Finding the Green in Greenspace: An Examination of Geospatial Measures of Greenspace for Use in Exposure Studies

by

Shani Pynn

A Thesis Presented to the
Faculty of the USC Graduate School
University of Southern California
In Partial Fulfillment of the
Requirements for the Degree
Master of Science
(Geographic Information Science and Technology)

August 2018

To my husband for his patience, my father for his drive, my mother for her understanding, and for the mountains, who may never read this, but have provided solace and inspiration for many years.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I am grateful to my adviser, Professor Sedano, for her quiet patience and gentle nudging which kept me on track while allowing me to walk on my own. I would like to thank everyone who served on my faculty juries for allowing me the leeway to go down this path while I was still discovering it. I am grateful for my committee for their enthusiasm and high standards for my work. These people have helped me to take something dear to my heart, which I was uncertain how to complement, and create a meaningful addition to the collective knowledge on the subject.

List of Abbreviations

GIS	Geographic information system
MODIS	Moderate Resolution Imaging Spectroradiometer
NASA	National Aeronautics and Space Administration
NDVI	Normalized Difference Vegetation Index
USGS	United States Geological Survey

Abstract

Over the years researchers have examined greenspace using definitions of varying breadth and various measures to capture that breadth. This study compared three such definitions and associated measures to assess their similarities and differences. It sought to spatially examine these measures and determine whether the results they produce are statistically interchangeable or different in various ecoregions and urbanization levels, as well as observe any notable nuances between types. For definitions and measures, the study used inventory (a polygon shapefile of parks), usage-based categorical (a raster data set of classified vegetated land covers), and biophysical (satellite imagery based Normalized Difference Vegetation Index (NDVI)) data. It tested these three types within neighborhoods in four different regions in the state of California. The regions chosen represented the north and south coast, inland desert, and central valley. Within each region, the data sets were tested in urban, suburban, and rural areas. The amount of greenspace represented by parks, vegetated land use classes, and spring Landsat 8 NDVI imagery within each tract was measured and the averages within each measurement type were compared to one another. It was expected that land use and NDVI data would show statistically greater amounts of greenspace cover in rural and suburban areas, but that parks data would show more in urban areas due to sensor resolution limitations. If true, this would suggest regional variation in measurement type comparability. It was also expected that additional type-specific strengths and weaknesses would emerge. This information will be useful in determining whether new combinations of greenspace measurements might prove fruitful. After analysis, the study found NDVI provided a statistically higher measure of greenspace overall, although there was some variation in the discrepancy between measures across area types.

Chapter 1 Introduction

To begin, this chapter describes the concept of greenspace and why that concept has captured the interest of epidemiologists and others concerned with the relationship between location and health. The chapter then describes differing definitions and methods of measuring greenspace found in the health literature. It explains the need to compare how these different definitions and measurements might yield different results and summarizes the methods this thesis has taken to provide that comparison.

1.1. What Is Greenspace and Why Do We Care About Measuring it in the First Place?

This section discusses the broad definitions of greenspace. Narrowing of focus to quantifiable definitions will be discussed later in chapter 2. Greenspace, while an intuitively simple concept, is defined differently by different researchers. It is therefore important to begin with a grasp on the basic root of this term. According to the Merriam Webster Dictionary (2018), "green space" is defined as "community space consisting of land (such as parks) rather than buildings" and the first use of this phrase was recorded in the year 1943. Breaking this down, the term space represents various concepts, many of which are associated with an area of defined or undefined size (Merriam-Webster Dictionary 2018). The word "green" defines a color, also "covered by green growth or foliage," and even "pleasantly alluring" (Merriam-Webster Dictionary 2018). Combining the two terms therefore conceptually captures the concept of a bounded or unbounded area largely consisting of green plants.

This exercise of tangibly defining greenspace illustrates the process of moving from a construct or a latent variable to a manifest variable. A construct, in this case greenspace, is an idealized classification (Montello and Sutton 2013). It is an imaginary box we use to categorize the world into greenspace and non-greenspace. Constructs like this one are quite common as defining and classifying the world around us helps us to study and understand it. Physically measuring these constructs, however, is technically impossible as the ideal breaks down when we try to find the physical line where the classification ends (Montello and Sutton 2013). To address this, we consider and define physical manifestations of the construct which can be measured. These measurable variables are termed manifest variables and the construct is considered a latent variable (Montello and Sutton 2013). In order to examine the manifest variables used to measure greenspace, it is therefore important to examine the makeup of the latent variable that is the construct of greenspace itself.

While green vegetation has been studied in other sciences such as ecology relating it to habitat value for wildlife, this study focused on greenspace as it relates to human effects and exposure. Past studies have correlated exposure to greenspace with various health benefits. Even with this narrowed focus for this study, there are several manifest variables we can use to represent greenspace. Since there are many different effects correlated with greenspace there are various ways of measuring it. Some are more applicable than others in each case.

It is important to touch on the motivation behind this drive to define and categorize the concept of greenspace for study. Greenspace is a term and concept that can bring to mind beautiful open areas with lush calming vegetation. This mental image, however, is a construct of greenspace, a latent variable which cannot be measured in itself. We as researchers endeavor to find representative terminology to define this concept and metrics and manifest variables to measure its qualities and quantities.

This study examines three widely used manifest variables representing greenspace and corresponding measurements and compares them to one another. The first of these manifest variables is inventoried area that was created or set aside with the intention of being greenspace. These are areas such as parks and nature preserves. These types of areas are often compiled into easily accessible data sets for large regions which can be quite useful for researchers. In California one such set is the California Protected Areas Database (CPAD). The second manifest variable used is a usage-based categorical measure of all greenspace exclusive types in a classified land use raster. The raster used for this study was the National Land Cover Database (NLCD). This database, first published in 2000, provides 30-meter comprehensive land cover data for the entire United States (Homer, Fry, and Barnes 2012). It is created by a group of governmental entities led by the U.S. Geological Survey and is a free to use data set used by land managers, researchers, and other interested persons. The third manifest variable used, biophysical, is a more explicit measure of greenspace plants themselves: remotely sensed imagery processed to show healthy living vegetation. In this case, this study used the Natural Difference Vegetation Index (NDVI) from Landsat 8 imagery. So, in summary, this study looked at compiled data sets of intentional inventory greenspace, classified usage-based categorical greenspace, and biophysical remotely sensed healthy vegetation itself.

It is important to note that these three valid manifest variables representing greenspace are not identical. Therefore, we need a Rosetta Stone of sorts to be able to compare them more effectively in scientific literature. This is the niche where this study fits in as a spatial comparison. This study aims to become a type of translation document helpful for combining research within and across fields of research pertaining to greenspace constructs. It seeks to ease extrapolation of findings if a researcher cannot find the specific data type he or she needs but does have access to a comparable one. This comparison study may also come in handy when trying to decide which type of data to use if multiple types are available since pros and cons of each could be determined. This study can also help when thinking about what question you are really asking of your data. Having this comparison study to refer to can help give backup reasoning for data choices.

1.2. Research Question

As the study of greenspace is a field of great breadth and depth, this subsection outlines the specific focus of this study within that field. Specifically, it briefly describes the main question and hypothesis which served as this study's main focus.

This main question addressed by this study is: What percent covers do inventory, usage-based categorical, and biophysical manifest variables of greenspace produce in various environmental settings and what do these results look like spatially? This question sought to identify statistical similarities and differences in percent cover produced between these three measures to determine if results can be reliably interrelated between studies using differing manifest variables.

The hypothesis addressing the above question was that the biophysical manifest variable using remotely sensed imagery data would provide the highest average cover. Usage-based categorical (classified) and inventory (parks data) manifest variables may provide more detail in urban settings where imagery is limited by sensor resolution. However, the biophysical variable should capture all discernable vegetation whether it has been identified and classified by the creators of the other two variables or not. Also, the usage-based categorical variable's data may have categories which include vegetation but are not exclusive to it. This potential combination may not allow all vegetation to be accurately measured.

The study also expected that the highest producing manifest variable will vary with urbanization. It expected imagery would capture the most area in rural settings, classification to be effective overall, and parks data sets to capture the most in urban settings. The reasoning for this is as follows. Imagery data is good but limited by sensor resolution. Classification data was created using multiple methods and has broad categories in terms of vegetation cover. It also has categories showing percentage ranges for urban cells. Since these are ranges, one cannot accurately identify values for these cells which causes problems. Parks data will not have adequate coverage in areas with existing, undescribed natural vegetation or farmland green areas, but may have a finer resolution for delineating green areas in urban settings.

Additionally, the study expected the highest producing manifest variable to vary with geographic location. In desert areas where vegetation is often sparser, more greenspace may be identified by the classification and parks data sets than the imagery as there may be less vegetation itself to be captured by the imagery. If this is the case, it brings up the possibility of overrepresenting vegetation in desert areas if the data is not ground-truthed in some manner.

Finally, this study aimed to provide a pros and cons list for each data type. This was done using these same categories to aid study method choices where amount is not the only greenspace related factor of interest.

1.3. How Will I Answer This Question?

To answer the question raised in the previous section, this study was designed to measure greenspace using three different manifest variables and compare their resulting percent covers in various environmental conditions. The following subsections outline this process. They begin by outlining the broader design of the study. Then they break it down into the stages taken and, finally, spend some time describing the manner in which the specific locations for the study areas were chosen and the reasoning behind those selections.

1.3.1. Design

The study compared a polygon data set of parks, exclusively vegetated areas within a classified set, and healthy vegetation from imagery. It conducted this comparison in neighborhoods selected within a broader study area containing spatial variability in environmental conditions. The study organized itself into four main stages: broad study area selection and data set identification, refinement of study areas, pertinent specific data collection for each area, and percent greenspace calculation and comparison between variables.

1.3.2. Study Area Selection and Refinement of Location

To generate broadly applicable results this study sought to capture a variety of vegetation types in a variety of environments. As vegetation and area are important components of greenspace, variations in those two variables were concerns for making the results of this study suitable for applicability in future research. To have a broad application, this study sought to include differing types of vegetation and environments within the sample set.

This study used the state of California as the broad sample area, effectively bounding the area for sample selection. With a span of nearly ten degrees of latitude and over 400,000 square kilometers of land this state includes a large variety of geographic areas from coastal rainforest to inland desert and encompasses area within ten ecologically distinct mapping zones used in the creation of the NLCD. It is also home to more than 1,200 vegetation communities (CNPS 2018). This high level of diversity was ideal for the purposes of this study.

To efficiently capture a degree of the diversity within the state, this study focused on a subset of four areas shown in Figure 1. These are urban centers and surrounding areas in the north and south coast, central valley, and southern desert regions of the state. They were chosen because these centers represent four separate ecoregions as described by both the NLCD and the

U.S. Environmental Protection Agency (U.S. Environmental Protection Agency, 2012; Homer, et al. 2004). This study selected these areas by overlaying a U.S. Census urban areas and urban centers layer over NLCD mapping regions and choosing centers which could accommodate enough (greater than five urban and five suburban) tracts for the study set.

Figure 1 Study area boundary and urban focus areas

Stage three of the study was data selection. The study selected data from reputable, freely available, and geographically vast compilations. Selecting free data was important as this study sought to compare practically applicable data sets. A large part of what makes data practical to use is finding data which is economically feasible to obtain. The necessity for a wider geographic area was to allow the study to compare data sources which are usable in a wide variety of areas or are of a type comparable to a potentially similar type existing in future potential study areas. These requirements help maintain the usefulness of this study as a possible tool in future researchers' decision-making toolkits.

The fourth stage of the study was the measuring and comparing of manifest variables representing greenspace. The study used percent cover of greenspace as the metric of comparison. The reasoning behind this is that this metric was more suited to a wide geographic scale data comparison than a method of active exposure such as distance to or time to greenspace as a destination. It was more practical to accurately measure spatial density than path distances along a network of roads at a state level. Also, since this study used both raster and vector data sets a density comparison was well suited to apply across data types.

For the comparison between manifest variables, this study used the analysis of variance between their means. This method is sufficient to compare averages and determine if they can be considered significantly interchangeable or not.

Chapter 2 Related Work

This chapter addresses the ways in which greenspace is measured. It first summarizes the work done by two sets of researchers to categorize the most common measures of greenspace found in the literature in recent years. It then goes into more detail on the three measures used in this study and how recent studies of greenspace have employed them.

2.1. A Closer Look at How We Define and Measure Greenspace

In the previous chapter we discussed the broader ways greenspace can be defined. Here we narrow our focus to look at defining it in quantifiable ways. More specifically, this section examines the various definitions for greenspace used by scientists correlating it to public health effects. Researchers have recently begun categorizing these concepts into taxonomies to better describe the work being done (Jorgensen and Gobster 2010; Hunter and Luck 2015). As greenspace can be broadly defined, some researchers describe their specific greenspace latent variables of interest clearly and some do not. Only works where greenspace measures were "sufficiently well described" were used in creating these taxonomies (Jorgensen and Gobster 2010).

Jorgenson and Gobster (2010) reviewed scientific literature on greenspace and public health from 1997 to 2010 and identified eight main ways that greenspace is measured. They termed these "codes" of measurements for greenspace characteristics as they were dividing and classifying them to create a taxonomy. More recently Hunter and Luck (2015) conducted a similar review and arrived at a slightly different list of codes. Table 1 provides a reference of codes identified in each of these studies, and the following paragraphs explain them.

Table 1 Greenspace measure codes identified by two recent literature reviews

Code	Jorgensen and Gobster 2010	Hunter and Luck 2015
None	X	-
Urban versus natural	X	-
Descriptive/narrative	X	X
Inventory	X	X
Area/distance	X	X
Biophysical	X	X
Human perceptual	X	X
Biodiversity	X	X
Ecosystem services	-	X

These “codes” used by Jorgensen and Gobster (2010) and Hunter and Luck (2015) describe the most common manifest variables currently employed to measure the latent variable of greenspace. As described in Chapter 1, the construct of greenspace is itself unmeasurable: Greenspace is an idea, and hence cannot be physically measured. A physically measurable approximation possessing the qualities of greenspace must be employed. Jorgensen and Gobster (2010) and Hunter and Luck (2015) used their codes to classify ways greenspace is measured into a taxonomy. As their purpose was to describe methods used to measure greenspace by various researchers these codes are roughly equivalent to latent or manifest variables of greenspace, depending on whether they are used to describe the construct we are attempting to measure or the type of measurement method being used to attempt to quantify that construct. For clarity, as the term code can have various meanings beyond describing a manifest variable, this text will refer to the codes used by these researchers as “measures” or “manifest variables” from this point forward.

The first manifest variable, “none,” was used to categorize studies that lacked an actual measure or description of greenspace (Jorgensen and Gobster 2010). They noted its use mainly with studies measuring attitudes about greenspace related concepts (Jorgensen and Gobster

2010). This variable would be akin to a study consisting of surveying participants on how they feel about "greenspace" without describing it further.

The second manifest variable, "urban versus natural," described fairly explicitly a measure consisting only of differentiating between "urban and natural settings." (Jorgensen and Gobster 2010). They included both real and simulated natural exposures of various immersivity's and noted its use mainly with studies on mental health (Jorgensen and Gobster 2010). This variable was found to be used in studies comparing activities such as running in areas with and without natural elements in the nearby environment (Jorgensen and Gobster 2010).

The third manifest variable, "descriptive/narrative," described a measure defined by participants rather than researchers (Jorgensen and Gobster 2010; Hunter and Luck 2015). Jorgensen and Gobster (2010) found this variable to be an erratic measure, and both they and Hunter and Luck (2015) noted its use with participants "'attitudes, meanings, and values.'" This can be a difficult measure as the creator of the study would have little control over what this variable ends up specifically representing.

The fourth manifest variable, "inventory," described lists of items and environmental characteristics including several specifically non-greenspace items (Jorgensen and Gobster 2010; Hunter and Luck 2015). Both studies found this to be a variably represented measure used for comparing greenspace with multiple health effects (Jorgensen and Gobster 2010; Hunter and Luck 2015). However, neither study found this variable to be frequently used although there was a small association with psychological health studies (Jorgensen and Gobster 2010; Hunter and Luck 2015).

The fifth manifest variable, "area/distance," described a "quantity or proximity" of greenspace within a defined area or distance from a reference point (Jorgensen and Gobster

2010; Hunter and Luck 2015). Both studies found this variable used with health studies associated with physical activity (Jorgensen and Gobster 2010; Hunter and Luck 2015).

The sixth manifest variable, “biophysical,” described a variety of more specific measures of lands and vegetation “falling short of biodiversity” (Jorgensen and Gobster 2010; Hunter and Luck 2015). Both studies found this variable to be often used with studies measuring preference (Jorgensen and Gobster 2010; Hunter and Luck 2015). A biophysical variable is the type to use data directly from aerial and satellite imagery systems. In this case the imagery provides a direct view of what exists in an area as measured by the light detectable by sensors. Collecting images in separate bands of the light spectrum allows individual object types to be distinguished by their spectral signature. Combining two of these bands, the red and near infrared, creates the NDVI. This results in an image where vegetation is distinct from non-vegetation, making it easier to extract for analysis. This type of manifest variable can be used to measure all vegetation in an area.

The seventh manifest variable, “human perceptual,” describes a dataset of land categorized by land use and land cover; these categories are derived from a collection of data types such as biophysical, elevation, and imagery data which are weighted and combined into a categorization scheme. Both studies found this variable, like the biophysical variable, is also often used with studies measuring preference (Jorgensen and Gobster 2010; Hunter and Luck 2015). This thesis describes this type of variable as “usage-based categorical” rather than “human perceptual.” The phrase human perceptual contains the implication that this data type is based on the perception of persons involved in a given study, but data of this type is also often compiled using sophisticated computer-based classification schemes. Usage-based categorical variables usually provide less detailed information about individual greenspace areas but provide a generalized categorical map covering an area. They categorize all land covers or uses in a given area. This provides categories representable as greenspace such as undeveloped, grassland, etc. (Lachowycz, et al. 2012). The NLCD is an example of the usage-based categorical variable type. It contains a biophysical component in the form of NDVI and other indices from imagery as well as other non-biophysical inputs such as elevation data (Homer et al. 2015). These components do not remain separate in the NLCD but are used to fit the land into categories. All of these inputs are integrated together and used to create the final land cover classes comprising the finished NLCD (Xian, G., et al. 2011).

The eighth manifest variable, “biodiversity,” describes a measure of the actual diversity of plants and animals in an area (Jorgensen and Gobster 2010; Hunter and Luck 2015). Both studies found this variable to be often used with health studies concerned with psychological well being (Jorgensen and Gobster 2010; Hunter and Luck 2015). Hunter and Luck (2015) expressed concern with this variable as it encompasses both participant perceived biodiversity and quantified biodiversity which may not be equivalent.

The ninth manifest variable, “ecosystem services,” describes the use of “various measures to evaluate the concept of ecosystem service provision” (Hunter and Luck 2015). These researchers added the manifest variable “ecosystem services” to the set created by Jorgensen and Gobster as the deemed research related to this topic had increased to a point where the separate manifest variable was warranted (Hunter and Luck 2015).

Hunter and Luck (2015) removed the “none” manifest variable from their list, judging it insufficient to be included, and took out “urban vs natural” presumably for the same reason. This revised their list to seven categories in which measurements of greenspace can be classified when used in association with health measures.

As stated in Chapter 1, the study herein compares how the percent cover of greenspace in an area varies with different manifest variables for greenspace. The manifest variables used by Jorgensen and Gobster (2010) and Hunter and Luck (2015) provide a useful starting point for this study in identifying manifest variables for comparison. Not all manifest variables employed by these authors were appropriate for comparison in this study.

The first three manifest variables listed, “none,” “urban vs natural,” and “descriptive/narrative,” are measures which cannot reasonably be analyzed in terms of percent cover, let alone compared in that form. These measures strongly relate to participants’ responses to greenspace concepts or elements rather than to measures those elements themselves. It was therefore not practical to compare neighborhood scale data from areas throughout the state for these measures without locating and surveying a reliable subject pool within each area.

The next manifest variable, “inventory,” can describe areas such as parks; and those areas can be measured spatially. This allows for greenspace to be evaluated by features which affect visitors even though they may not be vegetative in nature. The study herein used a data set of preserved lands to represent this measurement type.

“Area/distance” is a measure which somewhat overlaps the other variables as it addresses the way greenspace is quantified once it has been defined. This study measured greenspace cover within defined areas for all data sets run. Therefore, this measure is integrated into this study, but not as a variable being compared itself.

The “usage-based categorical” manifest variable, a measure for various classification criteria related to quality or type, is another variable which lends itself well to spatial measurement. This study used a national land use classification data set to represent the usage-based categorical variable of greenspace.

The “biophysical” manifest variable is perhaps the most literal measure of greenspace as it measures the “Presence/quantity of specific landscape elements (e.g. vegetation, % open land)” (Jorgensen and Gobster 2010). Of all measures, this is the one which would most explicitly measure the “greenness” of greenspace. It is also well suited to measuring spatially. This study used satellite imagery processed to show healthy vegetation as a representation of this variable.

The “biodiversity” manifest variable is a measure which is difficult to measure remotely and lacks broad standardized data sets. In California there is the California Natural Diversity Database, but this data set is not freely available with specific resolution down to a neighborhood level (State of California 2018). For this reason this study did not utilize a biodiversity variable of greenspace in its comparison.

The “ecosystem services” manifest variable is another measure lacking a free of charge broadly standardized data set. This, as well as the previous biodiversity manifest variable, describes a complex measure which deserves the attention of a full analysis in itself.

In the two studies described both Jorgensen and Gobster (2010) and Hunter and Luck (2015) compared these manifest variables to the types of studies using them to develop tables showing which variables were used for which type of study. This allowed them to determine which types of studies preferred which measures of greenspace. They both correlated greenspace measures to their uses but did not compare data sets from these manifest variables in shared areas to determine the variable’s relationships to one another. This study takes that step to further the understanding of these variables.

The next sections describe current work employing inventory, biophysical, and usage-based categorical manifest variables.

2.2. Greenspace Measurement Using the Inventory Manifest Variable

One of the main measures for greenspace is inventories of park lands. Public agencies maintain inventories of public park lands in the form of spatial vector data sets in geographic information systems (GIS). Such data allows for qualitative attributes of the lands to be recorded and spatial measure such as area and distance to be calculated. Studies comparing greenspace and public perception and physical health commonly use park data from inventory variables (Jorgensen and Gobster 2010).

Researchers often measure the area of, or distance to, greenspaces by limiting the spaces measured to a range of sizes and/or distances from a particular point in space. Some studies, such as Sugiyama, et al. (2016), only look at parks over a set size. Some, Sugiyama, et al. (2016), and Coombes, Jones, and Hillsdon (2010), use varying path distances along roadways to represent distances traveled to reach a park greenspace while other studies such as McCracken, Allen, and Gow (2016) use Euclidean, direct, distances.

At the time of Jorgensen and Gobster's study (2010), studies used the inventory type of greenspace measure in comparisons with a variety of health measures, albeit infrequently. This variable type is useful in studies with a lower geographic resolution as well. For example, Ambrey (2016) used it for area measures of greenspace within census districts in Australia when comparing the amount of greenspace near residents with physical activity and wellbeing.

A vector data format lends itself well to inventories as it accommodates multiple attribute fields and links well to databases. Inventory data representing greenspaces generally includes compilations by jurisdiction or land manager but can also include larger compilations of data sets. These represent managed public open spaces with varying amenities and amounts and types of landscaped vegetation. This type of variable may also include parks maintained in a natural state but will not generally include open areas used by communities which are not owned and managed for public use. For example, this type of variable often will not include open fields or hillsides under private ownership even when unofficial trails and bicycle tracks are used by the local community. It also may not include more scenic areas such as private beaches, retreats, golf courses, or smaller private garden areas, which would normally be thought of as greenspace.

The accuracy of this variable type is dependent on the frequency of updates to the data set. This can become problematic in temporal studies. If multiple inventories are combined for a study, this also introduces the possibility of varying update intervals which can reduce the temporal resolution of a study.

2.3. Greenspace Measurement Using the Biophysical Manifest Variable

We may use biophysical variables to examine greenspace in focused areas or over a time sequence. Researchers have used raster NDVI data from NASA MODIS satellites in spatial analyses of greenspace (Wu, et al. 2014) and in spatial measures over a time series (Younan, et al. 2016). Younan et al. (2016) utilized 250m resolution MODIS NDVI data at “250, 350, 500, and 1000 m” distances from residences in California.

While each of these measures of greenspace is useful for comparisons with some aspect of public health, a complete understanding of the effects of greenspace on health may require a combination of methods. For example, a study which focused on the actual movements of children through a landscape by tracking each child’s movements with a GPS unit utilized park, land use, and satellite imagery data to determine which areas were greenspace (Lachowycz, et al. 2012). This data was then used to determine timing and intensity of physical activity in greenspaces.

While some studies use multiple measures to get a good picture of greenspace exposure, this may not be possible in all cases. Without this, inventory variables, while excellent for delineating greenspaces with required elements, may sometimes not capture undocumented areas which would otherwise meet their criteria. Usage-based categorical variables, suited to and often used in greenspace buffer studies, can miss some necessary elements for a given study if these elements cross usage categories or have been classified within an unrelated and undesired category. Biophysical variables, often used in temporal studies and when living vegetation itself is the main manifest variable of interest, do not capture aspects such as access or usage type that the previous two types encompass. While it cannot hope to provide a formula for exact conversion in all situations, this study attempts to provide information useful for guiding estimations of what coverages these measures might make relative to one another if possible.

2.4. Greenspace Measurement Using the Usage-Based Categorical Manifest Variable

Another measure of greenspace is usage-based categories. Governmental and non-governmental entities compile usage-based categorical data for planning and/or research purposes. This variable often comes in the form of classified land use or land cover raster data sets. Researchers have used such raster data of categorical land uses to represent greenspace in exposure studies (Dalton, et al. 2016). Dalton and his colleagues took a land use map compiled by the United Kingdom and used a combination of its preset categories. They chose ones representing various vegetation and natural areas and used them to represent greenspace. They then used this greenspace representation to describe residential exposure to see if it had the benefit of reducing diabetes risk (Dalton, et al. 2016).

Another example of this variable in use in a greenspace study is in Wheeler et al. (2010). There the researchers tracked the movements of children and compared the time they spent in greenspace and non-greenspace areas (Wheeler, et al. 2010). Since this variable type provides a more exhaustive coverage of an area it can provide a better view of how much greenspace is in an area. The tradeoff is that it often does not have the supplemental data on amenities included in inventory measures. Researchers such as Van Den Bosch, et al. (2016) have also used these measures of greenspace for area distance measurements in buffers with population data as well as distance measures from individual addresses.

This variable does not necessarily always include all greenspaces. The land use data set used by Lachowycz and Jones (2014) covered some greenspaces but did not include private garden greenspaces. The variable may also include land uses other than natural spaces in greenspace categories. If it is necessary to truly capture all healthy happy vegetation in a study then the biophysical measure of greenspace may be necessary.

Chapter 3 Methods

The following text outlines the methods used in this study. It first describes the rationale behind the research design constructed. It then outlines the data sets used for the inventory, usage-based categorical, and biophysical measurement types. Finally, it details the analysis comparing the data types.

3.1. Research Design

This study aimed to examine variations of results given by three measures of greenspace, specifically: inventory of intentionally preserved green spaces such as parks, usage-based categorical categorized land use classifications representing green spaces, and areas shown to have biophysically healthy photosynthesizing “green” vegetation through the use of a normalized difference vegetation index determined from satellite imagery. As these are different measures which are all widely used, this study sought to examine their coverage relative to one another and examine them spatially.

3.1.1. Workflow

The methodology consisted of five steps. First, state level data was collected. Second, four representative study sites were selected. Third, study areas were selected in each of these areas in three separate urbanization levels. Fourth, percent greenspace cover was calculated in each of the study areas using each of the three measures being studied. Fifth, the cover of greenspace was compared between measures. This workflow is outlined in Figure 2 and is discussed in detail on the following pages.

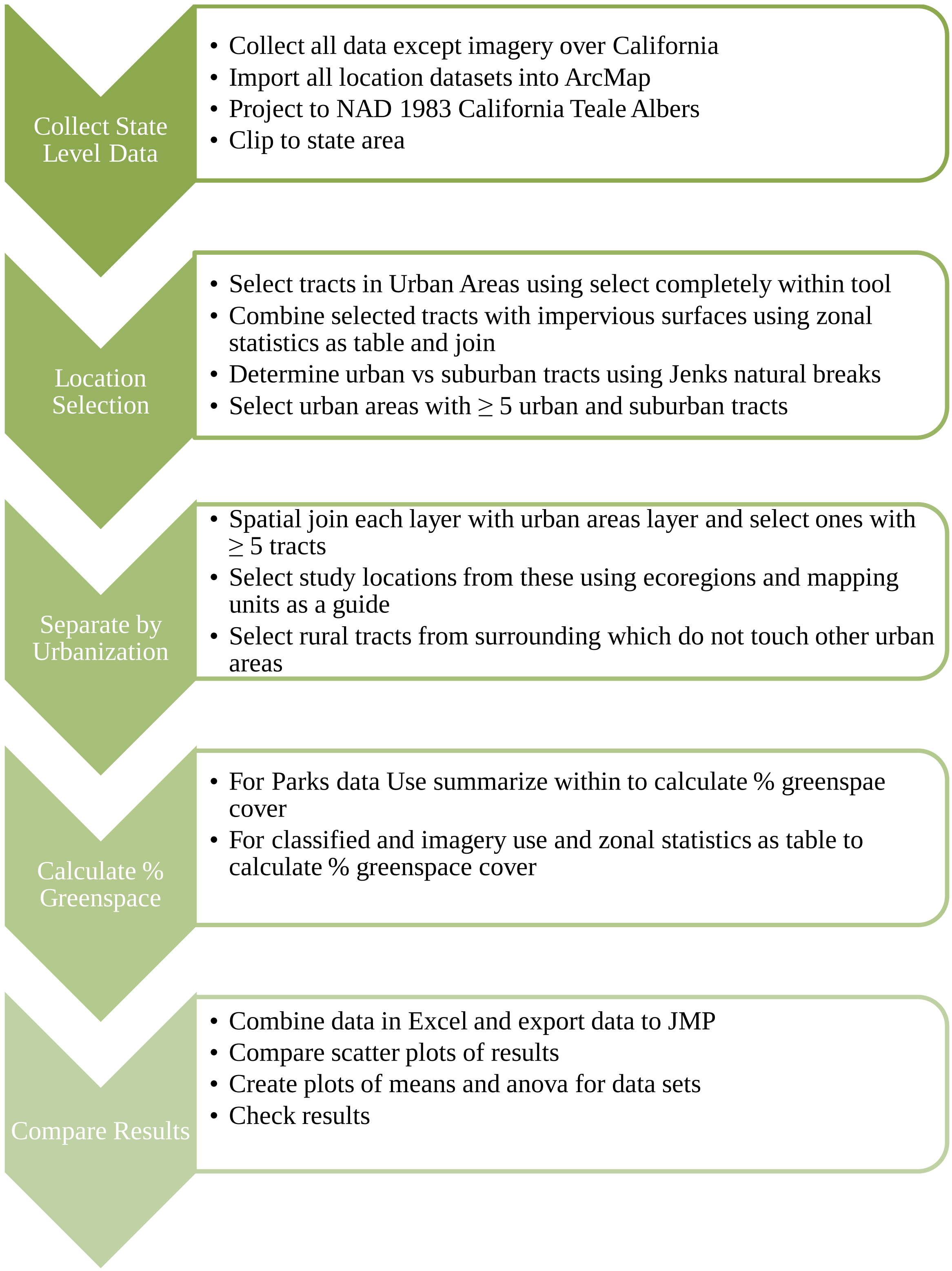

Figure 2 Thesis workflow

3.1.1.1. Software used

The study used Esri ArcGIS 10.5 Desktop for all data processing and geographic analysis. Practitioners in the GIS community frequently use this software, often for studies measuring geospatial area. Within ArcGIS, this study mainly used tools within the analysis, data management, and spatial analysis toolsets.

The study conducted some organization of statistical data and preliminary analysis using Microsoft Excel 2016. Specifically, this consisted of combining tables of summary, and zonal, statistics and calculating averages for each category.

The study conducted all other statistical analysis of results using JMP Pro 13 which is a widely used statistical analysis software.

3.2. Data Description

This study examined three types of data representing three manifest variables of greenspace. The first variable, inventory, used vector polygons of protected park lands from the CPAD. The second variable, usage based categorical, used an exhaustive mosaic land use data set from the NLCD. The third variable, biophysical, used NDVI to show the varying levels of green vegetation across a landscape, processed from Landsat 8 ARD satellite imagery tiles.

This study also required data to represent manifest variables for location factors. These included Census designated urban centers and areas and NLCD impervious surface data to represent urban areas, Census tracts to represent neighborhoods, and NLCD mapping units and California ecoregions to represent mapping units expected to contain vegetation with similar characteristics.

3.2.1. Type and Quality Requirements

This study required location data sufficient to select and delineate study areas representative of neighborhoods in multiple urbanization levels and ecological regions within the broader study area. Therefore, neighborhoods needed to be sub-city resolution, preferably small enough to have five of each type and maintain association with the core urban area. These neighborhood polygons also needed to be large enough to encompass greenspaces such as parks. The study required a data set representative of a variable which could reasonably be used to represent a level of urbanization to classify the neighborhoods into urban, suburban, and rural types. It needed a data set bounding the broader study area, in this case the state of California. It also needed a data set dividing that bounding study area into geographic areas of some similarity with relation to vegetation characteristics. This would allow for selection of varied geographic conditions to obtain a representative sample.

Parks (inventory greenspace) data needed to clearly represent the intentionality of greenspace use of land, i.e. public park, natural area, other intentional vegetated area to set it apart from other land. This data also needed to have defined areas, which suggested a vector data format. The spatial resolution needed to be small enough to have decent resolution within neighborhoods, preferably parcel level, but not so small that the resolution slows down processing times impractically.

Classified (usage-based categorical) green space data needed to consist of an exhaustive classification covering the entire bounding area. It needed to include at least one category representative of vegetation or vegetated area. It was preferred that this type of data not have any categories with mixed vegetation/non-vegetation classification, but a large reputable data set with mixed categories was acceptable. It needed to be small enough to have decent resolution within neighborhoods, but not so small that it slows down processing. It was preferred that it be a similar size to that used by current studies.

Imagery (biophysical) data needed to be unclassified satellite remote sensor data, preferably corrected for error and standardized. It needed to have a spatial resolution good enough to distinguish buildings from landscapes in an urbanized setting, or at least a resolution capable of having a variation in vegetation have an effect on the values measured. It was also ideal to have relatively low cloud cover and distortion (<30%) on imagery.

Additionally, all data sets needed to be geospatial, from near or comparable dates, free and publicly accessible, and applicable and consistent in makeup statewide. Similarities in spatial and temporal resolution were preferred, but variation was allowed for as it aided in addressing the potential of varying suitability of measurement types in differing circumstances.

3.2.2. Sources Used

Now that the ideal parameters for data have been established, the types used in this study and their suitability based on these criteria will be described. The following subsections describe the data sources used for this study.

To meet the data type requirements described in the previous section at a contiguous statewide level this study used data from national data sets. Specifically, it used data from the U.S. Census Bureau, the U.S. EPA, and the NLCD to determine project and study area boundaries. U.S. Census tracts from 2010 provided a representation of neighborhoods within and near cities. The study used U.S. Census Urban Areas to represent urbanization. The U.S. Census Bureau describes their urban areas as comprising of "a densely settled core of census tracts and/or census blocks…. along with adjacent territory containing non-residential urban land uses as well as territory with low population density included to link outlying densely settled territory with the densely settled core” (U.S. Census Bureau 2015). This study treats census tracts which are completely within an urban area as the core area which is then subdivided into urban and suburban. It treats the tracts partially within and surrounding urban areas as being rural since they contain land outside of the area. To represent the level of urbanization within the urban areas and make it divisible into suburban and urban categories, the study used NLCD percent impervious surface data.

To delineate the bounding study area the study used U.S. Census state shapes data. It used U.S. EPA ecoregion data to delineate geographic areas of similar vegetation types.

All data sets used for boundaries were in vector data format. The one data set in raster format was the NLCD impervious surface data. The data sets used for location selection are summarized in Table 2 Location data sources.

Table 2 Location data sources

Name	Type	Temporal Scale	Spatial Scale	Status	Source
2010 census state	Shapefile	2010	California	Free online	U.S. Census Bureau
2010 census designated urban centers and areas	Shapefile	2010	United States	Free online	U.S. Census Bureau
2010 census tracts	Shapefile	2010	tracts	Free online	U.S. Census Bureau
2011 NLCD impervious surfaces	Raster	2011	30m over the U. S.	Free online	Xian, G., et al. 2011.
NLCD mapping units landcover_bndry_022307	Shapefile	2001	United States	Free online	Homer, C., et al. 2007.
Level 3 ecoregions of California	Shapefile	2012	California	Free online	U.S. EPA Office of Research and Development (ORD) - National Health and Environmental Effects Research Laboratory (NHEERL)

To meet the selection criteria for parks data, this study used the California Protected Areas database (CPAD 2017) to represent intentional green space. This data set includes all intentionally protected lands within the state of California with a scale range of 1:5,000 to 1:150,000,000 and is summarized in Table 3 Parks data source. This data set is updated biannually. As this data is a compilation there is an attribute field showing the source of the original data (assessor's office, etc.). It is worth noting that, as this data is a compilation, some unreported green space may be underrepresented. Since this type of compiled data is used in exposure studies, this study treated this potential for missing data as a parameter of this type of greenspace measurement.

Table 3 Parks data source

Name	Type	Temporal Scale	Spatial Scale	Status	Source
CPAD_2017a_Holdings	Feature Class	2017	1:5,000 to 1:150,000,000	Free online	California Protected Areas Database

To meet the selection criteria for classified data, this study used the 2011 NLCD to represent land cover. This data set classifies all lands in the United States into 20 different land cover classes, 13 of which are vegetative classifications. This data is recorded at a 30m scale and is shown in Table 4 Land cover data sources, below.

Table 4 Land cover data sources

Name	Type	Temporal Scale	Spatial Scale	Status	Source
nlcd_2011_landcover_2011_edition_2014_10_10	Shapefile	2011	30m over the U. S.	Free online	Xian, G., et al. 2011.

The NLCD provides land cover classification mapping for the purpose of countrywide land cover classification and change detection (U.S. Department of the Interior, 2018). The methods used to create each iteration, roughly five years apart, uses more improved techniques than the last. For example, the 2011 data set preparation process used imagery data from multiple time periods within the data collection window to account for seasonal dormancies of vegetation (Homer et al. 2015). To preserve comparability between years the creators of the NLCD update prior year data sets with each new release (Homer et al. 2015).

The 2011 NLCD incorporated data from Landsat 5 TM imagery from leaf on and leaf off dates between late 2009 and late 2011. The creators, Jin et al. (2013), ran this imagery through two change detection models, multi-index indicated change analysis, and Zone. These models both used multiple indices to produce two change maps each. The MIICA model used "differenced Normalized Burn Ratio (dNBR), the differenced Normalized Difference Vegetation Index (dNDVI), the Change Vector (CV), and a new index called the Relative Change Vector Maximum (RCVMAX)" (Jin et al. 2013). The Zone model used dNBR and dNDVI. The data provided by these models created the "maximum potential spectral change map" used in the creation of the 2011 NLCD classifications. In addition to the imagery, the NLCD incorporated elevation and associated derived data, soil, crop, and wetlands data into its final map product (Homer et al. 2015).

As stated previously, the NLCD groups land covers into 20 classes which are grouped together (U.S. Department of the Interior, 2018). There are two water classifications, open water and perennial ice/snow, which signify a majority of water coverage or at least 25% cover or frozen water. There are four developed classifications showing various levels of intensity. The impervious surface data set was a main component of the creation of these classifications. There is one classification termed "barren," which, as it sounds, represents uncovered land with very little vegetation. There are three forest classifications representing seasonally dormant, evergreen, and a combination of both types of trees. There are two classifications representing shrublands, one of which is only used to describe a locally appropriate type within Alaska. The other classification represents vegetated areas with moderately high vegetation with at least 20% cover. There are four herbaceous classifications, three of which are, again, specific to Alaska and not pertinent to this study area. The one nationwide herbaceous category represents grass rangeland and other areas with a high percentage of grasses and herbs. There are two planted/cultivated classifications representing managed perennial grass and legume lands and annual crop production. Finally, there are two wetlands classifications representing marshy areas with woody and non-woody vegetative cover (U.S. Department of the Interior, 2018).

This study used nine of these classifications which are indicated in figure 3. These were: deciduous forest, evergreen forest, mixed forest, shrub/scrub, grassland/herbaceous, pasture/hay, cultivated crops, woody wetlands, and emergent herbaceous wetlands. The classifications for water, developed land, barren land, and those only used in Alaska were excluded.

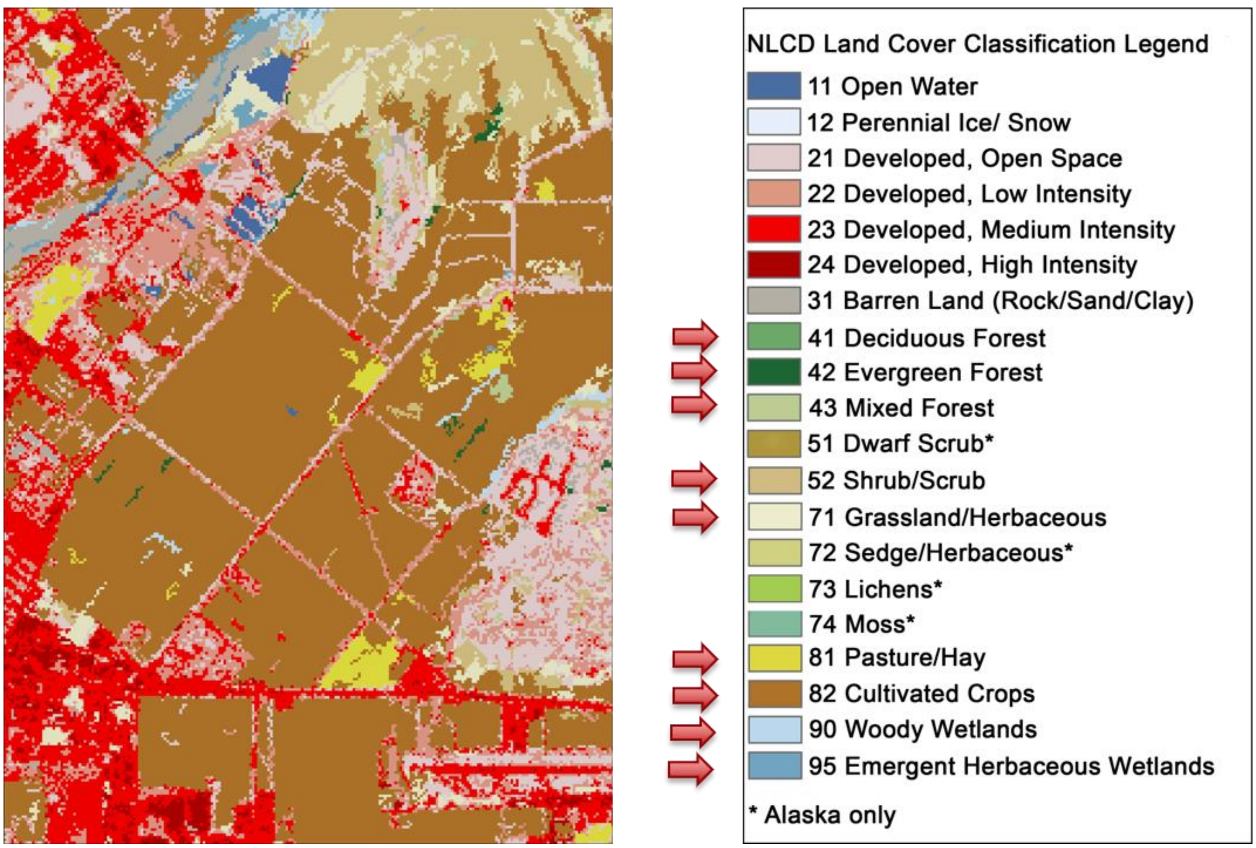

Figure 3 NLCD classifications

The developed classifications within this data set are comprised of a range of mixtures of green space and urbanized land. This study excludes those classes from the set representing green space as it is unknown exactly how much green space is in each cell of those types. While these types allow for inclusion of vegetation, visually they appear to mainly demarcate urban appearing elements such as roads. This can also be seen in figure 3. While excluding these areas technically creates an underrepresentation of green spaces within the area, this study treated this underrepresentation as a parameter of the type of greenspace measurement as data sets used for green space exposure studies are not necessarily created for that specific purpose.

To meet the selection criteria for imagery data, this study used imagery tiles obtained from the United States Geological Survey Land Satellite number 8 (USGS Landsat 8) Analysis Ready Data (ARD) to represent remotely sensed imagery. While imagery is used in the creation of the NLCD and therefore these sets are not mutually exclusive, imagery and land cover data sets are both often used separately as representations of greenspace (Dalton, et al. 2016; Wheeler, et al. 2010; Van Den Bosch, et al. 2016; Wu, et al. 2014; Younan, et al. 2016). Because of this, it is useful to compare them to one another to better understand how they each represent greenspace.

This study utilized imagery mainly from the months of February and March 2017, with some imagery from April used when the previous months had too much cloud cover or were otherwise unavailable. In doing this, the study was able to maintain all imagery data with less than 30% cloud cover. The seasonality of the data ensured the imagery captured vegetation within the spring growing season, which is often done when selecting imagery for NDVI greenspace representation. This selection highlights the difference in data product between greenspace represented in this manner and that represented by the NLCD which was created using imagery from multiple seasons. Another difference highlighted between these two representations of greenspace is the temporal element of the data. The study used the 2017 year as it was a good rainfall year after several drought years which would not have been as suitable for this study (Fresno State n.d California Protected Areas Database.). While this differs in time from the 2011 NLCD data, 2011 was also a year following good winter rainfall (Fresno State n.d.) so the data sets are as comparable as possible for the time difference. The rainfall from July through February of the 2011 and 2017 years, as well as the five years preceding each of them, is shown in table 5. The Fresno and Oxnard areas best exhibit the aforementioned trend, although the Indio and Vallejo areas are also similar between the two study years and the ones preceding them.

Table 5 Yearly rainfall from July through January in inches in the vicinity of each study area

Time Period	Fresno Area	Oxnard Area	Indio Area	Vallejo Area
7/2016 – 2/2017	9.48	10.28	0	17.4
7/2015 – 2/2016	8.92	0.67	0	10.18
7/2014 – 2/2015	4.67	5.9	0	14.98
7/2013 – 2/2014	0.73	0.78	0	1.32
7/2012 – 2/2013	6.29	4.65	0	12.56
7/2011 – 2/2012	3.3	5.16	0	7.07
7/2010 – 2/2011	10.2	11.96	0	12.64
7/2009 – 2/2010	6.14	7.16	0	17.06
7/2008 – 2/2009	4.14	2.64	0	5.13
7/2007 – 2/2008	6.32	6.6	0	15.18
7/2006 – 2/2007	2.42	4.48	0	7.12
7/2005 – 2/2006	6.19	4.76	0	16.37

To exhibit biophysical greenness, the study used the Red and Near Infrared bands to create the NDVI. This study used the NDVI as a single stand alone manifest variable for greenspace in this instance, as opposed to one of several indices used across multiple imagery sets, then combined with other land use data as in the NLCD. The details of the imagery used are summarized in Table 6 Imagery data sources, and Table 7 Individual Landsat tiles used, below.

Table 6 Imagery data sources

Name	Type	Temporal Scale	Spatial Resolution	Spectral Scale	Status	Source
Red imagery	ARD Surface reflectance tiles	2017 February through April	30 m	0.63 – 0.69 µm	Free online	USGS Landsat 8
Near Infrared imagery				0.76 – 0.90 µm		

Table 7 Individual Landsat tiles used

Vertical	Horizontal	Acquisition Date	Cloud Cover	Cloud Shadow	Snow, Ice
008	001	2017-03-17	0.4139	0.0130	0.0023
009	001	2017-04-27	26.3468	3.6698	0.0002
008	002	2017-04-27	8.8450	2.6003	0.0024
009	002	2017-03-03	25.6927	5.3403	0.0420
009	002	2017-04-27	13.3745	3.7185	0.0050
009	003	2017-03-03	6.7694	1.9488	31.9149
010	003	2017-03-12	9.9085	2.9452	18.0170
011	003	2017-03-12	5.3436	1.2512	0.1703
012	003	2017-02-01	6.1845	1.6016	02226
011	004	2017-03-14	1.3808	0.4657	0.0309
012	004	2017-03-14	2.9796	0.8592	0.3421
013	004	2017-03-07	1.2033	0.3449	0.0132
013	004	2017-03-14	1.3263	0.2523	0.0049
012	005	2017-03-07	1.1692	0.1263	0.0000
013	005	2017-03-23	0.3570	0.0358	0.0000

Since it was necessary to project some of the data, there was not a shared coordinate system, this study used the California Teale Albers projected coordinate system for the analysis. California Teale Albers is a projected coordinate system that covers the full extent of the state of California effectively. It is more localized than using a national system which increases geographic accuracy. It is also broader than the state plane system which allowed the analysis to be conducted in a shared system for all study areas.

3.3. Analysis

An outline of the analysis for this thesis is shown below. As stated in Software used, the study conducted this analysis using ArcMap 10.5.1 and JMP Pro 13. It calculated green space area within each study area tract using intentional, classified, and remotely sensed data in ArcMap and then compared the means using JMP Pro 13.

3.3.1. Location Selection

This study focused on four separate urban centers to look at greenspace within different geographic regions while still grouping sets within population centers. To do this, this study used the U.S. Census layer for urban centers and areas as it represents municipal areas which happen to be large enough that they might accommodate more census tracts within and near them associated with that municipal area. According to the file description, the areas are considered urban and everything outside of them is considered to be rural. This study therefore further subdivided the urban centers into urban and suburban areas by impervious surface cover (U.S. Census Bureau 2015).

In looking at municipal areas, this study focused on urban, suburban, and rural greenspace measurement as studies relating greenspace to public use or effect in some way are often separated in to one or more of those groups (Ambrey 2016; Lachowycz, et al. 2012; Van Den Bosch, et al. 2016; Wheeler, et al. 2010; Wu, et al. 2014; Younan, et al. 2016). It identified rural areas as ones outside the urban centers layer. It then identified urban and suburban areas as being areas within the layer as they are both, in some part, urbanized. The study separated these areas using impervious surface cover in tracts within all urban areas in the state for better representation of the split.

To do this, the study used a file of Census tracts within all urban areas within the state as zones for a zonal statistics function. As each census tract created a separate zone, the study was able to calculate the mean values for percent impervious surface cover within each census tract. It then identified the natural break between the two classes for the set of all tracts within urban areas using classification with the natural breaks Jenks method. The natural break was at 53.3 percent impervious surface cover. Therefore, tracts with over 53.3 percent mean impervious surface cover were considered urban and tracts below that threshold were considered suburban.

To maintain similarity with other exposure studies, this study used U.S. Census Tracts to represent neighborhoods (Ambrey 2016). In this case tracts were classified as urban, suburban, or rural. With rural tracts crossing in to the urban centers but including outside area and no area from neighboring centers. This was intended to capture rural area associated with the center.

Once the tracts within the urban centers layer were identified as either urban or suburban, centers with over 5 tracts in each category were selected. From these centers, ones in separate mapping units were identified and further narrowed down by selecting ones with more tracts to allow a better random subset to be selected for analysis.

Tracts were chosen randomly within the urban and suburban classes as no further categorization of neighborhoods was required for this study. Rural tracts were chosen by first selecting tracts crossing the urban boundary, and then going outward to choose more tracts until five were reached excluding tracts crossing into other urban centers.

3.3.2. Variables Measured

While the various variables and their representations are described throughout the analysis section of this study, it is prudent to compile them together for reference. This small subsection provides that quick reference, broken up in to location and analysis variables. For the analysis, Census tracts bounded the study areas where the percent cover of vegetated area represented greenspace. This study required variables describing properties of locations and types of greenspace measures. These variables and their representations are described below. The specific data used for these representations was outlined in the previous section on Data Description.

Table 8 below summarizes the overall list of variables observed within this study and the tangible factors used to represent them in this instance. This shows each variable of interest for determining suitable, representative locations for the study. Regional variability is represented by land cover mapping units and ecoregions. The exposure zone unit chosen was the neighborhood level. Urbanization is represented using designated urban centers, and at a finer grain using percent cover of impervious surfaces.

Table 8 Study variables for location selection

Variable	Representation
Urbanized areas	Classified urban areas and centers
Urbanization within areas	% cover of impervious surfaces
Exposure zones	Neighborhoods
Areas of similar vegetation types	National level land cover mapping units and ecoregions

Table 9 below summarizes the variables analyzed within this study and the data types used to represent them for this work. This shows the three greenspace measures examined and compared. Parks and designated open spaces represent inventory greenspaces. Greenspace related categories in a national land cover system represent usage-based categorical classification, and NDVI from remote sensed imagery represents biophysical greenspace.

Table 9 Study variables for measurement analysis

Variable	Representation
Inventory of intentional greenspace	Parks and designated protected open space
Usage-based categorical classified greenspace	Categories including exclusively greenspace in a national level land classification system
Biophysical greenspace, specifically vegetation itself	NDVI from remotely sensed imagery

3.3.3. Greenspace Analysis Using Inventory Data

The study intersected parks polygons with study census tracts. It then tallied the resulting areas of park land within each tract using the summary statistics tool, as this tool works with vector data. This tool calculated the average amount of park land in each tract for the full study area, as well as each city and each urbanization class.

3.3.4. Greenspace Analysis Using Usage-Based Categorical Data

The study selected land use polygons representing all vegetated land covers which were not urban or industrial and reclassified this set into a layer of green land covers. It then intersected the layer with the study tracts. For this part of the analysis, it used the zonal statistics as table tool. To do this, the study used the file of study area Census tracts as zones for a zonal statistics function. As each census tract created a separate zone, the study was able to calculate from the sums of the positive cells and the count of the total cells in the zones, percent cover values for vegetated land use cover within each census tract. The study compared these cover values with those resulting from the park data, and the NDVI data.

3.3.5. Greenspace Analysis Using Biophysical Data

The study downloaded 16 Landsat data scenes to completely cover all study tracts. It then calculated NDVI for each image tile by taking the difference over the sum of the near infrared and red bands of the imagery. It calculated this in ArcMap using the raster calculator tool with the following formula:

Equation 1 Calculating the Normalized Difference Vegetation Index

$$\text{NDVI} = \frac{(\text{Near Infrared} - \text{Red})}{(\text{Near Infrared} + \text{Red})}$$

It then combined all tiles into one larger NDVI image using the create mosaic tool. The study then projected the resulting raster once to a world grid and again to a local system, in this case NAD83 California Teale Albers. The projection was not done directly from the original Albers Equal Area to the California Teale Albers because of a repeating error in the program. Using this method of calculation before projection allowed the NDVI values to be calculated using the original data values rather than resampled values after projections. The study selected cells within the 0.2 to 0.9 range and reclassified them into a layer of green cover. The control set was checked first to verify reasonable values were being obtained.

The study intersected data from this layer with the study tracts using the zonal statistics as table tool as this tool works with raster data. It then averaged the resulting areas in each tract and compared them in the same groupings as the previous data sets.

3.3.6. Comparison of Resulting Data

The study compared the mean percent cover of greenspace measured by each type by plotting the values against each other and using an analysis of variance test (ANOVA). This test is a parametric test which assumes normally distributed data (Minitab 2017.). The ANOVA has some resilience to violation of the normality assumption when sample sizes are 15 or more (Minitab 2017). Therefore, this study used the ANOVA, and accepted the slightly elevated potential for false positive results. This study also used the Tukey-Kramer Honest Significant Difference (HSD) test which takes into account the number of means being compared to reduce false positives. This test, like the ANOVA, assumes normally distributed data but is considered resilient to violations of this rule.

The study repeated the comparison of methods in each urbanization level and geographic area represented to determine if there were any interesting observable patterns or relationships which could be useful for informing research decisions as to which method is best for measuring which situations.

Chapter 4 Results

The following text summarizes the results of the comparative analysis. The text first addresses the results of all study areas together. It then briefly summarizes the overall results of individual urbanization levels and geographic locations in one section where they can be viewed together. Following that, it shows the results of each of these locations in turn.

4.1. Statewide

NDVI from imagery data produced the highest mean percent cover of green space of 76%. CPAD parks data produced the least with a mean of 13%. NLCD produced a mean between the other two methods of 42% but had the highest standard deviation of 42%. The p-values in the ordered differences report using Tukey-Kramer HSD showed all pairs to be significantly different from one another. Figure 4 provides a visual representation of these results.

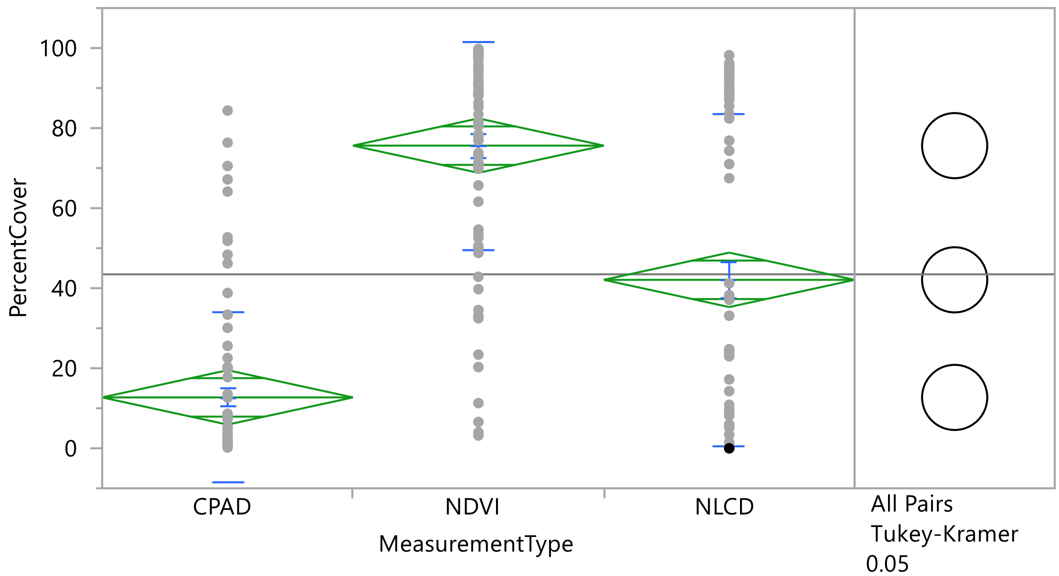

Figure 4 One-way ANOVA of statewide results

Figure 5 exhibits the geographical layout of these results. This provides a spatial view of all data sets used in this study. The main thing to observe in this view is that all data sets did appear in varying amounts throughout the study areas.

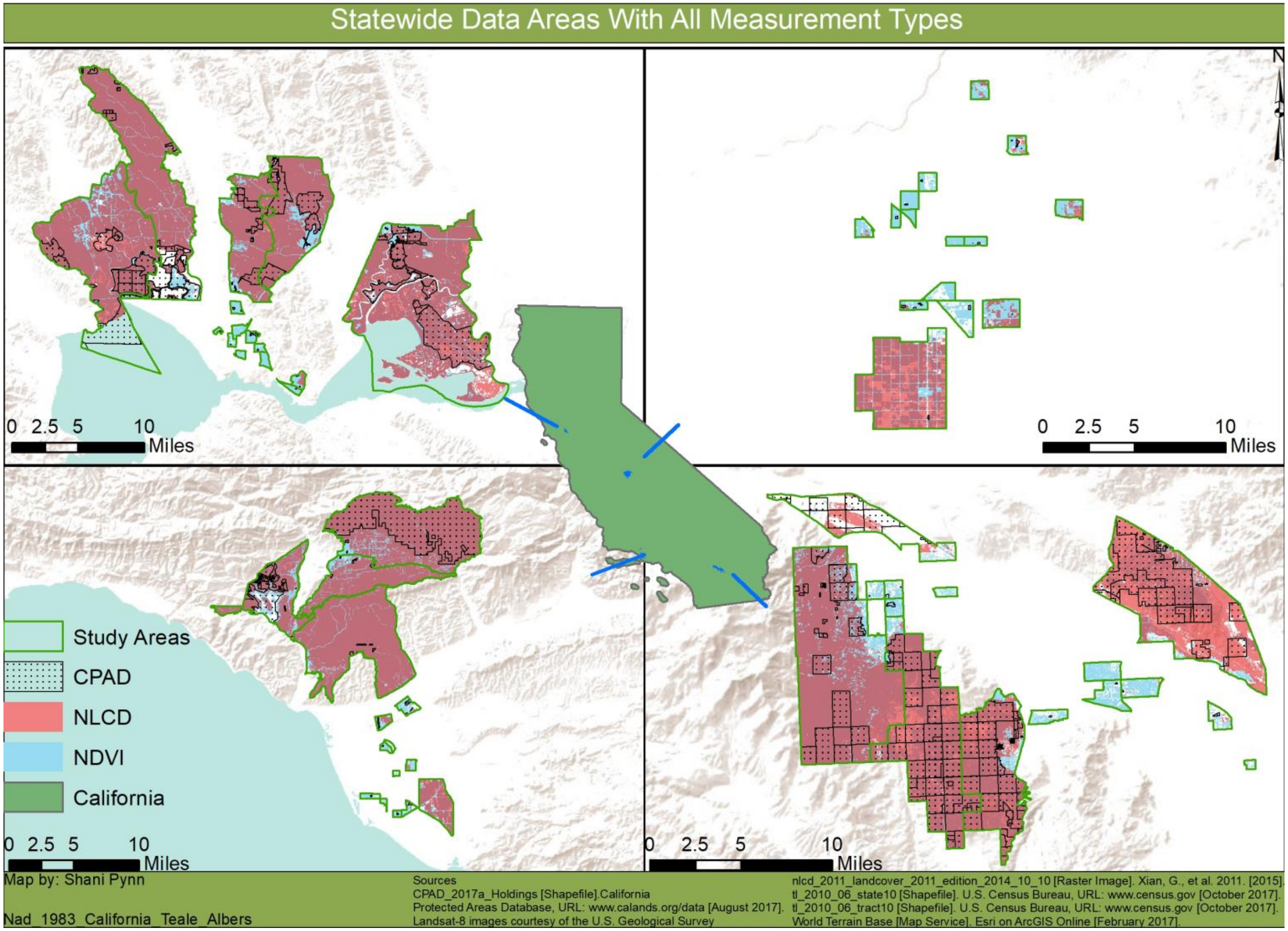

Figure 5 Statewide study area results

4.2. Overall Results

There was an overall trend showing NDVI as the method producing the highest greenspace cover. However, the NLCD method was significantly comparable to NDVI in rural and undeveloped areas. The following sections describe this in further detail.

Before delving in to the results of each area, it is useful to examine them side by side. Figure 6 shows the percent cover captured by each measurement method in the rural, suburban, and urban classes. Visually, the suburban and urban areas showed a marked increase in cover recorded using the biophysical NDVI method. The rural category showed higher levels using the other two methods as well.

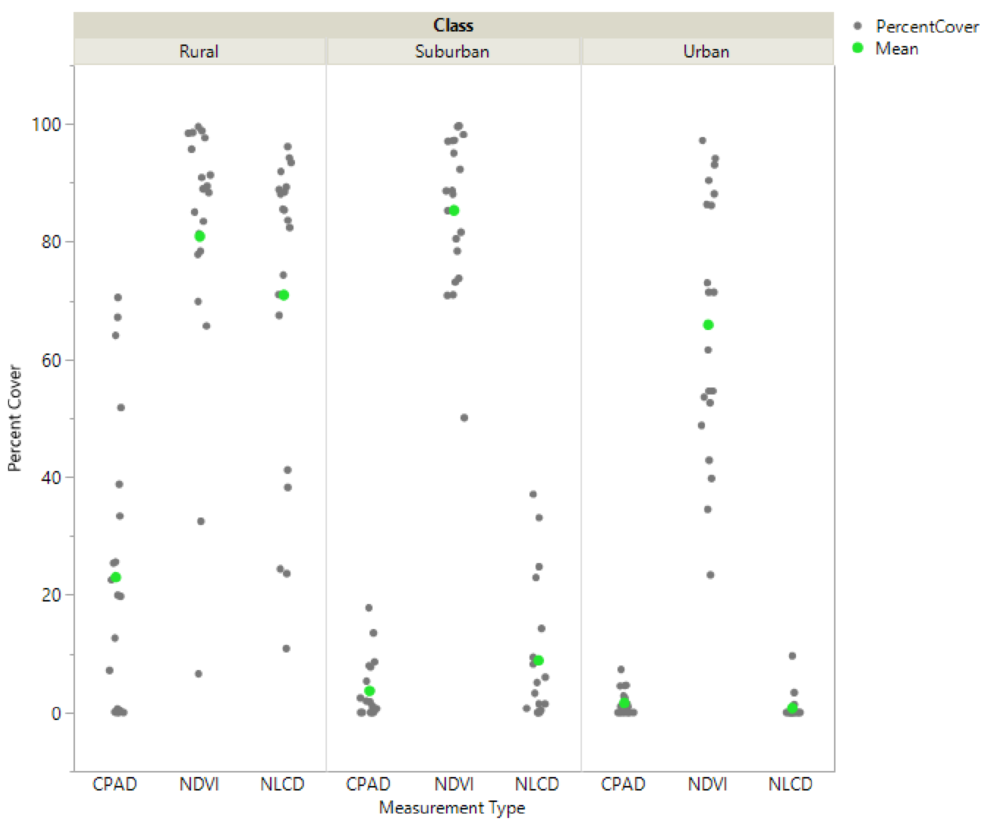

Figure 6 Resulting percent cover by measurement type and urbanization level

As with urbanization, the percent cover captured by each method varied across some of the geographic locations. Figure 7 shows the percent cover captured by each measurement method in the Fresno, Indio Cathedral City, Oxnard, and Vallejo area data sets. Visually, there was more variation exhibited within measurement types within the locations than when grouped by urbanization level in the previous figure. The NLCD captured less relatively in the Fresno area than the others. Also, CPAD and NDVI exhibited more of an even spread across measured percentages in Indio Cathedral City than in the other locations.

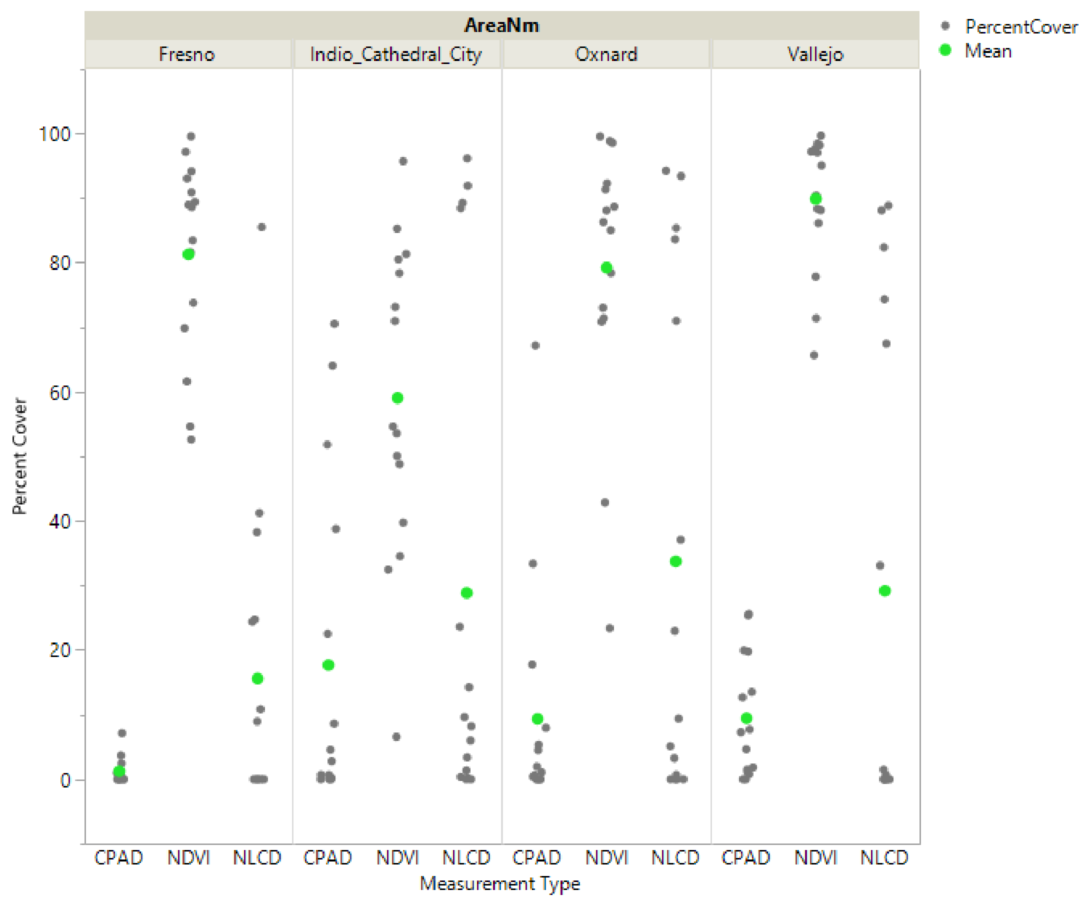

Figure 7 Resulting percent cover by measurement type and geographic area

4.3. Urban

NDVI from imagery data produced the highest mean percent cover of green space of 66% but had the highest standard deviation of 22. NLCD produced the least with a mean of 1%. CPAD parks produced a mean between the other two methods of 2%. However, the ordered differences p-values showed CPAD and NLCD to not be significantly different. All other pairs were significantly different from one another. Figure 8 provides a visual representation of these results.

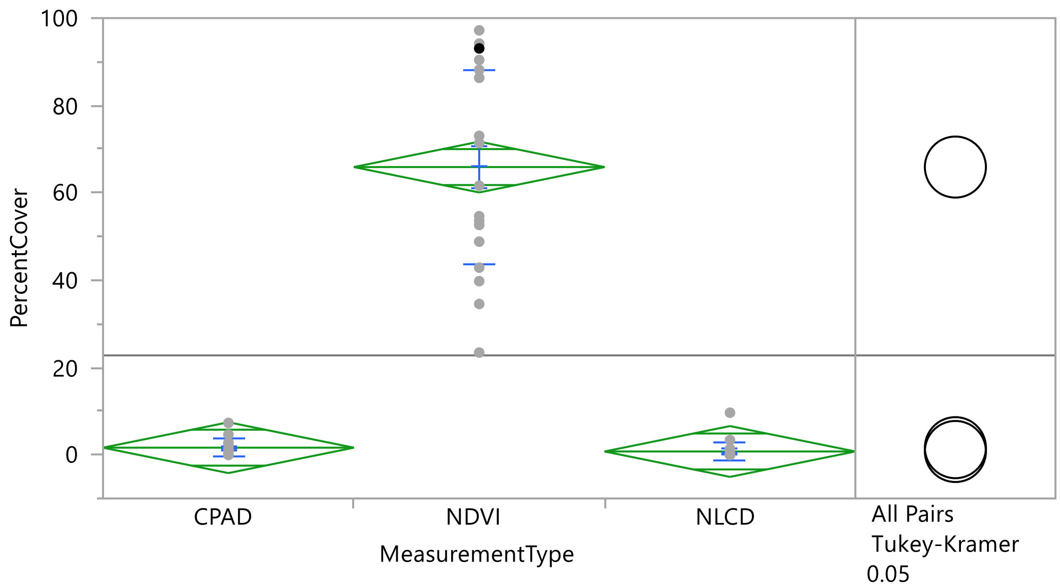

Figure 8 One-way ANOVA of urban results

Figure 9 exhibits the geographical layout of these results. NLCD coverage was lower in urban areas. This is potentially due to the excluded “developed” data types. NDVI coverage was high in these areas, and CPAD data had quite low cover in urban areas. This was fairly consistent across the study areas.

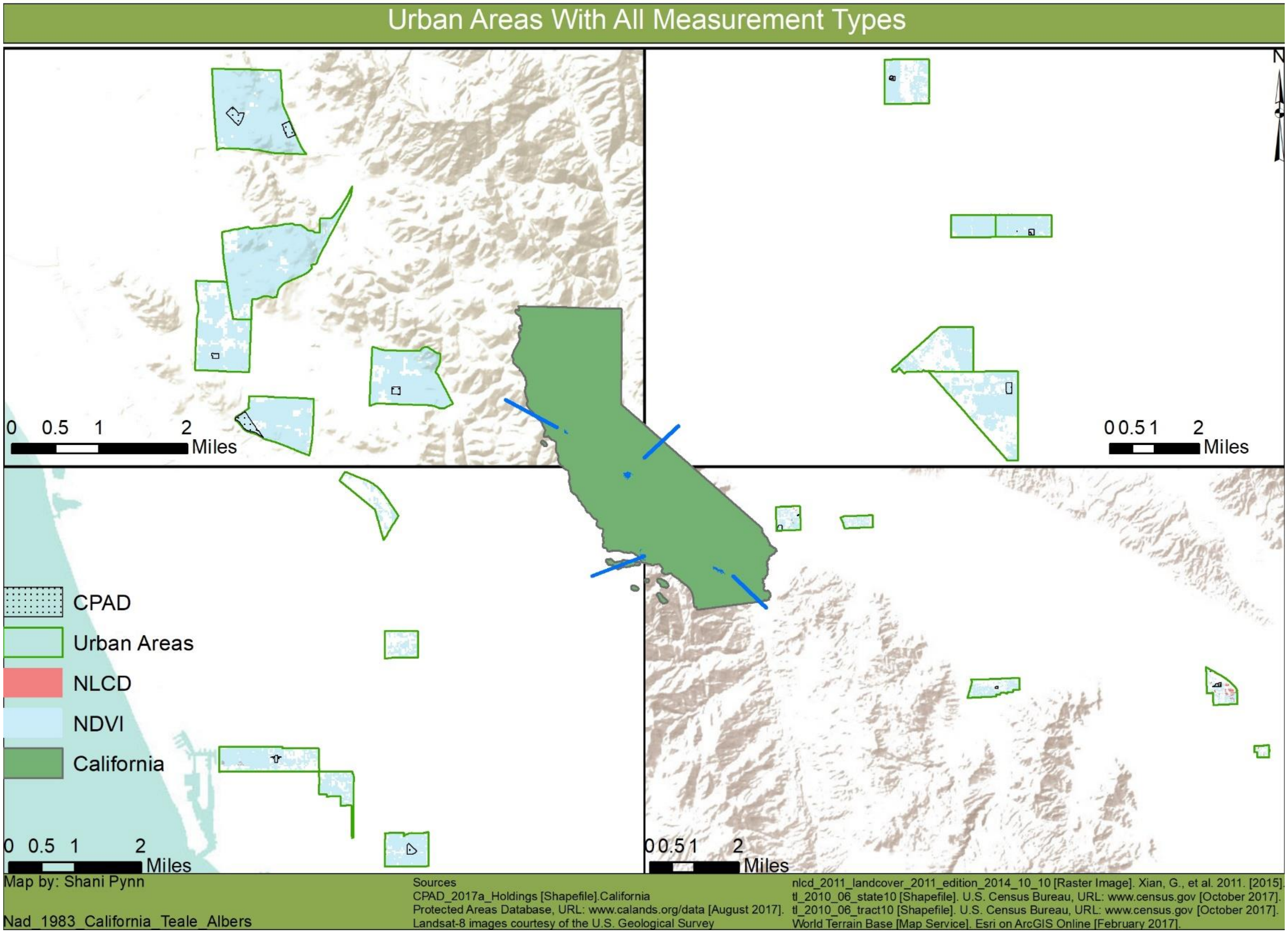

Figure 9 Urban Area Results

4.4. Suburban

NDVI from imagery data produced the highest mean percent cover of green space of 85% but had the highest standard deviation of 13. CPAD parks produced the least with a mean of 4%. NLCD produced a mean between the other two methods of 9%. However, the ordered differences p-values showed CPAD and NLCD to not be significantly different. All other pairs were significantly different from one another. Figure 10 provides a visual representation of these results.

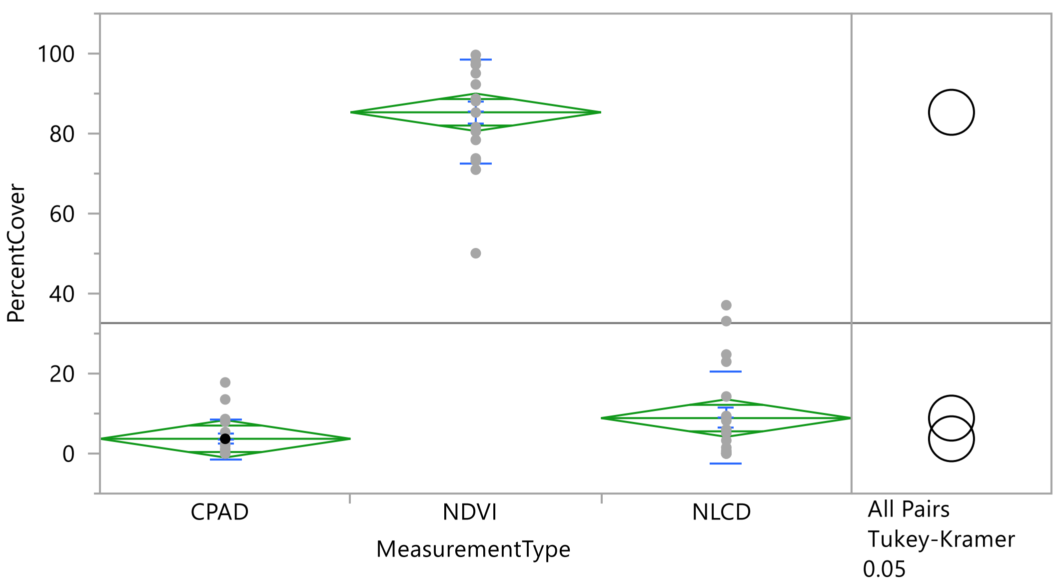

Figure 10 One-way ANOVA of suburban results

Figure 11Figure 9 exhibits the geographical layout of these results. NLCD data had low coverage compared to NDVI but was better represented than in the urban data set. The CPAD data coverage was, again, low.

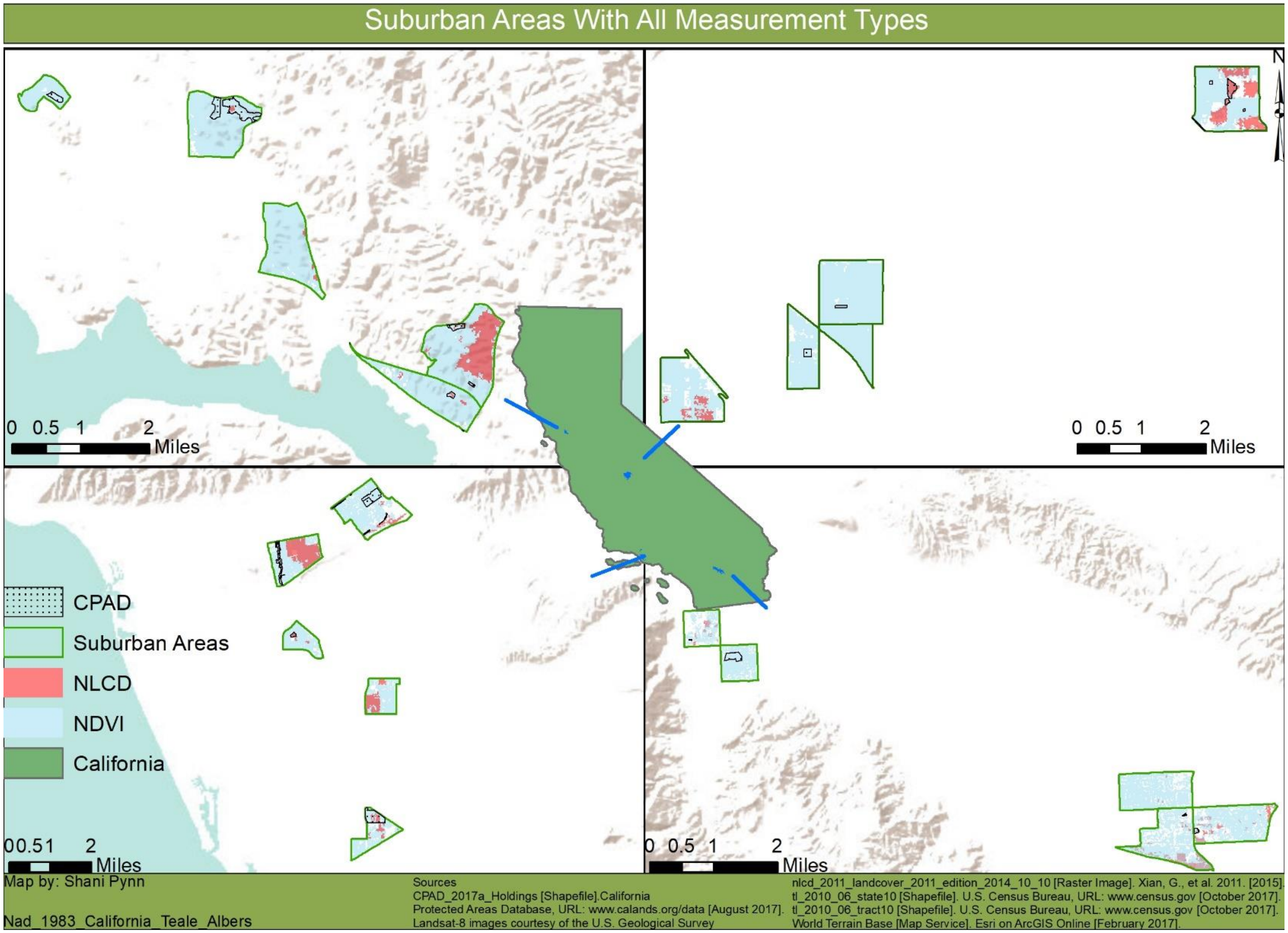

Figure 11 Suburban area results

4.5. Rural

NDVI from imagery data produced the highest mean percent cover of green space of 81%. NLCD followed with a mean of 71% and had the highest standard deviation of 27. CPAD parks produced the least mean percent cover with a mean of 23%. The ordered differences p-values showed NDVI and NLCD to not be significantly different. All other pairs were significantly different from one another. Figure 12 provides a visual representation of these results.

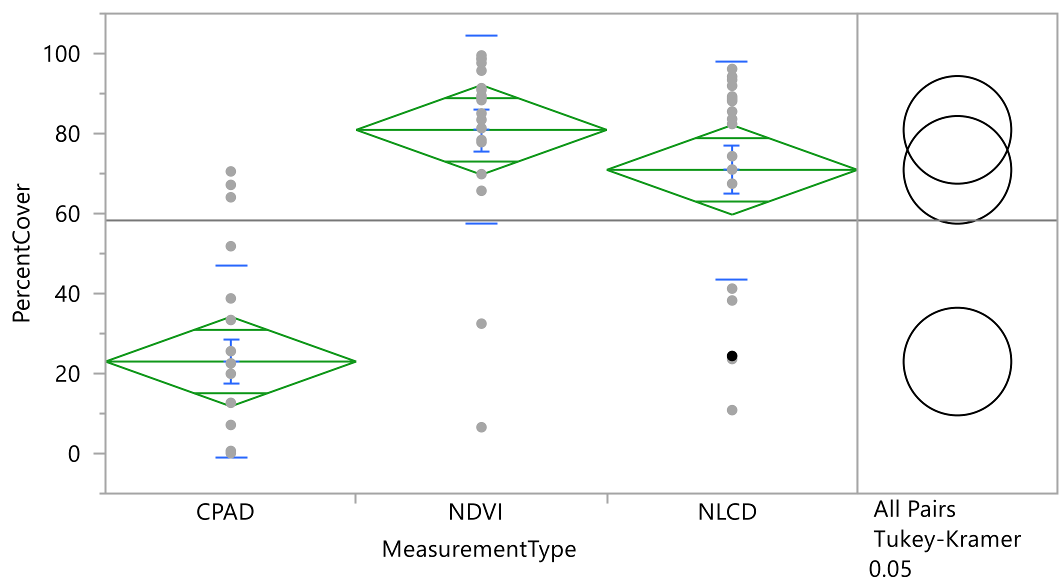

Figure 12 One-way ANOVA of rural results

Figure 13 exhibits the geographical layout of these results. One interesting observation in the rural areas is that there are patches the NLCD captures as greenspace which the NDVI does not and vice versa. Upon visual examination of the data, many of the areas captured as greenspace by the NLCD, but not NDVI, were classified by the NLCD as pasture/hay. Since these can be short lived annual vegetation types there may not have been enough living vegetation present in February and March when the Landsat scenes for the NDVI were taken.

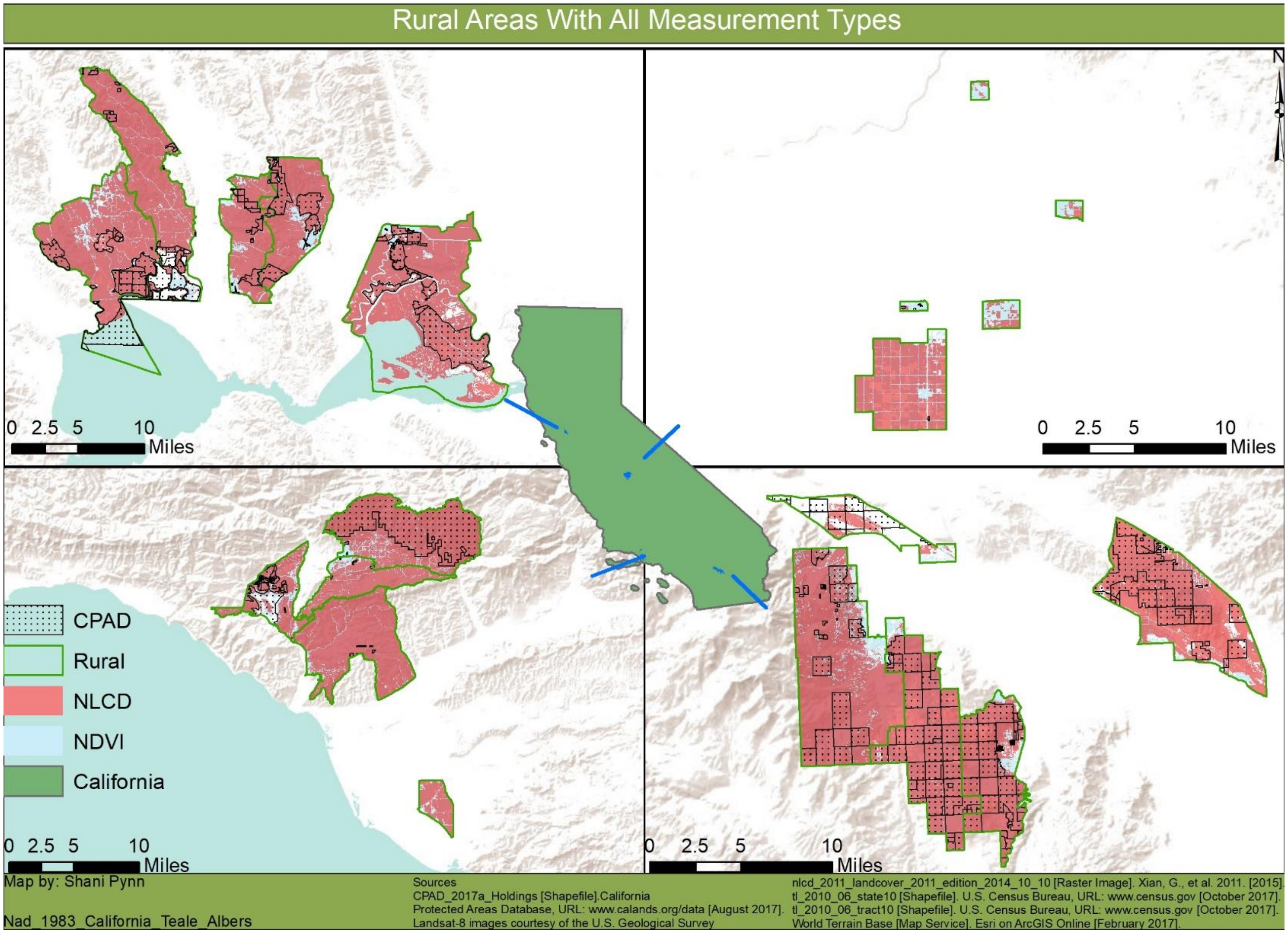

Figure 13 Rural area results

4.6. North Coast

NDVI from imagery data produced the highest mean percent cover of green space of 90%. CPAD parks produced the least with a mean of 9%. NLCD produced a mean between the other two methods of 29% but had the highest standard deviation of 39. However, the ordered differences p-values showed CPAD and NLCD to not be significantly different. All other pairs were significantly different from one another. Figure 14 provides a visual representation of these results.

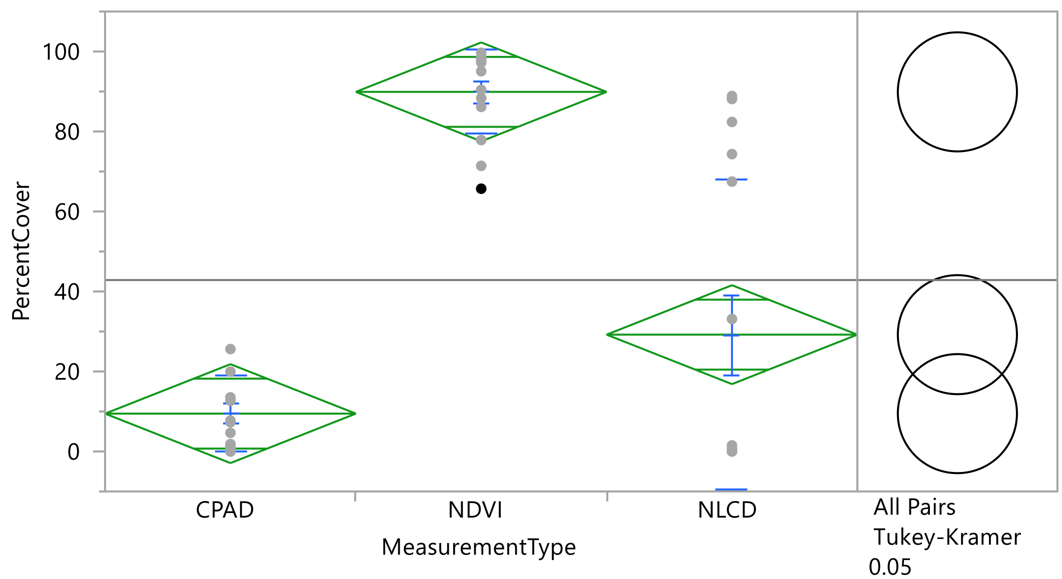

Figure 14 One-way ANOVA of north coast results

Figure 15 exhibits the geographical layout of these results. Here the CPAD data captures a large amount of open water. While the study excluded water from the NLCD and NDVI rasters, the CPAD data set was left intact. The NDVI is more visible than NLCD, without noticeable gaps.

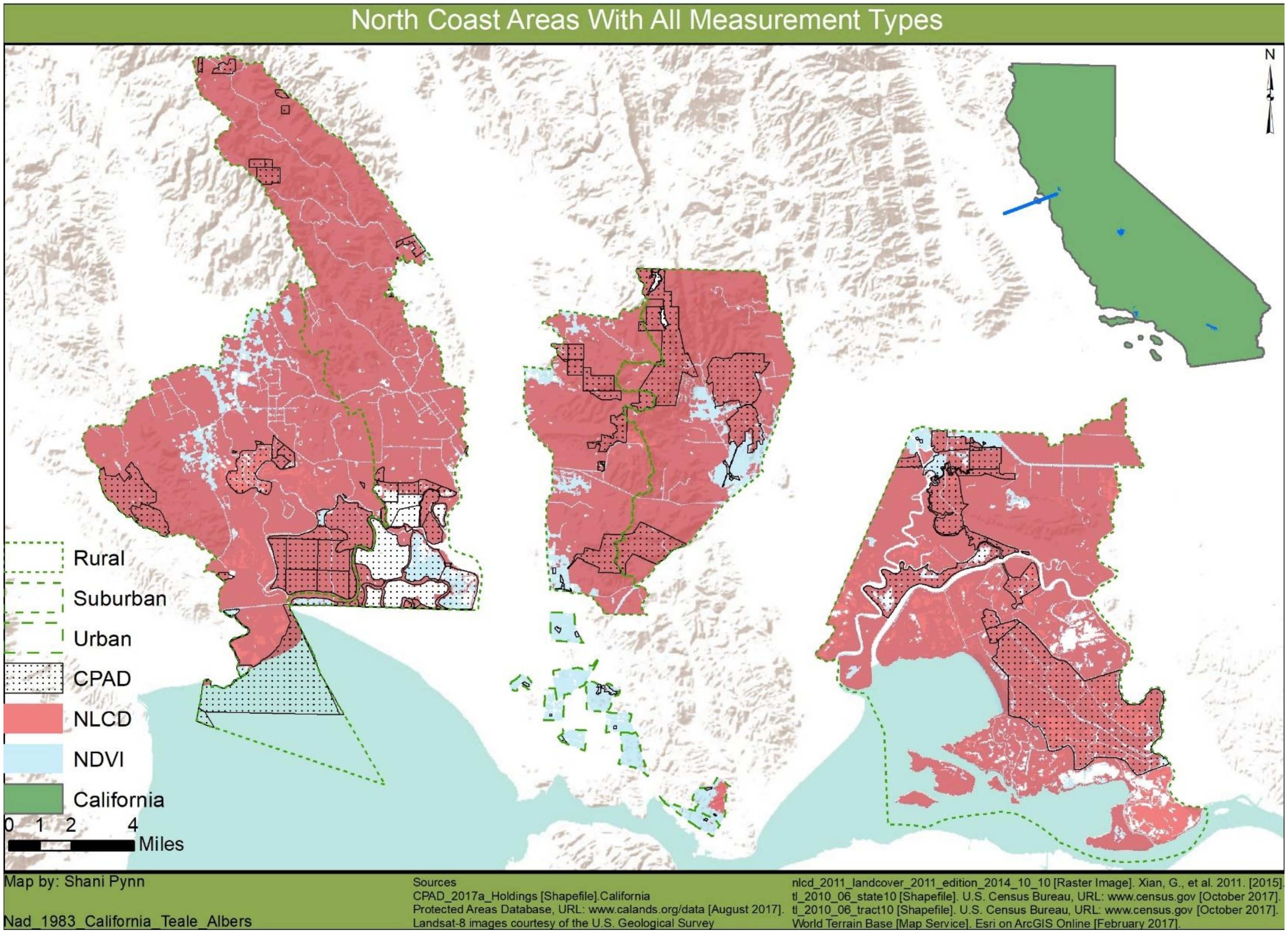

Figure 15 North coast area results

4.7. South Coast

NDVI from imagery data produced the highest mean percent cover of green space of 79%. CPAD parks produced the least with a mean of 9%. NLCD produced a mean between the other two methods of 34% but had the highest standard deviation of 40. However, the ordered differences p-values showed CPAD and NLCD to not be significantly different. All other pairs were significantly different from one another. Figure 16 provides a visual representation of these results.

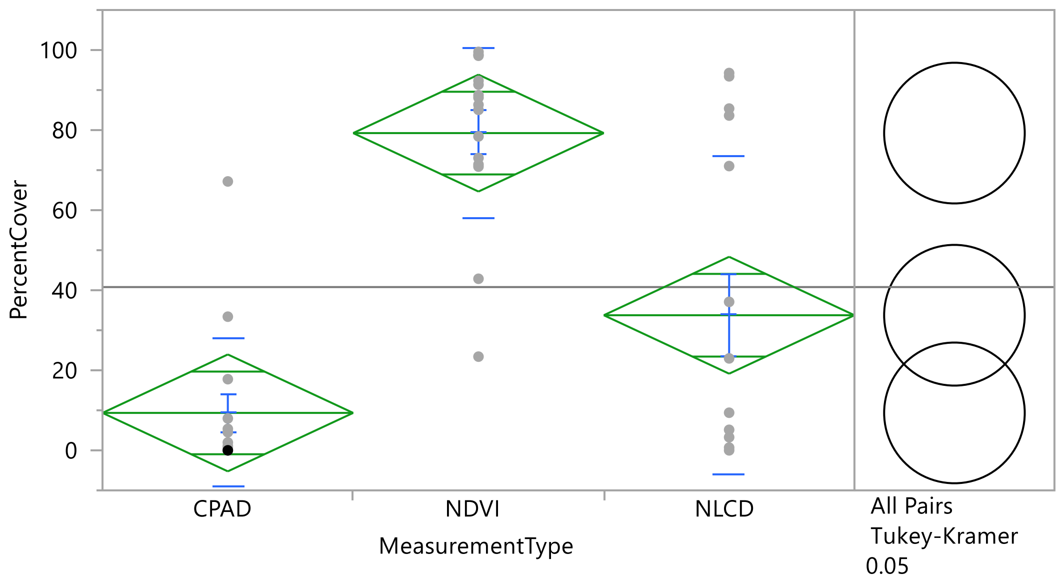

Figure 16 One-way ANOVA of south coast results

Figure 17 exhibits the geographical layout of these results. The three types seem to produce the most similar results in the rural areas. There is a lot of CPAD represented at the edges of the rural areas, farthest from the city and into the mountains. Suburban and urban areas mostly contain the NDVI data and have more sparse greenspace areas defined by the NLCD and inventory types.

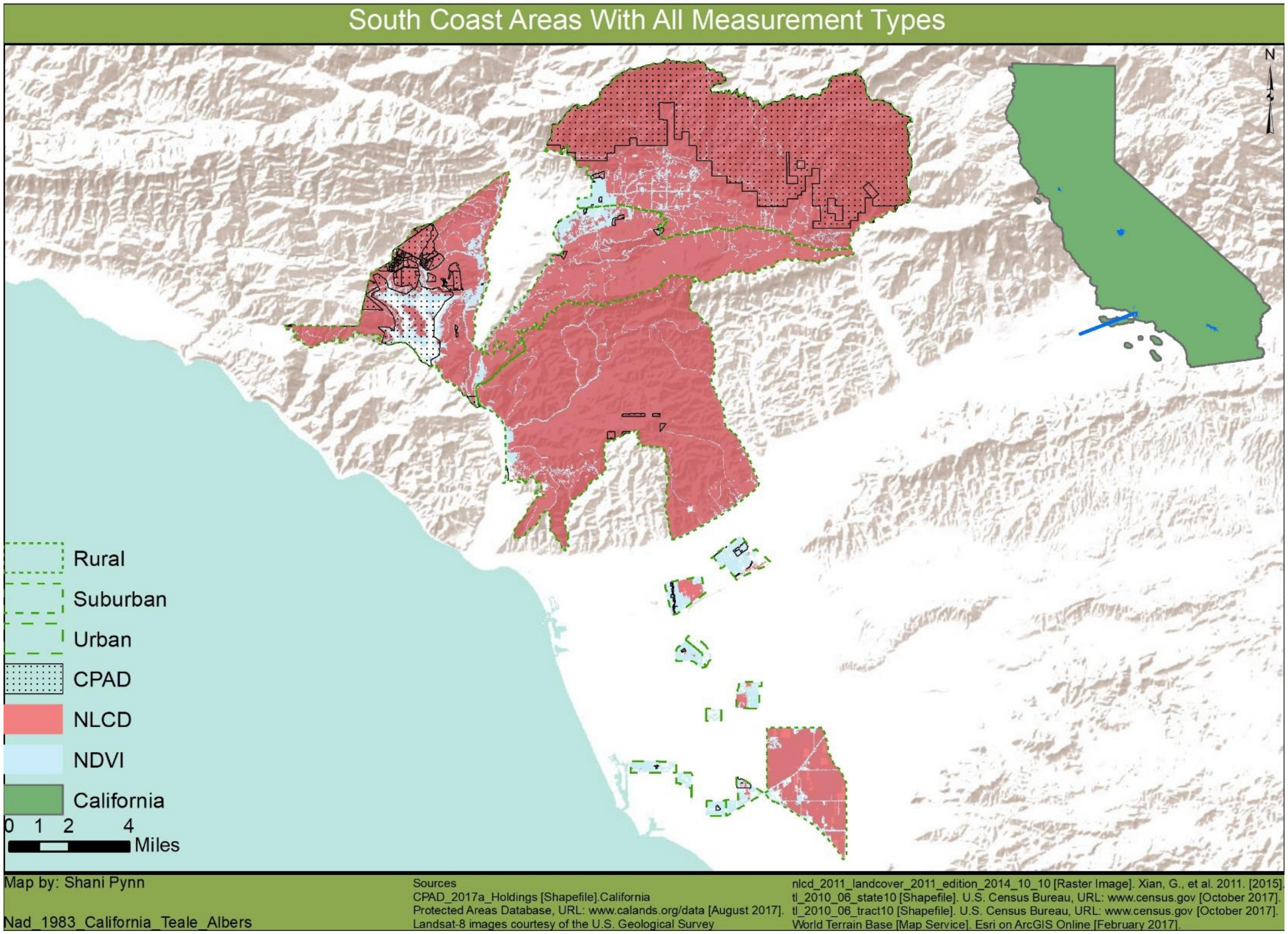

Figure 17 South coast area results

4.8. Inland Desert

NDVI from imagery data produced the highest mean percent cover of green space of 59%. CPAD parks produced the least with a mean of 26%. NLCD produced a mean between the other two methods of 29% but had the highest standard deviation of 40. However, the ordered differences p-values showed CPAD and NLCD to not be significantly different. All other pairs were significantly different from one another. Figure 18 provides a visual representation of these results.

Figure 19 exhibits the geographical layout of these results.

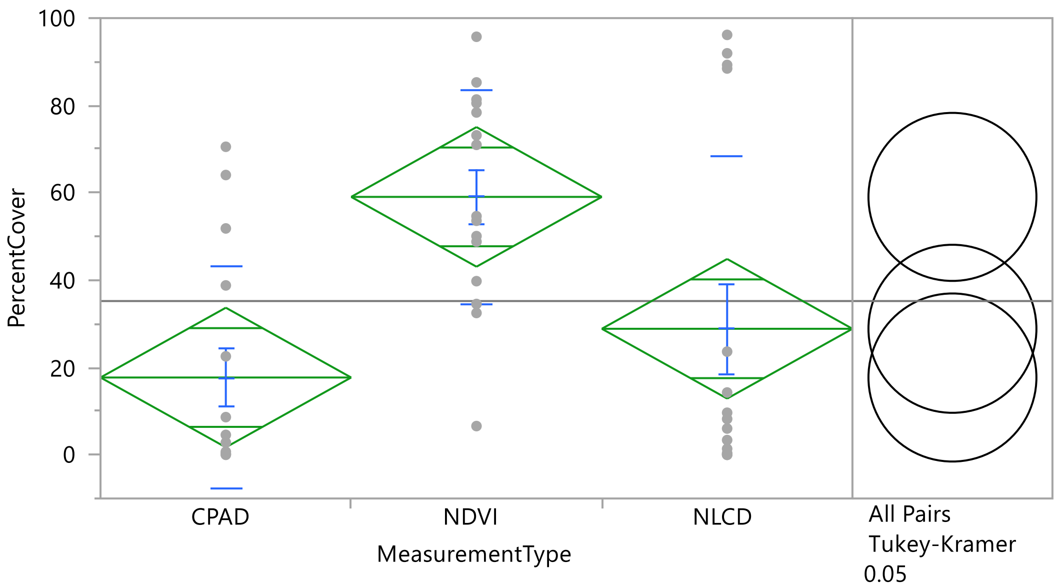

Figure 18 One-way ANOVA of inland desert results

Figure 19 exhibits the geographical layout of these results. In the inland desert there was a much lower rural representation of NDVI data than the NLCD. This may be due to a lower vegetation density in the drier climate. There was also a higher level of representation of the CPAD type of greenspace in these areas. Again, the NLCD data did not appear as well in the urban and suburban tracts.

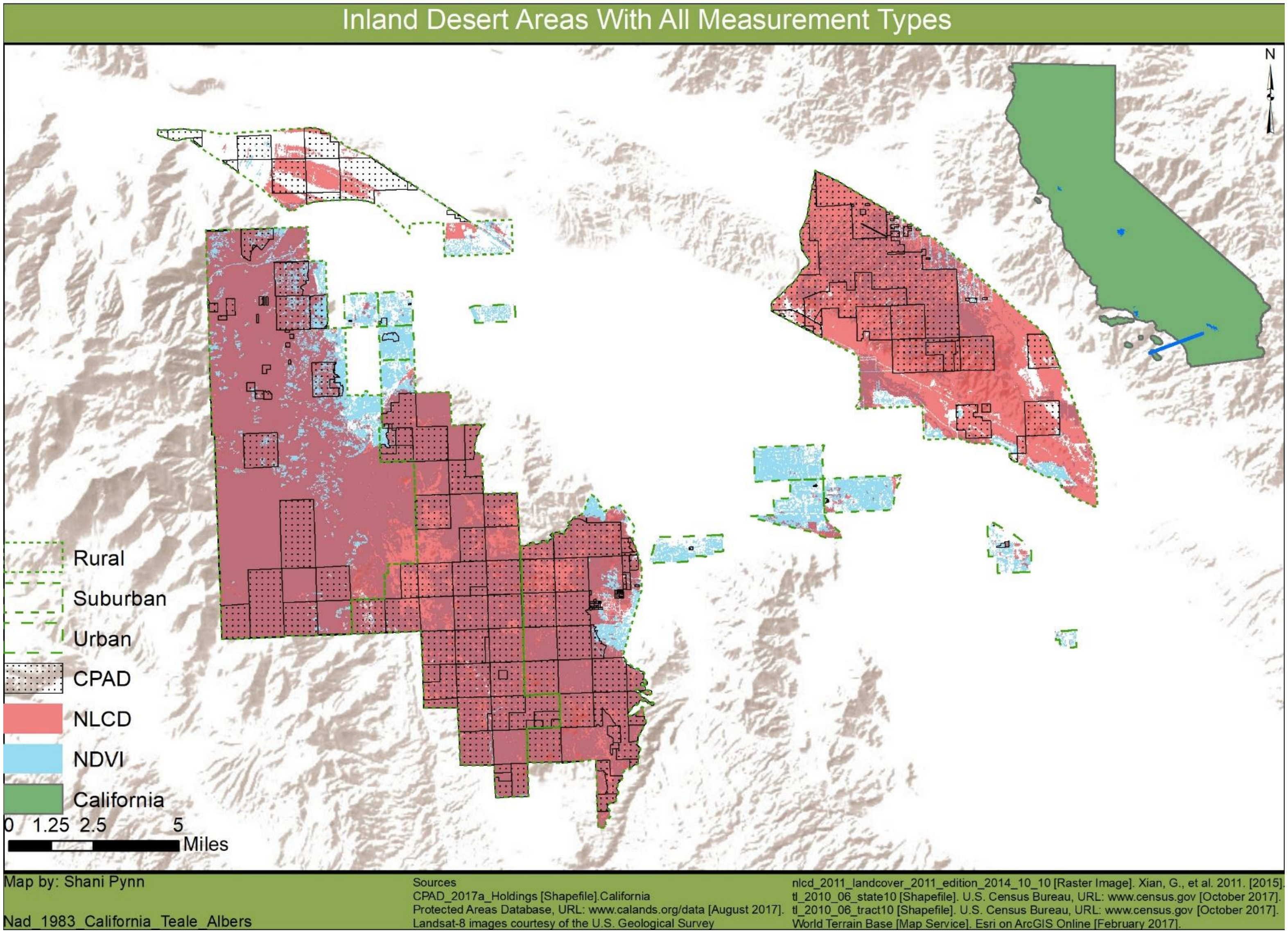

Figure 19 Inland desert area results

4.9. Central Valley

NDVI from imagery data produced the highest mean percent cover of green space of 81%. CPAD parks produced the least with a mean of 1%. NLCD produced a mean between the other two methods of 16% but had the highest standard deviation of 24. However, the ordered differences p-values showed CPAD and NLCD to not be significantly different. All other pairs were significantly different from one another. Figure 20 provides a visual representation of these results.

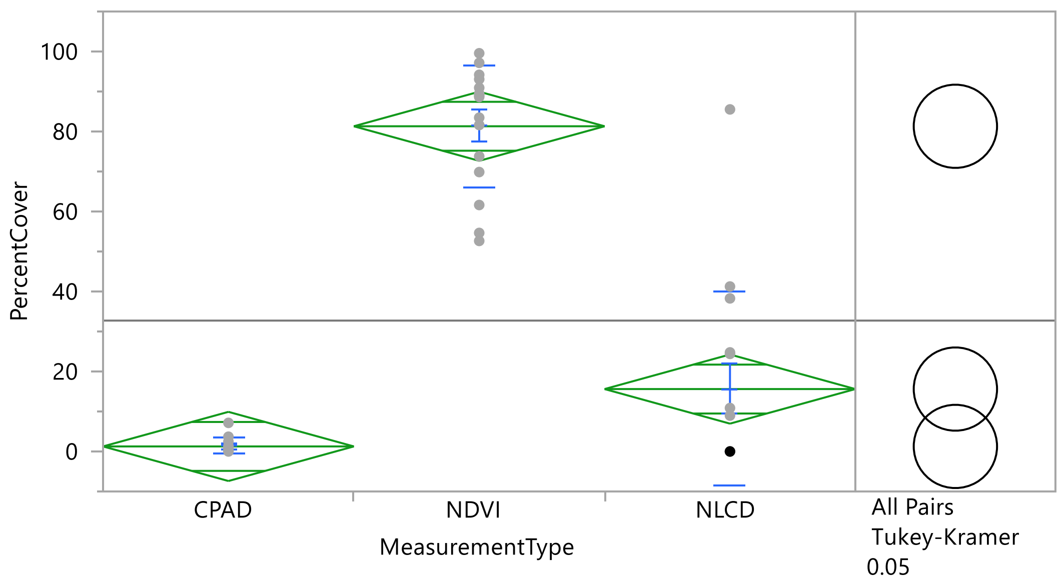

Figure 20 One-way ANOVA of central valley results

Figure 21 exhibits the geographical layout of these results. This provides a closer view of the difference between the NLCD and NDVI data. There is clear mottling in the rural areas where one or the other type indicates greenspace. Even though these two types were not significantly different in terms of total percent cover per area, they appear to be delineating different areas on the ground.

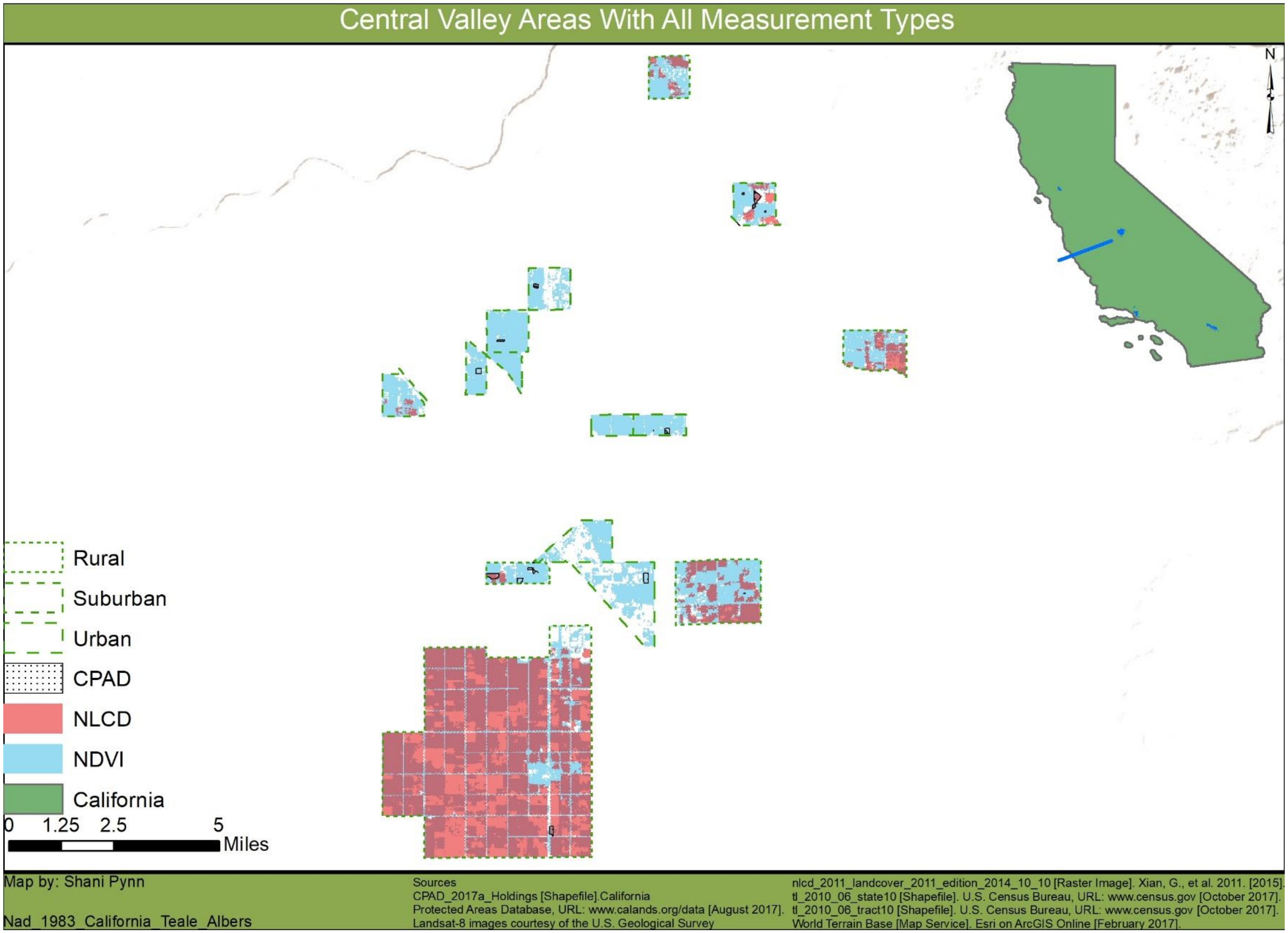

Figure 21 Central valley area results

Chapter 5 Discussion and Conclusion

This chapter discusses the results of the study in relation to the hypothesis made at its beginning. It then outlines some practical implications for data usage benefits and shortfalls of each type. Finally, it speculates at potential research questions which could be addressed in the future.

5.1. Revisiting the Hypothesis

The study provided one correlating and two contrary results to what was expected. As with most scientific inquiry, the contrary answers sparked new hypotheses about the data used.

To recall the main research question: What percent covers do inventory, usage-based categorical, and biophysical measures of greenspace produce in various environmental settings and what do these results look like spatially? The hypothesis was that the biophysical measure using remotely sensed imagery data would provide the highest average cover. The study showed this to be correct.

The study also expected that the highest producing method will vary with urbanization with imagery capturing the most in rural settings, classification effective overall, and parks data sets capturing the most in urban settings. However, the results showed the method producing the highest measure of greenspace did not vary with urbanization. However, in rural settings, the NLCD data was just as comprehensive as the imagery. This is different from the hypothesis, which predicted NDVI providing the highest values in rural, CPAD in urban, and NLCD fairly consistently overall.

The likely reason for the lower measurements of the NLCD in suburban and urban areas is the exclusion of the three classifications combining greenspace and non-greenspace urban elements. As these categories were range blocks of greenspace coverage, it was not possible to determine, for this analysis, the actual amount of greenspace within them. They were therefore excluded which lessened the comprehensiveness of the data set.

The low values for CPAD may be due to the nature of inventory compilation. This data set, unlike the other two, is not an exhaustive measure. CPAD is compiled of a list of parks by category, not the result of a search for every park in an area. Therefore, some parks and open areas may not make it on the list. Also, much of the greenspace in urban areas may come from urban landscaping elements that are not considered parks. It is possible these areas may have a larger percentage of the overall greenspace environment than expected.

NDVI performed surprisingly well in urban areas despite the resolution level of the imagery. This does, however, bring up the possibility that there may still be more greenspace within urban areas than was found in this study. It is possible another more ideal measure could exist for capturing the entirety of it.

Finally, the study expected the highest producing method to vary with geographic location. However, the study showed the method producing the highest measure of greenspace did not vary with the broader geographical location of the study areas. While NLCD still produced significantly less cover in the inland desert, the p value comparing NLCD and NDVI was much larger in this area than the others. It wasn't high enough, however, to show the two measures were comparable. This suggests there may be a degree of increase in greenspace representation in NLCD compared to NDVI in areas where vegetation may be less dense or less green itself. However, the level of significance of this when collecting measures of greenspace cover may be low.

5.2. Implications for Choosing Data Types

As stated at the beginning of this thesis, one of the goals of this study was to provide a brief summary of practical usability considerations for each of the measurement types studied. This section outlines the pros and cons of each variable in turn. It then briefly hypothesizes some suitability considerations for each type.

5.2.1. Inventory Data

The benefits of the inventory data used in this study were true to Jorgensen and Gobster's (2010) definition of the data type. Of the three measures used, the CPAD inventory was the only one with customized attribute information about each greenspace area. This data set also best described any intentionality for lands to be used as greenspaces, such as being installed and maintained as public parks. These qualities suggest this variable is suitable for measuring manipulation of greenspaces towards measured effect outcomes.

As an inventory, this datatype is susceptible to exclusions due to non-comprehensive compilation. It only includes greenspaces meeting the inventory criteria which may leave out measurable amounts of vegetation in larger study areas. It also may not include large groups of intentional greenspaces if the manager of those spaces does not know or care to submit their data for inclusion in the inventory. These caveats reinforce the need for careful consideration of data set suitability and allowable deviations from comprehensiveness associated with a given research project.

Given these pros and cons, the best suitability of this manifest variable may be for the creators of the inventory and the entities contributing to it. Since this type of data has the capability of recording detailed attributes, it can be quite useful to organizations who are able to have input on what those attributes should be. Also, unexhaustive compilation may be less of a concern for entities who know all of the lands of interest to them are included by way of having participated in the compilation process. That said, health studies may find this variable useful as it may catalog features applicable to specific health conditions such as graded walking pathways or exercise equipment which other variables may not capture.

5.2.2. Usage-Based Categorical Data

The benefits of the usage-based categorical data used in this study were in line with what would be expected of a classified variable and provided some of the benefits of both inventory and biophysical measures. The data set used included multiple greenspace and non-greenspace categories which, while not as detailed as the inventory data, gives the option of examining subtypes of greenspace. The data set also provided exhaustive coverage and roots in remotely sensed data. This gives more confidence to the assumption of capturing the on the ground conditions. This variable also has the benefit of a compilation background based on a nuanced blend of compilation factors.

As a classified variable, this data type is susceptible to user exclusions due to mixed categories. In this study, three categories were excluded for this reason which may have had a significant impact on the resulting levels of greenspace reported. The data set used in this study also had lower resolution than could have been possible with a smaller scale data set.

Given the pros and cons, the best suitability of this variable may be for measures of greenspace in areas where sub category distinction is not required. For the NLCD data set, this might be areas farther away from mixed urban landscaping. There may also be a benefit to using this type of variable in larger regional studies needing exhaustive measures or ones which can be broken down by type.

5.2.3. Biophysical Data

The benefits of the biophysical data used in this study were in line with the direct nature of this variable. The data set used provided a direct measure of existing vegetation within its resolution capabilities, producing a straightforward view of the “green” and where it was in the surrounding “space.” This method captured a large amount of vegetation, even in urban areas where resolution was expected to reduce accuracy.

One interesting caveat found were the gaps in the NDVI imagery compared to the NLCD data in rural areas. This suggests there may be a loss of comprehensiveness of this variable unless it is applied over a longer time period to compensate for possible seasonality of vegetation cover.

As biophysical data, this variable is susceptible to errors from collection and processing. In this case, cloud cover can interfere with satellite imagery as a data collection method. This necessitated an expansion of the temporal window for data tiles in this study. Fortunately, all necessary coverage was still obtainable within the spring season. This may not be the case in areas with more frequent rainfall or cloud cover.

Given the pros and cons, the best suitability of this variable may be for any study concerned with accurately capturing and accounting for all of the “green” in a space. Also, the raw nature of this data type lends well to uses related to analyzing properties of the greenspace vegetation itself.

5.3. Future Work

As with all scientific inquiry, investigating hypotheses unearths new questions. This section describes the limitations of this study as well as some potential avenues of investigation for future research.

This study had two main limitations, the first of which is that it examined data sets at available, not optimal, quality. To maintain practical usability of the results, each data set used was publicly available and provided broader coverage. This sacrificed the precision which could have been gained by using specifically detailed, best tailored to specific small areas types of data. While this makes it applicable for many studies, it does not accurately compare the data types at their optimal conditions.

The second main limitation was that this study measures cover of greenspace only. In doing so, it did not compare measurement types for use measuring the types of vegetation, such as trees, meadows, and water features. It also did not measure health benefits gained from greenspace. This decision traded a degree of breadth of application for focused information.

Both of these limitations lend themselves well to the possibility of becoming future research foci. Future opportunities for research expounded by this study are therefore comparison of data types using localized optimal conditions and data sets, and comparisons measuring specific health related values which may not be correlated to cover specifically. Comparisons of other measures of greenspace could also be beneficial as they could identify additional suitability concerns and benefits.

References

Ambrey, Christopher L. 2016. "An investigation into the synergistic wellbeing benefits of greenspace and physical activity: Moving beyond the mean." *Urban Forestry & Urban Greening* 19, (1 September): 7-12.

California Protected Areas Database. 2017 "CPAD_2016a_Holdings." *Caalifornia Protected Areas Database.* Accessed November 2016. www.calands.org.

CNPS. “California Native Plant Society Vegetation Program.” Accessed January 9, 2018. http://www.cnps.org/cnps/vegetation/.

Coombes, Emma, Andrew P. Jones, and Melvyn Hillsdon. 2010. "The relationship of physical activity and overweight to objectively measured green space accessibility and use." *Social Science & Medicine* 70, (8 January): 816-822.

Dalton, Alice M., Andrew P. Jones, Stephen J. Sharp, Andrew J. M Cooper, Simon Griffon, and Nicholas J. Wareham. 2016. "Residential neighbourhood greenspace is associated with reduced risk of incident diabetes in older people: a prospective cohort study." *BMC Public Health* 16, no. 1 (18 November): 1171-1181.

Eldeb, Abdelatif B. 2013. "Assessing the Equity Implications of Greenspace Distribution in an Arid Region." Master’s thesis, The University of Texas at El Paso.

Fresno State. n.d. *CIMIS Weather Data.* Accessed February 8, 2018. http://www.wateright.net/cimisweather.

Homer, Collin, Chengquan Huang, Limin Yang, Bruce Wylie, and Michael Coan. 2004. "Development of a 2001 National Land-Cover Database for the United States." *Photogrammetric Engineering & Remote Sensing* 70 no. 7 (July): 829-840.

Homer, Collin H., Joyce A. Fry, and Christopher A. Barnes. 2012. "The National Land Cover Database, U.S. Geological Survey Fact Sheet." 2012-3020 4p.

Homer, Collin G., Jon A. Dewitz, Limin Yang, Suming Jin, Patrick Danielson, George Xian, John Coulston, Nathaniel D. Herold, James D. Wickham, and Kevin Megown. 2015. “Completion of the 2011 National Land Cover Database for the conterminous United States-Representing a decade of land cover change information.” *Photogrammetric Engineering and Remote Sensing*, 81, no. 5: 345-354

Hunter, Ashlea J., and Gary W. Luck. 2015. "Defining and measuring the social-ecological quality of urban greenspace: a semi-systematic review." *Uran Ecosystems* 18: 1139-1163.

Jin, Suming, Limin Yang, Patrick Danielson, Collin Homer, Joyce Fry, and George Xian. 2013. “A comprehensive change detection method for updating the National Land Cover Database to circa 2011.” *Remote Sensing of Environment,* 132: 159-175.

Jorgensen, Anna, and Paul H Gobster. 2010. "Shades of Green: Measuring the Ecology of Urban Green Space in the Context of Human Health and Well-Being." *Nature and Culture,* 5 no. 3 (Winter): 338-363.

Lachowycz, Kate, Andy P. Jones, Angie S. Page, Benedict W. Wheeler, and Ashley R. Cooper. 2012. "What can global positioning systems tell us about the contribution of different types of urban greenspace to children's physical activity?" *Health & Place,* 18 (25 January): 586-594.

Lachowycz, Kate, and Andy P. Jones. 2014. "Does walking explain associations between access to greenspace and lower mortality?" *Social Science & Medicine,* 107 (15 February): 9-17.

McCracken, Deborah S., Deonie A. Allen, and Alan J. Gow. 2016. "Associations between urban greenspace and health-related quality of life in children." *Preventive Medicine Reports,* 3 (3 February): 211-221.

Merriam-Webster Dictionary. April 7. Accessed April 10, 2018.

Minitab, 2017 *One-Way ANOVA*. Minitab. Accessed May 19, 2018. http://support.minitab.com/en-us/minitab/17/Assistant_One_Way_ANOVA.pdf

State of California. 2018. *California Natural Diversity Database.* Accessed March 26, 2018. https://www.wildlife.ca.gov/Data/CNDDB.

Sugiyama, Takemi, Karen Villanueva, Matthew Knuiman, Jacinta Francis, Sarah Foster, Lisa Wood, and Billie Giles-Corti. 2016. "Can neighborhood green space mitigate health inequalities? A study of socio-economic status and mental health." *Health & Place,* 38 (1 February): 16-21.

U.S. Census Bureau. 2015. *2010 Census Urban and Rural Classification and Urban Area Criteria.* February 9. Accessed February 12, 2018. https://www.census.gov/geo/reference/ua/urban-rural-2010.html.

U.S. Department of the Interior. 2018. "Multi-Resolution Land Characteristics Consortium (MRLC)." Last modified April 16, 2018. Accessed May 15, 2018. https://www.mrlc.gov/index.php.

U.S. Environmental Protection Agency. 2012. "Level III Ecoregions of California." Corvallis: U.S. EPA Office of Research and Development (ORD) - National Health and Environmental Effects Research Laboratory (NHEERL), January 1.

Van Den Bosch, Matilda Annerstedt, Pierpaolo Mudu, Valdas Uscila, Maria Barrdahl, Alexandra Kulinkina, Brigit Staatsen, Wim Swart, Hanneke Kruize, Ingrida Zurlyte, and Andrey I. Egorov. 2016. "Development of an urban green space indicator and the public health rationale." *Scandinavian Journal of Public Health,* 44: 159-167.

Wheeler, Benedict W., Ashley R. Cooper, Angie S. Page, and Russell Jago. 2010. "Greenspace and children's physical activity: A GPS/GIS analysis of the PEACH project." *Preventive Medicine*, 51 (9 June): 148-152.

Wu, Chih-Da, Eileen McNeely, J. G. Cedeño-Laurent, Wen-Chi Pan, Gary Adamkiewicz, Francesca Dominici, Shih-Chun Candice Lung, Huey-Jen Su, and John D. Spengler. 2014. "Linking Student Performance in Massachusetts Elementary Schools with the ‘‘Greenness’’ of School Surroundings Using Remote Sensing." *Plos One,* 9 no. 10: 1-9.

Younan, Diana, Catherine Tuvblad, Lianfa Li, Jun Wu, Fred Lurmann, Meredith Franklin, Kiros Berhane, et al. 2016. "Environmental Determinants of Aggression in Adolescents: Role of Urban Neighborhood Greenspace." *Journal of the Academy of Child & Adolescent Psychiatry,* 55 no. 7 (July): 591-601.

Zerwig, Lorena M. 2016. "The Effects of Green Space Utilization among Nursing Students in Regards to Stress Level and Blood Pressure." DNP diss., Brandman University. Accessed May 24, 2017. ProQuest LLC.