Connecting People with Nature Through a Network of Large Natural Areas

by

William Norten and **Dan Plechaty**

Dr. Jennifer Swenson, Advisor

April 17th, 2015

Table of Contents

pg. 3 Introduction

pg. 4 Literature Review

pg. 8 Methods

pg. 11 Results

pg. 21 Conclusion

pg. 22 Acknowledgements

pg. 23 Appendix A. Regression Results

pg. 31 Appendix B. Geospatial Results

pg. 34 Appendix C. Geospatial Model

pg. 38 Appendix D. Model Python Scripts

Fig. 1. Extent of the South Atlantic Landscape Conservation Cooperative

Introduction

The South Atlantic region, extending from southern Virginia to northern Florida, is a diverse and growing area that includes large urban centers such as Atlanta, Jacksonville, Charlotte and Raleigh. With a large number of the region's population already residing in these metropolitan areas, the percentage urban residents in South Atlantic's population is expected to increase rapidly over the next few decades. The pressures from this expanding urbanization will result in the further loss of natural areas unless proactive land conservation actions are undertaken. As it stands, not all of these urban residents currently have access to a large natural area in which to recreate and enjoy nature, making timely conservation efforts in the region vital. It is against this background that the South Atlantic Landscape Conservation Cooperative (SALCC) is working to "inform resource management decisions" and facilitate conservation planning among "federal agencies, regional organizations, states, tribes, NGOs, universities and other entities" in the region[1].

The goal of this project is to work with the SALCC to perform a feasibility analysis for having a large natural area nearby every urban resident in the South Atlantic region. To accomplish this, we performed a literature review in order to determine how much people value large natural areas and how willing they are to travel to reach them. Then, we developed a novel, flexible and scalable geospatial toolset which calculates two measures of access: Euclidean distance and driving times. Finally, we employed an exploratory statistical analysis to see how access correlates with socio-economic, geographic and other variables. It is our hope that our work will both help the SALCC identify areas in need of future conservation efforts and inform them about what is valued in a large natural area and what groups' needs in the region are currently not being met.

The literature review will follow, along with an explanation of the various methods used to elicit the value people place on natural areas. Our methods section is after that, and is divided

[1]Strickland, J (2010). Welcome to the South Atlantic Landscape Conservation Cooperative. Last updated May 2013; accessed February 2015. http://www.southatlanticlcc.org/notes/Welcome_to_the_South_Atlantic_Landscape_Conservation_Cooperative

into two parts: a brief description of the geospatial toolset and an explanation of the regression model and the variables used in it. Finally, we present our results and discuss their implications.

Literature Review

In order to inform the subsequent portions of this analysis, we reviewed the non-market valuation literature in order to identify the factors that influence how much people value large natural areas. Non-market valuation can come in multiple forms, be it stated or revealed preference. These tools are utilized in order to place a price on objects and services that have no normal market by which to assign value. The South Atlantic Landscape Conservation Cooperation (SALCC) is interested in conserving large natural areas for all urban residents in the United States' southeast region. These parks and protected areas can provide many societal benefits. To better inform our client of the possible paths forward to correcting this lack of parkland we have undertaken a study of the relevant non-market valuation literature to establish price points for large natural areas. Two different valuation methods were investigated: travel cost and price per hectare. Additionally we wanted to find the average willingness to travel distance. With these three metrics, it is our hope that this information will aid in the SALCC's future land conservation activities.

Contingent Valuation

This method is focused on getting directly to the source of what the average consumer is willing to pay (WTP) for a non-market good. This is commonly done by surveying said consumers on what their WTP (or willingness-to-accept) for a nonmarket good. This can inform researchers on what the correct amount, in economic terms, the supply for the good should be. In our study we are concerned about what the "correct level" of natural area conservation is for urban areas. Contingent valuation (CV) also allows researchers to put a price on units of nonmarket goods, in this case, hectares of open space. We utilized Brander and Koetse's *The Value of Open Space: Meta-Analyses of Contingent and Hedonic Pricing Results*, which aggregates "over 90 studies dealing with open space valuation… published over the past 30

years" into one range of WTP for urban open spaces[2]. In essence Brander and Koetse have provided a literature review of the last 30 years for CV studies, which informs our own study.

The results from Brander and Koetse are telling. They found that "The mean value is US$ 13,210 per hectare per annum, and the median value is US$ 1,124". Obviously there is a huge range between these two numbers, representing a skewed distribution of prices. This large range of values makes us hesitant to recommend using these prices when trying to determine the correct price to pay for natural area conservation. That being said the median price of $1,124 per hectare does not appear to be exorbitant, thus it would be our recommendation to SALCC to stay closer to this number than mean.

It should be noted that there are a few key differences between our research question and what Brander and Koetse used in their paper. First and foremost, the SALCC is looking at conserving space specifically for recreational value, while Brander and Koetse defined open space as, "forest, park, green space, undeveloped land, and agricultural land". Obviously agricultural and undeveloped land may not provide recreational value, yet the paper does not provide enough information to parse out individual WTPs for each of these categories. Brander and Koetse also included studies from places outside the SALCC region, including other U.S. regions and European countries. Although all prices were reported in U.S. dollars, it is likely that other regions and especially other countries have different valuations for open space.

Travel Cost

The travel cost method is quite simple; a researcher surveys a population and asks them directly how much they paid for their trip along with any associated costs they may have incurred during their activities. This can be a wide range of dollar points to add up, for example how far the respondent drove takes into account not only the price of gas to get to and from, but also the wear and tear on the vehicle. Another example of how complicated this methodology can be is that it takes into account the opportunity costs of the time spent on the trip versus other activities like working and thus not earning a wage. Many economists prefer the travel cost

[2] Brander, Luke and Koetse, Mark (2011). The Value of Urban Open Space: Meta-Analysis of Contingent Valuation and Hedonic Pricing Results. Journal of Environmental Management, 92, 2763-2773. doi: 10.1016/j.jenvman.2011.06.019

method to the contingent valuation as it is based on real dollar amounts that people have spent, not what they theoretically would pay. In essence what the travel cost method reports is the WTP, yet grounded in real numbers. This methodology allows us to inform the SALCC on what their service population is actually WTP for these open spaces.

For our study we relied on Zawacki et al.'s analysis[3], which draws upon the U.S. Fish and Wildlife Service's annual National Survey of Fishing, Hunting and Wildlife-Related Recreation (henceforth FHWAR), specifically the 1991 FHWAR[4]. The 1991 FHWAR survey was selected as it is the last year that the survey contained questions regarding the average distance traveled on day trips. The FHWAR is helpful as a data source as the results are reported at both the state and national levels, so we are able to concentrate on the six states within the SALCC. It should be noted that the travel cost numbers are reported as the consumer surplus at the individual trip level. Consumer surplus is "the difference between individual willingness to pay and actual expenditure for a good or service"[5]. In slightly simpler terms consumer surplus is the difference between what was paid and the average amount spent on trips. They find that the average national consumer surplus for hunters and anglers is $37.40, while for nonconsumptive consumer surplus is $63.20. It should be noted that these numbers are in 1991 US dollars and are reported as untruncated, meaning the entire population was used, not just those who took trips. Utilizing these numbers SALCC is able to translate the consumer surplus as individual WTP per park trip. Thus this valuation could then be used to estimate the annual economic impact of adding natural areas to any urban areas missing ones.

Just as with the CV method, Zawacki et al.'s numbers do not represent the SALCC region, but they may be used to better inform what the average consumer surplus for the region may be. This can be important for SALCC as the numbers could "be used in benefit-cost analyses as a first approximation for the benefits of providing wildlife viewing access" in

[3] Zawacki, W., Marsinko, A. & Bowker, J.M. (2000). A Travel Cost Analysis of Nonconsumptive Wildlife-Associated Recreation in the United States. Forest Science, 46, 496-506. Retrieved from http://search.proquest.com/docview/19927303?accountid=10598

[4] U.S. Department of the Interior, U.S. Fish and Wildlife Service, and U.S. Department of Commerce, U.S. Census Bureau (1991). National Survey of Fishing, Hunting, and Wildlife-Associated Recreation.

[5] Zawacki, W., Marsinko, A. & Bowker, J.M. (2000).

addition other types of recreation like hunting and fishing.[6] These are also real numbers representing people's WTP for access to natural areas that have been found across the country, albeit some time ago.

Average Willingness-to-Travel Distance

The U.S. Fish & Wildlife Service's FHWAR survey was also utilized to calculate the average distance in-state park goers would be willing to travel (WTT). Unfortunately, the FHWAR survey stopped asking questions regarding visitors' travel distance with the 1991 survey[7]. Thus the reported numbers are not recent, and without further study it is unknown if they represent the average WTT today. One possibility for translating the WTT would be to look at how vehicle miles traveled (VMT) and income levels have changed since 1991. The other interesting factor regarding the WTT is that the FHWAR also reported individual state travel statistics. SALCC will be able to make a more nuanced decision depending on what state they are working in. The state average WTTs are listed in Table 1 as follows:

Table 1. Non-consumptive In-State Travel

State	Distance (miles)
Virginia	17
North Carolina	20.7
South Carolina	12.3
Georgia	17.8
Florida	25
Alabama	22.2
Average	19.7

These numbers confirm that the SALCC's definition of nearby access is reasonable, as they suggested using a 20 mile buffer around metropolitan statistical areas (MSAs) when investigating park access.

[6] Zawacki, W., Marsinko, A. & Bowker, J.M. (2000).
[7] U.S. Department of the Interior, U.S. Fish and Wildlife Service, and U.S. Department of Commerce, U.S. Census Bureau (2011). National Survey of Fishing, Hunting, and Wildlife-Associated Recreation.

Conclusion

These three metrics, price per hectare, consumer surplus and average willingness-to-travel, will help SALCC inform their future conservation efforts by putting real dollar values on what people say they are willing to pay, what they have actually spent to enjoy these natural areas, and how far they are willing to travel to access these parks. While it must be said these numbers are approximations, we recommend them as a starting place for SALCC in their future conservation efforts. We plan to continue our efforts to provide and narrow price points for our client.

Methods

There are two principal methods by which we hope to accomplish our project objectives: a geospatial analysis and a regression analysis. The geospatial analysis consists of the development of a toolset to quantify access to large natural areas from urban areas in the South Atlantic region through two measures. The first measure is a Euclidean distance calculation from each MSA's boundaries to the nearest qualifying large natural area, and the second is a network analysis that calculates driving times to large natural areas. Our exploratory statistical was executed in order to analyze what factors might explain why some MSAs in the SALCC region lack access to large natural areas. This regression analysis incorporates two measures of park accessibility, the Euclidean distance calculations from our GIS analysis, and the total qualifying parkland as a percentage of the total land area in each of the MSAs and buffers. These two variables are utilized as proxies for park access from which we built two statistical models using socioeconomic and geographic data.

I. Geospatial Toolset

One of the principal deliverables for our client is to identify those urban areas in the South Atlantic region that currently are lacking a nearby large natural area. In order to accomplish this goal, we developed a novel, flexible and scalable toolset using Python and the ArcGIS software suite[8]. This involved defining for our analysis a) what constitutes an urban area, b) what constitutes a large natural area, and c) what constitutes nearby access. We used

[8] ESRI (Environmental Systems Resource Institute). ArcMap 10.2. July 31, 2013. Redlands, California. http://www.esri.com/software/arcgis/arcgis-for-desktop

Metropolitan Statistical Area boundaries from the U.S. Census Bureau to define urban areas[9], a Conservation Biology Institute dataset on protected areas using those sites larger than 5000 acres with public access to define large natural areas (also called 'parks' for short in this text)[10], and we considered large natural areas within 20 miles of the border of an urban area or within an hour's drive from the MSA centroid to be nearby. The toolset features the ability to allow the client to specify their own parameter values in order to evaluate how these choices affect our results on access; these results are saved to disk in order to facilitate future comparisons.

There are three components to the toolset: a data preparation script, a Euclidean distance script, and a network analysis script. The data preparation script takes as an input the name of the LCC to prepare the data for; while the data is initially subset for the SALCC, the hope is that this toolset will be distributed to LCCs across the country in order to help inform conservation decisions in other regions as well. After this step, one is ready to run either of the Euclidean distance or network analysis scripts. The Euclidean distance script takes as inputs the minimum park size and the maximum distance from the edge of the Metropolitan Statistical Area; the default parameters, chosen by the SALCC, are 5000 acres and 20 miles. The script creates a buffer around each of the MSAs equal to the distance threshold, and finds the nearest qualifying park for each MSA (if any). The network analysis script takes as inputs the minimum park size and the maximum driving time from the centroid of the MSA for which to search; the default parameter is three hours, although any large natural areas greater than one hour away are considered inaccessible. The script uses the Network Analysis extension to ArcGIS to calculate driving times from the centroid of each MSA to its nearest park, and saves the routes generated. For a more detailed explanation of the scripts in the toolset, please refer to Appendix C.

II. Park Accessibility Analysis

Upon completion of our geospatial analysis, our results show that only two MSAs were in fact missing nearby large natural areas. As only two MSAs were indeed lacking access, it would be nearly impossible to glean any sort significant analysis for the lack of parkland from

[9] U.S. Census Bureau. TIGER/Line® Shapefiles: Metropolitan Statistical Areas. August 22, 2013. ftp://ftp2.census.gov/geo/tiger/TIGER2013/CBSA/tl_2013_us_cbsa.zip
[10] Conservation Biology Institute. Protected Area Database-US (CBI Edition), Version 2. October 31, 2012. http://consbio.org/products/projects/pad-us-cbi-edition

such a small sample size. Instead of focusing solely on why these two specific MSAs lack large parks, we utilized proxy variables to explain park access: distance from the centroid to nearest park edge, and percentage of the total MSA and buffer area that consists of qualifying parkland. Previous park availability research has identified two forms of access, availability and accessibility. The availability method measures "the rate of the supplies vs. the demands within a pre-defined region", while the accessibility approach measures "the nearest neighbor—the distance to the closest green space using simple Euclidean distance".[11] The proxy variables we chose, centroid distance and total park percentage, fit well within both these park accessibility measures. In order to select which explanatory variables to use in park access models we relied on past research. We focused on socioeconomic indicators, specifically, poverty and ethnic demographic data as other researchers have found that green and open spaces are often lacking in minority racial and ethnic communities[12], while other studies report "consistently that neighborhoods with higher SES [socioeconomic status] levels enjoy greater accessibility to green spaces".[13] We chose to focus on these factors when selecting our variables. Yet, we also acknowledge that there are other dynamics that may explain differing levels of park access. Past research in this area of study has indicated that "A further dimension that influences the spatial relationships between green spaces and the urban built environment is the topographic landscape, especially… elevation".[14] Thus we were not only interested in explaining access through the social and economic makeup of each MSA, but whether the physical makeup of the surrounding land was influencing access as well.

Finally, we chose to complete three models for comparison. This was done for validity, since it is unlikely that one model can tell the true story of park access. We sought optimal models for both the access and availability metrics and then ran these models' variables with the opposing response variable to see whether they corresponded in significance.

[11] Dai, Dajun (2011). Racial/ethnic and socioeconomic disparities in urban green space accessibility: where to intervene. Landscape & Urban Planning, 102, 234-244. doi:10.1016/j.landurbplan.2011.05.002
[12] Dai, Dajun (2011).
[13] Wen, Ming et al. (2013). Spatial Disparities in the Distribution of Parks and Green Spaces in the USA. Annals of Behavioral Medicine, 45 (supplement 1), 18-27. doi: 10.1007/s12160-012-9426-x
[14] Davies, Richard, et al. (2008). City-wide relationships between green spaces, urban land use and topography. Urban Ecosystems, 11, 269-287. doi: 10.1007/s11252-008-0062-y

Results

I. Geospatial Results

Our geospatial analysis indicates that 2 of the 42 MSAs do not have access to a nearby large natural area, as calculated by the buffer analysis script using the default parameters (within 20 miles of the edge of MSA, and of at least 5000 acres). These are Albany, GA and Goldsboro, NC. As shown in Figures 2 and 3, these MSAs also have the longest driving times to their closest parks, as calculated by the network analysis script using the default parameters. Two additional MSAs have driving times greater than an hour, the SALCC's threshold for accessibility: Florence, SC and Winston-Salem, NC. Figure 4 below illustrates how the two measures of access compare; there is a strong and positive pairwise correlation between them, with a correlation coefficient of 0.905 and a p-value of 0.0000. This increases our confidence in the validity of our results, and identify Albany and Goldsboro in particular as important areas to target with future conservation efforts.

While it is important to single out those areas that are currently lacking access to large natural areas, it should also be noted that many of the urban areas in our sample had sufficient access, both in terms of driving times and distance. Of the 42 MSAs, a full 31 of them had a qualifying large natural area that was at least partially within the boundary of the MSA, while a further 9 of them had a qualifying large natural area at least partially within 20 miles of the boundary. Likewise, 22 of the MSAs had driving times of less than 45 minutes, and 11 MSAs had driving times of less than 30 minutes. Some of the communities with the best access to large natural areas using both metrics of access include Virginia Beach-Norfolk-Newport News, VA-NC, Brunswick, GA, Jacksonville, FL, Spartanburg, SC, Columbus, GA-NC, Wilmington, NC, and Hinesville, GA.

Fig. 2. Access to Large Parks from Urban Areas

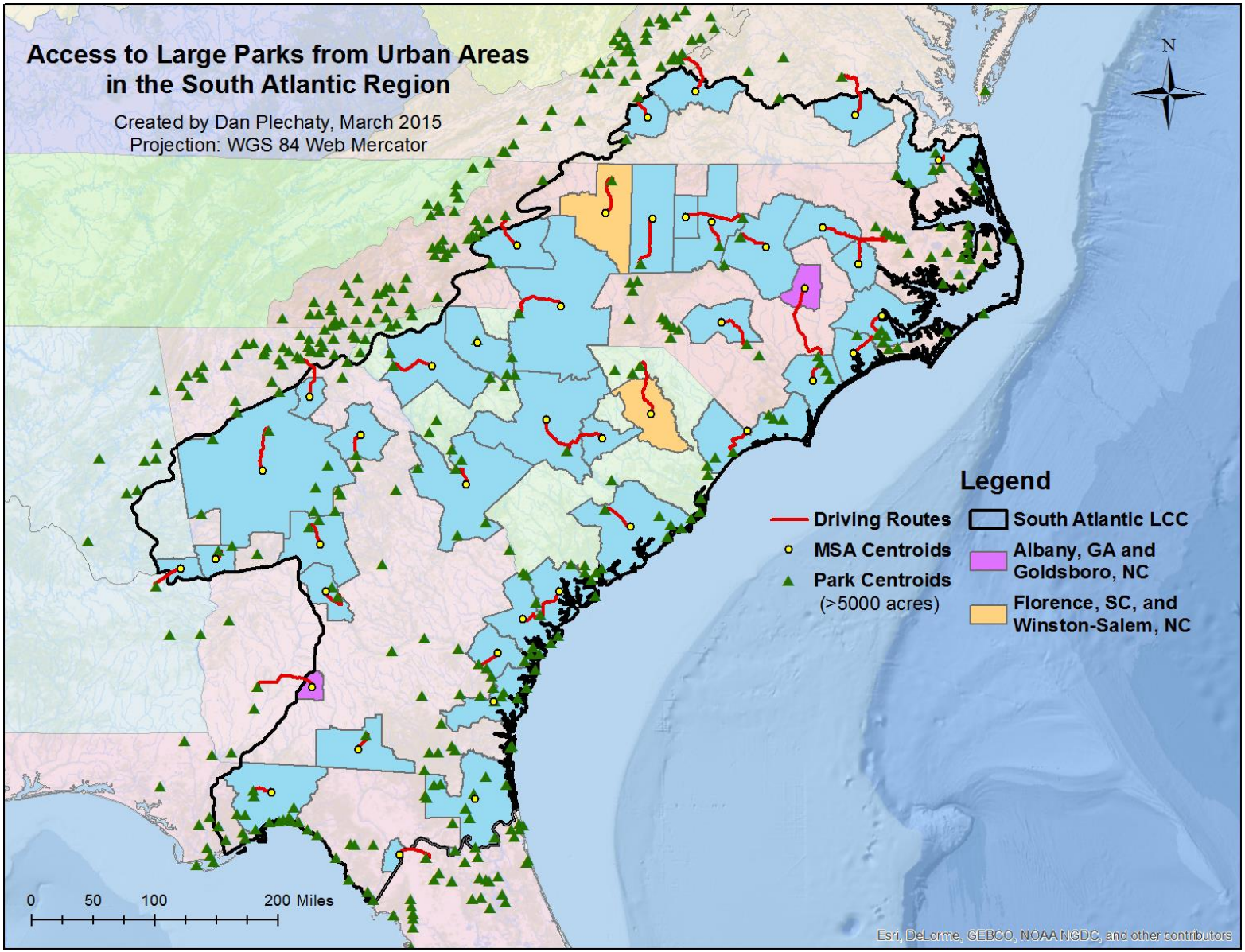


Fig. 3. Detailed View of Durham-Chapel Hill and Albany

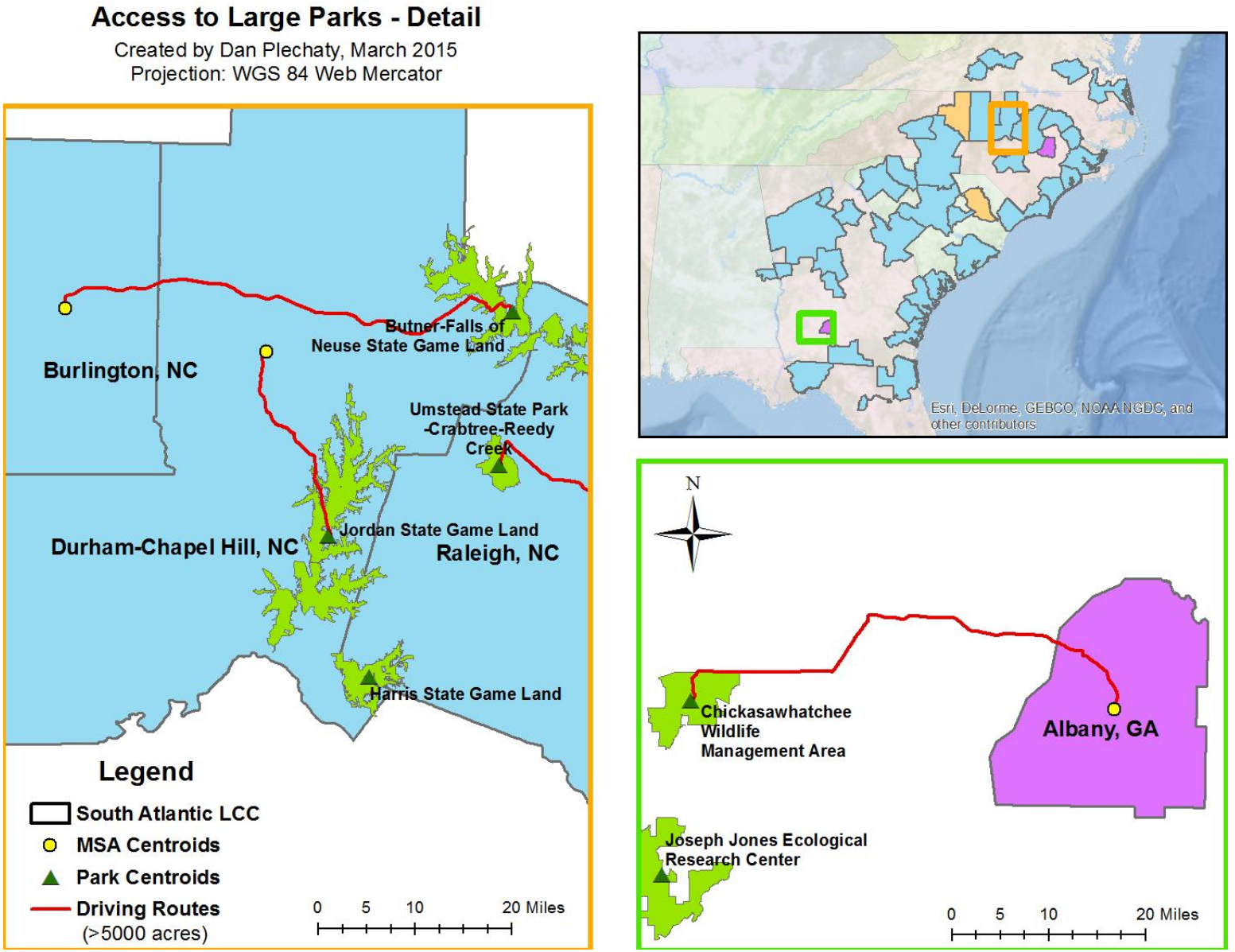
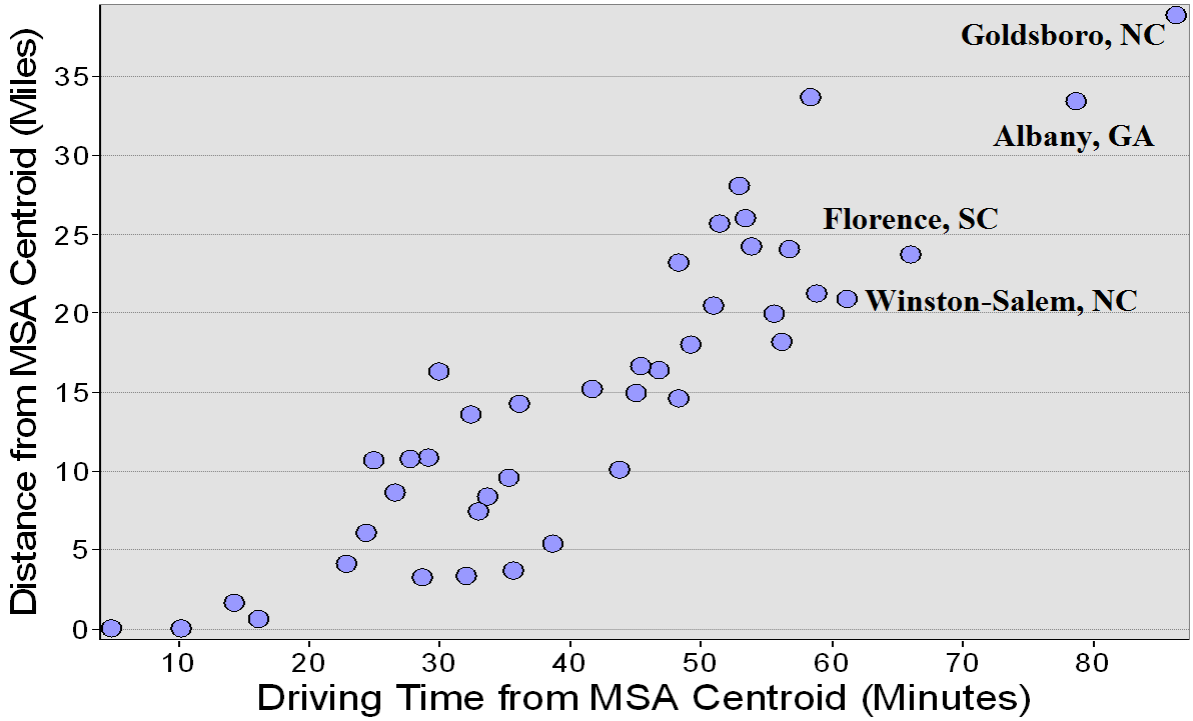


Fig. 4. Euclidean Distance and Driving Times



Data considerations and possible improvements

The PAD-US, MSA and street network datasets are all updated periodically by their respective owners; as subsequent versions become available, they can be substituted in for the versions of these datasets that are currently provided. New parks added to the PAD-US dataset might change whether or not a MSA has park nearby, new MSA boundaries will affect the distance and driving time estimates, and new streets added to the network dataset may affect driving time estimates as well. Of these, the possibility that there are missing (or misclassified) parks is the most serious, and probably the most likely to result in inaccuracies in the analysis. While care was taken to filter out records in the PAD-US database that did not fit our criteria (for example, we have removed private lands, military bases and wastewater treatment plants from the database), it is possible that some were missed (or removed erroneously) or that the user wishes to use more conservative (or more liberal) definitions of what constitutes a park with recreational value to the public. Currently, there is no way to do this save by editing the where clause in the script that controls which parks are filtered out, but one way to potentially improve this tool would be to give the user more options at the beginning on how they wish to filter out parks.

One might also question using MSA boundaries as a proxy for urban areas. MSAs are a statistical construct of the U.S. Census Bureau; many of them are not traditional urban areas, and often have low-density development in the vicinity of the boundary. If the boundaries of the MSA are more expansive than that of the core urban area, we may be overstating the accessibility of large natural areas to urban residents. Due to these concerns, care should be taken when making comparisons between MSAs in terms of the distance/driving times to parks. This analysis is envisioned as the starting place for making comparisons, but more localized knowledge is useful for further interpretation.

While the above concerns apply to the whole of the analysis, there are additional concerns with the Network Analysis script in particular. The PAD-US dataset does not have the entrance locations to parks, and as such we simply found the centroid of parks and then joined this to the nearest road segment and used this as the end point. The sheer size of these parks means that this could introduce substantial inaccuracies to the analysis – driving around the park to the 'entrance' point may add on 15 minutes to the driving time unnecessarily. On the other side of this, using the MSA centroid as a starting point might also obscure differences in driving times to the park from one side of a city as opposed to another. One way in which this could be made more interactive would be to allow the user to specify a starting location within a specific MSA, but as this tool was designed for conservation planners it was chosen not to focus on the specific circumstances of individuals.

Finally, the current tool only works by specifying Landscape Conservation Cooperatives within the United States. Theoretically, this could be done with any other regional specification, or by directly specifying the MSAs that one is interested in. The current version of the tool does not do this because our client is interested in the South Atlantic region in particular; the option to specify other regions was only included because it was realized that the analysis would work just as easily for them. If a user wishes to use a different region (other than a LCC), there is no way for them to do so without modifying the Data Preparation script. Should this be desired, please contact the author at the email address specified for additional assistance.

II. Regression Results

Summary Statistics

Please refer to Table 2 for a detailed breakdown of the descriptive statistics of each of the explored variables, and the resulting log transformations. For reference, all tables and figures relevant to this section are presented in Appendix A. We used data from the U.S. Census' 2007 American Community Survey for all 42 MSAs from the SALCC region. Violent crime statistics were gleaned from the Federal Bureau of Investigation's Crime in the U.S. database. Geography data was sourced from a mix state websites and Land Scope America[15]. Additionally the U.S. protected areas database, which we utilized in the GIS analysis, is also being used for the geographical variables. Unfortunately not all 42 MSAs had GIS data that was available or applicable for our chosen variables, thus this resulted in some variables reporting less than 42 observations for each of the 23 variables.

Due to the elevated skewness and kurtosis of the following variables we choose to log transform these datasets: total area, total park area, total park percentage, total population, unemployment, mean household income, poverty, percentage with bachelors or higher, percentage of vacant housing, renter percentage, no vehicle available, per capita income, and population density. This was done in order to meet the model assumption that our residuals would be normally distributed. To include the physical makeup of the geography surrounding the MSAs we chose to create dummy variables based on whether they were in one of four geographic types, mountain, piedmont, coastal plain and coastal. The difference between coastal plain and coastal in this case is whether the city is located on the ocean, or on an inland plain. It

---

[15] Crime sourced from http://www.fbi.gov/about-us/cjis/ucr/crime-in-the-u.s/2011/crime-in-the-u.s.-2011/tables/table-6
New Bern crime statistics sourced from http://www.newbern-nc.org/files/5414/0327/5704/2011-IntelligenceCrimeReport-Updated.pdf
Greenville NC crime data sourced from http://www.ncdoj.gov/getdoc/7225a87f-1838-4f97-a559-4c441d317249/2011-Annual-Summary.aspx
GA geography sourced from: http://www.georgiaencyclopedia.org/articles/geography-environment/geographic-regions-georgia-overview
NC geography sourced from:
http://www.ncpublicschools.org/curriculum/socialstudies/elementary/studentsampler/20geography
SC geography sourced from: http://sc.gov/Government/Local/Pages/localGovRegionOne.aspx
FL geography sourced from: http://www.landscope.org/florida/natural_geography/
VA geography sourced from: http://www.virginiaplaces.org/regions/physio.html

should also be noted that there were four cities that qualified for both the piedmont and coastal plain dummy variables as they exist on what is known as the fall line, or where the piedmont shifts quickly to coastal plain. We listed these MSAs under coastal plain as it is slightly more likely that the city started below the fall lines so products could continue down the river without interruption from the numerous waterfalls that gave name to the fall line.

Statistical Analysis

As previously mentioned, our overall hypothesis is that park access in the SALCC region can be explained by our proposed explanatory variables. In relation to this, our statistical analysis will test the general joint hypothesis about the slope regression coefficients is as follows: H_o: $\beta_1 = \beta_2 = \beta_j = 0$. Our alternative hypothesis is as follows, H_a: at least one β does not equal zero. In addition, we hypothesize for each explanatory variable that, H_o: β_o = no significant relationship between each of the explanatory variables and centroid distance. Alternatively, H_a: β_o = a significant relationship between each of the explanatory variables and centroid distance exists.

Park Accessibility Analysis

Initially, we collected data for the 24 explanatory variables that we presumed could help explain variation in centroid distance in the SALCC MSAs and ran an exploratory analysis by using different combinations to come up with the best model. The model that had thus far proven most significant contained the following variables: log of population density, percentage of high school or higher graduates, log of unemployment, log of no vehicle, mean commute time, public transit, log of percent renter, median age, coastal, piedmont, mountain, percent non-white and violent crimes. We chose the best model by looking at the following factors: 1) how well the assumptions were being met, 2) the highest R^2 and adjusted R^2 values, and 3) how significant each variable was. In the multiple linear regression model, the average relationship between the thirteen explanatory variables and the response variable (centroid distance), is given by the linear function below:

$$E\,(centroiddistance \mid X_j) = -194.69 + 3.06(logpopdensity) - 0.77(HSorHigher) - 4.54(logunemployed) - 11.96(lognovehicle) + 1.01(meancommutetime) - 2.63(publictransit) + 57.17(logpercentrenter) + 2.01(medianage) - 11.45(coastal) + 0.07(piedmont) - 1.03(mountain) - 0.02(percentnonwhite) + 0.01(violentcrimes)$$

As shown in Table 3a, this model gives an R^2 of 0.4985, which means that 41.94 percent of the variation in centroid distance is captured by the model. The associated adjusted R^2 is 0.2657, which is a relatively low number.

Based on the significance test presented in Table 3a, there are three explanatory variables that are found to be statistically significant: median age and coastal (at the $p<0.05$ level) and log of percent rent (at the $p<0.01$ level). Therefore, we can reject the null hypothesis in favor of the alternative and conclude that there is significant relationship between each of the three explanatory variables and centroid distance. From the F-test we conducted on the 42 observations, the result shows $F(13,28) = 2.14$, $p<0.05$; which means that at least one of the coefficient βs does not equal zero. We can thus reject the null hypothesis in favor of the alternative hypothesis (at least one β does not equal zero).

The interpretation of the log percent renter coefficient is that a one percentage increase in percentage of renters is associated with a 57.17 mile increase in centroid distance, holding all other variables constant. This large number and the fact that log of percent rent is significant at the $p<0.01$ level raises a number of questions and concerns. In all likelihood this variable is capturing unexplained variation in the model. The median age coefficient is interpreted that a one year increase in median age is associated with a 2.01 mile increase in centroid distance, holding all other variables constant. The coastal coefficient can be interpreted as: coastal cities are 11.45 miles further from the nearest qualifying park, holding all other variables constant.

According to the above significant variables the largest predictor of whether a large natural area exists nearby a MSA is the percentage of the population that rents as opposed to owns their dwelling. The median age of the population is also an interesting factor since it means

that the older the city's population is the more likely that there will be qualifying parkland nearby. This could be capturing the fact that larger cities are typically younger while outlying rural areas, which some MSAs capture, are greyer. Lastly the coastal cities having less parkland makes sense as well because these cities have less land to establish parks on. Based on this model, though there are many caveats, we suggest that SALCC focus its future conservation efforts on young coastal cities that have low ownership rates.

Park Availability Analysis

As previously mentioned we sought to explain the differing levels of available parkland in the MSAs by approaching the analysis with two measures, access and availability. In the previous model we demonstrated that the most significant explanatory variables were renter percentage, median age and whether the cities were on the coast or not. In order to see if there was a difference between the access measurements we ran the same explanatory variables, the results of which will be discussed later. The optimal model to explain the variation in park availability involved 13 variables: log of total population, log of population density, log of bachelor's degree, log of unemployed, log of per capita income, log of no vehicle, total drive time, log of percent rent, median age, coastal, mountain, piedmont, percent non-white. The linear expression produce by this multiple linear regression is as follows:

$$E\,(\textit{logparkpercent} \mid X_j) = -4.42 + 0.95(\textit{logtotalpop}) - 1.11(\textit{logpopdensity}) - 0.99(\textit{logbachelors}) + 0.43(\textit{logunemployed}) + 0.43(\textit{logpercapincome}) + 0.56(\textit{lognovehicle}) - 0.17(\textit{totaldrivetime}) + 0.01(\textit{logpercentrent}) - 0.09(\textit{medianage}) + 0.21(\textit{coastal}) - 0.42(\textit{piedmont}) + 0.11(\textit{mountain}) - 0.05(\textit{percentnonwhite})$$

As shown in Table 3b, this model gives an R^2 of 0.6398, which means that 63.98 percent of the variation in centroid distance is captured by the model. The associated adjusted R^2 is 0.4725. In comparison to the park access analysis, this model explains far more of the variation within park availability.

Based on the significance test presented in Table 3b, there are four explanatory variables that are found to be statistically significant: log of total population and percent non-white at the

$p<0.05$ level and log of population density and total drive time at the $p<0.01$ level. Therefore, we can reject the null hypothesis in favor of the alternative and conclude that there is significant relationship between each of the four explanatory variables and the response variable, log of percent park area. From the F-test we conducted on the 42 observations, the result shows $F(10,29) = 3.82$, $p<0.005$; which means that at least one of the coefficient βs does not equal zero. We can thus reject the null hypothesis in favor of the alternative hypothesis (at least one β does not equal zero).

The interpretation of the log population density coefficient is that a one percentage change in population density is associated with a 1.11% mile decrease in the percentage of MSA area that is parkland, holding all other variables constant. The log of total population coefficient is interpreted as a one percent increase in total population is associated with a 0.95% increase in in the percentage of MSA area that is parkland, holding all other variables constant. The total drive time coefficient can be interpreted as in the percentage of MSA area that is parkland is associated with decrease of 2.3% for each minute of additional total drive time holding all other variables constant. Lastly, the percent non-white coefficient can be interpreted by the percentage of parkland is associated with a decrease of 5% for each additional percent of non-white population, holding all other variables constant.

From this model we can expect that denser and less white cities likely have less parkland, while larger populations slightly offset this trend. Also the further one must drive to a qualifying park the less likely that there is a nearby qualifying park. This intuitively makes sense, thus the three significant variables are far more interesting from an analysis standpoint. It is our recommendation that SALCC focus future conservation efforts to combat these social injustices. The most effective way to close gaps in parkland availability would be to establish parks nearby nonwhite neighborhoods that are located in densely populated MSAs.

Availability and Accessibility Model Comparison

As mentioned previously we also compared our access analysis model with our availability model using the same explanatory variables to see if they match up. Refer to Table 3c for the breakdown of the comparison model. 50.8% percent of the variation in response

variable, log of park percentage is explained by this model. There is only one significant variable at the $p<.05$ level, log of population density, while median age is borderline in its significance. The interpretation of the log of population density coefficient is that, a one percent increase in population density is estimated to correspond with a 0.62% decrease of percent that parkland makes up of the total MSA area. This is interesting as log of population density has been significant in all three models, meaning it is the most likely the most significant variable when explaining both park access and availability.

Park Access & Availability Analysis Conclusions

Park access, through the independent variable centroid distance, can be explained by two significant factors: log of percent renter and median age of population. Generally this could mean that MSAs which have higher renting populations and are younger on average are likely to have less access to large parkland. Variation in park availability can generally be explained through four explanatory variables log of total population, percent non-white, log of population density and total drive time. Our park availability model indicates that MSAs with denser populations and larger populations of nonwhite residents will likely have less parkland available for recreation. The social justice aspects of these findings are not insignificant.

Conclusion

Our analysis is unique in that it provides a comprehensive list of urban areas in the SALCC region alongside two measures of their access to qualifying large natural areas – Euclidean distance and driving time. This was done through the development of a novel ArcGIS toolset that allows the user to tailor their analysis by specifying key parameters, such as the park size or the maximum distance to the park. These measures of access are complemented by key findings from the literature review and regression analysis, which together will help inform the SALCC as they consider future conservation actions to address the gaps identified in this report. We found that Albany, GA and Goldsboro, NC did not have access to a nearby large natural area, a result that was reflected in both the Euclidean distance and driving time metrics. Furthermore, our two park access models also supported and expanded these findings for our client.

Acknowledgments

We would like to acknowledge our advisor, Dr. Jennifer Swenson of the Nicholas School of the Environment at Duke University, and our client, Dr. Rua Mordecai of the South Atlantic Landscape Conservation Cooperative, for their invaluable expertise and assistance. Additional assistance with the GIS toolset was provided by John Fay, an Instructor and Research Associate at Duke University. Our statistical analysis is indebted to Dr. Elizabeth Albright, who donated hours of her day to make our regressions more robust. No financial assistance was received for this project, and we declare no competing interests.

Appendix A. Regression Results

Model Assumptions

Multiple linear regression models require that the data and outcomes meet these five assumptions: normality of residuals; linearity; residuals display a heteroskedastic pattern; that there is no perfect multicollinearity; and observations are independent. Meeting each of these assumptions is important for the overall validity of the model results. Should an assumption fail to be met, the results may be affected by bias and high standard error rates, while the subpopulations and coefficients could be less robust in the face of changes in the model.

Normality of Residuals: Meeting the normality of the residuals assumption is important because failing to meet it leads to bias in the data. Examining the residual versus fitted (RVF) plot in Figure 5, it is clear to see that there are few outliers, therefore they are unlikely to affect the outcomes. Confirming that the assumption has been met is done by running the Shapiro-Wilk's test, which is shown in Table 4. The null hypothesis for this test is that the data is normally distributed, and with a reported p-value of 0.75 we fail to reject the null hypothesis. It can be safely said that there is little to no bias in the model due to non-normal residuals.

Linearity: To meet the assumption of linearity, the residuals should be randomly and evenly scattered above and below the fitted value line. This assumption is tested by studying the RVF plot, please refer to Figure 5 for the referenced RVF plot. Per the RVF plot there appear to be no non-linear patterns present and the residuals do in fact appear to be randomly scattered, indicating the linearity assumption is met by the model. It would make little sense to use multiple linear regression to predict values if the patterns in the RVF plot were quadratic for example.

Heteroskedasticity: The assumption of heteroskedasticity means that there is an equal spread of the residuals around the fitted line, with little or no tapering at either end of the RVF plot. As shown in Figure 5, the RVF plot in this model results in a largely random scattering with little or no tapering. As a result, it can be said that there is likely little to no homoskedasticity present. The test for heteroskedasticity, shown in Table 6 has a null hypothesis that the spread is heteroskedastic, and with a reported p-value of 0.4427, we fail to reject the null hypothesis.

Should this test have failed homoskedasticity, it would affect the standard error in the model, and thus would manipulate the t-tests and confidence interval.

Multicollinearity: This assumption states that no two explanatory variables should be perfectly correlated with each other. Violating this assumption can lead to the coefficient estimates jumping around when variables are either removed or added to the model. Even before running any of the models we knew that there might be problems with there being high levels of multicollinearity amongst the explanatory variables. Due to the nature of park access as determined by distance from a central point in a MSA, variables like total area and total drive time are likely closely correlated with each other. Yet, our variance inflation factor (VIF) test reported an average VIF of 2.8, as shown in Table 5, lower than the agreed upon cutoff of four,. Thus the assumption is met in the broadest sense.

Independence of Observations: The independence of the observations is important for distinguishing whether or not spatial and temporal autocorrelation is present in the results. In this particular model the geospatial autocorrelation is most likely present due to the varying geographic and climatic regions across the SALCC region. In other words MSA data will appear more similar to each other if they are from the same region, and in this case they are all from one region in the world. In turn the error rates among MSAs in each region would more closely resemble other MSAs in their similar region. By adding a geography dummy variable we helped to diffuse the similarities between regions by grouping MSAs according to elevation, thus capturing the variation in this way. Regardless of the dummy variable, it must be acknowledged that there is likely spatial autocorrelation in the presented model. As there is only one year of data measured in this case there is no temporal autocorrelation.

Table 2. Summary Statistics

Variable:	Observations:	Mean:	Standard Deviation:	Minimum:	Maximum:	Variance:	Skewness:	Kurtosis:	Units:
Centroid Distance	42	14.89	9.77	0.01	38.88	95.36	0.39	2.53	Miles
Total Area	42	18,584.95	10,317.70	6921.15	59,348.67	106,000,000	1.93	7.64	Square Kilometers
Total Park Area	42	1228.58	1360.72	0	6313.76	1,851,570	2.16	7.83	Square Kilometers
Total Park Percentage	42	6.14	5.73	0	26.14	32.84	1.73	5.8	Percent
Total Drive Time	42	41.6	17.36	4.9	86.41	301.38	0.2	3.04	Minutes
Total Population	42	530,217.9	850,535.50	74,160	5,271,550	723,000,000,000	4.38	24.33	People
Unemployment	42	4.02	0.97	2.7	7.4	0.94	1.18	4.77	Percent
Public Transit	42	0.92	0.93	0	4.2	0.87	1.77	6.1	Percent
Mean Commute Time	42	22.29	2.2	18.4	30.7	4.82	1.52	6.61	Minutes
Median Household Income	42	45,125.29	5,862.19	36,278.00	58,111.00	34,400,000	0.66	2.44	2007 USD
Per Capita Income	42	23,726.55	3,144.66	16,593	30,072	9,888,867	0.16	2.31	2007 USD
Below Poverty Line	42	10.9	2.8	6.6	16.5	7.86	0.45	2.24	Percent
HS or Higher	42	83.89	3.94	72.5	89.3	15.52	-0.78	3.05	Percent
Bachelors Degree or Higher	42	25.05	7.17	14.1	41.8	51.36	0.61	2.58	Percent
Vacant Housing	42	13.17	2.51	7.7	34.8	25.2	2.51	10.33	Percent
Rented Housing	42	34	5.31	25.4	49.3	28.19	0.77	3.42	Percent
No Vehicle Available	42	6.89	1.61	4.3	11.5	2.58	0.74	3.41	Percent
Median Monthly Rent	42	701.88	100.06	544	921	10,010.94	0.29	2.17	2007 USD
Median Age	42	35.4	3.41	24.8	41.6	11.66	-1.07	4.01	Years
Percent White	42	67.57	10.25	47	87.5	105.04	0.01	2.22	Percent
Percent Non-White	42	32.43	10.25	12.5	53	105.04	-0.01	2.22	Percent
Population Density	42	124	98.52	30	448.9	9,706.33	1.91	6.4	People/Sq Km
Violent Crimes	42	407.52	145.68	135.9	716.9	21,222.11	0.11	2.4	Crimes per 100,000 Residents
Mean Rent	42	701.88	100.05	544	921	10,010.94	0.29	2.17	2007 USD
Log Transformations:									
Log of Total Area	42	9.71	0.48	8.84	10.99	0.23	0.43	2.99	Log of Sq Km
Log of Total Park Area	42	6.55	1.17	3.94	8.75	1.37	-0.35	2.49	Log of Sq Km
Log of Total Park Percentage	42	1.41	0.97	-0.97	3.26	0.95	-0.35	2.82	Log of Percentage
Log of Total Population	42	12.65	0.92	11.21	15.48	0.84	0.93	3.65	Log of People
Log of Unemployment	42	1.37	0.22	0.99	2	0.05	0.55	2.96	Log of Percentage
Log of Below Poverty Line	42	2.36	0.26	1.89	2.8	0.07	0.05	2.11	Log of Percentage
Log of Bachelors Degree or Higher	42	3.18	0.28	2.65	3.73	0.08	0.14	2.14	Log of Percentage
Log of Vacant Housing	42	2.53	0.3	2.04	3.55	0.09	1.26	5.3	Log of Percentage
Log of Rented Housing	42	3.52	0.15	3.24	3.9	0.02	0.38	2.87	Log of Percentage
Log of No Vehicle Available	42	1.9	0.23	1.46	2.44	0.05	0.17	2.65	Log of Percentage
Log of Per Capita Income	42	10.07	0.13	9.72	10.31	0.02	-0.12	2.57	Log of 2007 USD
Log of Population Density	42	4.58	0.68	3.4	6.11	0.47	0.32	2.74	Log of People/Sq Km

Table 3a. Centroid Model Results

Centroid Distance Model			
Source	Sum of Squares	DF	Mean of Squares
Model	1949.44	13	149.96
Residual	1961.03	28	70.04
Total	3910.47	41	95.38

Number of Observations	42
F (13,28)	2.14
Prob > F	0.0446
R-Squared	0.4985
Adjusted R-Squared	0.2657
Root MSE	8.3688

Centroid Distance	Coefficient	Standard Error	t	P>|t|	95% Confidence Interval	
Log of Population Density	3.058	3.283	1.08	0.289	-2.739	8.854
HS or Higher	-0.771	0.529	-1.46	0.156	-1.854	0.311
Log of Unemployed	-4.535	8.695	-0.52	0.606	-22.345	13.276
Log of No Vehicle	-11.595	8.875	-1.31	0.202	-29.774	6.584
Mean Commute Time	1.013	0.983	1.03	0.311	-1	3.026
Public Transit	-2.626	2.342	-1.12	0.272	-7.424	2.171
Log of Percent Rent	57.173	19.58	2.92	0.007	17.065	97.281
Median Age	2.01	0.723	2.78	0.01	0.529	3.491
Coastal	-11.449	4.603	-2.49	0.019	-20.879	-2.019
Coastal Plain	OMITTED					
Piedmont	0.073	4.865	0.01	0.988	-9.893	10.038
Mountain	-1.026	8.011	-0.13	0.899	-17.437	15.385
Percent Non-White	-0.023	0.255	-0.09	0.929	-0.545	0.499
Violent Crimes	0.011	0.011	1.06	0.3	-0.011	0.033
Constant	-194.694	103.503	-1.88	0.07	-406.709	17.322

Table 3b. Park Percentage Model Results

Total Park Percentage Optimal Model			
Source	Sum of Squares	DF	Mean of Squares
Model	26.03	13	2
Residual	14.66	28	0.52
Total	40.69	41	0.99

Number of Observations	42
F (10,29)	3.82
Prob > F	0.0014
R-Squared	0.6398
Adjusted R-Squared	0.4725
Root MSE	0.72355

Total Park Percentage	Coefficient	Standard Error	t	P>|t|	95% Confidence Interval	
Log of Total Population	0.95	0.35	2.72	0.011	0.233	1.667
Log of Population Density	-1.113	0.391	-2.85	0.008	-1.914	-0.312
Log of Bachelors or Higher	-0.988	1.13	-0.87	0.391	-3.312	1.335
Log of Unemployed	0.427	0.815	0.52	0.604	-1.242	2.096
Log of Per Capita Income	0.43	2.626	0.16	0.871	-4.949	5.809
Log of No Vehicle	0.56	0.705	0.79	0.434	-0.885	2.005
Total Drive Time	-0.023	0.008	-2.75	0.01	-0.04	-0.006
Log of Percent Rent	0.007	1.809	0	0.997	-3.7	3.713
Median Age	-0.091	0.069	-1.31	0.201	-0.233	0.051
Coastal	0.211	0.428	0.49	0.626	-0.665	1.087
Piedmont	-0.42	0.401	-1.05	0.304	-1.241	0.401
Mountain	0.106	0.622	0.17	0.866	-1.168	1.38
Percent Non-White	-0.05	0.2016	-2.3	0.029	-0.094	-0.006
Constant	-4.422	25.213	-0.18	0.862	-56.068	47.224

Table 3c. Comparison Model Results

Total Park Percentage Comparison Model			
Source	Sum of Squares	DF	Mean of Squares
Model	20.67	13	1.59
Residual	20.02	28	0.72
Total	40.69	41	0.99

Number of Observations	42
F (10,29)	2.22
Prob > F	0.0373
R-Squared	0.508
Adjusted R-Squared	0.2795
Root MSE	0.84561

Total Park Percentage	Coefficient	Standard Error	t	P>|t|	95% Confidence Interval	
Log of Population Density	-0.621	0.286	-2.17	0.038	-1.207	-0.035
HS or Higher	0.071	0.053	1.33	0.194	-0.038	0.181
Log of Unemployed	1.079	0.879	1.23	0.229	-0.72	2.879
Log of No Vehicle	1.116	0.897	1.24	0.224	-0.721	2.952
Mean Commute Time	0.138	0.099	1.39	0.175	-0.065	0.342
Public Transit	0.056	0.237	0.24	0.816	-0.429	0.54
Log of Percent Rent	-2.931	1.978	-1.48	0.15	-6.984	1.121
Median Age	-0.142	0.073	-1.94	0.062	-0.291	0.008
Coastal	0.727	0.465	1.56	0.129	-0.226	1.68
Piedmont	-0.344	0.492	-0.7	0.49	-1.351	0.663
Mountain	0.255	0.81	0.31	0.755	-1.404	1.913
Percent Non-White	-0.043	0.026	-1.69	0.103	-0.096	0.009
Violent Crimes	0.0003	0.001	0.26	0.801	-0.002	0.003
Constant	7.018	10.459	0.67	0.508	-14.405	28.441

Table 4a. Centroid Distance Shapiro-Wilk's Test Results

Shapiro-Wilk's Test	Normality Assumption	Results			
Variable	Observations	Weight	Value	Z-Statistic	Prob > Z
Residual	42	0.982	0.728	-0.669	0.74813

Table 4b. Total Park Percentage Shapiro-Wilk's Test Results

Shapiro-Wilk's Test	Normality Assumption	Results			
Variable	Observations	Weight	Value	Z-Statistic	Prob > Z
Residual	42	0.984	0.672	-0.839	0.79921

Table 4c. Comparison Model Shapiro-Wilk's Test Results

Shapiro-Wilk's Test	Normality Assumption	Results			
Variable	Observations	Weight	Value	Z-Statistic	Prob > Z
Residual	42	0.984	0.659	-0.879	0.81043

Table 5a. Centroid Distance Multicollinearity Test Results

Variance Inflation Factor (VIF) Test:	Multicollinearity Assumption	
	VIF	1/VIF
Log of Percent Rent	5.13	0.195
Percent Non-White	3.99	0.251
Median Age	3.57	0.28
Piedmont	2.9	0.345
Public Tranist	2.78	0.36
Mean Commute Time	2.73	0.367
Mountain	2.55	0.392
HS or Higher	2.54	0.394
Log of No Vehicle	2.38	0.42
Log of Unemployed	2.23	0.449
Coastal	2.14	0.467
Log of Population Density	2.11	0.474
Violent Crimes	1.38	0.723
Mean VIF	2.8	

Table 5b. Total Park Percentage Multicollinearity Test Results

Variance Inflation Factor (VIF) Test:	Multicollinearity Assumption	
	VIF	1/VIF
Log of Per Capita Income	9.6	0.104
Log of Total Population	8.04	0.124
Log of Bachelors or Higher	7.94	0.126
Log of Percent Rent	5.86	0.171
Log of Population Density	5.39	0.186
Median Age	4.4	0.227
Percent Non-White	3.82	0.262
Piedmont	2.63	0.38
Log of Unemployed	2.62	0.382
Coastal	2.47	0.405
Mountain	2.06	0.486
Log of No Vehicle	2.01	0.497
Total Drive Time	1.63	0.613
Mean VIF	4.5	

Table 5c. Comparison Model Multicollinearity Test Results

Variance Inflation Factor (VIF) Test:	Multicollinearity Assumption	
	VIF	1/VIF
Log of Percent Rent	5.13	0.195
Percent Non-White	3.99	0.251
Median Age	3.57	0.28
Piedmont	2.9	0.345
Public Transit	2.78	0.359
Mean Commute Time	2.73	0.367
Mountain	2.55	0.392
HS or Higher	2.54	0.394
Log of No Vehicle	2.38	0.42
Log of Unemployed	2.23	0.449
Coastal	2.14	0.467
Log of Population Density	2.11	0.474
Violent Crimes	1.38	0.723
Mean VIF	2.78	

Table 6a. Centroid Distance Heteroskedasticity Test Results

Breusch-Pagan / Cook-Weisberg Test	Heteroskedasticity Assumption
Chi^2 Score	0.59
Prob > Chi^2	0.4427

Table 6b. Total Park Percentage Heteroskedasticity Test Results

Breusch-Pagan / Cook-Weisberg Test	Heteroskedasticity Assumption
Chi^2 Score	2.06
Prob > Chi^2	0.1512

Table 6c. Comparison Model Heteroskedasticity Test Results

Breusch-Pagan / Cook-Weisberg Test	Heteroskedasticity Assumption
Chi^2 Score	0.01
Prob > Chi^2	0.9135

Figure 5a. Centroid Distance RVF Plot

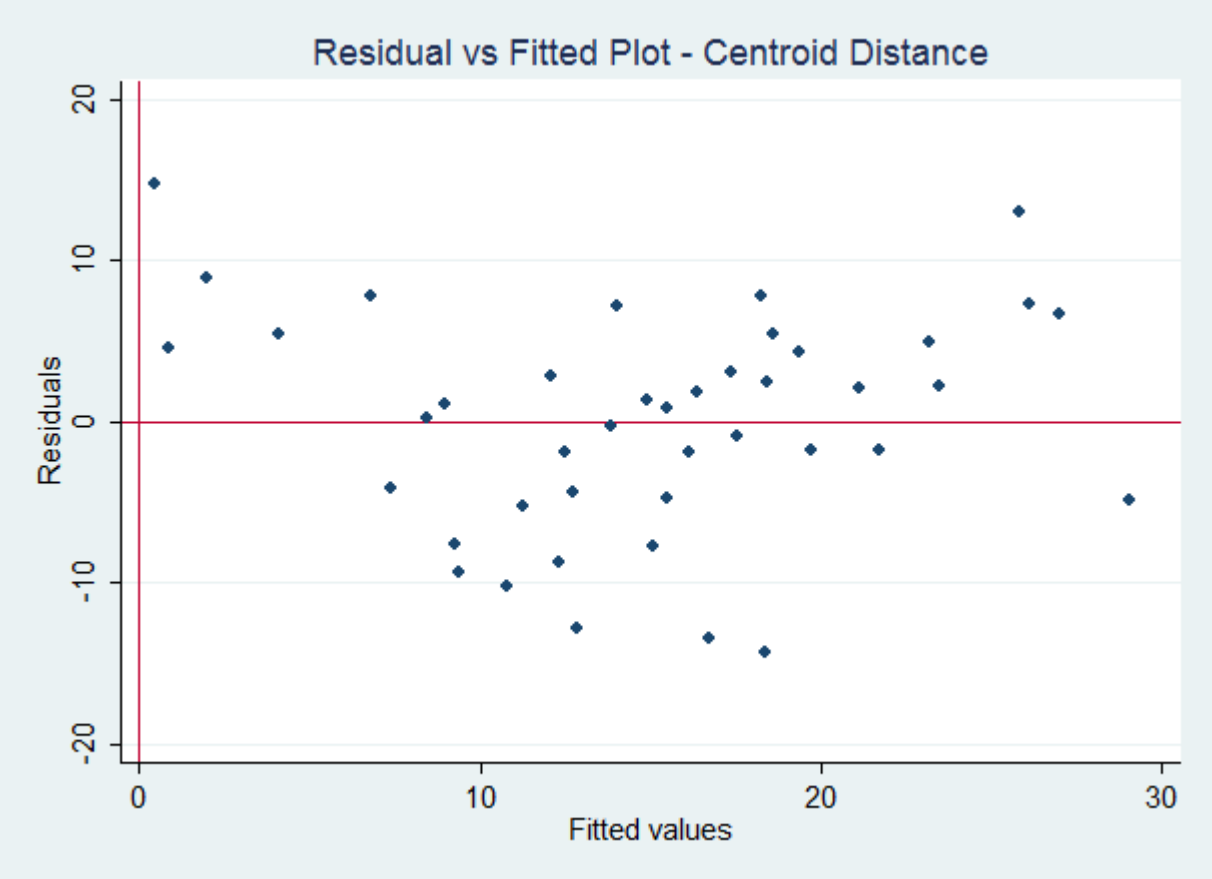
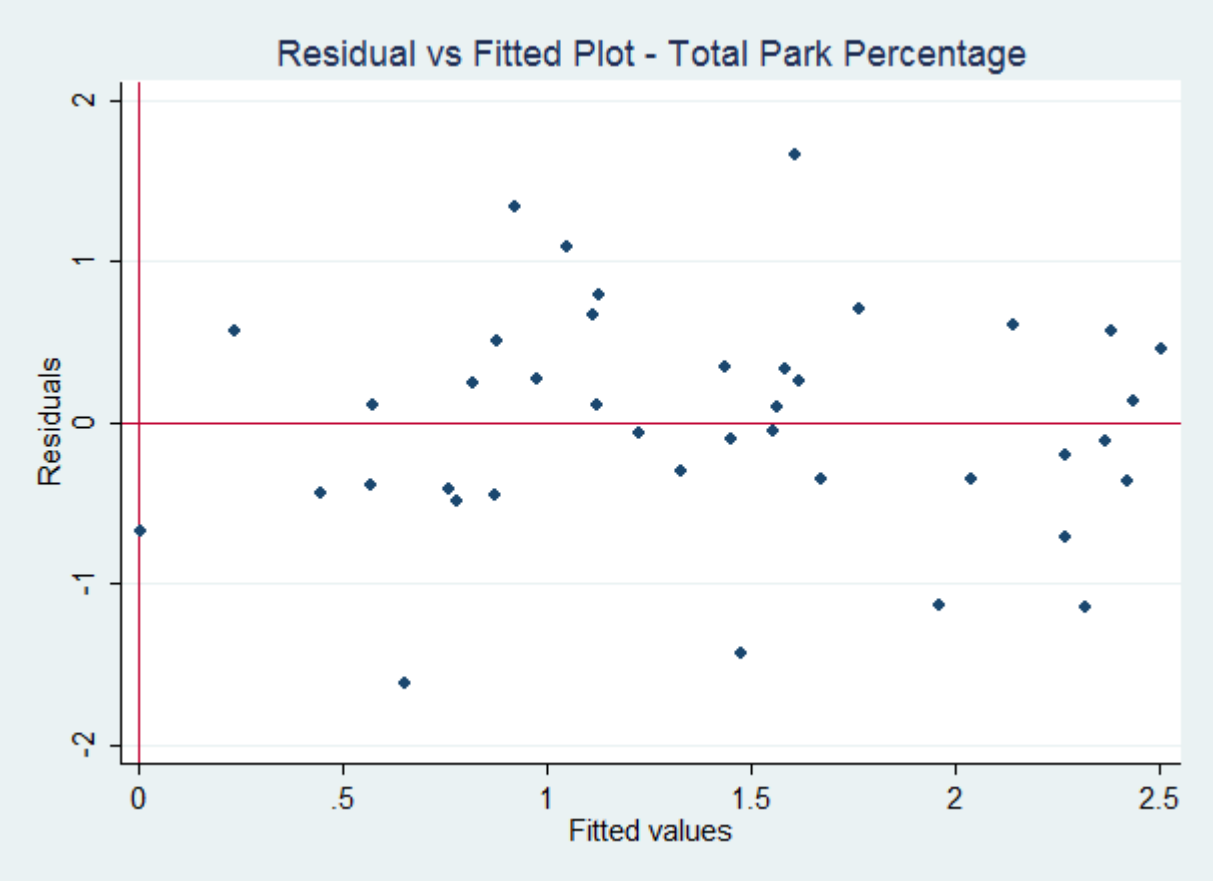


Figure 5b. Total Park Percentage RVF Plot



Appendix B. Geospatial Results

Table 7. Euclidean Distance to Nearest Park by MSA

Metropolitan Statistical Area	Miles from Edge	Miles from Centroid
Albany, GA	23.2	33.4
Athens-Clarke County, GA	0	10.1
Atlanta-Sandy Springs-Roswell, GA	0	16.4
Auburn-Opelika, AL	4.6	16.3
Augusta-Richmond County, GA-SC	0	8.4
Brunswick, GA	0	0
Burlington, NC	10.3	24.2
Charleston-North Charleston, SC	0	9.5
Charlotte-Concord-Gastonia, NC-SC	0	26.0

Columbia, SC	0	20.5
Columbus, GA-AL	0	4.1
Durham-Chapel Hill, NC	0	7.4
Fayetteville, NC	0	18.2
Florence, SC	0	23.7
Gainesville, FL	0	14.9
Gainesville, GA	1.1	18.0
Goldsboro, NC	23.5	38.9
Greensboro-High Point, NC	0	28.1
Greenville, NC	10.8	25.7
Greenville-Anderson-Mauldin, SC	0	23.2
Hickory-Lenoir-Morganton, NC	0	16.6
Hilton Head Island-Bluffton-Beaufort, SC	0	15.2
Hinesville, GA	0	10.6
Jacksonville, FL	0	1.6
Jacksonville, NC	0	5.4
Lynchburg, VA	0.2	21.2
Macon, GA	0	10.8
Myrtle Beach-Conway-North Myrtle Beach, SC-NC	0	14.6
New Bern, NC	0	3.6
Raleigh, NC	0	14.2
Richmond, VA	7.8	24.0
Roanoke, VA	0.1	13.6
Rocky Mount, NC	8.5	33.7
Savannah, GA	0	10.8
Spartanburg, SC	0	0.6
Sumter, SC	4.6	19.9
Tallahassee, FL	0	3.3
Valdosta, GA	0	8.6
Virginia Beach-Norfolk-Newport News, VA-NC	0	0

Warner Robins, GA	0	3.3
Wilmington, NC	0	6.1
Winston-Salem, NC	0	20.9

Table 8. Driving Time to Nearest Park by MSA

Metropolitan Statistical Area	Minutes from Centroid
Albany, GA	78.7
Athens-Clarke County, GA	43.8
Atlanta-Sandy Springs-Roswell, GA	46.9
Auburn-Opelika, AL	30.0
Augusta-Richmond County, GA-SC	33.7
Brunswick, GA	10.2
Burlington, NC	54.0
Charleston-North Charleston, SC	35.3
Charlotte-Concord-Gastonia, NC-SC	53.5
Columbia, SC	51.0
Columbus, GA-AL	22.9
Durham-Chapel Hill, NC	33.0
Fayetteville, NC	56.2
Florence, SC	66.1
Gainesville, FL	45.1
Gainesville, GA	49.3
Goldsboro, NC	86.4
Greensboro-High Point, NC	53.0
Greenville, NC	51.5
Greenville-Anderson-Mauldin, SC	48.3
Hickory-Lenoir-Morganton, NC	45.4
Hilton Head Island-Bluffton-Beaufort, SC	41.7
Hinesville, GA	25.0
Jacksonville, FL	14.3

Jacksonville, NC	38.6
Lynchburg, VA	58.9
Macon, GA	27.7
Myrtle Beach-Conway-North Myrtle Beach, SC-NC	48.3
New Bern, NC	35.6
Raleigh, NC	36.1
Richmond, VA	56.8
Roanoke, VA	32.5
Rocky Mount, NC	58.4
Savannah, GA	29.2
Spartanburg, SC	16.2
Sumter, SC	55.6
Tallahassee, FL	28.7
Valdosta, GA	26.6
Virginia Beach-Norfolk-Newport News, VA-NC	4.9
Warner Robins, GA	32.1
Wilmington, NC	24.4
Winston-Salem, NC	61.3

Appendix C. Geospatial Model

Toolset Overview

Within the toolset, users begin by running a tool that prepares the data for a specified area of the country if they are not analyzing the South Atlantic LCC. Next, they can choose either the Buffer Analysis tool or the Network Analysis tool. In the Buffer Analysis tool, they specify the minimum park size and the maximum distance from the edge of a MSA; in the Network Analysis tool, they specify the minimum park size and the maximum driving time from the MSA centroid. After running one of these tools, a table is saved with distance (in miles or minutes) to the closest qualifying large natural area for each MSA alongside additional descriptive statistics.

Workflow

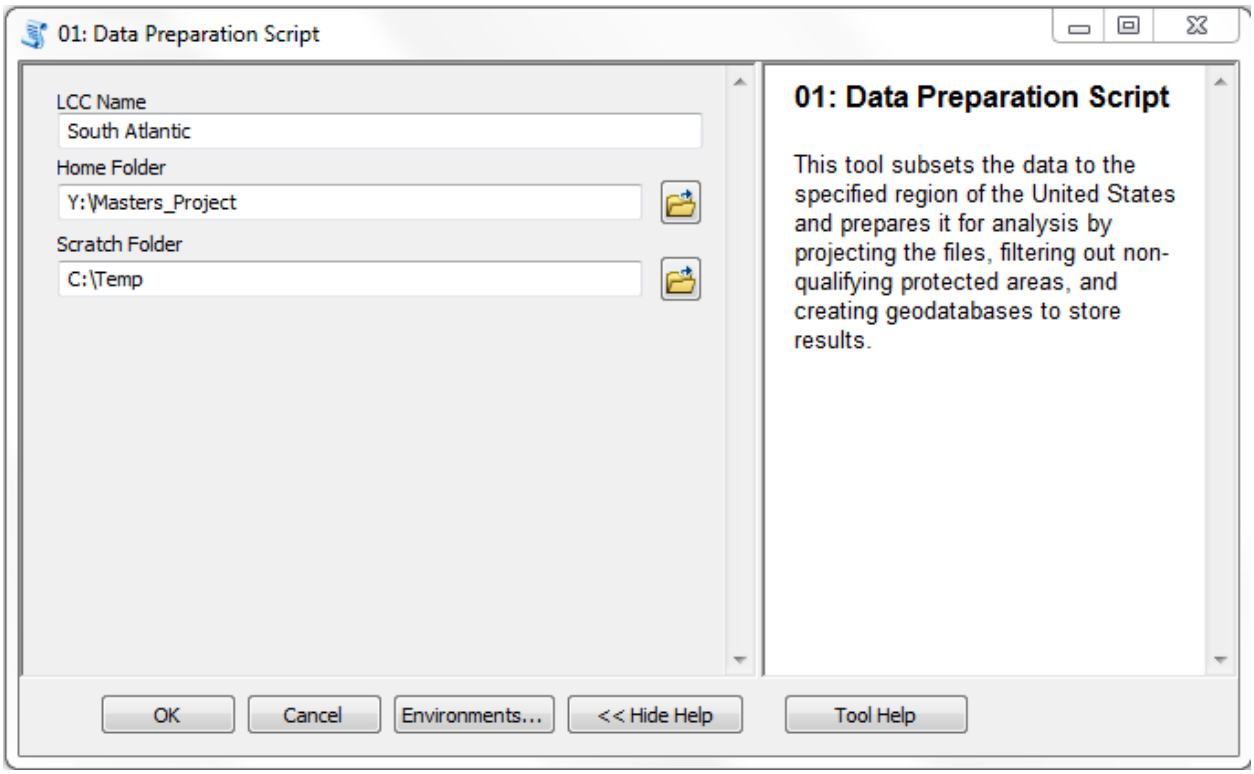
This toolset was designed for the South Atlantic Landscape Conservation Cooperative in order to inform their long-term conservation planning by identifying gaps in access to large natural areas from urban areas. As such, the tool is distributed with the data prepared for analysis in this region. Data is distributed for the entire United States, however, so users may first run the Data Preparation script and indicate one of the other 18 LCCs that are entirely or partially within the nation's borders[16]. After this, all users have the option of running either the Buffer Analysis or Network Analysis tools; they do not build on each other and do not need to be run sequentially, but hopefully complement each other by offering different measures of accessibility. Tools can be accessed by opening the SALCC_Map.mxd map document and then by double-clicking on the desired tool within the SALCC_Tools.tbx toolbox. Before running the Network Analysis tool, users should ensure that they have enabled the Network Analyst extension to ArcMap.

01: Data Preparation Script

This tool subsets the data to the specified region of the United States and prepares it for analysis by projecting the files, filtering out non-qualifying protected areas, and creating geodatabases to store results. As stated previously, it does not need to be run before analysis begins unless if one is planning on analyzing a different region of the United States.

[16] U.S. Fish and Wildlife Service. Land Conservation Cooperatives. March, 2013. http://www.fws.gov/GIS/data/national/index.html#LCC

Figure 6. Data Preparation Script

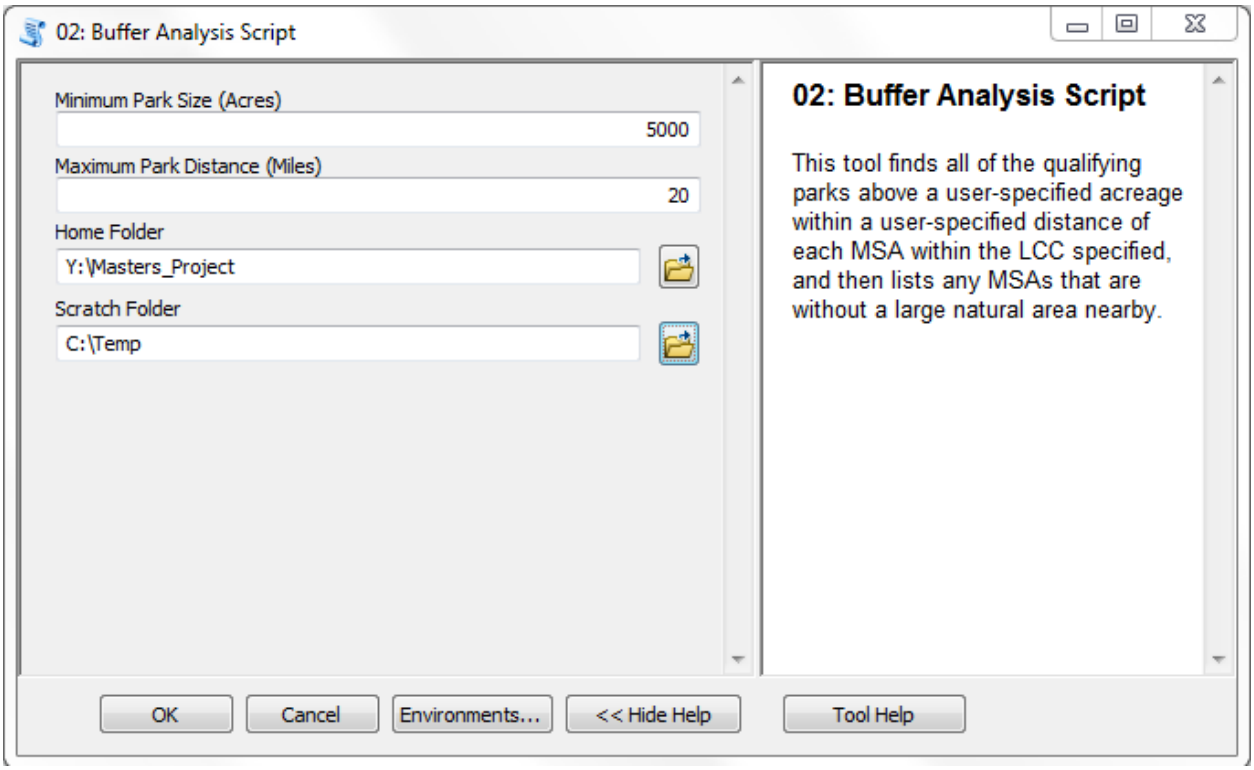


There are three user inputs to this tool: the name of the Landscape Conservation Cooperative (LCC), the home folder in which the toolbox (and data) resides, and a scratch folder in which to store temporary files. 'South Atlantic' is the default LCC; for other options, please consult the LCC names in the shapefile, while noting that as the PAD-US dataset is limited to the United States, an analysis on any LCC that is not entirely within the U.S. will contain errors.

02: Buffer Analysis Script

This tool finds all of the qualifying parks above a user-specified acreage within a user-specified distance of each MSA within the LCC specified, and then lists any MSAs that are without a large natural area nearby. First, a buffer is created around each MSA boundary based on the user-specified distance threshold, and any protected areas smaller than the user specified acreage are filtered out. Next, the script finds the intersection of the buffered MSAs and any qualified protected areas, and this information is saved to a table. Finally, any MSAs that do not have a qualifying protected area nearby are printed to the screen which serves as a first step so that one can look more closely at why these gaps in access exist for particular MSAs.

Figure 7. Buffer Analysis Script

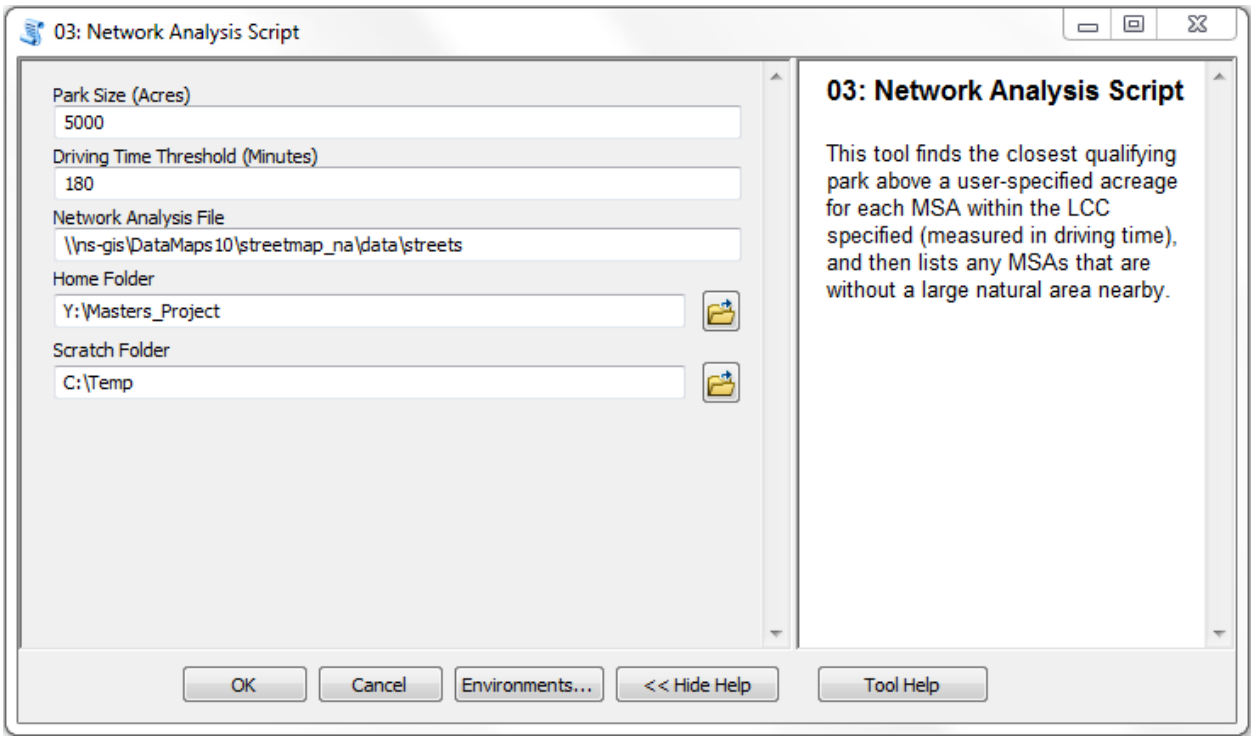


There are four user inputs to this tool: the distance threshold (in miles), the park area threshold (in acres), the home folder in which the toolbox (and data) resides, and a scratch folder in which temporary files will be stored. The default values for the first two variables are 20 miles and 5000 acres, which were the parameters specified by our client, the SALCC. Users are encouraged to explore other parameter values in order to see how they influence accessibility.

03: Network Analysis Script

This tool finds the closest qualifying park above a user-specified acreage for each MSA within the LCC specified, as measured by driving time. First, any protected areas smaller than the user specified acreage are filtered out, and the centroids of the MSAs and the filtered protected areas are generated. Next, a Closest Facility layer is created, with the MSA centroids loaded as incidents and the filtered protected area centroids as facilities. Finally, the network analysis is solved and the driving times (and routes) are generated; routes are stored in the Results database, and may be visualized by adding them to the map. These driving times complement the distances calculated previously, with the hope that together they provide a more nuanced picture of how access to large natural areas differ between different urban areas.

Figure 8. Network Analysis Script



There are five user inputs to this tool: the park area threshold (in acres), the driving time threshold (in minutes), the location of the street network dataset, the home folder in which the toolbox (and data) resides, and a scratch folder in which temporary files will be stored. 5000 acres is again used as the default park area threshold, while 180 minutes is used as the maximum trip time after which the program no longer investigates a route. While the rest of the data is provided along with the tool, the street network dataset is not due to its prohibitive size. Users operating from Duke's Nicholas School of the Environment can access the dataset listed at the tool's default location; other users will have to specify the file path of the street network dataset provided to them by ESRI.

Appendix D. Python Scripts

Script 1. Data Preparation

```
# ---------------------------------------------------------------------------
# Data_Prep.py
# Created by Dan Plechaty, updated December 2014
# Description: Prepares data for further analysis; the user chooses which
#           LCC to analyse, and the data is projected and subset to this
area.
# ---------------------------------------------------------------------------
```

```
# Import system modules
import arcpy, sys
from arcpy import env

# Prompt for user inputs
lcc_name = sys.argv[1]       # default is 'South Atlantic'; would need to
rerun this for a different region
home_folder = sys.argv[2]    # default is "Y:\SALCC"
scratch_folder = sys.argv[3] # default is "C:\Temp"

# Local variables:
data_folder = home_folder + "\Data"
base_data = data_folder + "\Base_Data.mdb"
scratch_data = scratch_folder + "\Scratch.mdb"
result_data = data_folder + "\Results.mdb"
protected_areas = base_data + "\Protected_Areas"
msa_areas = base_data + "\MSA_Areas"
fws_lcc = data_folder + "\\fws_lcc.shp"
tl_2013_us_cbsa = data_folder + "\\tl_2013_us_cbsa.shp"
PADUSCBIEdition_v2 = data_folder + "\\PADUSCBIEdition_v2.shp"
arcpy.AddMessage("Warning: This tool is processor intensive, and can take
upwards of 20 minutes to run...")

# Create geodatabases
arcpy.AddMessage("Creating geodatabases in which to store results...")
arcpy.Delete_management(scratch_data)
arcpy.Delete_management(base_data)
arcpy.Delete_management(result_data)
arcpy.CreatePersonalGDB_management(scratch_folder, "Scratch.mdb")
arcpy.CreatePersonalGDB_management(data_folder, "Base_Data.mdb")
arcpy.CreatePersonalGDB_management(data_folder, "Results.mdb")

# Set environment settings
env.workspace = scratch_data
env.overwriteOutput = True

# Create shapefiles of the selected LCC and the MSA boundaries
arcpy.AddMessage("Creating LCC and MSA shapefiles...")
where_clause = "area_names = '" + lcc_name + "'"
arcpy.Select_analysis(fws_lcc, "lcc_mask", where_clause)
arcpy.Select_analysis(tl_2013_us_cbsa, "msa_select", "\"MEMI\" = '1'")

# Project the MSAs, LCC mask and Protected Areas
spatial_reference = arcpy.SpatialReference(3857) # WGS 1984 Web Mercator
(Auxiliary Sphere) - used for all data (except NA Layer)

dropFields = ["CSAFP", "CBSAFP", "GEOID", "NAMELSAD","LSAD", "MEMI", "MTFCC",
"ALAND", "AWATER", "INTPTLAT", "INTPTLON"]
arcpy.DeleteField_management("msa_select", dropFields)
```

```
arcpy.AddMessage("Projecting MSAs...")
arcpy.Project_management("msa_select", "project_MSAs", spatial_reference)

dropFields = ["Area_Num", "Shape_Leng", "Shape_Le_1"]
arcpy.DeleteField_management("lcc_mask", dropFields)
arcpy.AddMessage("Projecting LCC mask...")
arcpy.Project_management("lcc_mask", "project_LCC", spatial_reference)

arcpy.AddMessage("Projecting protected areas...")
arcpy.Project_management(PADUSCBIEdition_v2, "project_PAD",
spatial_reference)
dropFields = ["gis_scr", "scr_date", "comments", "gis_acres", "Shape_Leng",
"Shape_Le_1"]
arcpy.DeleteField_management("project_PAD", dropFields)

# Clip out subsets of the MSAs using the LCC mask, and protected areas using
a buffered version of the LCC mask
arcpy.AddMessage("Clipping subsets of protected areas and MSAs...")
arcpy.Clip_analysis("project_MSAs", "project_LCC", msa_areas, "")
arcpy.Buffer_analysis("project_LCC", "project_LCC_buffer", "50 Miles",
"FULL", "ROUND", "NONE", "")
arcpy.Clip_analysis("project_PAD", "project_LCC_buffer", "pad_multi", "") # a
buffer is used in case an MSAs closest park is outside of the LCC's boundary

# Split multipart polygons and calculate their areas
arcpy.AddMessage("Splitting multipart polygons...")
arcpy.MultipartToSinglepart_management("pad_multi", protected_areas)
arcpy.AddMessage("Calculating acreage...")
arcpy.AddField_management(protected_areas, "Acreage", "DOUBLE", "", "", "",
"Contiguous Park Area in Acres", "NULLABLE", "REQUIRED", "")
arcpy.CalculateField_management(protected_areas, "Acreage",
"!shape.geodesicArea@acres!", "PYTHON", "")

# Clean up and present final instructions
arcpy.Delete_management(scratch_data)
arcpy.AddMessage("Data Preparation Complete!")
arcpy.AddMessage("You may now proceed to using either the buffer or network
analysis scripts.")
```

Script 2. Buffer Analysis

```
# ---------------------------------------------------------------------------
# Buffer_Analysis.py
# Created by Dan Plechaty, Updated December 2014
# Description: After the initial data processing is performed, this script
#              queries the user for inputs and based off of these responses performs
#              a buffer analysis around the MSAs and finds the protected areas that
#              intersect these MSAs after filtering.
# ---------------------------------------------------------------------------
```

```
# Import system modules
import arcpy, sys
from arcpy import env

# Prompt for user inputs
park_size = sys.argv[1]      # default is 5000 (Acres)
park_distance = sys.argv[2]  # default is 20 (Miles)
home_folder = sys.argv[3]    # default is "Y:\SALCC"
scratch_folder = sys.argv[4] # default is "C:\Temp"

# Local variables:
data_folder = home_folder + "\Data"
base_data = data_folder + "\Base_Data.mdb"
result_data = data_folder + "\Results.mdb"
scratch_data = scratch_folder + "\Scratch.mdb"
protected_areas = base_data + "\Protected_Areas"
msa_areas = base_data + "\MSA_Areas"
arcpy.CreatePersonalGDB_management(scratch_folder, "Scratch.mdb")

# Set environment settings
env.workspace = scratch_data
env.overwriteOutput = True

# Create euclidean buffer based on user input
park_distance_miles = park_distance + " Miles"
buffer_output = scratch_data + "\\buffer_" + park_distance
arcpy.AddMessage("Buffering MSAs...")
arcpy.Buffer_analysis(msa_areas, buffer_output, park_distance_miles, "FULL", "ROUND",
"NONE", "")

# Filter protected areas based on user input
where_clause = "[Acreage] >= " + park_size + " AND NOT [own_type] = 'Private Land' AND
NOT [s_des_tp] = 'Military Reservation' AND NOT [p_des_tp] = 'Private - Unprotected' AND
NOT [p_des_tp] = 'Military Reservation' AND NOT [p_des_tp] = 'Army Corps of Engineers
Land/Water' AND NOT [p_loc_ds] = 'Federal Hydroelectric Plant' AND NOT [p_loc_ds] =
'State Hydroelectric Project'"
arcpy.AddMessage("Filtering protected areas...")
arcpy.Select_analysis(protected_areas, "pad_filtered", where_clause)

# Find the intersection between the buffered MSAs and the filtered protected areas
arcpy.AddMessage("Finding intersection of MSAs and protected areas...")
arcpy.SpatialJoin_analysis(buffer_output, "pad_filtered", "msa_pad_intersection",
"JOIN_ONE_TO_MANY", "KEEP_ALL", "", "INTERSECT", "", "")
dropFields = ["Join_Count", "BUFF_DIST", "Shape_Length_1", "Shape_Area_1",
"ORIG_FID", "ORIG_FID_1"]
```

```
arcpy.DeleteField_management("msa_pad_intersection", dropFields)

# Export tables to the Results geodatabase and an Excel file
out_table = "Buffer_Results_" + park_distance + "Miles_" + park_size + "Acres"
out_table_path = result_data + "\\" + out_table
arcpy.AddMessage("Saving output table to " + out_table_path)
arcpy.TableToTable_conversion("msa_pad_intersection", result_data, out_table, "", "", "")
input_table = result_data + "\\" + out_table
output_excel = home_folder + "\\Docs\\Buffer_Results_" + park_distance + "Miles_" +
park_size + "Acres.xls"
arcpy.AddMessage("Saving excel file to " + output_excel)
arcpy.TableToExcel_conversion(input_table, output_excel, "ALIAS", "CODE")

# Display which MSAs (if any) do not have a qualifying large natural area
rows = arcpy.SearchCursor(out_table_path)
row = rows.next()

lacking_msas = 0
msa_list = []
while row:
  join_id = row.getValue("JOIN_FID")
  if join_id == -1:
    if lacking_msas == 0:
      arcpy.AddMessage("There are no qualifying large natural areas in:")
    arcpy.AddMessage(row.getValue("NAME"))
    lacking_msas = lacking_msas + 1
  row = rows.next()

if lacking_msas == 0:
  arcpy.AddMessage("All of the MSAs had a qualifying large natural area.")

# Clean up and present final instructions
arcpy.Delete_management(scratch_data)
del row, rows
arcpy.AddMessage("Analysis Complete!")
```

Script 3. Network Analysis

```
# ---------------------------------------------------------------------------
# Network_Analysis.py
# Created by Dan Plechaty, Updated December 2014
# Description: After the initial data processing is performed, this script
#           queries the user for inputs and based off of these responses
performs
#           a network analysis from MSA centroids and finds the protected
areas that
#           are within a specified driving time after filtering.
```

```
# ---------------------------------------------------------------------------

# Import system modules
import arcpy, sys
from arcpy import env

# Prompt for user inputs
park_size = sys.argv[1]      # default is 5000 (Acres)
park_minutes = sys.argv[2]   # default is 180 (driving time cutoff, in
minutes)
network_data = sys.argv[3]   # default is "\\ns-
gis\DataMaps10\streetmap_na\data\streets"
home_folder = sys.argv[4]    # default is "Y:\SALCC"
scratch_folder = sys.argv[5] # default is "C:\Temp"

# Local variables:
data_folder = home_folder + "\Data"
base_data = data_folder + "\Base_Data.mdb"
result_data = data_folder + "\Results.mdb"
scratch_data = scratch_folder + "\Scratch.mdb"
protected_areas = base_data + "\Protected_Areas"
msa_areas = base_data + "\MSA_Areas"
arcpy.CreatePersonalGDB_management(scratch_folder, "Scratch.mdb")

# Set environment settings
env.workspace = scratch_data
env.overwriteOutput = True
arcpy.CheckOutExtension("Network")

# Filter protected areas based on user input
arcpy.AddMessage("Warning: Driving times are approximate, and do not include
the effects of construction or traffic.")
where_clause = "[Acreage] >= " + park_size + " AND NOT [own_type] = 'Private
Land' AND NOT [s_des_tp] = 'Military Reservation' AND NOT [p_des_tp] =
'Private - Unprotected' AND NOT [p_des_tp] = 'Military Reservation' AND NOT
[p_des_tp] = 'Army Corps of Engineers Land/Water' AND NOT [p_loc_ds] =
'Federal Hydroelectric Plant' AND NOT [p_loc_ds] = 'State Hydroelectric
Project'"
arcpy.AddMessage("Filtering protected areas...")
arcpy.Select_analysis(protected_areas, "pad_filtered", where_clause)

# Calculate MSA and park centroids to use in network analysis
arcpy.AddMessage("Calculating MSA and park centroids...")
arcpy.FeatureToPoint_management(msa_areas, "centroids", "CENTROID")
arcpy.FeatureToPoint_management("pad_filtered", "pad_points", "CENTROID")

# Construct network analysis object
arcpy.AddMessage("Constructing network analysis dataset...")
arcpy.MakeClosestFacilityLayer_na(network_data, "Closest Facility", "Time",
"TRAVEL_TO", park_minutes, "1", "", "ALLOW_DEAD_ENDS_AND_INTERSECTIONS_ONLY",
```

```
"'Non-routeable Segments';OneWay", "USE_HIERARCHY", "",
"TRUE_LINES_WITHOUT_MEASURES", "", "")

# Load locations of MSA and park centroids
arcpy.AddMessage("Loading locations of MSA and park centroids...")
arcpy.AddLocations_na("Closest Facility", "Incidents", "centroids", "Name
Name #", "5000 Meters", "", "'SDC Edge Source' SHAPE", "MATCH_TO_CLOSEST",
"APPEND", "NO_SNAP", "5 Meters", "INCLUDE", "'SDC Edge Source' #")
arcpy.AddLocations_na("Closest Facility", "Facilities", "pad_points", "Name
p_des_nm #", "15 Kilometers", "OBJECTID", "'SDC Edge Source' SHAPE",
"MATCH_TO_CLOSEST", "APPEND", "NO_SNAP", "5 Meters", "INCLUDE", "'SDC Edge
Source' #")

# Solve network analysis
arcpy.AddMessage("Performing network analysis...")
arcpy.Solve_na("Closest Facility", "SKIP", "TERMINATE", "")

# Export tables to the Results geodatabase and an Excel file
out_table = "NA_Results_" + park_minutes + "Minutes_" + park_size + "Acres"
out_table_path_sort = result_data + "\\" + out_table
out_table_path = out_table_path_sort + "_presort"
arcpy.AddMessage("Saving output table to " + out_table_path_sort)
arcpy.CopyFeatures_management("Closest Facility\\Routes", out_table_path, "",
"0", "0", "0")
dropFields = ["FacilityID", "FacilityRank", "IncidentCurbApproach",
"FacilityCurbApproach", "IncidentID"]
arcpy.DeleteField_management(out_table_path, dropFields)
arcpy.Sort_management(out_table_path, out_table_path_sort, [["Total_Time",
"DESCENDING"]])
output_excel = home_folder + "\\Docs\\NA_Results_" + park_minutes +
"Minutes_" + park_size + "Acres.xls"
arcpy.AddMessage("Saving excel file to " + output_excel)
arcpy.TableToExcel_conversion(out_table_path_sort, output_excel, "ALIAS",
"CODE")

# Display which MSAs (if any) do not have a large natural area within an hour
drive
rows = arcpy.SearchCursor(out_table_path_sort)
row = rows.next()

lacking_msas = 0
msa_list = []
while row:
  time = row.getValue("Total_Time")
  name = row.getValue("NAME")
  if time >= 60:
    if lacking_msas == 0:
      arcpy.AddMessage("There are no large natural areas within an hour drive
from:")
    message = name + " is a %s minute drive away."
```

```
    arcpy.AddMessage(message % time)
    lacking_msas = lacking_msas + 1
  row = rows.next()

if lacking_msas == 0:
  arcpy.AddMessage("All of the MSAs had a large natural area within an hour
drive.")

# Clean up and present final instructions
arcpy.Delete_management(scratch_data)
arcpy.Delete_management(out_table_path)
del row, rows
arcpy.AddMessage("Analysis Complete!")
```