5-2013

Census Data Extractor A toolset to select and save U.S. Census data

Mairead Siobhan Rauch

Census Data Extractor
A toolset to select and save U.S. Census data

by

Mairead Siobhan Rauch
Candidate for Bachelor of Science
Department of Environmental Resources Engineering
With Honors
May 2013

Approved

Thesis Project Advisor: _____

Charles N. Kroll, P.E., Ph.D.

Second Reader: _____

Stewart A. Diemont, Ph.D.

Honors Director: _____

William M. Shields, Ph.D.

Date: _____

Abstract

To help communities better understand and manage urban forests, the USDA Forest Service has developed the i-Tree software suite (www.itreetools.org). To include demographic data in future versions of i-Tree, the Forest Service sought a simpler way to obtain geographically-linked data from the U.S. Census Bureau. At present, the Census Bureau has not provided a way to efficiently download specific statistics for large areas. Data from the Census Bureau website is typically spread across many data files, none of which have headers or other ways to quickly reference the data.

To minimize errors from manually retrieving and collating the data, I designed the Census Data Extractor, a toolset that uses the R and Microsoft Visual Basic for Applications (VBA) programs. In the VBA portion, the user selects the statistics that they want from a graphical user interface, and several output files are generated. In the R portion, the VBA output files are accessed and the selected statistics for each data report are consolidated into a single text file. The toolset can be used to obtain data from the 2010 Decennial Census and the 2006-2010 American Communities Survey 5-Year Estimates, which contain the most up-to-date, detailed information available. A case study is presented where demographic data is obtained and compared to tree health data within the City of Syracuse, NY.

Table of Contents

Acknowledgements

I'd like to give special thanks to the USDA Forest Service Northern Research Station located in Syracuse, NY for inspiring and funding the initial work on this project. I'd like to especially thank Allison Bodine and Dr. Dave Nowak for their guidance and encouragement.

I would also like to thank my advisor, Dr. Chuck Kroll, for his patience, advice, and encouragement throughout this project.

I'm very appreciative of the testing and feedback that Ethan Bodnaruk, Danielle Kaveney, and Kevin Hennigan provided.

Finally, I want to thank the countless friends and family members who served as sounding boards every step of the way.

Introduction

As a part of continuing efforts to expand and refine the i-Tree software suite, a toolset that helps communities better understand and manage urban forests, the USDA Forest Service plans to integrate demographic data within certain software packages. In the United States, the U.S. Census Bureau is the clear choice for demographic data. However, the two major ways in which the Census Bureau makes this data available are not well-suited to the Forest Service's needs. The American Factfinder, a search application on the Census Bureau website, is helpful for collecting data on specific locations, but not for creating a database. The ftp server, a file directory on the Census Bureau website, contains hundreds of zip files containing the desired information, but none of the data files have headers.

To retrieve and collate the data, I designed the Census Data Extractor (CDE). This toolset uses a graphical user interface (GUI) that runs on Microsoft Visual Basic for Applications (VBA) to help the user select and name statistics and the R program to find and consolidate those statistics into a single text file.

U.S. Census Bureau geography

The U.S. Census Bureau divides the United States by over 50 different geographic types, which are described on the U.S. Census website (U.S. Census Bureau, 2012b). For most purposes, the types that are most important to understand are the Census Block, Census Block Group, and Census Tract. Generally, Census Blocks correspond to individual blocks as determined by road networks. Census Block Groups are groupings of Census Blocks that contain between 600 and 6,000 people. Census Tracts are groupings of Census Block Groups that contain between 1,200 and 8,000 people. These geographic

types do not cross county lines, but can cross city lines. Figure 1 shows an example of Census Blocks and Block Groups contained within a Census Tract in Syracuse, NY.

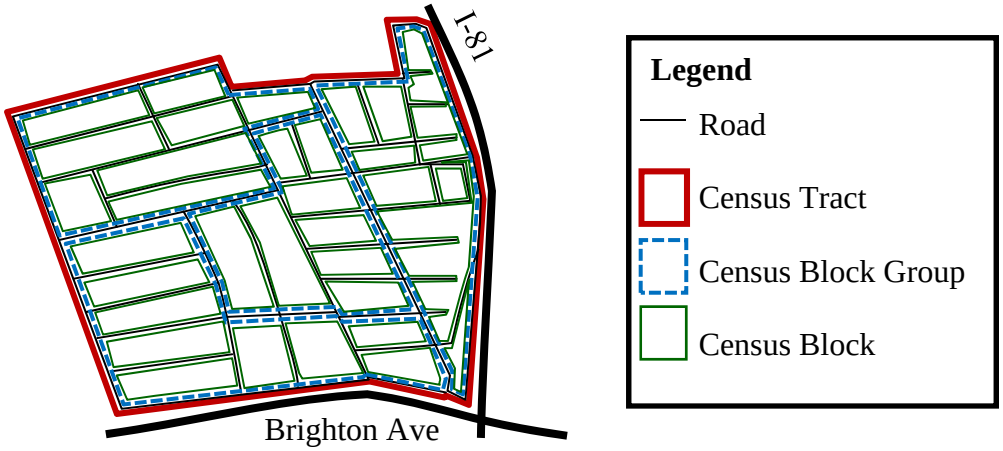

Figure 1. Census Tract 54 in Syracuse, NY with Census Block Groups, Census Blocks, and roads.

Datasets

In the United States, the national decennial Census, hereafter referred to as the Census, is performed once every 10 years, with the goal of gathering information on every person living in the country (U.S. Census Bureau, 2011b). In 2010, the data collected included basic information such as sex, age, and race, as well as household characteristics such as family size and whether a home was rented or owned (U.S. Census Bureau, 2012a). From participant responses, the U.S. Census Bureau created the Summary File 1 (SF1) database, which contains 235 data tables with information down to the Census Block level, 86 down to the Census Tract level, and 10 down to the County level (U.S. Census Bureau, 2012a).

The American Communities Survey (ACS) is an annual survey sent to approximately 3.5 million addresses each year (U.S. Census Bureau, 2011a). In addition to the basic information covered by the Census, the ACS has a broader set of information, with statistics related to such things as income, the age of buildings, and commuting time (U.S. Census Bureau, 2013). The CDE uses the 2006-2010 5-year ACS estimates, which

are a compilation of 5 ACS surveys. This dataset contains over 900 data tables, many of which have information down to the Census Block Group level (U.S. Census Bureau, 2011a). Since the ACS is only sent to a small portion of homes each year, ACS data is not available for Census Blocks.

Census Data Extractor Features

The CDE has two parts: the first employs Visual Basic for Applications (VBA) within a Microsoft Excel framework, and the second employs the R statistical and computing package, a freeware programming package available online (R Project, 2006).

In the VBA part, the user is directed through several menus in a graphical user interface (GUI) to select the Census and ACS information that is wanted. First, the user can look through the available data in the ACSFields and SF1Fields worksheets. When the user selects a table containing data of interest, a list of the statistics contained in that table is shown, and the user can then save them for use later in the process (Figure 2).

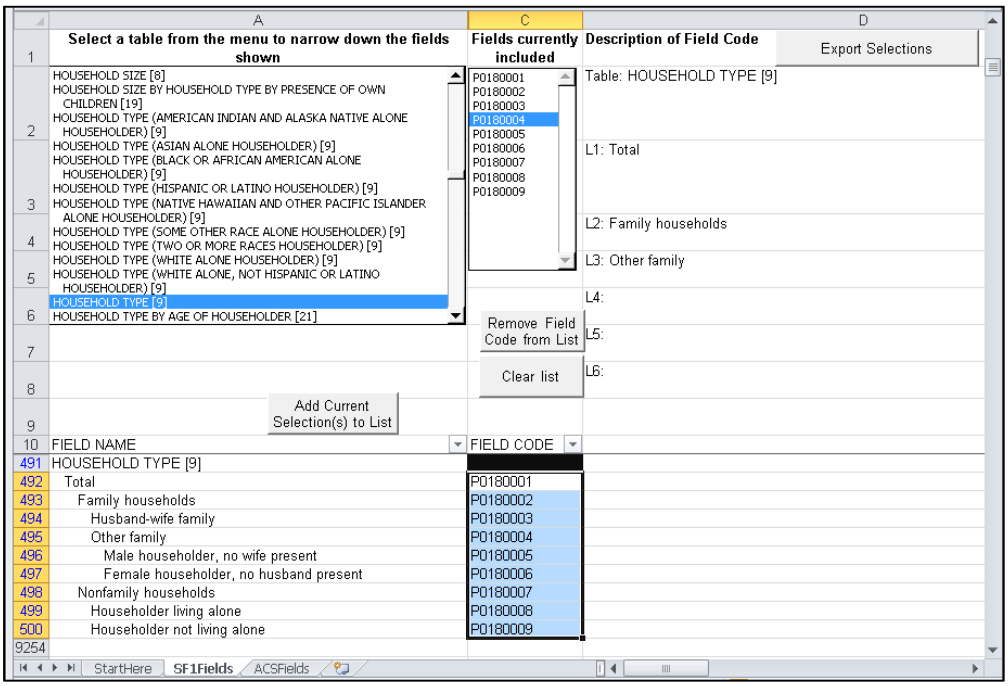

Figure 2. Form for selecting statistics of interest from VBA tool.

Next, the user selects the States being considered and whether data from the Census, the ACS, or both is being used. The user can then assign new names to the data fields (Figure 3), and combine data fields to create new fields (Figure 4). The user can also select which fields from the Geoheader file, which contains geographic data, to include. The final result is a series of output files which contain information regarding the data to obtain. The user can view the settings they selected in the main menu (Figure 5).

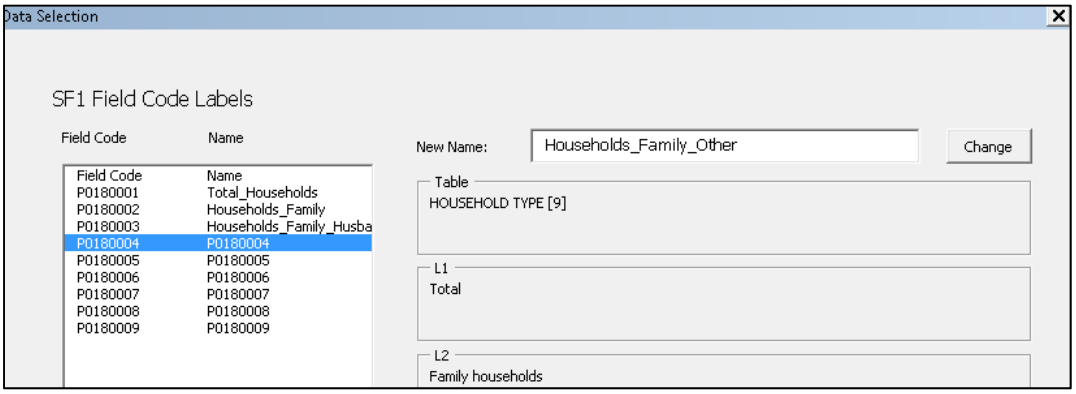

Figure 3. Section of SF1 Field Code Labels page in Data Selection userform.

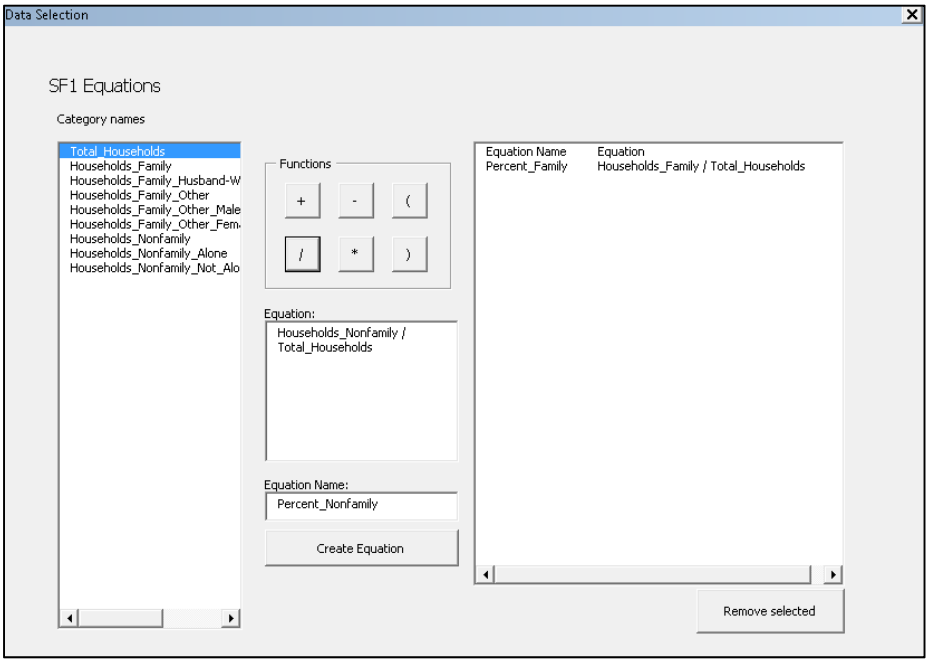

Figure 4. Selection from SF1 Equations page of Data Selection userform.

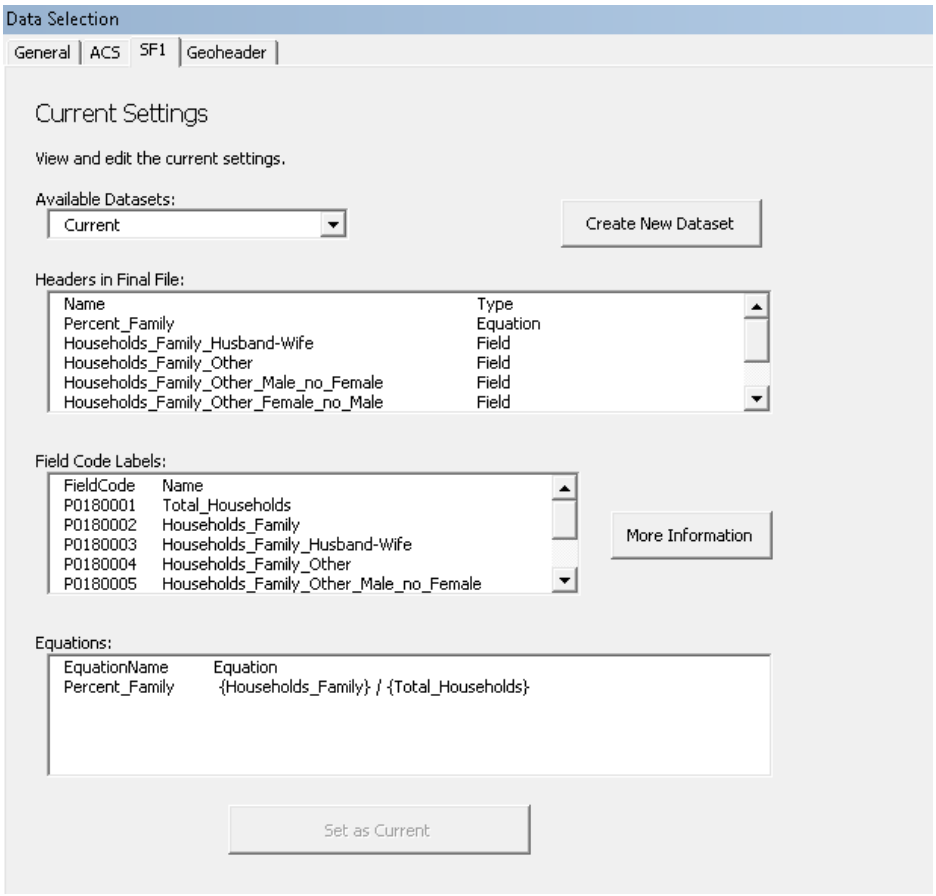

Figure 5. Menu showing current settings for SF1 data.

The user is then guided to the ftp server on the U.S. Census Bureau website, where they can download the appropriate zip files with the desired data. This data is used in the R portion of the CDE, which reads the VBA output files to extract the desired data and create the final data files. The user can then link that data to shapefiles provided by the U.S. Census Bureau, and thus use the data in Geospatial Information System (GIS) programs.

Example Study

As a demonstration of possible applications for the CDE toolset, I conducted a simple comparison of demographic factors and characteristics of the urban forest in Syracuse, NY. Of interest is whether the health of urban trees is correlated with specific

demographic factors. Such information could be useful to foresters and urban planners and managers as they allocate resources to maintain and enhance urban forests. This is only a preliminary study and should not be considered at all conclusive.

Introduction

Collectively, all the trees in a city are referred to as the urban forest. That includes street trees and trees in gardens, yards, and parks (Corona et al., 2012). While often looked at as decorative, the urban forest performs a number of important ecosystem functions. By blocking wind and radiation, trees can reduce the need for heating and air conditioning (Grove et al., 2006). They absorb atmospheric gases and provide a surface for dry deposition of atmospheric particles, thus improving air quality (Nowak and Dwyer, 2007). Urban trees can also reduce stormwater runoff via increased infiltration and evapotranspiration and improve the quality of stormwater by uptaking nutrients and retaining particulate matter (Nowak and Dwyer, 2007). In addition, trees can also help reduce urban heat island effects by utilizing energy for evapotranspiration processes (Akbari, Pomerantz, and Taha, 2001). The growth of urban trees can sequester carbon, reducing atmospheric carbon dioxide levels (McPherson, 2006). Urban forests also provide a number of psychological benefits, such as a greater sense of community and reduced stress (Kuos, Sullivan, Coley, and Brunson, 1998).

Considering the benefits of the urban forest, understanding the factors at play in the health and diversity of the urban forest can help planners take actions that will ensure a healthy and sustainable urban forest. To that end, understanding which socioeconomic factors, if any, affect the health and diversity of the urban forest is useful. In Chicago, Iverson and Cook (2000) found that "wealthy regions have higher tree cover while poor

regions have lower [tree cover]". Of interest is whether similar patterns are present in Syracuse, NY.

Methods

I used the CDE to obtain five socioeconomic characteristics for the state of New York: the percentage of households renting, the vacancy rate, the percent below the poverty line, the percent of people over 25 who graduated high school, and the percent of homes built before 1950. I downloaded the TIGER/Line shapefile, a map produced by the U.S. Census Bureau that shows Census area delineations, for block groups in Onondaga County. I clipped this to the boundaries of Syracuse in ArcGIS, and attached the socioeconomic data to the shapefile.

The Forest Service provided data from the 2009 i-Tree survey of Syracuse, in which trained volunteers measured and recorded the characteristics of each tree found in 200 0.1-hectare plots distributed throughout Syracuse. In ArcGIS, I linked the i-Tree data to the Syracuse Block Groups (Figure 6).

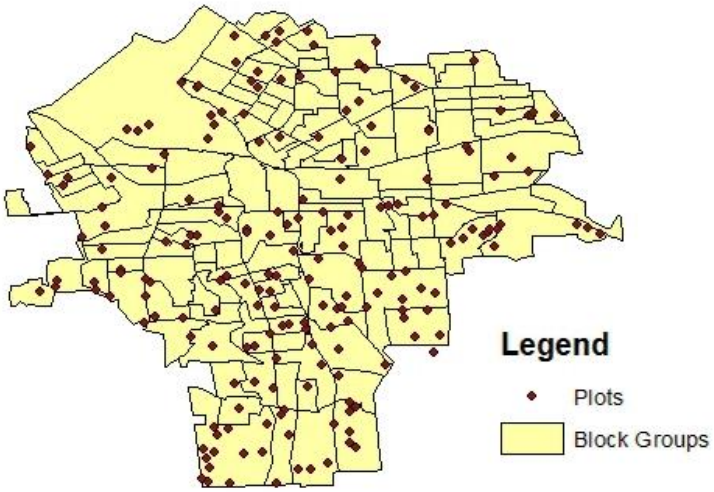

Figure 6. Map of Syracuse showing Block Groups and sample plot locations.

Using the two datasets together, I calculated three statistics for each Block Group that contained a sample plot. The number of trees per person and the crown area per person give an impression of canopy cover and tree population as it relates to the human population. The percent dominant species demonstrates how vulnerable a tree population is to new pests and diseases.

Results and Discussion

There was no obvious relationship between the selected socioeconomic factors and forest characteristics (Figure 7). The correlation coefficients for the two datasets are very low (Table 1). Since this region was originally forested, it would make sense that the number of trees would not be well correlated with socioeconomic factors. However, there are many factors which would add uncertainty to any conclusions. This data included commercial districts and other areas that are not strictly residential. Additionally, the effects of population density and land use were not taken into account. This study focused on the population and homogeneity of the tree population, but it did not factor in the relative desirability of certain species in an urban setting or the placement of trees on lots. I did not consider the density of the crowns of individual trees. Clearly, this study has only scratched the surface of this subject.

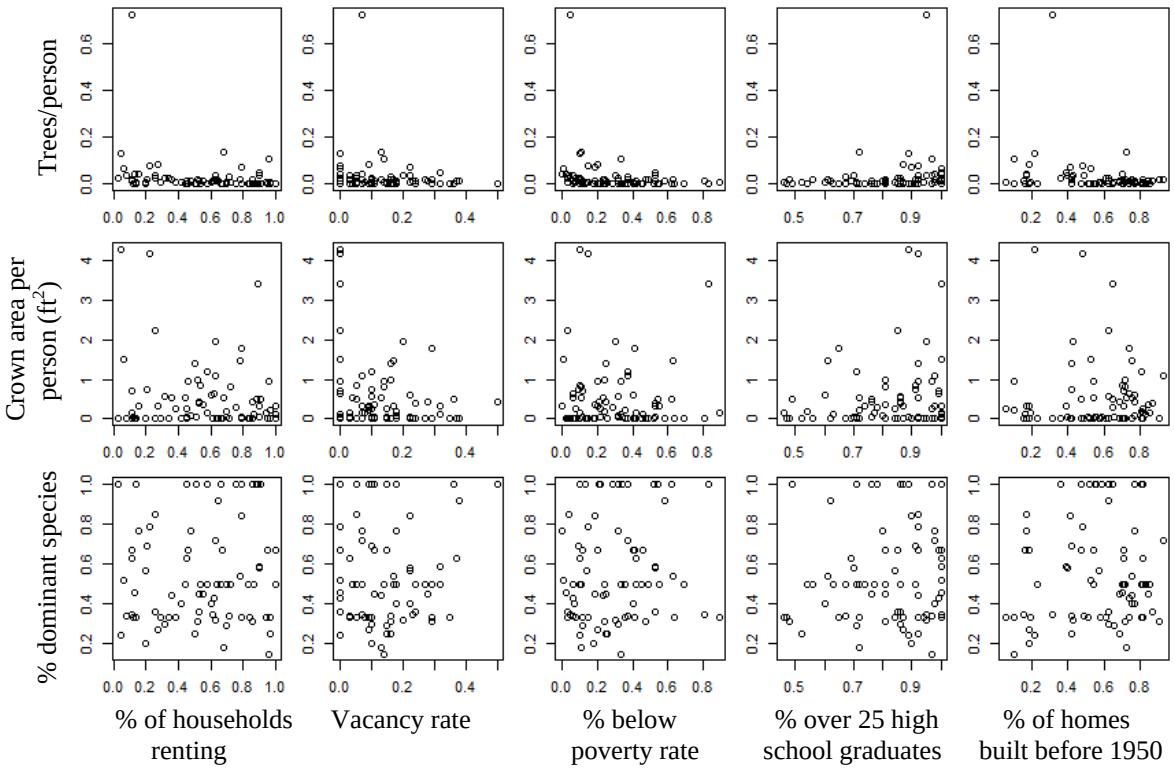

Figure 7. Scatter plots of socioeconomic statistics and urban forest characteristics. Each point represents a Census Block Group.

Table 1. Correlation coefficients for relationships between socioeconomic statistics and urban forest characteristics.

	% of households renting	Vacancy rate	% below poverty rate	% over 25 high school graduates	% of homes built before 1950
Trees/person	-0.24	-0.11	-0.21	0.15	-0.21
Crown area/ person (ft^2)	-0.12	-0.26	-0.03	0.16	-0.02
% Dominant species	0.06	-0.00	0.13	0.14	0.10

Conclusion

The Census and ACS databases are a rich source of useful statistics about the American population. At present, these databases do not have a convenient way to quickly download specific statistics, especially for large areas or many unrelated statistics. As such, the purpose of the CDE is to provide a simple tool to help people access this data. The resulting data files can easily be linked to GIS data, making it simpler for planners and researchers to incorporate it into their work.

Works Cited

Akbari, H., Pomerantz, M., & Taha, H. (2001). Cool surfaces and shade trees to reduce energy use and improve air quality in urban areas. *Solar Energy*, 295-310.

Corona, P., Agrimi, M., Baffetta, F., Barbati, A., Chiriacó, M. V., Fattorini, L., et al. (2012). Extending large-scale forest inventories to asses urban forests. *Environmental Monitoring and Assessment*, 1409-1422.

Grove, J., Troy, A., O'Neil-Dunne, J., Burch Jr., W., Cadenasso, M., & Pickett, S. (2006). Characterization of households and its implications for the vegetation of urban ecosystems. *Ecosystems*, 578-597.

i-Tree. (2013). *About Us*. Retrieved March 9, 2013, from i-Tree Website: http://www.itreetools.org/about.php

Iverson, L. R., & Cook, E. A. (2000). Urban forest cover of the Chicago region and its relation to household density and income. *Urban Ecosystems*, 105-124.

Kuos, F. E., Sullivan, W. C., Coley, R. L., & Brunson, L. (1998). Fertile ground for community: inner-city neighborhood common spaces. *American Journal of Community Psychology*, 823-851.

McPherson, E. G. (2006). Urban forestry in North America. *Renewable Resources Journal*, 8-12.

Nowak, D. J., & Dwyer, J. F. (2007). Understanding the benefits and costs of urban forest ecosystems. In J. E. edited by Kuser, *Urban and community forestry in the northeast, 2nd ed.* (pp. 25-46). New York: Springer.

R Project. (2006). *What is R?* Retrieved April 4, 2013, from R Project Website: http://www.r-project.org/

U.S. Census Bureau. (2011a, December 8). *The 2006-2010 ACS 5-year summary file technical documentation.* Retrieved March 24, 2013, from U.S. Census Bureau ftp server: http://www2.census.gov/acs2010_5yr/summaryfile/

U.S. Census Bureau. (2011b). *What is the Census?* Retrieved March 20, 2013, from U.S. Census Website: http://www.census.gov/2010census/about/

U.S. Census Bureau. (2012a, March). *2010 Census Summary File 1 Technical Documentation.* Retrieved March 25, 2013, from U.S. Census Website: http://www.census.gov/prod/cen2010/doc/sf1.pdf

U.S. Census Bureau. (2012b). *Appendix A. Geographic Terms and Concepts.* Retrieved April 26, 2013, from u.S. Census Bureau website: http://www.census.gov/geo/reference/pdfs/GTC_10.pdf

U.S. Census Bureau. (2013, February 1). *About the American Community Survey: Questions on the form and why we ask.* Retrieved March 24, 2013, from U.S. Census Bureau Website: http://www.census.gov/acs/www/about_the_survey/questions_and_why_we_ask/