in Study	Source
Nice Ride	71	niceridemn.org
Fitness/Spa	48	ESRI Community Analyst
Farmers Market	8	Hennepin Co.
Community Garden	88	Hennepin Co.
Brew Pub	26	ESRI Community Analyst
Major Roads	197*	MN Geospatial Commons
Colleges	9	ESRI Community Analyst
Retirement Facility	3	ESRI Community Analyst
Parks	68	Hennepin Co.

Dependent Variable Preparation

The Ordinary Least Squares (OLS) and Exploratory regression methods require a single dependent variable that can be used to test the predictive ability of independent variables. For a successful analysis, the cultural amenities all needed to exist as a point feature class. Nice Ride, farmers markets, and community gardens were all acquired in this format. Fitness/spas, brew pubs, colleges, and retirement facilities were acquired as addresses. To facilitate the creation of point feature classes, an address locator was created to geocode the addresses.

Parks were converted to a point feature using the feature to point tool. All points were weighted equally with a designated factor of one so that no single amenity would have more influence than the others. Utilizing a spatial join between the study area tracts and each point feature yielded a total number of amenities located in each of the 72 tracts (Figure 2). Since major roads ran through multiple tracts, the roads count reflects the number of major roads passing through the tract (or creating the border with another tract). This caused the total number of roads in the study to appear large. The same road may create the border between several tracts, and was therefore counted as an amenity for each tract that it bordered.

Independent Variable Data Selection

Most of the demographic data for this study was acquired from the American Community Survey’s 2013 five year estimates at the census tract level. Parcel zoning data were derived from parcel data available to the public on the Hennepin County open GIS website.

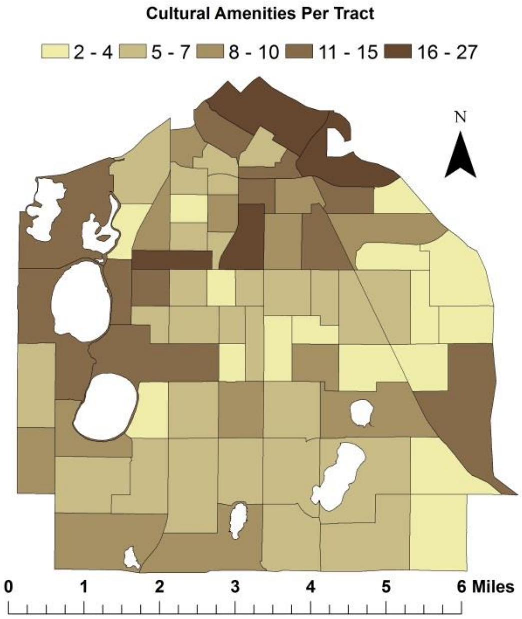

Figure 2. Total number of cultural amenities by census tract.

Land cover data were derived from Minnesota Land Cover Classification System data available on the Minnesota Geospatial Commons. Table 2 provides a complete listing of the final group of possible predictors along with their data source.

Specific demographic variables were selected based on their significance in previous studies and their inclusion in the census. Education data were included because of the strong ties between education and population distribution over the past 60 years (Frey, 1993). Racial percentages were included in the analysis due to the strong minority grouping observed by Anderson and Sternberg in 2012. Unemployment and household income figures were selected in the study as an indicator of the job market as well as possible pockets of poverty within the city (Frey, 1993).

Population and population density were included for their logical necessity when analyzing demographic patterns. The tract boundaries included the lakes area, but when calculating population density, lake area was excluded from the total tract area to avoid the misrepresentation of those tracts as less dense than the residential areas actually are.

Parcel zoning figures were included as a result of visual interpretation of initial exploratory regression findings.

Table 2. List of selected predictors and data source used in OLS/Exploratory regression analysis.

Predictor	Data Source
% No High School	ACS 2013
% High School	ACS 2013
% Some College	ACS 2013
% Bachelors	ACS 2013
% Asian	ACS 2013
% Black	ACS 2013
% White	ACS 2013
91-100% Impervious Cover	MLCCS
76-90% Impervious Cover	MLCCS
# Industrial Parcels	Hennepin County
# Commercial Parcels	Hennepin County
# Residential Parcels	Hennepin County
Population Density	ACS 2013
% Unemployed	ACS 2013
Mean Household Income	ACS 2013
Population	ACS 2013
Median Age	ACS 2013

Independent Variable Preparation

Most of the data from the American Community Survey (ACS) were delivered in a normalized form. Those that did not list data as a percentage were converted by dividing the provided number by the overall population of the tract.

The three parcel categories (industrial, commercial, and residential) were determined by filtering parcels to the desired zoning category and then running the feature to point geoprocessing tool. Performing a spatial join with the census tracts produced the necessary totals for use in the regression analysis. Figure 3 shows an example output of the spatial join process that was performed on the parcel layers.

The Minnesota Land Cover Classification system (MLCCS) impervious surface figures were determined by first filtering the category desired and then calculating geometry to determine area for each segment. A spatial join with a summary field produced the total area per tract. The total area of each category was then divided by the total area of the tract to determine what percentage of each tract fit into each MLCCS category.

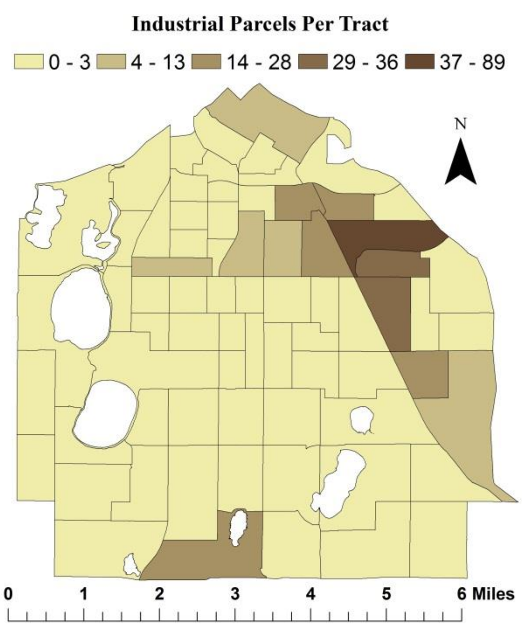

Figure 3. Number of industrial parcels per census tract.

OLS Exploratory Regression

The exploratory regression geoprocessing tool that uses the method of Ordinary Least Squares Exploratory Regression was developed by Esri's Lauren Rosenshein, Lauren Scott, and Monica Pratt (Rosenshein, Scott, and Pratt, 2011). The tool tests selected explanatory or predicting variables against a single dependent variable and produces an output diagnostic. In this study the dependent variable (Y) was the cultural amenity totals. The independent variables (X) were the demographic, zoning, and impervious surface data in Table 2.

OLS comprises five statistical checks that test for redundancy, completeness, significance, bias, and performance (Figure 4). The output diagnostic assesses which (if any) combinations of predicting variables best models or explains the independent variable. The initial results of the tool can often be used to identify explanatory variables that should be considered in another round of testing.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_n X_n + \epsilon$$

Figure 4. The OLS regression equation where Y is the dependent variable, β are the coefficients, X are the independent (explanatory) variables, and ϵ is the random error.

Geographically Weighted Regression

According to spatial statistics best practices it is important to consider the use of Geographically Weighted Regression (GWR) when using the OLS Exploratory Regression method (Rosenshein *et al.*, 2011). GWR can provide a better model by creating a regression equation for each individual feature based on the shape and size of the bandwidth defined by the user. The GWR tool indicates "where non-stationarity is taking place on the map, that is where locally weighted regression coefficients move away from global values" (Bivand, 2014). The intent of the tool is to find and improve on local areas where coefficients may not fit the global model. The models produced in this study, however, did not display improvement when using GWR, indicating that non-stationarity was not taking place with enough significance to change the model.

Results

The model variables that best explain the occurrence of cultural amenities in South Minneapolis are shown along with their variance inflation factors (VIF) in Table 3.

Table 3. The best fit model resulted from OLS exploratory regression and their associated VIF values.

Variable	VIF
% Bachelor's Degree	1.18
Population	1.32
Commercial Parcels	1.57
Household Income	1.69
% Black	2.08

The VIF values are all well under the Esri defined threshold of 7.5 which confirmed that these variables are not redundant. All five variables also had significant coefficient p values less than 0.05 to indicate strong relationships between individual explanatory variables and the dependent variable (Table 4).

Table 4. P values show statistically significant coefficient variables.

Variable	P Value	Coefficient
Commercial Parcels	0.0000	0.0557
% Black	0.0005	12.0409
% Bachelor's Degree	0.0020	5.3598
Population	0.0129	0.0010
Household Income	0.0201	0.0000

The adjusted R-Squared value for this model was 0.68, which means that the model explained 68% of the occurrences of cultural amenities. The Akaike's Information Criterion (AIC) was the lowest for this model at 346.47. Similar to the adjusted R-Squared value, the AICc is an indicator of good model performance, but only when compared to other models.

The Koenker (BP) statistic returned a p value of 0.032585, which was not statistically significant and means that the model showed consistent relationships across the geographic area of the study. The Jarque-Bera statistic of 0.58 was also non-significant which indicated a normal distribution of residuals (Figure 5).

Viewing the residual standard deviations of each tract offered the opportunity to see where model predictions were greater or lesser than reality (Figure 5). It is important to note that even the few tracts that were predicted high or low remained less than 2.5 standard deviations from the mean. For under predictions, the largest standard deviation was 1.9 and for over predictions, a single tract was 2.48 deviations while the next largest was 1.58.

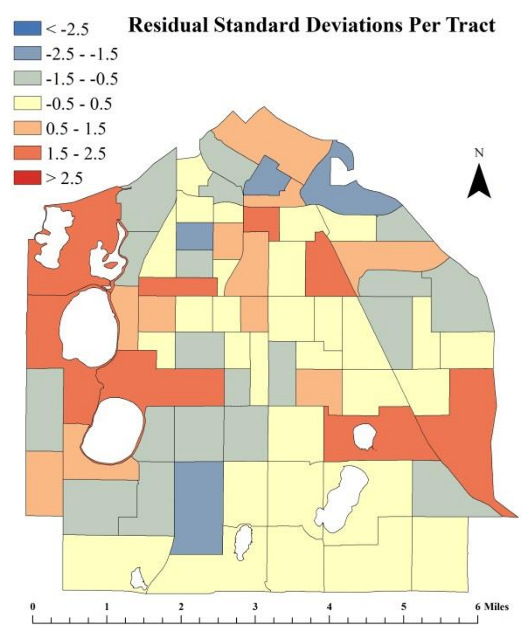

Figure 5. Residual standard deviations per census tract reflects a statistically normal distribution with no tracts exceeding 2.5 deviations from the norm.

The final assumption of the OLS model requires a minimum of 0.10 for the

spatial autocorrelation p value. With 0.57 the selected model is shown not to be excessively clustered nor dispersed.
One of the considerations for selecting this model was the lack of multicollinearity between variables. The GWR tool was also used but did not improve the model. The GWR model had an AIC value over 347 and an Adjust R-Squared value of 0.68. The only difference between the OLS and GWR model was a higher AIC, which confirmed that GWR would not improve the model.

Discussion

OLS Findings

The results of the OLS Exploratory method reveal a model that confidently predicts 68% of the cultural amenities in the study. Realistically, a higher R-Squared value would have provided even higher confidence in the model, but OLS specifies a minimum of 50% to pass, which means this model is well above that goal.

The chosen model represents the best of 6,871 trials and resulted in a combination of variables that includes race, education, income, commercial composition, and overall population. Within each of these categories it is important to note that while the final model reflects one variable within each category, the significance of other variables may have been more prevalent in other models. The percentage of Asian race, for example, was significant in more models than Black race, but in combination with other explanatory variables, the inclusion of Black percentage yielded the better model (Table 5).

Table 5. Explanatory variables and the percentage of models in which they were found significant.

Variable	**% Significant**
Commercial Parcels	100
Population	99.95
% Asian	72.16
% Bachelor's Degree	62.55
% Black	60.36
% White	35.41
Residential Parcels	24.9
Mean Household Income	24.23
Unemployment	22.06
Population Density	20.72
Industrial Parcels	16.96
% High School	12.03
% Some College	11.77
% Impervious 91-100%	6.96
% Impervious 76-90%	5.1
% No High School	1.71

The significance values were also useful to eliminate variables from subsequent modeling attempts. While the visual clustering indicated the amount of impervious surface may have been a predicting factor (Figure 6), it is clear from model significance that it was among the least likely predictors. Alternatively, population and commercial parcels proved to be the two most significant explanatory variables and thus appeared in the majority of passing models.

It is also worth noting while the five modeled variables provide the highest Adjusted R-Squared value, commercial parcels alone met all OLS assumptions, but with lower R-Squared (57%) and higher AICc values (363.4). Each variable added into the model improved the R-Squared percentage by roughly 3% and reduced AICc numbers by 3.

The coefficient values displayed in Table 5 show the percentage of black population and the percentage of bachelor degree holders have the strongest coefficients while household income has the lowest value. Commercial parcels did not have as strong of a coefficient value as bachelor degree holders and black percentage, but still proved to predict the highest amount of the study area.

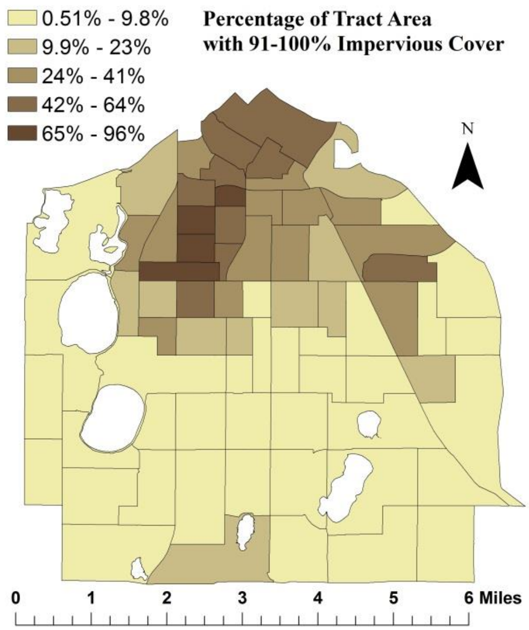

Figure 6. Percentage of tract area categorized by the MLCCS with 91-100% impervious cover.

A significant result of the model predictions is evident in the mapping of residual standard deviations. There appears to be a pattern of under predictions for tracts that had lakes (Figure 7). With so much of the tracts area consumed by the lakes it would have been conceivable that the model would have made over predictions due to the lakes inclusion as one or more park amenities. The county data had divided the areas around the lakes as separate parks which is the most logical method because a beach across the lake from a soccer field is similar to a playground down the block from a soccer field. The result of more parks would conceivably cause over predictions for the tracts with lakes, but that was not the case. The model produced under predictions in those tracts instead. It is likely other factors related to the parks could help to provide a better explanation of the results in those particular tracts.

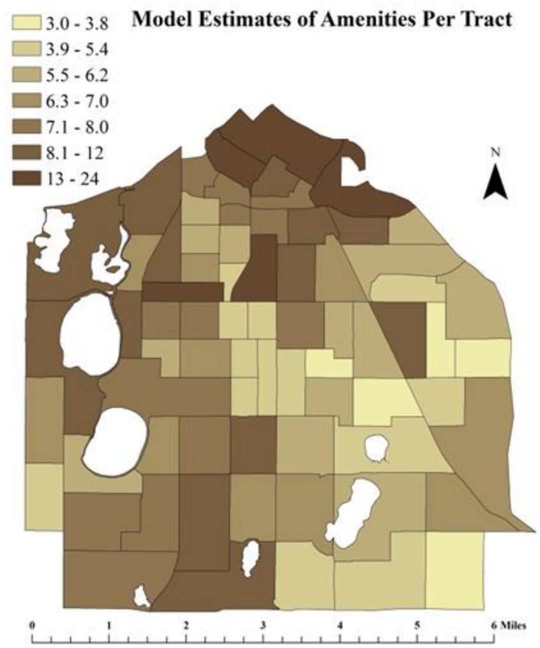

Figure 7. The OLS models predictions of cultural amenity occurances.

Potential Sources of Error

The first and perhaps most likely source of error in the study is the determination of the cultural amenities that comprised the dependent variable. While supported by previous literature, the nature of defining culturally significant locations in a community is certainly a debatable practice. The methods employed in the selection process attempted to mitigate any one selection from dominating the variable group, and several variables (coffee shops and bus stops for example) were left out for that reason.

The second source of potential error is in the selection of explanatory variables. With the continually expanding process of demographic data collection, it is not feasible to test all demographic topics. There may have been a demographic variable that would serve to improve the model. Even more difficult is finding an appropriate method of developing and incorporating non-demographic data as explanatory variables. Two such methods were attempted in this study, the incorporation of parcel zoning and percentage of impervious cover. Commercial parcels proved to be the most significant variable in the study while impervious surfaces appeared to have the least significance. A more robust breakdown of the impervious surface percentages would perhaps improve the significance in similar models.

Suggestions for Future Research

Future analysis would benefit from temporal comparisons. The inherent difficulties with such a study would be the acquisition of comparable spatial data across time periods as was the initial intent of this study. Due to the nature of changing political boundaries, and altered survey methods for census, finding an appropriate method is a difficult task. Perhaps the next iteration of the ACS survey or the next decennial census will have more consistent figures to perform such a time lapse study.

Still another area of beneficial study is in the comparison of multiple urban areas utilizing the same dependent variables. Even within the Twin Cities it would be beneficial to see what models prove to be the stronger predictors.

Conclusion

The first conclusion from the results of the OLS model is that the composition of commercial parcels in a tract is the single most significant predictor of cultural amenities in this study. This is a logical finding when considering the dependent variable inclusion of brew pubs, and fitness/spas. More difficult to assess is whether the commercial parcels lead to more amenities, or if more amenities lead to the development of more commercial parcels. It is likely a mutually progressive relationship so that as one increases so does the other. City planners may identify this model as a tool to justify actions to encourage commercial and amenity growth in an effort to increase the prevalence of higher education and incomes in select neighborhoods.

The lack of residential parcel composition significance in the model means that cultural amenities are not necessarily tied to the locations where there are more numerous parcels such as single family neighborhoods.

The remaining four independent variables all provided incremental increases to the confidence of the model. All predictors had a positive correlation which means in essence that higher population, higher education, higher income, and larger composition of the black population proved to be factors that served to better predict the cultural amenities of this study.

The model reveals valuable insights for the demographic of people who reside in the study area, as well as the community and business planners. Based on the correlations between higher education and higher income in socioeconomic status, it is feasible to assess the phenomena modeled as a form of gentrification (Wyly and Hammel, 1998). While gentrification implies the intent to change the demographics, the results of this study may indicate a generally unplanned cultural gentrification. It is too difficult to classify the modeled patterns as traditional gentrification primarily because of the amenity selection. With such a wide variety of sources and the lack of temporal analysis, this study is intended to explain relationships as they exist.

Acknowledgements

My thanks go to Mr. John Ebert and Dr. David McConville for their instructional assistance throughout the Saint Mary's University of Minnesota graduate GIS program. A special thanks to Mrs. Greta Poser for her coursework instruction as well as her introduction of the OLS method, without which this project would not have taken place.

Additionally I would like to thank the general efforts of the Minnesota State agencies, and Hennepin County, for their efforts to make spatial data easily accessible.

References

Anderson, M. B., and Sternberg, C. 2012. "Non-White" Gentrification in Chicago's Bronzeville and Pilsen: Racial Economy and the Intraurban Contingency of Urban Redevelopment. Urban Affairs Review, 49(3), 435-467. doi:10.1177/1078087412465590

Beale, C. L. 1975. The Revival of Population Growth in Nonmetropolitan America. Economic Research Service (DOA), ERS-605.

Bivand, R. 2014. "Geographically Weighted Regression". Applied Spatial Data Analysis with R, 305-308.

Elwood, S., and Leitner, H. 2003. GIS and Spatial Knowledge Production for Neighborhood Revitalization: Negotiating State Priorities and Neighborhood Visions. Journal of Urban Affairs, 25(2), 139-157. doi:10.1111/1467-9906.t01-1-00003

Frey, W. H. 1987. Migration and Depopulation of the Metropolis: Regional Restructuring or Rural Renaissance?. American Sociological Review, (2). 240.

Frey, W. H. 1993. The New Urban Revival in The United States. Urban Studies, (4-5), 741.

Johnson, K. M., and Beale, C. L. 1994. The Recent Revival of Widespread Population Growth in Nonmetropolitan Areas of the United States. Rural Sociology, 59(4), 655-667. doi:10.1080/13574809.2011.646246

Reynolds, M. 2012. A Glamorous Gentrification: Public Art and Urban Redevelopment in Hollywood, California. Journal of Urban Design, 17(1), 101-115.

Rosenshein, L., Scott, L., and Pratt, M. 2011. Exploratory Regression: A Tool for Modeling Complex Phenomena. Retrieved May 5, 2015 from www.esri.com/news/ arcuser/ 0111/ files/exploratory.pdf

Vias, A. C. 2012. Micropolitan Areas and Urbanization Processes in the US. Cities, 29(Supplement 1), S24-S28. doi:10.1016/j.cities.2011.12.005

Wyly, E. K., and Hammel, D. J. 1998. Modeling the Context and Contingency of Gentrification. Journal of Urban Affairs, 20(3), 303.