5-2015

A Study of Colloquial Place Names through Geotagged Social Media Data

Yuan Liu
University of Tennessee - Knoxville, yliu89@vols.utk.edu

To the Graduate Council:

I am submitting herewith a thesis written by Yuan Liu entitled "A Study of Colloquial Place Names through Geotagged Social Media Data." I have examined the final electronic copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Geography.

Shih-Lung Shaw, Major Professor

We have read this thesis and recommend its acceptance:

Lee F. Han, Hyun Kim

Accepted for the Council:
Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

A Study of Colloquial Place Names through Geotagged Social Media Data

A Thesis Presented for the
Master of Science
Degree
The University of Tennessee, Knoxville

Yuan Liu
May 2015

.

Acknowledgements

I am grateful for the continuous and patient guidance from my advisor, Dr. Shih-Lung Shaw. The research idea for my thesis originally came from Dr. Shaw, and I further explored the idea over the past two years with his dedicated instructions. Although my work is still not perfect, Dr. Shaw has put me on the right track of academic research. His passion for research and rigorous academic attitudes are impressive, and will influence me my entire life.

Many thanks go to Dr. Hyun Kim, not only for the knowledge I have gained from him during the past two years, but also for always being patient, encouraging, and helpful with my questions. I appreciate Dr. Han Lee for taking his valuable time to serve on my committee. His wise insights and inspiring suggestions greatly helped me in my thesis research.

I would also like to thank my beloved fiancé, Jianjiang Yang, and my kitten, Maomao, for their unfailing company and support.

Abstract

Place is a rich but vague geographic concept. Much work has been done to explore the collective understanding and perceived location of place. The last few decades have seen rapid expansion in the use of online social media and data sharing services, which provide a large amount of valuable data for research of colloquial place names. This study explored how geotagged social media data can be used to understand geographic place names, and delimit the perceived geographic extent of a place. The author proposes a probabilistic method to map the perceived geographic extent of a place using Kernel Density Estimation (KDE) based on the geotagged data uploaded by users. The author also used spatio-temporal analysis methods in GIS to explore characteristics, hidden patterns, and trends of the places. Flickr, a popular online social networking service that features image hosting and sharing, was selected as the main data source for this project. The results show that outcomes of KDE with different functions and parameters differ from each other; therefore, it is crucial to select the proper KDE bandwidth in order to obtain appropriate geographic extents. Official boundaries and reference boundaries can be used to assess the geographic extents. Google Maps Street View is another useful source to examine the visual characteristics of places. Spatio-temporal analysis of the geographic extents over time reveals significant location changes of the places composed of man-made structures. Besides names and variations of place names, related colloquial terms, like Cades Cove of the Great Smoky Mountains National Park, are also useful sources when delimiting a place. Several examples are analyzed and discussed. Studies like this research can improve our understanding of geotagged Online Social Network (OSN) data in the study of colloquial place names as well as provide a temporal perspective to the analysis of their perceived geographic extents.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Research Background

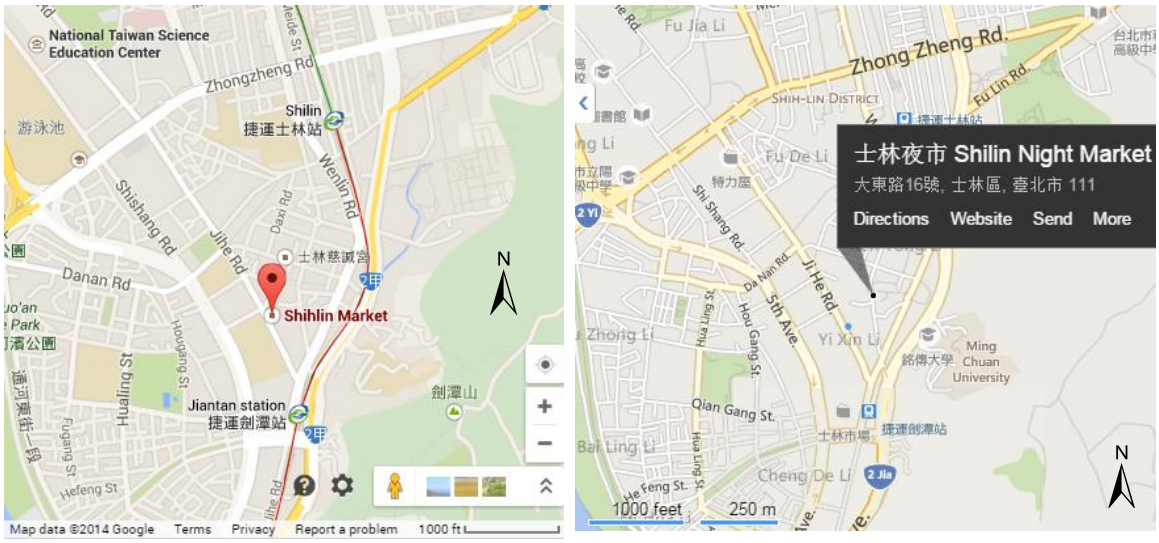
Digital technology has become the dominant information transmission approach in the twenty-first century, where rigorously structured information, translated into binary data, can be easily shared, analyzed, and effectively used in various ways [1]. In particular, the development of geographic information systems [2] has had profound effects on all aspects of geographic data production and research. It offers geographical measurements and presentations based on coordinates. However, *place* is a vague concept developed through peoples' interactions and experiences with surrounding environments. According to Goodchild [1], place indicates "the space within which humans carry out habitual aspects of their lives, such as shopping, work, recreation, and sleeping" (p.28) for geographers. Thus, the definitions and geographic extents of places are specific to individuals, and depend on time, which is subjective and fuzzy. People may use a variety of colloquial terms to describe or identify a place; however, even though people have an idea of where a place is located and whether a certain area is part of the place, it is hard to sharply define that place's boundary. Furthermore, notions and descriptions of places may evolve over time. Methods and rules are required to translate informally generated information about places into a geographic information system (GIS).

For example, the Taipei Shilin Night Market is a street market, famous with tourists, that starts at around 2 p.m. [3]. It is a traditional market that contains hundreds of food vendors and small restaurants as well as surrounding businesses and shops. However, in informal conversation it is sometimes referred to as "Shilin Traditional

Market." "Shilin Temporary Market" and "Xilin Market" are also commonly used names for the night market in the vernacular. The market has no definite location or explicit boundary. Most of the time, people use an approximate location in the vicinity of the Ji He Road and 6th Avenue to locate and represent the market (Figure 1).

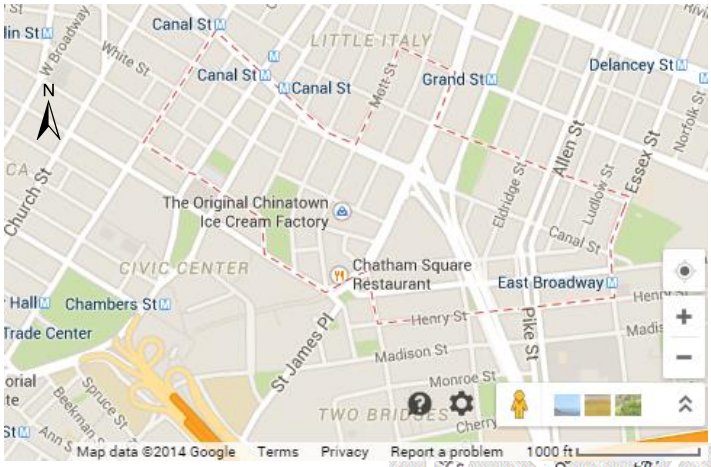
Similarly, Manhattan Chinatown is in the borough of Manhattan in New York City. It is usually described as being bounded by the Lower East Side to the east and Little Italy to the north. However, no authoritative geographic extent is defined for it. It is indicated differently by different map services such as Google Maps (https://maps.google.com) and Zillow (http://www.zillow.com), as shown in Figure 2. More interestingly, descriptions of its general extent in Wikipedia (https://www.wikipedia.org) and Wikitravel (http://wikitravel.org) are neither the same as the map services, nor do they agree with each other. Manhattan Chinatown has several names among the Chinese community such as "唐人街" and "华埠." People also refer to it as "NYC Chinatown" or simply "Chinatown" in daily conversation.

Even for places that have official names and administrative boundaries, publicly perceived information of the place can be useful. A case in point is the Great Smoky Mountains National Park. "Great Smoky Mountains National Park" is the official, administrative name of the national park. However, people usually refer to it as some abbreviation of its official name such as "Smokies," "Smoky Mountains," and "Great Smoky NP." "Cades Cove," one of the most famous tourist attractions in the national park, is sometimes used to refer to the national park itself. Its official boundary is used mainly for administration purposes, and is not of much use for ordinary visitors'

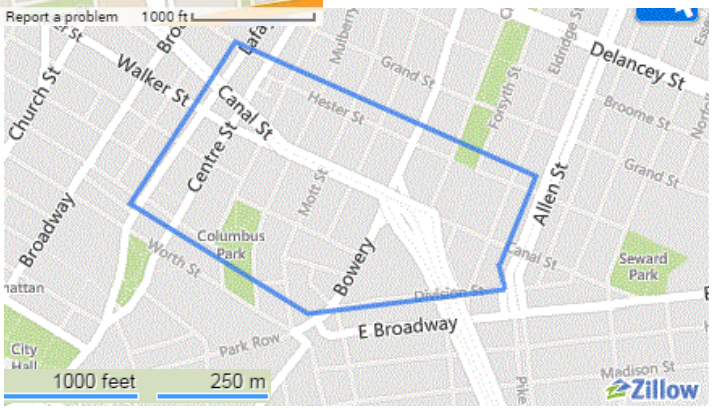


Google Bing Map

Figure 1 Location points of Taipei Shilin Night Market in Google Map versus Bing Map



Google Map



Zillow

Figure 2 Geographic extents of Manhattan Chinatown in Google Map versus Zillow

activities. The geographic extent that visitors care about is the locations of accessible tourist attractions within and around the national park area.

Thus, *colloquial place names* are terms that may be official or informal, which are used by people in daily life to refer to a place. Some colloquial place names may not be used in written language nor be formally recognized. They are even not correct sometimes. But colloquial place names may offer much more data about the place than the formal name does, especially in user-generated geographic information. Studying colloquial place names may yield a more thorough understanding of a place and its human activities, and help researchers apply GIS functions in the research of places.

This study defines the geographic extent of places based on user-generated geographic data as *perceived geographic extent*. The perceived geographic extent of colloquial place names delimits the place according to geographic data that people have uploaded to Flickr and other such services. It may be different from the official boundary, if one exists. Knowing the perceived geographic extent of colloquial place names would be valuable for tourists, human behavioral research, and commercial location-based services.

Great Smoky Mountains National Park, Manhattan Chinatown, and Taipei Shilin Night Market have been selected as cases for this study. Both Shilin Night Market and Manhattan Chinatown have a variety of colloquial place names, and neither place has an official boundary. Manhattan Chinatown is a good example of a place composed of man-made, highly accessible structures with a relatively large amount of data. Street vendors and small storefronts populate the marginal areas of Shilin Night Market; thus, its perceived geographic extent is more flexible and likely to change over time. While Great

Smoky Mountains National Park has an official name and boundary, it is more often referred to by its colloquial names, and its perceived geographic extent may differ from the official boundary. Exploring these three cases, the author hopes to understand how user-generated geographic data can be used to delimit specific place-scale geographic objects, and propose a method to map perceived geographic extents at various probability levels.

Large collections of geotagged online social network (OSN) data are publicly accessible nowadays, and they present an opportunity to acquire information about the locations of places as well as their colloquial names. An increasing number of OSN users are able to share geographical information online because of the popularity of portable smart devices and the development of location-based services. By February 2009, more than one hundred million geotagged images were uploaded to Flickr, which is about 3% of all images uploaded to the Internet [3]. According to a 2013 Verge report, more than 3.5 million new images are uploaded to the Internet daily [4]. On Twitter, nearly one in five tweets reveals the user's location and timestamp through geotagging or metadata [5]. A 2013 Pew Research Center survey [6] shows that about 30% of adult social media users have automated location tagging on at least one of their accounts.

Geotagged OSN data are contextually rich, combining geographic information with text information about the place where an image is taken. Using geotagged images as an example, the geographic information is mainly reflected by an image's geotagged information and metadata, while the text information includes titles, tags, and descriptions given by the person uploading the image, as well as comments by people viewing the image. The author does not consider image recognition in this study, even

though it is also a practical method to extract information of geotagged images. Geotagged OSN data is a useful resource to analyze places and human behavior, and distinguishes geotagged OSN data from traditional geographic information supplied by authoritative or commercial mapping agencies such as United States Census Bureau and Google Maps. Several interesting studies have been conducted using this data resource: detecting unusual global and local events [7, 8], spatial pattern and demographic indicators of event-related social media content [9], geographical visualization of geotagged OSN data [10], discovering preferences of certain areas [11], and offering personalized travel recommendations [12].

User-generated text information represents the perceived cognition of people uploading the images, and it contains various colloquial terms that refer to the place. By combining these fuzzy colloquial expressions with explicit geographical information, researchers are able to collect, handle, store, and analyze information about the place with computer-based technology. If sufficient information is available, spatio-temporal trends and patterns will be revealed for colloquial place names. Using the three aforementioned places, this study explores perceived identification and location of places through geotagged Flickr images, and proposes a probability method to map the perceived geographic extent of the place. The probability means, in this case, percentage of the images of a place geotagged within the perceived geographic extent of this place. Spatio-temporal characteristics and patterns in terms of geographic extents of case study places are analyzed.

1.2 Research Objectives

This study demonstrates how geotagged OSN data can be used to explore colloquial terms for a place, generate perceived geographic extents of colloquial place names, and analyze spatio-temporal characteristics of perceived geographic extents.

A proper data source for this study must be selected from the many OSN services available. The service must have abundant users, extensive influence, and large amounts of data that cover a long span of time. Processing and filtering methods need to be conducted on the massive, complex data in order to obtain reliable, representative data subsets that indicate colloquial place names of each case study, which will be used to generate perceived geographic extents. Using the data subsets, the author will propose a probabilistic method to map the perceived geographic extent of places according to people who uploaded the images. Considering the fuzziness and complexity of colloquial terms for places, the author will analyze the distribution of some related colloquial terms about each case study place, and compare them with their generated probable geographic extents. While these colloquial terms are far from thoroughly utilizing the collected information of places, they offer an inspiring example for further studies. This study will demonstrate and interpret the spatio-temporal characteristics, hidden patterns, and trends of the case study places, which are mainly composed of man-made structures.

1.3 Organization of the Thesis

This thesis is organized into five chapters. The next chapter reviews the literature relevant to this study, including place and colloquial place names, extent generation from fuzzy point sets, kernel density estimation, and spatio-temporal analysis. Chapter 3 is devoted to the methodology of data acquisition, data processing, and perceived geographic extent

generation. In this chapter, the author defines the methods used to remove data redundancy to produce more reliable analysis and results. The process designed to extract preliminary research boundaries and target image data subsets is also discussed in Chapter 3. Further, this chapter addresses the method used to generate perceived geographic extent of colloquial place names, based on probability levels, as well as the routine used to conduct a spatio-temporal analysis. Chapter 4 presents results and discusses perceived geographic extent generation and analysis. In this chapter, the author verifies and selects appropriate density estimation results of colloquial place names, then maps the geographic extent of probability levels accordingly. It further explores expansion of colloquial terms of places as well as their contribution to geographic extent generation. This chapter also demonstrates spatio-temporal analysis results. Chapter 5 concludes this thesis with limitations and opportunities for future research.

Chapter 2

Literature Review

2.1 Place and Colloquial Place Names

Place is a rich but vague concept. Goodchild [1] discusses the definition of place under several different circumstances and shows how GIS techniques can be used to operationalize place in specific areas of research. Studies in the field of common-sense geography, also known as "Naïve Geography" [13], provide a framework to capture and reflect peoples' perceived understandings of geographic space and time. Naïve means instinctive or spontaneous [13]. Formal models created to represent, manage, and analyze common-sense geographic concepts in GIS could help deliver temporal and spatial information to the public, especially user communities with little or no training [14]. The National Center for Geographic Information and Analysis approved a research initiative entitled "Formal Models of Common-Sense Geographic Worlds" in 1996; its purpose was to identify basic elements of common-sense conceptualizations of geographic space, entities, and processes, and develop an integrating framework [15]. Formalizing the representation and analysis of common-sense geographic concepts is drawing more interests, and some place-based versions of well-known GIS functionalities, such as join and buffer, are being developed and applied [16]. Yao and Jiang [17] proposed the concept "qualitative location," which means the spatial location that is referred to using linguistic terms such as "town center," "southeast region," and "nearby," and a method to visualize the qualitative locations in GIS. This is an inspiring study that aims to formalize the naïve geographic concepts in GIS. However, little research has been specifically conducted for colloquial place names.

The rise of volunteered geographic information provides a great opportunity for research about colloquial place names. Pasley et al. [18] extracted geographic information from the World Wide Web to generate informal places that are referred to in daily life, but that have no entry in official geographical resources such as gazetteers. Jones et al. [19] believes that vague place names are frequently accompanied by the names of more precisely defined co-located places, which lie within certain distances to the case study place, therefore they proposed a model to generate approximate crisp boundaries of the place through co-occurrence frequency. Based on these studies, the author proposes methods to automatically obtain geographic features and associated footprints from public Internet sources and generate regional gazetteers [20, 21].

The last few decades have seen a rapid expansion in the use of social media and data sharing services. They have generated large volumes of geotagged social media data, which are potentially a valuable and more accurate source of knowledge about social phenomena [22]. As the first online social networking service that features image and video sharing, Flickr has provided a large integrated data source for research of colloquial place names. Flickr's research and development team analyzed the photo tags and generated aggregate knowledge in the form of prominent tags for arbitrary areas all over the world [23]. Keßler et al. [24] proposed a bottom-up gazetteer building approach by clustering and filtering Flickr geotagged photos. Thomee and Rae [25] applied scale-space theory to generate boundaries for locally characterized regions, though the regions were also extracted by generally detecting prominent tag occurrences. Hollenstein and Purves [26], narrowing the research objects to finer granularity, explored the terms used in vernacular language to describe city core areas. They also looked at the nature of error

and imprecision in tagging and georeferencing. Li and Goodchild [27] generated collective views of places' geographic extents and inclusion relations using a similar methodology, in order to reveal the hierarchic relationship among them and further develop place-based concepts and applications. Along with the rapid development of supercomputing, high performance geoprocessing workflow was developed to harvest crowd-sourced gazetteer entries for automatic, general, bottom-up gazetteer construction [28].

Much work has been done to describe and define the geographic extent of colloquial place names defined by user generated geographic data, especially geotagged online social media data. Nevertheless, place is usually vague and diverse, especially places that appear in the vernacular, created by informal consensus. Most studies have focused on large-scale general extraction of identification and geographic extent from the clusters of geotagged data records, instead of on distinguished individual places. More specific and precise analysis for place-scale colloquial geographic objects, such as city core areas [26, 29], is necessary to further understand applications of user generated geographic data on the collective understanding and locating of places.

2.2 Extent Generation from Fuzzy Point Sets

Functionalities for processing fuzzy geographic feature sets in GIS arose out of the need to handle uncertainty and give soft computing technology, which uses inexact solutions to computationally hard tasks such as the solution of NP-complete problems, the ability to support vague information processing [30]. One large challenge of using fuzzy sets in GIS problems is to specify membership and then define the border. Defining a vague object using any kind of threshold always has weakness, because vague geographic

objects are Sorites susceptible, a definition of vagueness derived from Sorites Paradox, and there is no single threshold value that genuinely distinguishes the object from non-object [31]. However, mapping geographic extent of colloquial place names can be very useful for both scientific research purposes and daily life.

Several methods of generating polygonal representations of places using point sets have been published. For example, several membership functions of fuzzy-set-based classification method [28, 30] were introduced to extract geographic footprints of certain polygonal places. Liu et al. [32] proposed similar point-set-based region (PSBR) models to approximate areal objects, especially vague areal objects, and manage their spatial relationships. These two approaches offer practical and effective methods to estimate polygonal geographic extent using fuzzy point set data, but several issues remain when applying them in this research. First, these methods expect the point set to be well clustered and unimodal, which means only one central hotspot exists. This will yield an outline of all points within the point set or its subset, which will not reveal any multimodal features. Nevertheless, in real-world cases, areal places often appear to be multimodal, which means they have more than one hotspot, and it is important to demonstrate subtle distribution features in order to generate more precise perceived geographic extents. Second, these are not convenient methods for analyzing probability levels of perceived boundary. Researchers must calculate and extract a corresponding subset using the threshold membership score, and then they must generate a boundary for each result of a certain probability level. Third, these methods do not perform smoothly on sparse point sets or sparse areas of the point distribution. Though continuous threshold scores can be obtained, the corresponding point subsets are discrete, as are the

corresponding generated boundaries. In sparse point sets or sparse areas of the point distribution, each individual data record may greatly affect the resulting boundary or even determine the geometry. Thus, the result has very low tolerance for error of each data record and one data point with semantic or geographic error, which is quite common for volunteered geographic information data; this may greatly change the result. When the point set is not very densely populated, reasonable corresponding geographic extents for every probability level may be difficult to obtain.

There is another approach beside the point-set-based methods discussed above. Keßler, C., et al. [24] demonstrated a method using Delaunay triangulation to find clusters within point clouds. This method does not restrict the geometry of places. To classify the points and edges into polygonal regions of density value (edge length), an edge length threshold was selected. Kernel density estimation combined with a density threshold filter is a method widely used to generate approximate polygonal extent [19, 26, 27, 33]. These two methods share certain similarities and both solve the first two issues that arise from the aforementioned point-set-based methods. The multimodal feature of point sets is revealed. These methods also simplify the generation process of perceived geographic extents at different probability levels into a single step once the triangular network or kernel density surface is created. However, for the Delaunay triangulation method, the third issue still remains. The boundary of this method is discrete and directly affected by each point location. Thus, the author chooses the Kernel density estimation method with density threshold filter to construct perceived geographic extents from user generated geographic data sets.

2.3 Kernel Density Estimation

Density estimation is the construction of an estimate, based on observed data, of an unobservable underlying probability density function [34]. Applied in geographic research, it interpolates a point data set into a continuous density surface, which is usually represented as a raster. The basic density estimation method is to sum all point values within a certain circle, then divide by the area of this circle [35]. Subsequently, kernel density estimation (KDE) is proposed to generate a smoother density estimation surface and estimate probability function [36]. KDE is applied to studies in various disciplines [37], such as social and economic study [38], agriculture [39] and public health [40]. In geographic studies, it is a commonly used spatial analysis technique to transform a geographically distributed set of points into a density surface in a GIS environment [41]. Given a set of independent points, $s_1 \ldots s_n$, the point distribution probability density function can be estimated using kernel density estimator:

$$f(x,y) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{d_i(x,y)}{h})$$

where $f(x,y)$ is the estimated density value at location(x,y), n is the total number of input points, h is a measure of the window width and is called bandwidth (e.g., for a circular kernel it is the radius of the circle), $d_i(x,y)$ is the distance between event point i and location (x,y), and K is a density function characterizing how the contribution of point i varies as a function of $d_i(x,y)$ [42]. Based on KDE results, density maps are proposed as a mean to summarize selected image data so that the distribution of image records can be examined without visual cluttering and occlusion issues [43].

To get fitting KDE result, the appropriate kernel function, grid cell size, and bandwidth must be selected. Some literature states that kernel function is less important than bandwidth for determining the most appropriate density surface [36, 44]. Choice of bandwidth controls the degree of smoothness for the resulting density surface. When smoothing is insufficient, the resulting density is too rough to fully analyze the data set; when smoothing is excessive, small but important features of the density distribution may be smoothed out [45]. Several studies have discussed bandwidth determination issue [45-47] and concluded two basic approaches: a fixed bandwidth for the whole distribution and an adaptive bandwidth, which in the end may become another dimension of KDE [37]. For the fixed bandwidth selection method, the most important task is to define the right value. Jones et al. defined main methods for choosing fixed bandwidth into two generations: the first group such as least performance rules of thumb, least square cross validation, and biased cross validation; the second group such as superior performance, solve equation plug-in, and smoothed bootstrap methods [45].

Least-squares cross-validation (LSCV) bandwidth matrix selector [48, 49] is used to minimize the expansion of the Integrated Square Error (ISE) and is a commonly used algorithm [50]. Biased cross-validation (BCV) bandwidth matrix selector is only available for bivariate data [51]. To some extent, it is a hybrid of cross-validation and plug-in [50]. Two types of BCV algorithm with slight differences is demonstrated by Sain, Baggerly and Scott [51]. The optimal asymptotic choice of the bandwidth can be obtained from minimizing the Mean Integrated Square Error (MISE) [52]. Smoothed cross-validation (SCV) bandwidth selector [53] is based on explicit estimation of the exact integrated squared bias and the asymptotic integrated variance [54]. Plug-in

bandwidth selector [55] uses pilot bandwidths to estimate some values of the unknown density, and the estimates are plugged in to the equation for the ideal bandwidth [56]. LSCV, BCV, SCV, CVh, and plug-in bandwidth selecting algorithms are used and compared for optimized estimation results in this research.

2.4 Spatio-temporal Analysis

Most of these studies consider and handle colloquial identifications and geographic extent of places as static. However, place is the space within which people carry out habitual aspects of their lives that are largely unique to the individual. The identification and geographic extent of places are likely to vary through time as habits change, spaces are learned, or people migrate [1]. Achievements in spatio-temporal analysis offer powerful analytical approaches to explore their temporal features. In 1970, Hägerstrand [57] first introduced the concept of time geography, where time is considered to be a third dimension. Despite that, most current GIS analyses are based on static modeling. Nadi and Delavar [58] believe "a growing number of researches in temporal GIS are being performed, which may dominate GIS market in the near future." (p.1). Langran [59] described a taxonomy of all available access methods of spatio-temporal data in temporal GIS through partitioning and indexing. Cheylan and Lardon [60] further discussed the conceptual and practical problems in a more systematical manner when constructing spatio-temporal data and formalizing spatio-temporal research questions, such as concepts and formalization of temporal factors and dynamic behaviors as well as spatial and temporal analysis methods. The efficient management of continuously changing geographical data and the discovery of hidden patterns in the change of objects in large data sets are challenging but popular research topics [61, 62]. Erwig et al. [63] proposed

an approach to extract changing regions into digital representation as 3D (2D space and 1D time) entities. With increasing abilities to represent, visualize, and manage geographic data with time dimension, spatio-temporal analysis has been widely applied to research in GI Science. Yu and Shaw [64] designed a space-time GIS to represent and analyze spatio-temporal activity data in both physical and virtual space at the individual level. In a study of crowd behavior and special social events using cell-phone data, Calabrese et al. [65] performed spatio-temporal analysis of time series of stops to detect users' moving pattern around event time, and predict their destinations. Versichele et al. [66] conducted spatio-temporal analysis to crowd distribution at festivities, though it still uses short time intervals such as hour and day. Spatio-temporal visualization and analysis can be applied to explore urban expansion during certain transitional time periods [67, 68], and to map and interpret land cover/land use transitions and landscape changes [69, 70] using GIS and satellite images. While these studies address various topics, they supply valuable concepts and approaches for spatio-temporal analysis that can be adopted to explore perceived geographic extents of colloquial place names.

Chapter 3

Data and Methods

3.1 Data Acquisition

3.1.1 Data Source

Flickr is a popular image and video hosting website that was created in 2004. It encourages the use of Flickr data by outside developers through well supported APIs, developer mailing lists, and the App Garden for showcasing applications created with Flickr data. Following a different business model, Snapchat (https://www.snapchat.com/), another popular image and video hosting website, does not permit legal access to the data it hosts [71]. Flickr started to support the image geotagging feature in 2004, pre-dating corresponding features of most other popular online social networking services such as Twitter, Instagram (http://instagram.com/) and Sina Weibo (http://weibo.com/). For this reason, Flickr's data covers a longer span of time. Unlike the approximate location information captured by Facebook, Flickr images are geotagged at precision levels ranging from 1 to 16, where 1 denotes country level and 16 denotes street level. Sixteen is the default level. Though accuracy of this spatial information cannot be guaranteed, owing to the natural characteristics of volunteered geographic information, the vast majority of images are geotagged at the highest location precision level 16 [72]. Flickr had a total of 87 million registered members by 2013 and more than 3.5 million new images are uploaded daily [3]. Hundreds of millions of images in their 10-year image data archive are geotagged, creating an appropriate data resource for this research [73].

Flickr images are stored with information that includes mandatory fields such as the original Exchangeable Image File Format (Exif) files, image ID, contributing user ID,

Table 1 Geotagged social media data quantity and quality from several most popular online social networking platforms

OSN Platform	Time Range	Data Quality
Flickr	Since 2004	Large amount, high precision level of spatial information
Facebook	Since 2004	Large amount, approximate spatial information
Twitter	Since 2006	Large amount, high precision level of spatial information
Foursquare	Since 2009	Medium amount, high precision level of spatial information
Sina Weibo	Since 2009	Launched and mainly used by Chinese, high precision level of spatial information
Instagram	Since 2010	Relatively large amount, quickly increasing

and the time upload occurred, as well as optional information set by users such as title, tag, description, and usage restrictions. Exif is a "standard that specifies the formats for images, sound, and ancillary tags used by digital cameras (including smartphones), scanners, and other systems handling image and sound files recorded by digital cameras" (http://en.wikipedia.org/wiki/Exchangeable_image_file_format). Spatial reference information is stored in coordinates of latitude and longitude. It is either extracted from the Exif file of the image or manually located by the user through a Flickr map interface called Organizr. Then, a precision level is automatically assigned to the image, depending on the precision of GPS coordinates in the Exif file or the zoom level of Flickr Organizr when the image is uploaded. Image "tags," which are a set of unstructured text-based annotations provided by the person uploading the image, are used to reveal the cognition on the image along with its location. On Flickr, tag is a keyword or category label of the image that helps users find images that have something in common [75]. Since tags of an image are usually a set of case-insensitive words or short phrases that are easy to understand, semantic analysis is not necessary for this research.

For this research, Flickr APIs were used to download publicly available image data. Images downloaded for this research were uploaded to Flickr with an accuracy level of 16 between January 1st, 2004 and March 1st, 2014. As discussed in the Introduction, the author chose the Great Smoky Mountains National Park, Taipei Shilin Night Market, and Manhattan Chinatown for this study. In case the results deviated due to incomplete data, the author downloaded images from much larger geographic extents covering all possible area of these three places for further trimming and processing.

Besides Flickr image data, common boundary or location information of the three places were needed. For the Great Smoky Mountains National Park, whose official administrative boundary is already defined, the author downloaded the park boundary in shapefile format from the United States Geological Survey (http://www.usgs.gov/). For Manhattan Chinatown and Taipei Shilin Night Market, the author used the boundary maps, location points, or descriptions of some of the most commonly-used online resources such as Google Maps, Bing Maps, Zillow and Wikitravel. The author also downloaded the World Street Map from ESRI as a base map for better visualization.

3.1.2 Preliminary Boundary

In order to reduce computational expenses, a preliminary boundary was selected for each place according to its approximate location, physical barriers, and distribution of images tagged with colloquial place names. Then the image data located outside the preliminary boundaries are trimmed to obtain the preliminary data set, which will be used as the foundation for further analysis. Since only rough preliminary boundaries and data set are selected at the current stage, some approximate values, features, and text descriptions are considered. Different standards are applied to select the preliminary boundaries for different types of places.

For places with official boundaries, which is the case for the Great Smoky Mountains National Park, that boundary is fully utilized. Because the boundary is a long polygon (see Figure 3), the scale reference of its size was obtained by measuring its width in the approximate middle section, which measures about 30 km. To select the preliminary boundary of the Great Smoky Mountains National Park, the author considered this scale reference, the distribution of image data tagged as

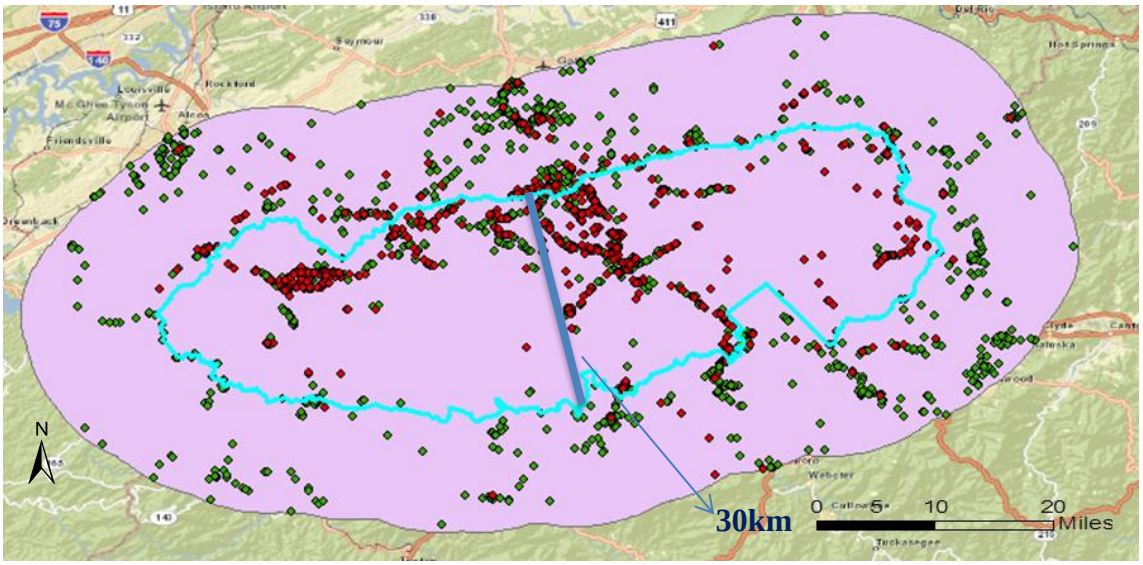


Figure 3 Preliminary boundary of the Great Smoky Mountains National Park

“greatsmokymountainsnationalpark” (the red dots shown on Figure 3), and the location of other places with obvious image clusters such as the city of Knoxville. Ultimately, a buffer polygon of the official boundary with a buffer distance of 15 km was chosen as the preliminary boundary.

Different resources such as Google Maps, Zillow, and Wikitravel give different boundaries for Manhattan Chinatown. The author considered its approximate extent as defined by several of the most commonly used resources, the distribution of image data tagged as “chinatown” (red dots shown in Figure 4), and the physical barriers of rivers when selecting the preliminary boundary. The area extending from the southern end of Manhattan north to 14th Street was chosen as the preliminary boundary.

Because Taipei Shilin Night Market has only a few vague location points and no official boundaries, the preliminary boundary was selected according to these rough location points, distribution of images, and physical barriers of an elevated road and a river. Thus, the area between Keenlung River, Xinsheng Elevated Road, and Zhongzheng Road was selected. This area excludes interferential places, such as other market-type places, or other significant image hotspots such as Shilin Presidential Residence, which hosts annual flower expos, as shown in Figure 5.

3.1.3 Data Processing

There are tens of hundreds of images with exactly the same coordinates and dates, uploaded by the same user in the data set. This is caused by Flickr’s batch geotagging feature. For instance, a user with the ID “****5900@N00” uploaded 17 images with the coordinates “35.564925, -83.331062” in the Great Smoky Mountains National Park area, at approximately 9:00 am on 2006.11.23. All were tagged with “camping,” “smokies,”

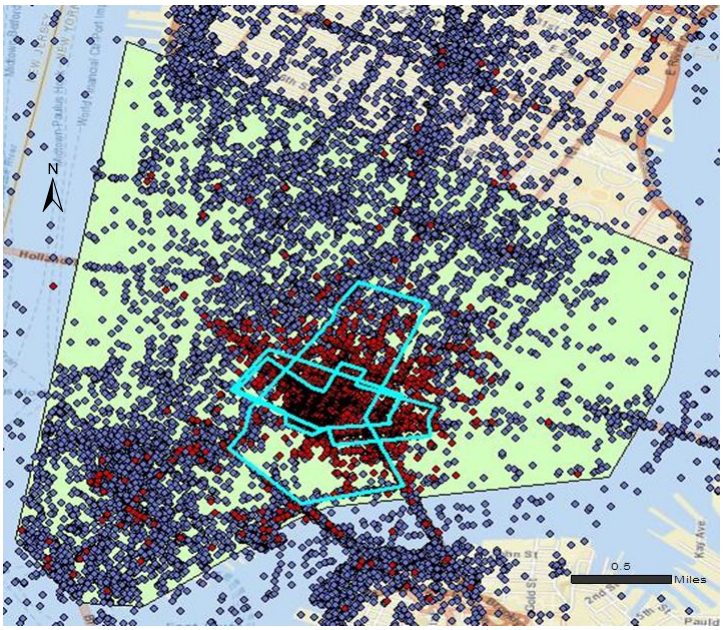


Figure 4 Preliminary boundary of Manhattan Chinatown

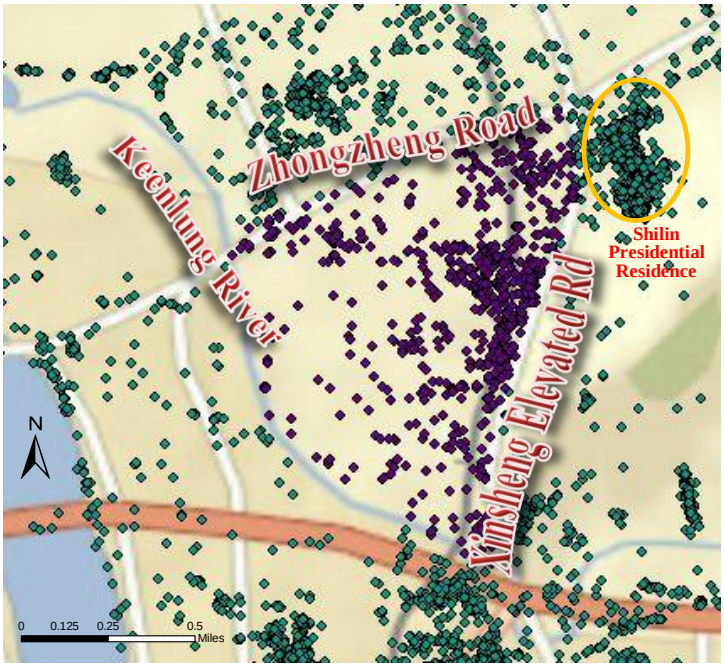


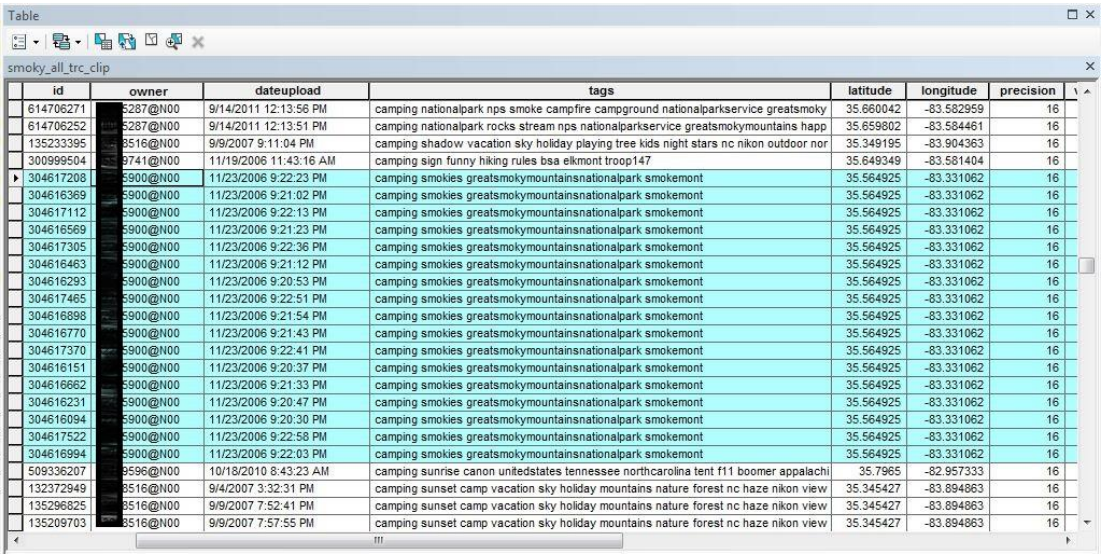
Figure 5 Preliminary boundary of Taipei Shilin Night Market

“greatsmokymountainsnationalpark,” and “smokemont,” as shown in Figure 6. Many interesting research topics concerning spatio-temporal human activity patterns and user generated geographic information may be conducted using these data. However, the research for this thesis addresses publicly perceived geographic extent of colloquial place names by using the input of as many people as possible. These duplicated image records from a single user will inaccurately increase the weight of several individual users and lead to bias in the research results. These should be identified as redundancies and removed before conducting an analysis. Therefore, when more than two images from the same person with the same coordinates and same tags are uploaded within six hours of each other, those images are considered redundant and only the first image record is kept. The final *effective* image set of the place is composed of images geotagged in the preliminary boundary of the place without redundancy.

3.2 Perceived Geographic Extent

3.2.1 Data Selection

In order to explore the perceived geographic extent of colloquial place names, tags from images geotagged within preliminary boundaries are counted. There are 3,776 different tags from the non-redundant images geotagged in the preliminary study area of Taipei Shilin Night Market, though 1,224 images do not have any tags. When analyzing images of this place, tags in the local native language, Traditional Chinese, are critical. To illustrate, 87 images are tagged with name of Taipei Shilin Night Market as “shilinnightmarket,” while 286 images are tagged with its Chinese name “士林夜市.” Besides Chinese and English names, some tags are aliases, such as “士林臨時市場”



id	owner	dateupload	tags	latitude	longitude	precision
614706271	[redacted]5287@N00	9/14/2011 12:13:56 PM	camping nationalpark nps smoke campfire campground nationalparkservice greatsmoky	35.660042	-83.582959	16
614706252	[redacted]5287@N00	9/14/2011 12:13:51 PM	camping nationalpark rocks stream nps nationalparkservice greatsmokymountains happ	35.659802	-83.584461	16
135233395	[redacted]8516@N00	9/9/2007 9:11:04 PM	camping shadow vacation sky holiday playing tree kids night stars nc nikon outdoor nor	35.349195	-83.904363	16
300999504	[redacted]9741@N00	11/19/2006 11:43:16 AM	camping sign funny hiking rules bsa elkmont troop147	35.649349	-83.581404	16
304617208	[redacted]5900@N00	11/23/2006 9:22:23 PM	camping smokies greatsmokymountainsnationalpark smokemont	35.564925	-83.331062	16
304616369	[redacted]5900@N00	11/23/2006 9:21:02 PM	camping smokies greatsmokymountainsnationalpark smokemont	35.564925	-83.331062	16
304617112	[redacted]5900@N00	11/23/2006 9:22:13 PM	camping smokies greatsmokymountainsnationalpark smokemont	35.564925	-83.331062	16
304616569	[redacted]5900@N00	11/23/2006 9:21:23 PM	camping smokies greatsmokymountainsnationalpark smokemont	35.564925	-83.331062	16
304617305	[redacted]5900@N00	11/23/2006 9:22:36 PM	camping smokies greatsmokymountainsnationalpark smokemont	35.564925	-83.331062	16
304616463	[redacted]5900@N00	11/23/2006 9:21:12 PM	camping smokies greatsmokymountainsnationalpark smokemont	35.564925	-83.331062	16
304616293	[redacted]5900@N00	11/23/2006 9:20:53 PM	camping smokies greatsmokymountainsnationalpark smokemont	35.564925	-83.331062	16
304617465	[redacted]5900@N00	11/23/2006 9:22:51 PM	camping smokies greatsmokymountainsnationalpark smokemont	35.564925	-83.331062	16
304616898	[redacted]5900@N00	11/23/2006 9:21:54 PM	camping smokies greatsmokymountainsnationalpark smokemont	35.564925	-83.331062	16
304616770	[redacted]5900@N00	11/23/2006 9:21:43 PM	camping smokies greatsmokymountainsnationalpark smokemont	35.564925	-83.331062	16
304617370	[redacted]5900@N00	11/23/2006 9:22:41 PM	camping smokies greatsmokymountainsnationalpark smokemont	35.564925	-83.331062	16
304616151	[redacted]5900@N00	11/23/2006 9:20:37 PM	camping smokies greatsmokymountainsnationalpark smokemont	35.564925	-83.331062	16
304616662	[redacted]5900@N00	11/23/2006 9:21:33 PM	camping smokies greatsmokymountainsnationalpark smokemont	35.564925	-83.331062	16
304616231	[redacted]5900@N00	11/23/2006 9:20:47 PM	camping smokies greatsmokymountainsnationalpark smokemont	35.564925	-83.331062	16
304616094	[redacted]5900@N00	11/23/2006 9:20:30 PM	camping smokies greatsmokymountainsnationalpark smokemont	35.564925	-83.331062	16
304617522	[redacted]5900@N00	11/23/2006 9:22:58 PM	camping smokies greatsmokymountainsnationalpark smokemont	35.564925	-83.331062	16
304616994	[redacted]5900@N00	11/23/2006 9:22:03 PM	camping smokies greatsmokymountainsnationalpark smokemont	35.564925	-83.331062	16
509336207	[redacted]9596@N00	10/18/2010 8:43:23 AM	camping sunrise canon unitedstates tennessee northcarolina tent f11 boomer appalachi	35.7965	-82.957333	16
132372949	[redacted]8516@N00	9/4/2007 3:32:31 PM	camping sunset camp vacation sky holiday mountains nature forest nc haze nikon view	35.345427	-83.894863	16
135296825	[redacted]8516@N00	9/9/2007 7:52:41 PM	camping sunset camp vacation sky holiday mountains nature forest nc haze nikon view	35.345427	-83.894863	16
135209703	[redacted]8516@N00	9/9/2007 7:57:55 PM	camping sunset camp vacation sky holiday mountains nature forest nc haze nikon view	35.345427	-83.894863	16

Figure 6 Duplicate images example at the Great Smoky Mountains National Park area

(Shilin Temporary Market) and “士林觀光市場” (Shilin Tourist Market). This preliminary study area excludes any other types of market places, and some users may split up the name, so tags like “market,” “nightmarket” and “傳統市場” (Traditional Market) should also be considered as referring to the Shilin Night Market, along with other versions (e.g., “shilinnightmarket 士林夜市” and “xilinnightmarket”). By filtering for images that have at least one of these tags, 595 images of Taipei Shilin Night Market were extracted, which is about seven times greater than the results obtained by using only its complete English name.

In the study area of the Great Smoky Mountains National Park, 4,979 of the 24,225 non-redundant images have no tags. The remaining images contain 17,252 unique tags. If searching for images using “greatsmokymountainsnationalpark,” 2,116 images are filtered out. However, the Great Smoky Mountains National Park is referred to by many abbreviations and aliases in daily language: “Smokey,” “Smokies,” and “Smoky Mountains.” Moreover, according to the U.S. National Park Service (http://www.nps.gov/findapark) there are no other national parks within 100 km of zip code 37738, which is the location point of the Great Smoky Mountains National Park in Google Maps. Variations of key words such as “thesmokies,” “nationalpark,” “greatsmokymountainsnp,” “grsm,” and “greatsmokynp” can all be considered tags for the Great Smoky Mountains National Park. In addition, some tags with combinations of the name and a certain time like “greatsmokymountainsinthefall” are also considered to refer to this place, based on their meanings. By applying these rules to select tags for filtering, the number of images for the Great Smoky Mountains National Park expands to 7,834.

The selection of tags for Manhattan Chinatown are similar to the other two study areas. Manhattan Chinatown has 49,496 images with no tags and as many as 130,430 unique tags, among which are tags like “chinatown,” “chinatownnyc,” “manhattanchinatown,” “neighborhoodmanhattanchinatown.” They were selected for analysis of Manhattan Chinatown. It is noteworthy that Manhattan, as a well known international tourist attraction, has a large number of image tags in languages other than English. There are not only Chinese tags like “纽约市唐人街” (Chinatown of New York City), “华埠” (Chinatown), and “中国城” (Chinatown), but also “中国のクオーター” (“Quarter of China” in Japanese), “quartierchinois” (“Chinatown” in French), “차이나타운” (“Chinatown” in Korean) and so on. Translating and understanding these tags increases the complexity of adequate analysis. Additionally, tags like “chinaown” are obviously typographical errors that can also be used to select images about Manhattan Chinatown. It is interesting to note that 6,300 images have at least one of the tags selected above, whereas 6,129 images are tagged with “chinatown.” Most of the images that are about Manhattan Chinatown and geotagged for this area use the tag “chinatown,” including images using mainly non-English tags.

However, colloquial names for a specific place—how people refer to it in daily speech—are much fuzzier and more complicated. For example, Taipei Shilin Night Market is famous for food vendors and small restaurants. Though surrounding businesses and shops selling nonfood items are also part of the night market, the distribution of images tagged with information about all kinds of food provides a perspective into Shilin Night Market. Another good example is the Great Smoky Mountains National Park. Cades Cove, one of the most widely known tourist attractions in the park, is sometimes

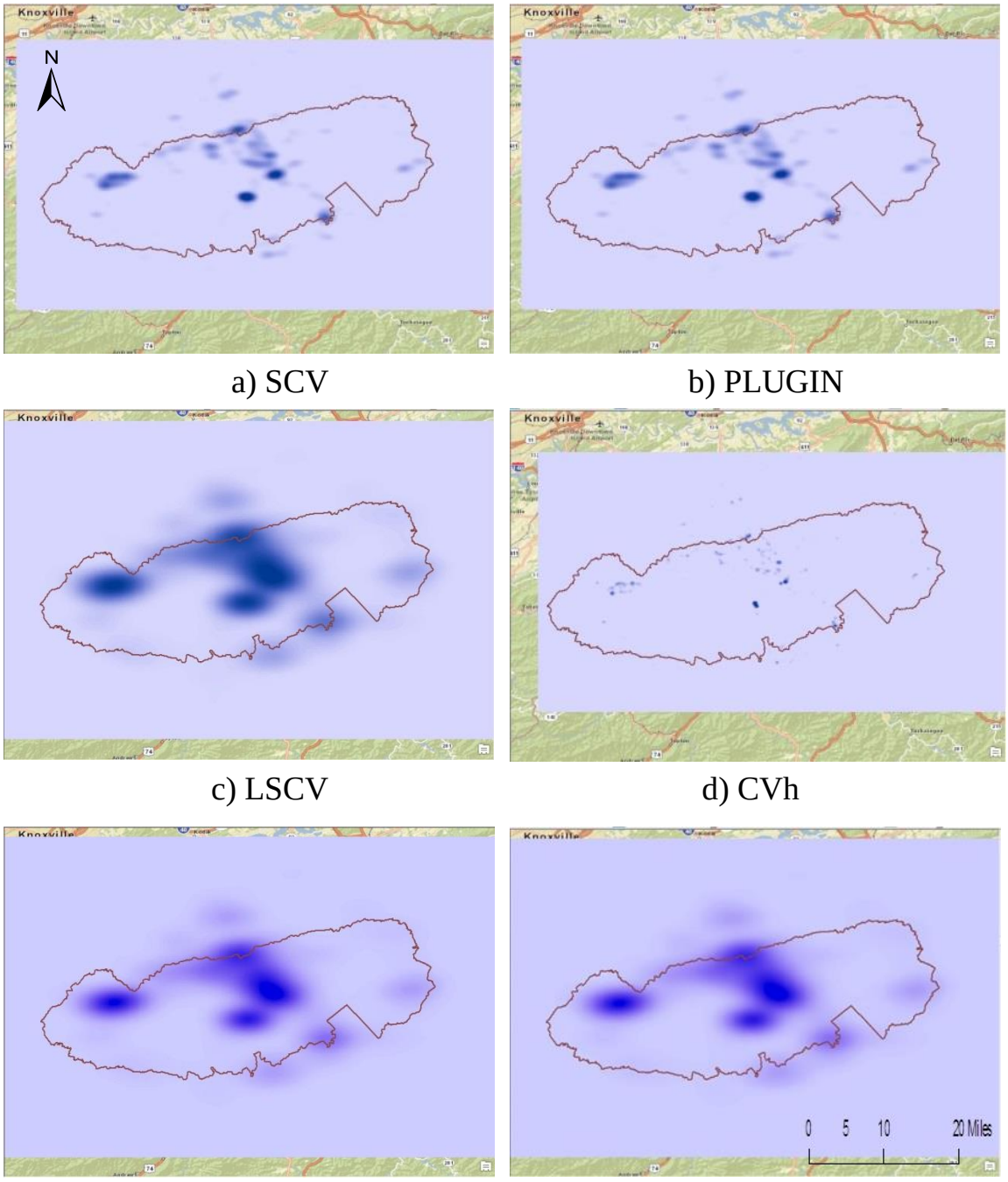
used to refer to the national park itself. Sometimes users also tag an image with the name of a trail or cabin they visited instead of name of the national park. Smoky Mountain was historically an important habitation of the Cherokee, and there are still many sites and activities about Cherokee culture in that area. In addition, the town of Cherokee, North Carolina is located at the east entrance of the national park [76]. Because of this, one expects that the geographic extent of images tagged with "cherokee" is related to the national park.

3.2.2 Density Estimation and Probability Map

We use a spatial density estimation method to represent the distribution of geotagged images and generate perceived geographic extent of the case study place. Kernel density estimation, a widely used spatial analysis technique, was selected for this study to generate a smoother density estimation surface. Selection of kernel function, grid cell size, and bandwidth is important for appropriate KDE analysis. Some easy-to-use KDE tools integrated with bandwidth selection methods are already available in both widely-used GIS software such as Spatial Analyst Extension for ArcGIS (ESRI®) and spatial analysis and modeling platforms such as Geospatial Modeling Environment (GME) [77]. The KDE tool in GME links ArcGIS to the "ks" [78] library in statistical software package R. If the grid size of the output estimation surface is too large, the details of the characteristics of the estimation result will not be properly displayed, while a grid size that is too small will waste computational efforts. Therefore, a series of grid sizes were tried and compared. When the grid size reaches a value that output raster surfaces with smaller grid size are not significantly more detailed for the specific case study, this value is selected. After testing and comparing, the grid size selecting method of ArcMap was

selected, which is the shorter of the width or height of the output extent in the output spatial reference, divided by 250 [2] is used. For its default and recommended kernel type, Gaussian, several bandwidth estimation algorithms are available including the plug-in estimator (PLUGIN), smoothed cross validation (SCV), likelihood cross validation (CVh), biased cross validation (BCV), a second BCV algorithm (BCV2), and least squares cross validation (LSCV) [77]. All these bandwidth selecting methods were tried and the researcher chose the most appropriate one among them by visualizing and comparing the results. Detailed discussion of selecting a bandwidth is in Chapter 4. Figure 7 demonstrates resulting density surfaces of KDE for points of images tagged with names or name variations of Manhattan Chinatown using different bandwidth selecting algorithms including SCV, plug-in, LSCV, CVh, BCV and BCV2. The boundary of Manhattan Chinatown on Zillow map is also displayed as reference for visualization.

The density value of each grid cell in the resulting density estimation surface can obviously be used as a measurement of probability level of which this area unit can be considered as part of the case study place. In other words, density value of the study area can be used to delimit the probable geographic extent of the case study place at different probability levels, according to the opinion of corresponding image data providers. However, the raw density raster file is not convenient for visualizing or for further analysis. If the density surface is reclassified using an interval of 5% of the maximum density value, areas of different percentage value ranges, 0-5%, 5%-10% … 95%-100%, are generated. That is to say, probable geographic extents of the place are generated at a series of probability levels.



a) SCV b) PLUGIN

c) LSCV d) CVh

e) BCV f) BCV2

Figure 7 KDE results of image points tagged as the Great Smoky Mountains National Park using SCV, plug-in, LSCV, CVh, BCV or BCV2 as bandwidth selecting algorithm

3.2.3 Spatio-temporal Analysis

Based on the perceived geographic extents of the case study places, we conduct a spatio-temporal analysis. As discussed in the Introduction, the perceived geographic extent of colloquial place names usually changes over time. This is especially true of places formed by aggregating human activities that are based on built structures like Manhattan Chinatown and Taipei Shilin Night Market. Migration, reconstruction, and extension of infrastructure and other man-made structures are important reasons. Images of each place taken during the entire time period (2004–2014) are divided into several subsets using an appropriate time interval, which is one year for this research.

The longitudinal consistency over time of these colloquial place names at certain probability levels can be examined in terms of location, geographic extents, and approximate area. To reduce possible bias from datasets that are too small and improve the reliability of the probability maps of each data subset, time windows may need to be adjusted to assign proper amounts of data images into data subsets. For example, there are either no or very few images of Shilin Night Market from 2004 to 2008 and 2013 to 2014. The images from 2005 to 2008 are combined into one data subset, and those from 2013 to 2014 are combined into another subset, creating a time series of data subsets of 2005-2008, 2009, 2010, 2011, 2012 and 2013-2014. Also, images from 2004 and 2005 are combined for Manhattan Chinatown, keeping the remaining data as single year subsets. A time series of geographic extent snapshots is created after conducting density estimations and generating a probability map from each data subset. Preliminary visualization revealed that the geographic extent of Manhattan Chinatown at a low probability level was expanding and migrating northeast, although the area of high

Table 2 Annual numbers of effective images and images tagged with place name variations of each case study

	The Great Smoky Mountains National Park		**Taipei Shilin Night Market**		**Manhattan Chinatown**	
	Effective	Tagged	Effective	Tagged	Effective	Tagged
2004	9	2	0	0	555	29
2005	234	172	8	5	3,189	89
2006	529	222	39	23	11,336	311
2007	1,454	589	91	24	14,835	518
2008	1,670	518	134	44	22,293	646
2009	2,147	657	473	100	26,986	849
2010	2,571	818	649	139	32,699	852
2011	4,513	1,534	681	83	52,582	838
2012	4,568	1,526	802	104	59,587	1,058
2013	5,787	1,695	1,470	52	62,039	957
2014	563	124	378	21	8,731	153

probability level did not change much. Changes were visualized, detected, and interpreted in this manner for Manhattan Chinatown and Shilin Night Market in order to reveal characteristics, hidden patterns, and trends of the collective cognition of these places.

Chapter 4

Results and Discussion

For each of the three case studies, the *Effective* column in Table 3 shows the number of non-redundant images geotagged within the preliminary boundary of the place. However, the amount of redundancy does not differ much if duplicated tags (those with more than two images from the same user with the same coordinates that are uploaded within six hours with each other) are ignored, as shown in the column *Redundancy Ignoring Tags*. Many redundant images uploaded within a short time period have exactly the same tags for the Great Smoky Mountains National Park (84.6%), for Manhattan Chinatown (81.4%) and for Shilin Night Market (87.7%). That is to say that users uploading and geotagging a batch of images for one location tend to give them the same tag. This phenomenon agrees with our experience that images geotagged at the same location usually represent the same place, same activity, and similar stories. When users give different tags to some of the images in a batch, these images usually have different contents, which should be considered in this study. Thus, a duplicate tag is one of the rules that define redundant data.

4.1 Geographic Extent of Colloquial Place Names

4.1.1 The Great Smoky Mountains National Park

The Great Smoky Mountains National Park lies within the Blue Ridge Mountains, and it is full of old growth forests. Elevations in the national park range from 876 feet to 6,643 feet [79]. Most areas in the national park are difficult for visitors to access. The park's activity areas mainly include tourist attractions, trails, cabins, and roads. Some

Table 3 Numbers of total images and effective images of the three places

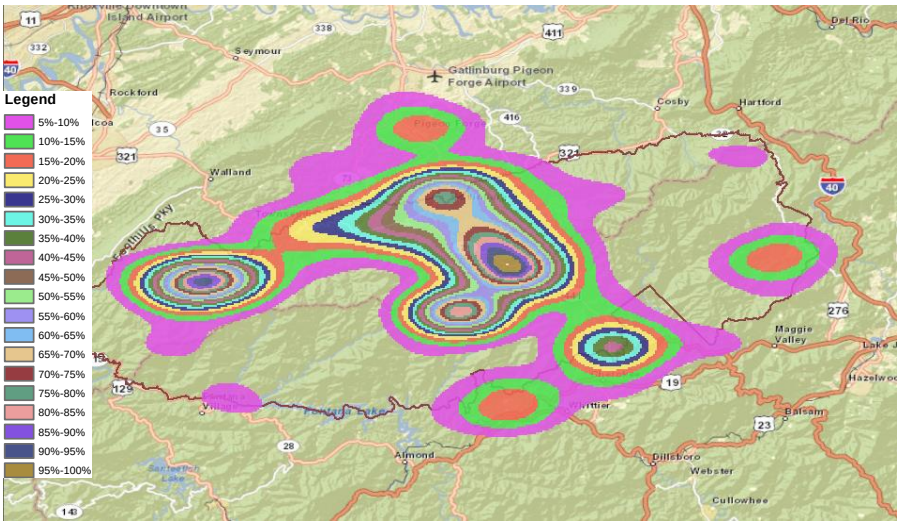
	Total	Effective	Redundancy	Redundancy Ignoring Tags
The Great Smoky Mountains National Park	46,318	24,225	22,093	26,109
Manhattan Chinatown	509,928	294,959	214,969	264,152
Taipei Shilin Night Market	10,428	4,725	5,704	6,500

visitors may not even consider these difficult-to-reach areas as part of the national park, even though these areas are within the official park boundary. Most of these areas have no cell phone signals or wireless Internet connection for portable electronics. Few images are geotagged within these areas. Thus, perceived geographic extent of colloquial place names of the national park may be located at roads, trails, and tourist attractions, and are different from the park's official administrative boundary.

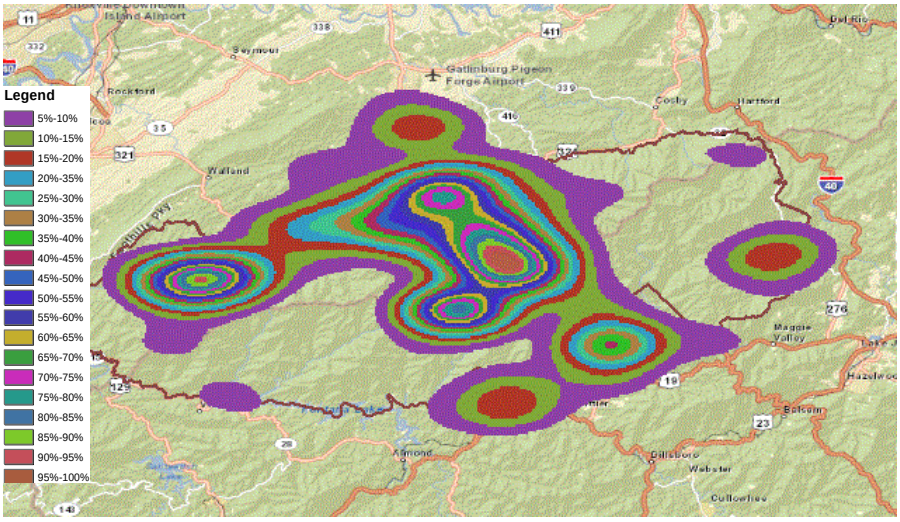
As discussed in Data and Methods, an output grid cell size of 300m is selected based on the approximate area of the national park's official boundary. This is used to obtain results with sufficient detail while controlling computational cost. Consequently, the official geographic extent in the resulting raster covers 34,707 grid cells. After conducting KDE using different bandwidth selecting algorithms, six resulting density estimation raster files were generated with the percentage value of probability for each grid cell. The density surface is reclassified using an interval of 5% of the maximum density value for better visualization and interpretation. Figure 8 shows the geographic extent of the Great Smoky Mountains National Park with probability greater than 5% using the SCV, plug-in, LSCV, CVh, BCV, and BCV2 as KDE bandwidth selecting algorithm. The national park's official boundary is a useful reference when selecting the most appropriate KDE result for generating a reliable probability map.

The geographic extents derived from the SCV and plug-in bandwidth selectors cover some of the most popular and accessible areas, and clearly show the characteristics and hotspots from the geotagged images. These results show a hotspot of images tagged with colloquial place names of the national park at Pigeon Forge, which is outside of the official boundary. However, these geographic extents are small and scattered. For

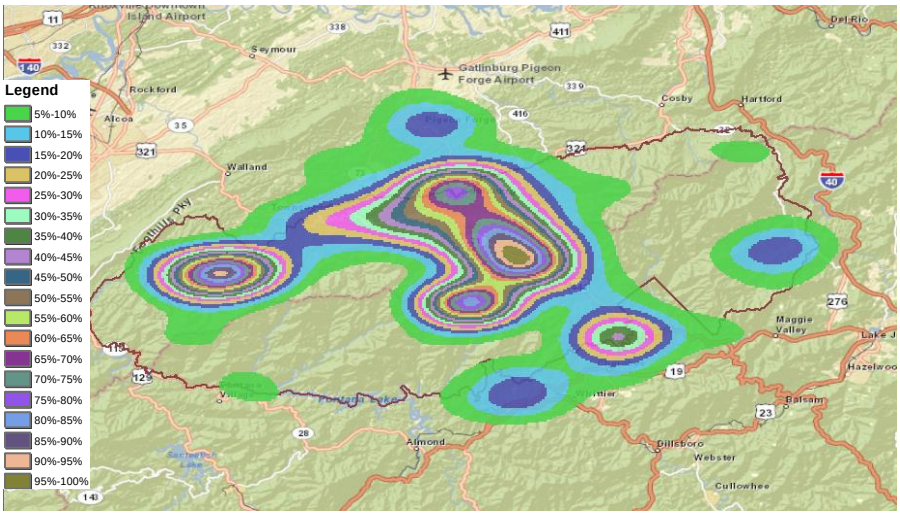
Figure 8 Geographic extents of the Great Smoky Mountains National Park of probability >5% using LSCV, BCV, BCV2, SCV, plug-in, or CVh as KDE bandwidth selectors



LSCV

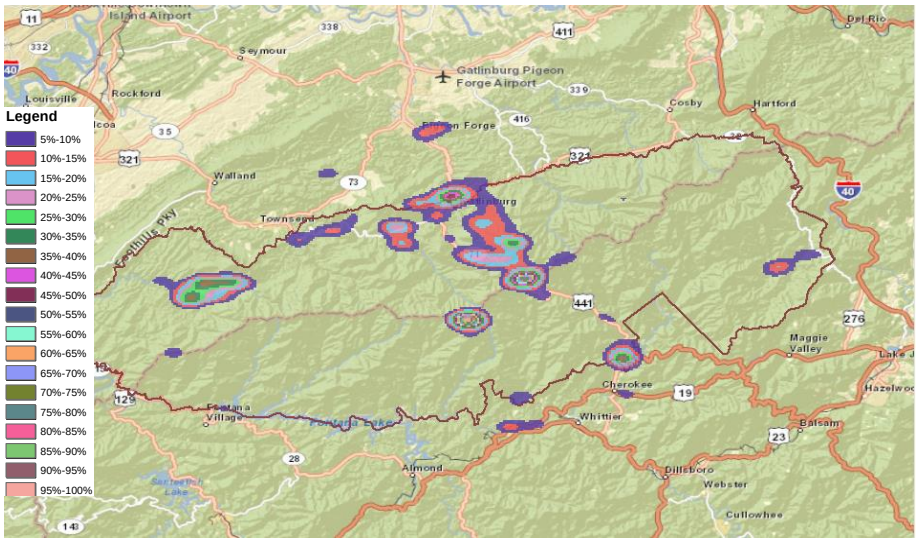


BCV

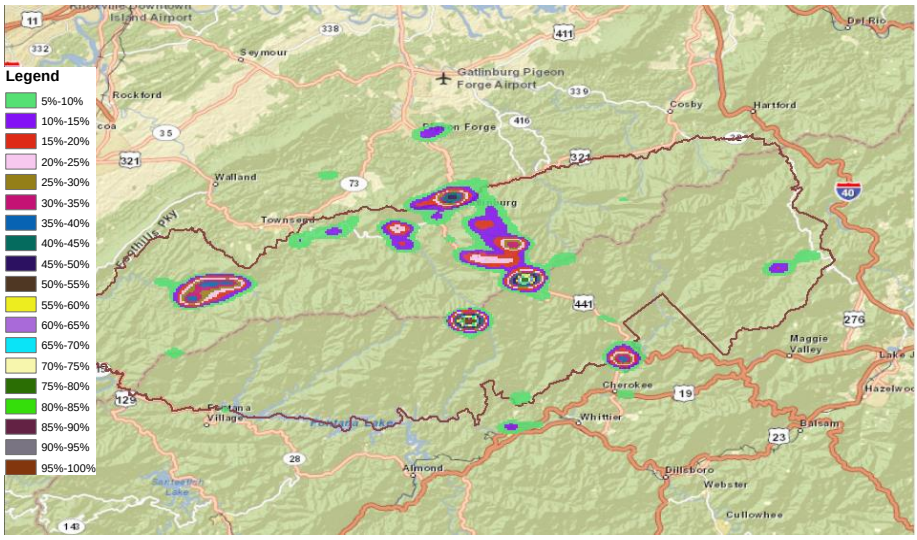


BCV2

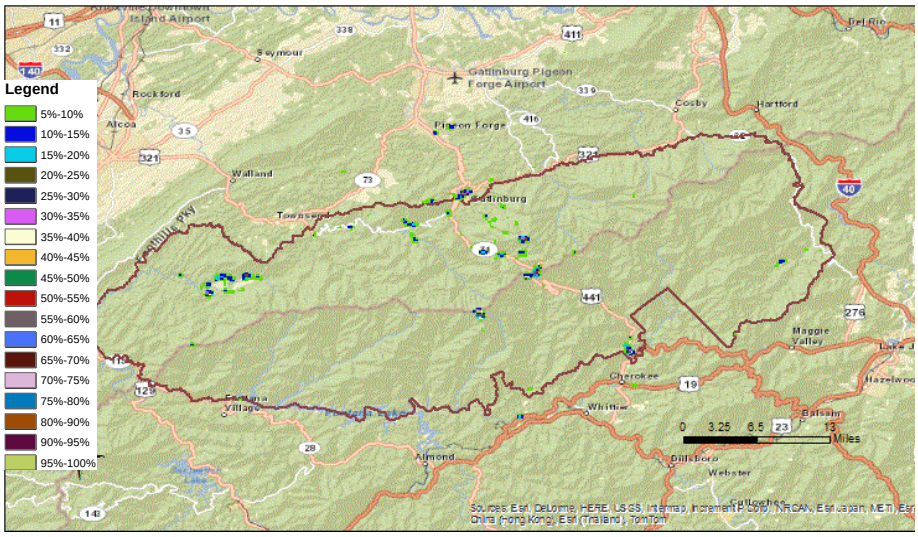
Figure 8. Continued



SCV



Plug-in



CVh

Figure 8. Continued

example, these extents only cover several segments of the road from Interstate Highway 40 to Cades Cove, even though it is clear that the missing segments are also considered by visitors to be within the national park. The CVh extent is even smaller. Contrary to these results, the results from BCV, BCV2, and LSCV smooth the boundaries and cover much larger areas. The areas of high probability levels present features that are similar to the other three results, although many details are missing. The geographic extents from BCV, BCV2, and LSCV include most areas that should be considered as being within the national park, but may also cover some extra areas. By taking the geographic extents of higher probability levels in these three results, more accurate extents may be created. Thus, the geographic extents from BCV, BCV2, or LSCV may be more appropriate.

Two indices are defined in order to verify the consistency between official boundary and perceived geographic extents,. The parts of perceived geographic extent that fall within the official boundary can be considered correct, and the proportion this occupies in the entire official extent is coverage rate:

$$C = \frac{N_c}{S}$$

where C is coverage rate, N_c is the number of grid cells in the resulting geographic extent that fall within the official boundary, and S is the number of grid cells in the official boundary. The ratio of the part that falls outside of the official boundary in the entire resulting geographic extent is out rate:

$$O = \frac{N_e}{(N_e + N_c)}$$

where O is out rate, N_e is the number of grid cells in the resulting geographic extent that fall outside of the official boundary, and N_c is the number of grid cells in the resulting

extent that fall within the official boundary. The geographic extent with a high coverage rate and a low out rate is closer to the official boundary. Extract by Mask tool in ArcGIS is used to extract parts that are within and outside of resulting geographic extents.

If the area of probability value greater than the lowest level, 5%, is selected as the resulted geographic extent, the number of grid cells in the geographic extent falling within and outside of the official boundary, as well as the coverage rates and out rates, are listed in Table 4.

Perceived geographic extents that result from using BCV and BCV2 as the bandwidth selectors have the same coverage rates and out rates, which are very similar to the results from using LSCV, especially the out rates. The coverage rate is slightly greater than LSCV. On the other hand, both coverage rates and out rates from BCV/BCV2 and LSCV are much greater than the other three results. Increased coverage rates lead to higher out rates as well.

Therefore, the areas that result from using BCV/BCV2 are used to map the perceived geographic extent of the Great Smoky Mountains National Park. BCV uses a biased cross validation criterion to minimize the estimation of asymptotic MISE, and is considered a hybrid of cross validation and plug-in. Comparing to plug-in estimators and unbiased cross validation, such as LSCV, BCV gives the largest bandwidth and the smoothest density estimation result [50]. But it loses some detailed information at areas of high probability. Thus BCV performs better at capturing the general geographic extents of the Great Smoky Mountains National Park, which has a relatively large area

Table 4 Grid cells in resulting geographic extents of probability level 5% with different bandwidth selecting algorithms falling within and outside of the official boundary of the Great Smoky Mountains National Park

	LSCV	BCV	BCV2	SCV	Plug-in	CVh
Out	9,923	10,168	10,168	759	799	107
In	18,655	19,130	19,130	3,546	3,670	524
Coverage Rate	53.75%	55.12%	55.12%	10.22%	10.57%	1.51%
Out Rate	34.72%	34.71%	34.71%	17.63%	17.88%	16.96%

and unevenly distributed image points. For better visualization, probability levels are merged into five classes: 0-5%, 5%-10%, 10%-20%, 20%-50% and 50%-100%; the area of 0-5% is ignored. Figure 9 shows the geographic extent map of the Great Smoky Mountains National Park at different probability levels.

A place's official boundary is useful when justifying the quality of geographic extent. The perceived geographic extent with a high probability level generally distributes along roads and tourist attractions, especially areas with relatively strong cell phone signals. This pattern reveals the crucial effect of accessibility, both physical and digital, on perceived geographic extents using geotagged OSN data.

4.1.2 Manhattan Chinatown

Unlike the Great Smoky Mountains National Park, there are several definitions and maps of the geographic extent of Manhattan Chinatown from various sources. While these sources have many differences, they do have the central core area in common. There is no well-defined, standard boundary to evaluate these resulting extents. The New York City borough of Manhattan is home to the largest enclave of Chinese people in the Western Hemisphere[80]. One characteristic that distinguishes it from surrounding areas is the concentration of Chinese stores and residential buildings with Chinese signs. In this case, evaluating the visual characteristics of the ambiguous areas can be an effective method to verify the geographic extent of Manhattan Chinatown. We used Google Maps Street View to examine the area's visual characteristics.

Figure 10 shows geographic extents of Manhattan Chinatown from different KDE bandwidth selectors where probability value is greater than 5%. Figure 11 shows the six resulted geographic extents overlaid according to ascending orders of size. The

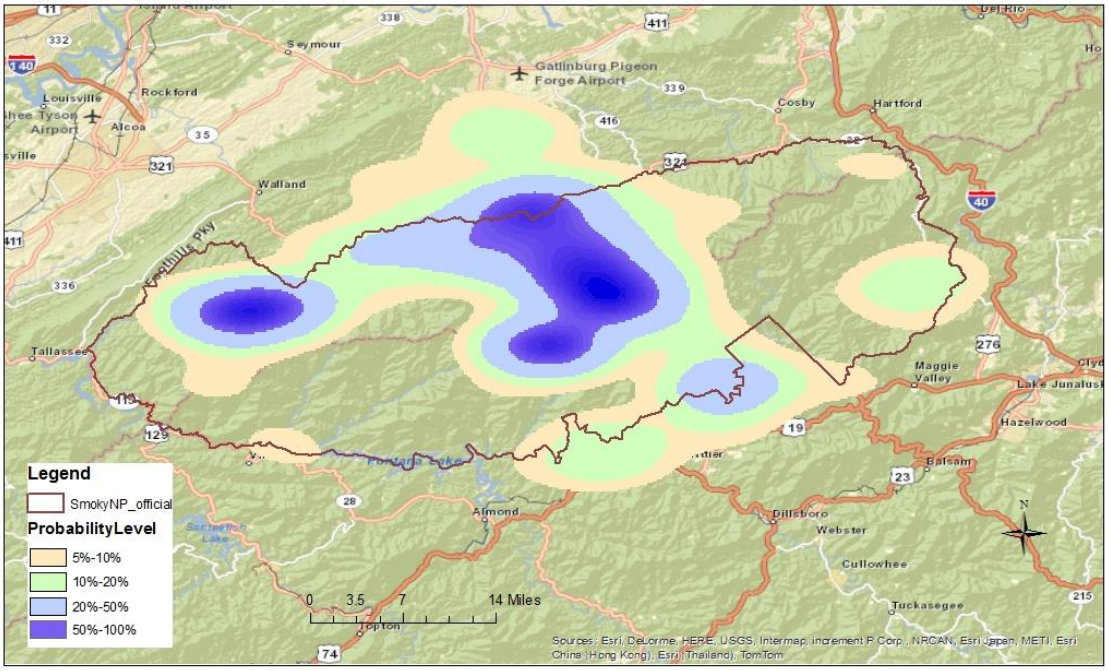
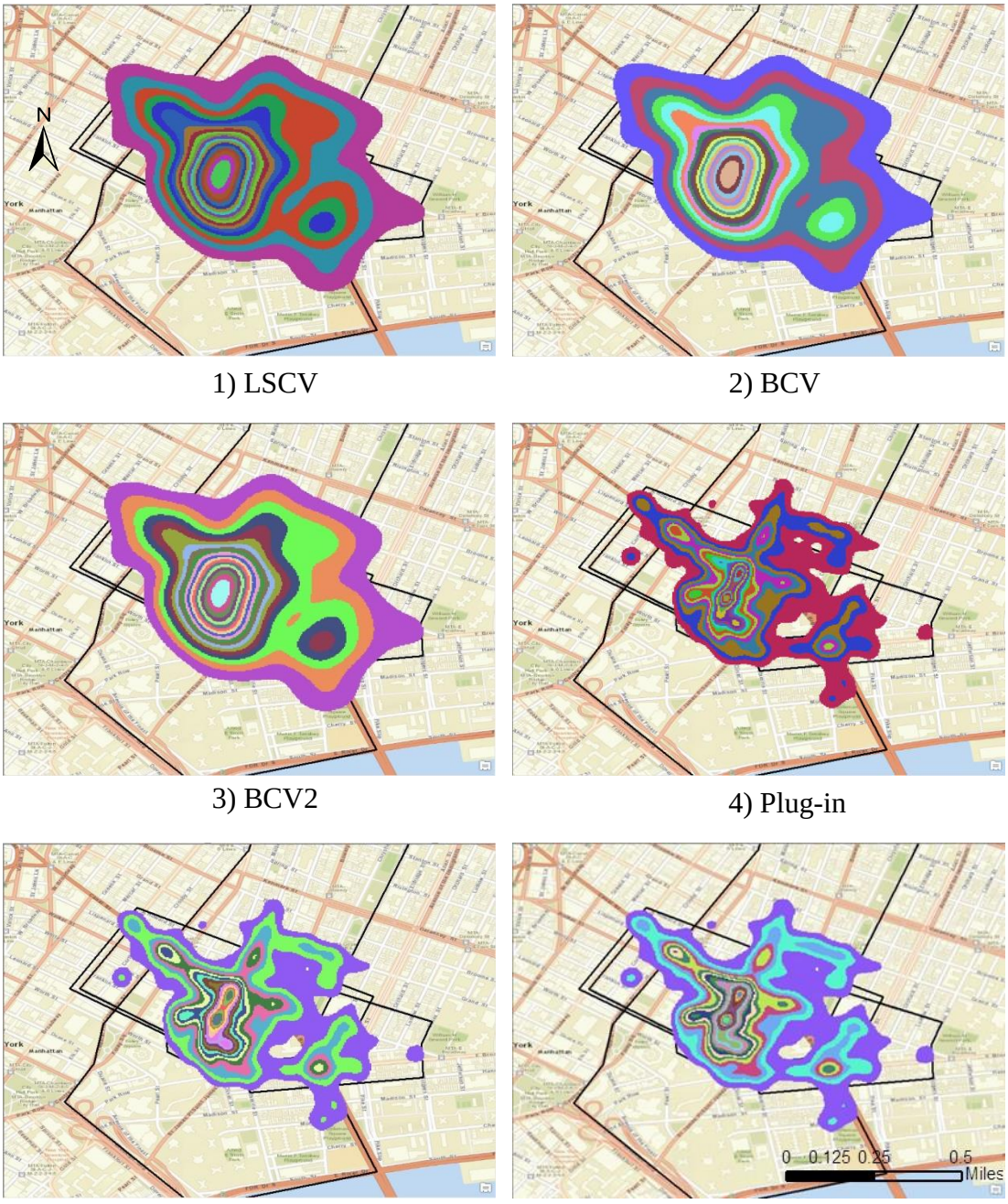


Figure 9 Probability map of the Great Smoky Mountains National Park based on images tagged with names and name variations



1) LSCV 2) BCV

3) BCV2 4) Plug-in

5) SCV 6) CVh

Figure 10 Geographic extents of Manhattan Chinatown of probability >5% using SCV, plug-in, LSCV, CVh, BCV or BCV2 as KDE bandwidth selecting algorithm

Figure 11 Comparing perceived geographic extents of Manhattan Chinatown from bandwidth selector plug-in, CVh, SCV, LSCV and BCV using Google Maps Street View at six sample areas

geographic extents from BCV/BCV2 and LSCV are smoother and cover a larger area, while the results from SCV, CVh, and plug-in produces better details. The last three results display not only more detailed boundaries but also capture the discontinuous areas with low probability values within the Manhattan Chinatown area such as the area bounded by the Manhattan Bridge, the Bowery, and Division Street (marked as ⑦ in Figure 11).

BCV and BCV2 yield the same results, which are similar to the results from LSCV. Shapes of the remaining three results resemble each other, but the SCV shape is slightly larger than CVh and plug-in shapes. To verify the accuracy of these geographic extents, seven sample areas were selected: ① Baxter Street between Grand Street and Hester Street, ② Elizabeth Street, south of Broome Street, ③ the intersection of Ludlow Street and Division Street, ④ Madison Street close to Catherine Street, ⑤ the intersection of Saint James Place and James Street, ⑥ Canal Street between Broadway and Cortlandt Alley, and ⑦ the block bounded by the Manhattan Bridge, the Bowery, and Division Street. The geographic extents from bandwidth selector CVh, SCV, and plug-in cover part of sample area ⑥ but do not cover areas ① to ⑤ and ⑦; the geographic extents from BCV/BCV2 and LSCV are similar and cover all seven areas. The south edge of the extent from LSCV is located right at Saint James Place and James Street, while the one from BCV/BCV2 slightly exceeds this cross area ⑤.

According to Google Maps Street View, sample areas ① to ⑤ (shown by green arrows in Figure 11) appear to have more Chinese signboards compared to other streets in New York, and most are in residential areas. Sample area ⑦ is a large complex of

Confucius Plaza that attracts fewer visitors. Even though there are fewer geotagged images of this area, it is considered part of Chinatown. However, there are few Chinese signs or Chinese stores in sample area ⑥ on Canal Street (shown by a red arrow in Figure 11), even though it is a prosperous street with many varied stores along it. In fact, even though the area of Canal Street between Mercer Street and Cortlandt Alley are classified within the geographic extent of Manhattan Chinatown using all six bandwidth selectors, there is no clear indication that it is part of the Chinese enclave. The westernmost appearance of Chinese signs starts at the intersection of Canal Street and Cortlandt Alley. The concentration of Chinese signs extends southwest to the area around Saint James Place and James Street, which locates at the resulting boundary when using LSCV. Nevertheless, Chinese signs seem to extend north to East Houston Street and east to the Pitt Street area, which are two or three blocks further than the largest resulting geographic extent of LSCV. It is difficult to define a clear boundary for Chinatown at its marginal areas where there are few Chinese signs, so not all these areas should be considered as part of Chinatown. While the resulting area from LSCV shows less detail of the boundary, on the whole it better represents the geographic extent of Manhattan Chinatown.

Based on the mapping method discussed above, the KDE result yielded by bandwidth selector LSCV is considered to be the geographic extent of Manhattan Chinatown. LSCV attempts to minimize ISE and select bandwidth that adapts to the smoothness of data [48]. The density estimation result from LSCV tends to under smooth data and makes changes of density prominent [51]. Though LSCV has high variability [45], it performs better at Manhattan Chinatown, which has a large amount of data, a

relatively small area, and possibly sharp peaks or valleys. Figure 12 shows the perceived geographic extent of Manhattan Chinatown at various probability levels.

For places with obvious physical characteristics, Google Maps Street View is a useful source to justify the quality of perceived geographic extent. Though Chinatown is in an urban area and has generally good accessibility and full cell phone signal, the perceived geographic extent still tends to distribute along roads at the edges of the area. When comparing reference boundaries from different sources, the perceived geographic extent with a probability level greater than 50% mainly falls in the shared area of all reference boundaries. The area with a probability level less than 50% and greater than 5% exceeds the boundaries specified by Google Maps and Zillow on the south and east, but does not completely cover the extent on the west. Thus, the perceived geographic extent derived from geotagged Flickr images shares the core area with other reference boundaries but does not agree with any of them.

4.1.3 Taipei Shilin Night Market

Using the same methods, Figure 13 shows the geographic extent of Taipei Shilin Night Market yielded by six different bandwidth selecting algorithms. All these resulting extents have a smaller, separate area several blocks away from the major area around the Shilin MRT Station, indicated by a red circle. There are several storefronts and food stands for passengers around the station. Although the station is obviously far from the general area of the night market and separate from it, there are still some geotagged images tagged with the name or name variations of Shilin Night Market. This may be because of its name and large passenger flow as a traffic node. This area was removed from further geographic extent analysis in this study.

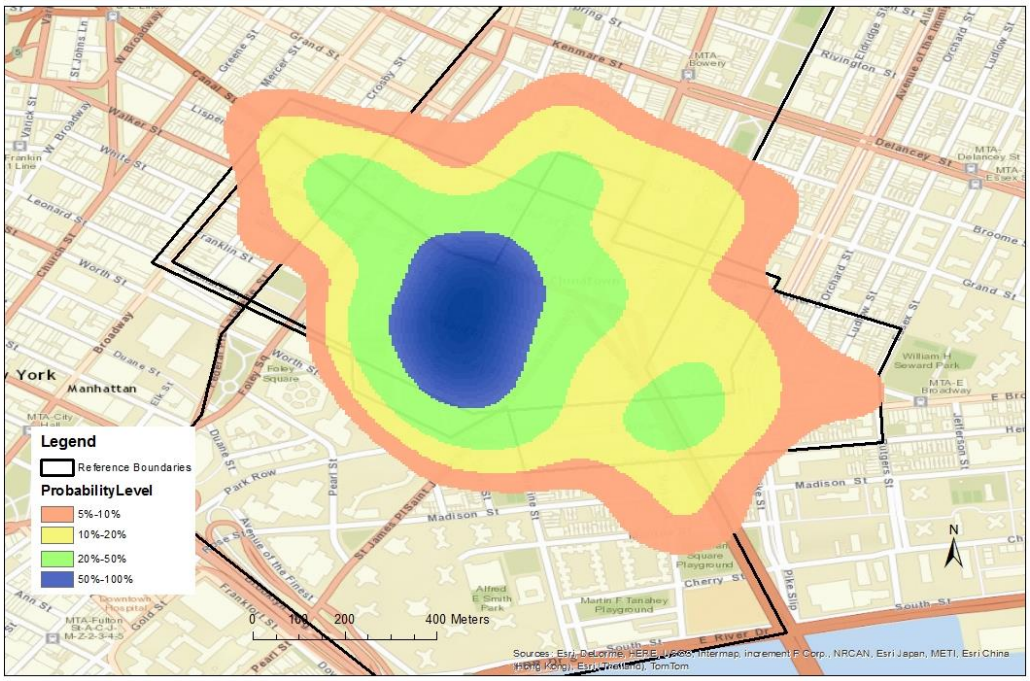


Figure 12 Probability map of Manhattan Chinatown based on images tagged with names and name variations

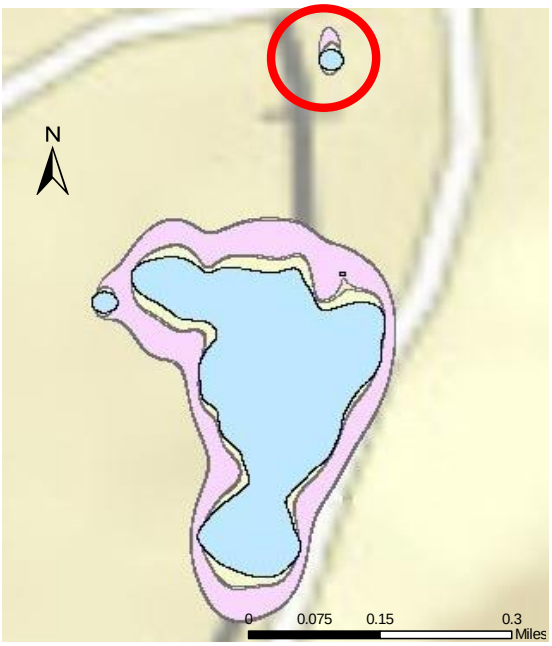


Figure 13 Geographic extent results of Taipei Shilin Night Market at probability >5% using six different bandwidth selecting algorithms.

Taipei Shilin Night Market does not have any reference boundaries from commonly used online sources available, such as Google Maps, Bing Maps, and Wiki Travel. The night market encompasses a food court, some storefronts, and roadside stands distributed in the surrounding area, as well as cinemas, video arcades, and karaoke bars. Some small side streets feature retractable roofs and vendors. Larger open streets are usually full of storefronts on both sides of the road that form the night market. The formation of night markets depends on high pedestrian accessibility, so streets crowded with roadside automobile or bicycle parking are not good potential areas for night markets. Google Maps Street View can be used to verify these geographic extents of Shilin Night Market. Symmetric Difference and ArcGIS's Merge tool were conducted to find the controversial areas of these resulting geographic extents, as shown in Figure 14. As in the case of Manhattan Chinatown, the results of BCV and BCV2 were the same and only BCV was considered. Due to the lack of detailed data about this area on the World Street base map, Google Fusion Tables (https://www.google.com/drive/using-drive/#fusiontables) was used instead as the base map and location reference. The Shape Escape (http://shpescape.com/) tool was used to directly import shapefiles to Google Fusion Tables. Six sample areas were selected: ③ and ⑥ were included in all the results, ① and ⑤ were in the results of using SCV or plug-in, but not in LSCV and BCV, ② and ④ were only in the results of LSCV and BCV.

At ① on Dabei Road, Google Maps Street View shows many storefronts with food, clothes, and small items as well as traces of temporary vendors, showing that this area has potential for a night market area. However, area ② on Xiaobei Road is a mostly residential area with only a clinic, an alteration and tailor shop, and a barber shop.

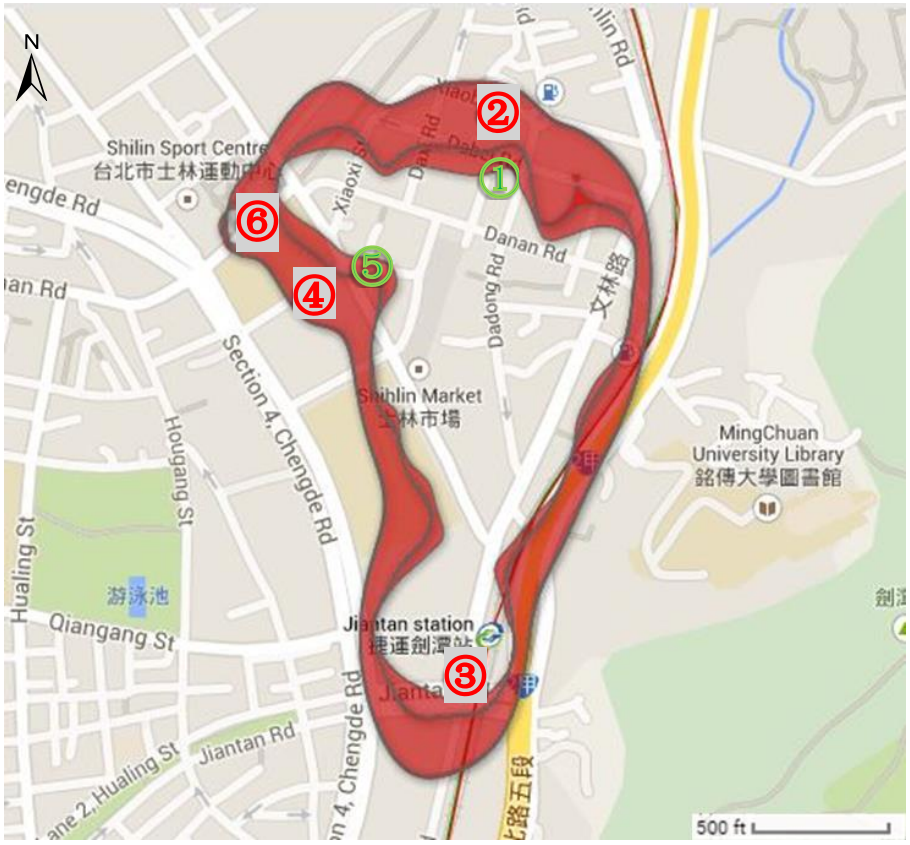


Figure 14 Comparing differences of resulting perceived geographic extents of Teipei Shilin Night Market with six different bandwidth selectors on Google Maps

Parked motorcycles and cars occupy both sides of the streets, leaving no space for street vendors. Its potential as a night market area is low.

Similarly, as shown in Figure 16, area ④ on Lane 195, Chengdu Road is a relatively large road with an aquarium and tall residential buildings on both sides. There appears to be no space for street vendors. However, area ⑤ across Jihe Road from ④ is on a small side street full of small food stores. Even though they appear to be closed during the day on Google Maps Street View, they are very likely part of the night market.

Google Maps Street View shows that the side of Jihe Road between Jiantan Road and Wenlin Road is a construction site, and there seems to be no space for street vendors. A few posters outside the construction site advertise Shilin Public Market. By observing this area using Google Maps Street View, it does not appear to have high potential as a night market area. However, this area is covered by all resulting extents of the aforementioned bandwidth selectors as part of Shilin Night Market. The most current street views of this area were taken in January 2012. One speculation is that this area had previously been a night market that was demolished before 2012. If so, this area should have a great number of geotagged images before 2012, but the number should diminish after that. This illustrates the uncertainty of using Google Maps Street View to verify the geographic extent of places.

Area ⑥ is located on Danan Road between Jihe Road and Shishang Road, and it is within the geographic extents of all six KDE bandwidth selectors. The remainder of this road segment is mostly covered by extents of only BCV and LSCV. However, based on Google Maps Street View, it is a wide and open road with a tall sport center and residential buildings on one side and a park on the other. The whole road segment does

Snapshots of street views at ① on Dabei Rd

Snapshots of street views at ② on Xiaobei

Figure 15 Snapshots of Google Maps Street View at ①on Dabei Road and ②on Xiaobei Road

Snapshot of street view at ⑤ Snapshot of street view at ④

Figure 16 Snapshot of Google Maps Street View at ④on Lane 195, Chengdu Road and at ⑤on a side street across Jihe Road from Chengdu Road

Figure 17 Snapshots of Google Maps Street View at ③ on cross of Jiantan Road and Jihe Road

Figure 18 Snapshot of Google Maps Street View at ⑥ on Danan Road between Jihe Road and Shishang Road

not seem to have any potential as a night market. One possible reason for the dense geotagged images for this location is the parking lot with a large visitor flow.

Hence, the geographic extent that results from using the plug-in bandwidth selecting algorithm is the most accurate among all six results, despite of some minor inaccuracies. Plug-in bandwidth selectors seek for a bandwidth that minimizes MISE [55]. Jones et al. [45] believe that the solve-the-equation plug-in method is most reliable in terms of overall performance after doing real data examples, asymptotic analysis, and simulations. Plug-in bandwidth selectors are usually tuned by arbitrary pilot estimation. They select larger bandwidths comparing to classical estimators [50], such as LSCV, and smaller bandwidths comparing to hybrid new estimators, such as BCV. In this case, the plug-in bandwidth selector performs better than any other methods for the density estimation of images of Taipei Shilin Night Market. Using the methods discussed above, Figure 19 shows the probability map of Taipei Shilin Night Market. With an outer area composed by street vendors and small businesses, Taipei Shilin Night Market is a good example of a place whose geographic extent has high flexibility and hardly any reference boundary. The probability mapping method of perceived geographic extent using geotagged OSN data provides a practical and interesting way to delimit this kind of place. Though there is no reference boundary, Google Maps Street View can be used to verify the quality of perceived geographic extents based on selected characteristics such as parking along the side of the street.

4.2 Related Terms of Colloquial Place Names

Colloquial names for a place are usually fuzzy. As discussed in Data and Methods, the names of some famous tourist attractions in the Great Smoky Mountains National Park

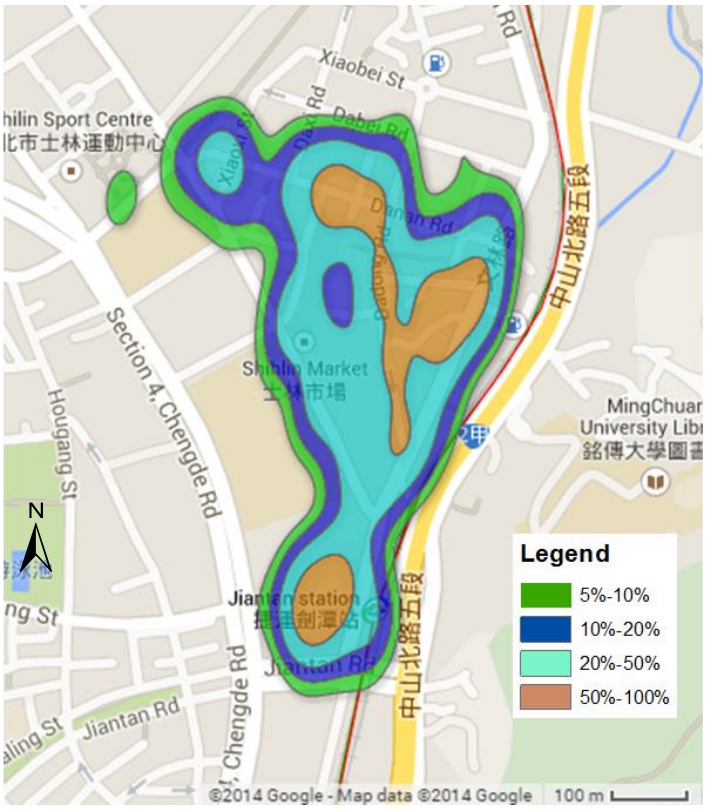


Figure 19 Probability map of Taipei Shilin Night Market based on images tagged with names and name variations

are sometimes used to refer to the national park itself, such as Cades Cove, Mont Le Conte, and Laurel Falls. Because of the historical habitats of the Cherokee Indians [81], images tagged with Cherokee are assumed to be highly related to the national park. Since cabins and trails are important components of tourist activities, those terms also appear frequently in image tags of the national park, and sometimes can be considered colloquial terms for the park.

Images tagged "trail," "cabin," "cherokee," or "cadescove," along with their variations, were separated from the preliminary dataset. The distribution of these images was mapped using the KDE method discussed above with bandwidth selector BCV, then the boundary of probability with a value greater than 5% was taken as the geographic extent. The geographic extent of images tagged "trail" (Figure 20) was even more consistent with the official boundary than images tagged with names. The out rate was 33.46% and coverage rate was 84.74%, while for the resulting extent of name variations using BCV bandwidth selector, out rate was 34.71% and coverage rate 55.12%. These tags were considered to be colloquial terms for the national park when filtering geotagged images. However, the Great Smoky Mountains National Park is part of the Blue Ridge Mountains and adjacent to several forests and parks that attract tourists. The area is also home to some beautiful trails, such as the protruding area on the northeast and the two smaller areas on the southwest. Thus, redundancy may be an issue when using variations and phrases for "trail" as a colloquial term for the Great Smoky Mountains National Park.

On the other hand, images tagged "cabin" are mainly outside the national park in the populated areas, such as Gatlinburg, Pigeon Forge, and Cherokee. This is expected because the Great Smoky Mountains National Park does not permit commercial

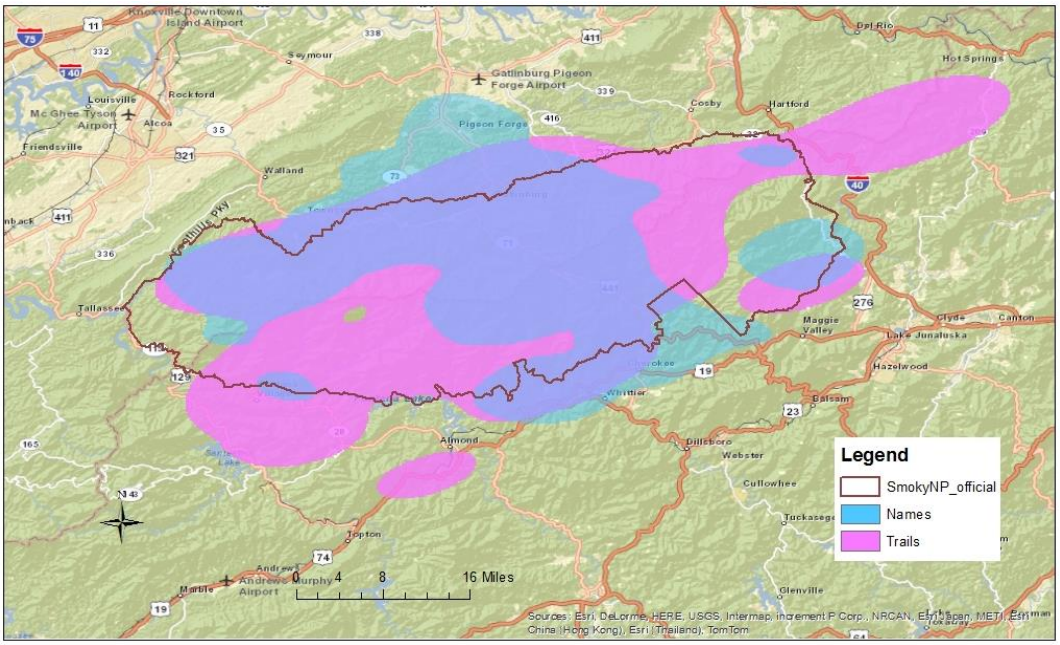


Figure 20 Geographic extent of images tagged about "trail" at the Great Smoky Mountains National Park

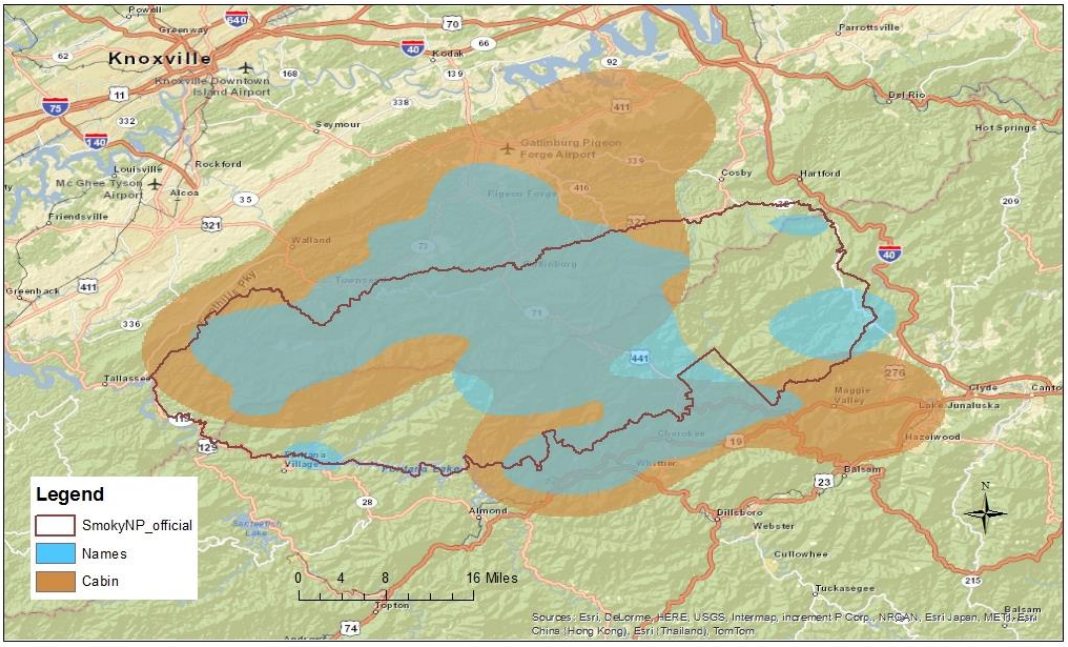


Figure 21 Geographic extent of images tagged about "cabin" at the Great Smoky Mountains National Park

activities. Mountain cabins are mostly constructed on relatively flat, hilly areas with high accessibility—except for a few featured cabin hotels—while a large part of the national park is located in the dense forests and high mountains. The coverage rate of the geographic extent of images tagged "cabin" was 62.89%, but the out rate was as high as 57.27%. Thus, even though cabin is an important colloquial term used by people, especially tourists, to indicate their activities at the Great Smoky Mountains National Park, the distribution of related image tags is much different from the national park boundary.

Similarly, it is commonly known that the term "cherokee" is highly related to the Great Smoky Mountains National Park's tourist activities in and around this area. However, the distribution of images tagged "cherokee" is not consistent with the official boundary. As shown in Figure 22, these images are mainly distributed near the town of Cherokee, N.C. with its famous museum and casino. A small portion of the images is geotagged at some historic sites of Cherokee culture within the national park.

Cades Cove, one of the most famous tourist attractions in the park, is sometimes used to refer to the national park itself. Images tagged "cadescove" mainly distribute at the Cades Cove area and the road leading to it. The geographic extent of these images is completely within the national park boundary. Thus some tourist attractions within the Great Smoky Mountains National Park like Cades Cove can supplement useful colloquial terms when analyzing colloquial place names of the national park.

Shilin Night Market is famous for its eateries and street vendors selling authentic Taiwanese snacks. Among the most famous snacks are deep-fried chicken breasts, pan-fried dumplings, grilled Taiwanese sausages, and pearl milk tea. To a great extent, these

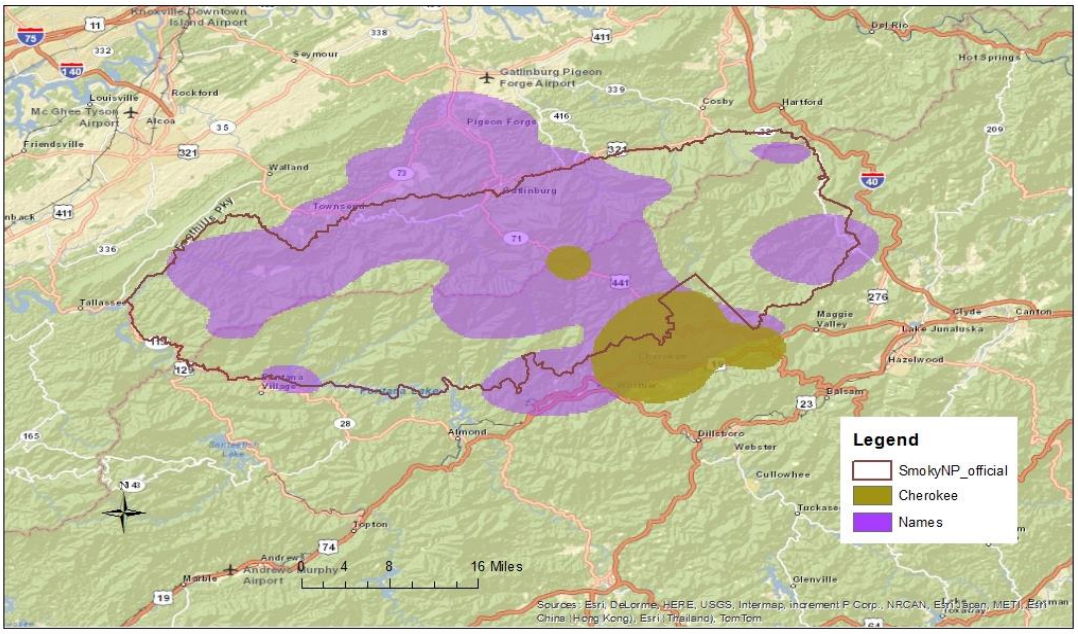


Figure 22 Geographic extent of images tagged about "cherokee" at the Great Smoky Mountains National Park

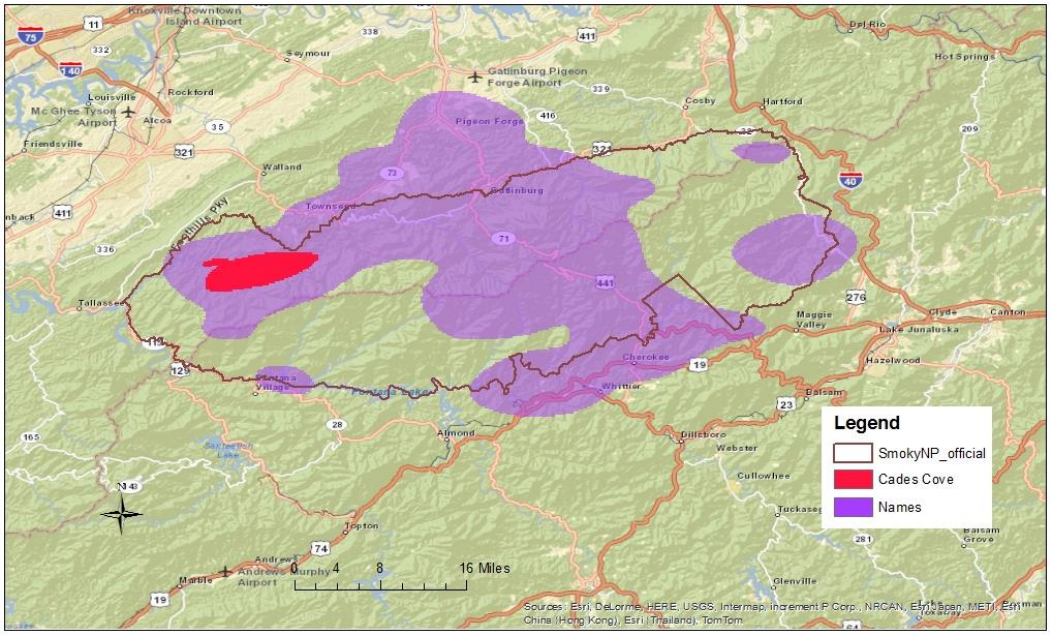


Figure 23 Geographic extent of images tagged about "cadescove" at the Great Smoky Mountains National Park

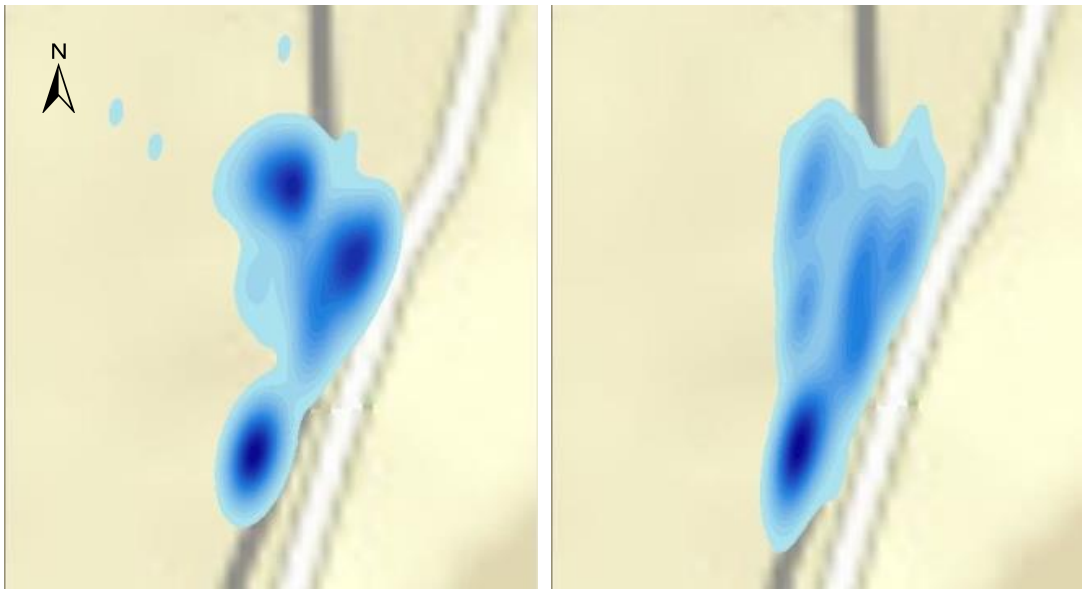
snacks can be considered trademarks of night markets in Taiwan. Though there are over one hundred night markets in Taiwan, Shilin Night Market is one of the most famous ones [82]. It also is the only night market within and around our preliminary study boundary. The distribution of images tagged snack names is consistent with the geographic extent of Shilin Night Market but covers a larger area, as shown in Figure 24. This comparison agrees with common sense. Those snack vendors and storefronts are not located only in night markets but may also be scattered throughout other locations such as areas around the Shilin MRT Station. Even though featured snacks are commonly used colloquial terms for night markets, using images with these tags to generate the geographic extent of the night market brought in certain redundancy that was hard to eliminate.

4.3 Spatio-temporal Analysis

Shilin Night Market utilizes the Shilin Public Market, which was purposely built to be a marketplace, and it occupies sidewalks adjacent to streets or entire streets that are normally thoroughfares by day. The extent greatly depends on locations of storefronts and mobile vendors. Based on the discussions above, the resulting geographic extent from plug-in bandwidth selector with probability value greater than 5% was used as the night market boundary. The number of selected images for each of the 11 years (from 2004 to March 1st, 2014) is: 0, 5, 23, 24, 44, 100, 139, 83, 104, 52 and 21. In order to get more reliable density estimation results, images from 2005 to 2008 and from 2013 to 2014 were combined, respectively. Figure 25 shows the resulting geographic extents of each time period. The geographic extents from 2005 to 2012 were similar but with small differences, while the extents of 2013 and 2014 were not

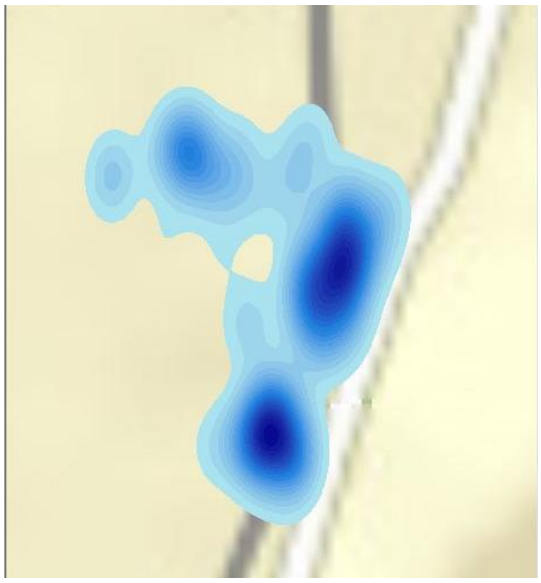
Figure 24 Geographic extent of images tagged about food at Taipei Shilin Night Market

Figure 25 Geographic extents of Taipei Shilin Night Market at different times

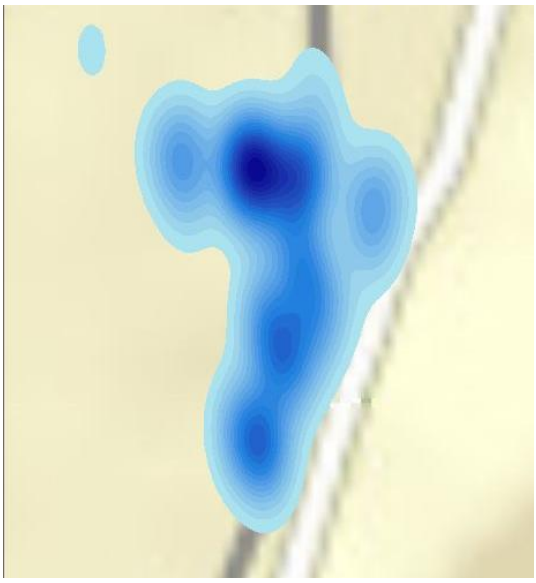


2005 - 2008

2009

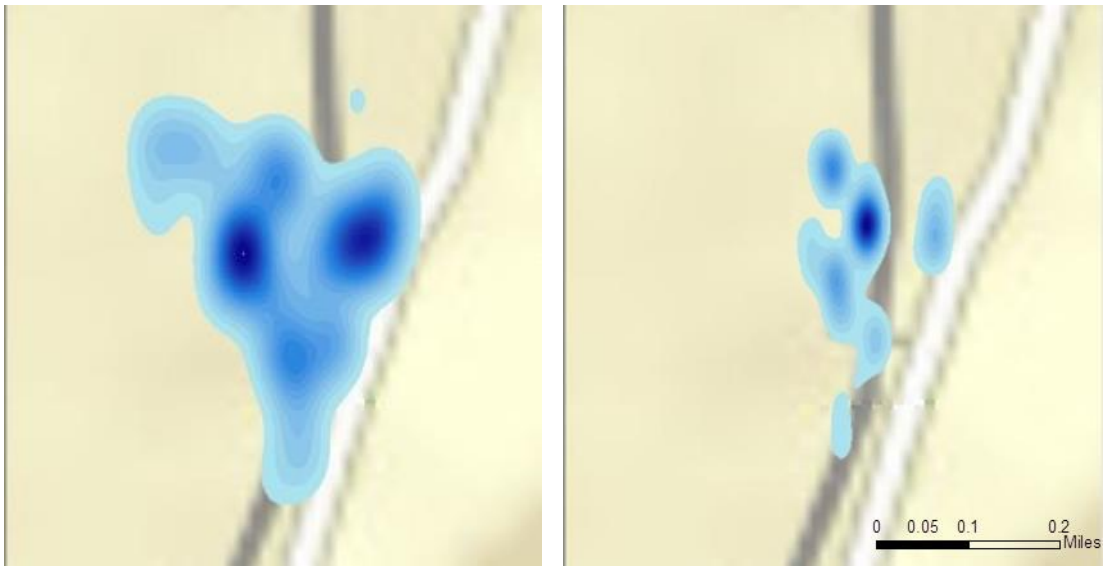


2010



2011

Figure 25. Continued



2012 2013 - 2014

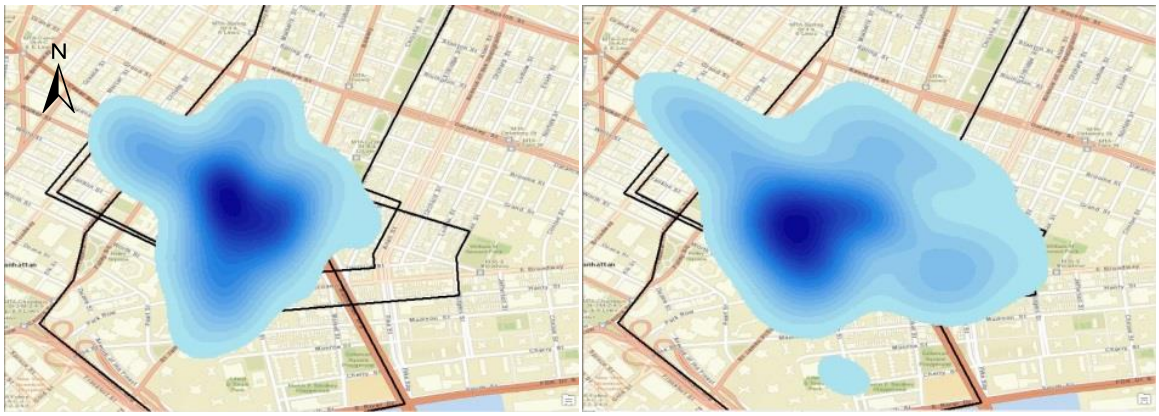
Figure 25. Continued

consistent. The reason for the unusual result of the last two years is not clear. According to these resulting extents, the night market has been expanding northwest since 2010. Even though there was already a relatively small number of images geotagged at the northwest before 2008, the expansion trend began from 2010 to 2012.

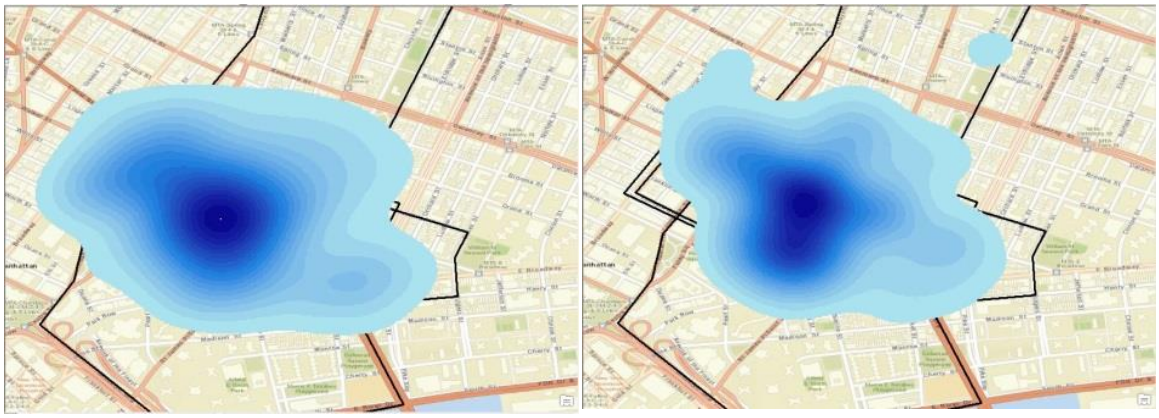
It is interesting to note the density change of the area on Jihe Road between Jiantan Road and Wenlin Road, which is south of the night market. This area was obviously part of the geographic extent with high probability from 2005 to 2010, and it gradually expanded. Its extent reached a peak in 2010. However, it began to shrink in 2011, and almost disappeared by 2013 and 2014. As shown in Google Maps Street View in Figure 17, this area was a construction site in January 2012. This result confirms the speculation that this area used to be an important part of Shilin Night Market, but was torn down sometime in 2011.

Manhattan Chinatown is composed of man-made infrastructure. Reconstruction and extension of the infrastructure complex, as well as migration of residential and business activities, can change its geographic extent according to peoples' cognition. The KDE result of images tagged with Chinatown's name and name variations, using the LSCV bandwidth selecting algorithm with a probability level greater than 5%, was used as the geographic extent of Manhattan Chinatown. Since the number of images in 2004 is too small to generate a reliable geographic extent, the images from 2004 to 2005 were combined and the rest of the images were grouped into one-year time windows. The numbers of images from 2004 plus 2005 to 2014 are 118, 311, 518, 646, 849, 852, 838, 1058, 957 and 153. The resulting geographic extents for each year are shown in Figure 26.

Figure 26 Geographic extents of Manhattan Chinatown at different times



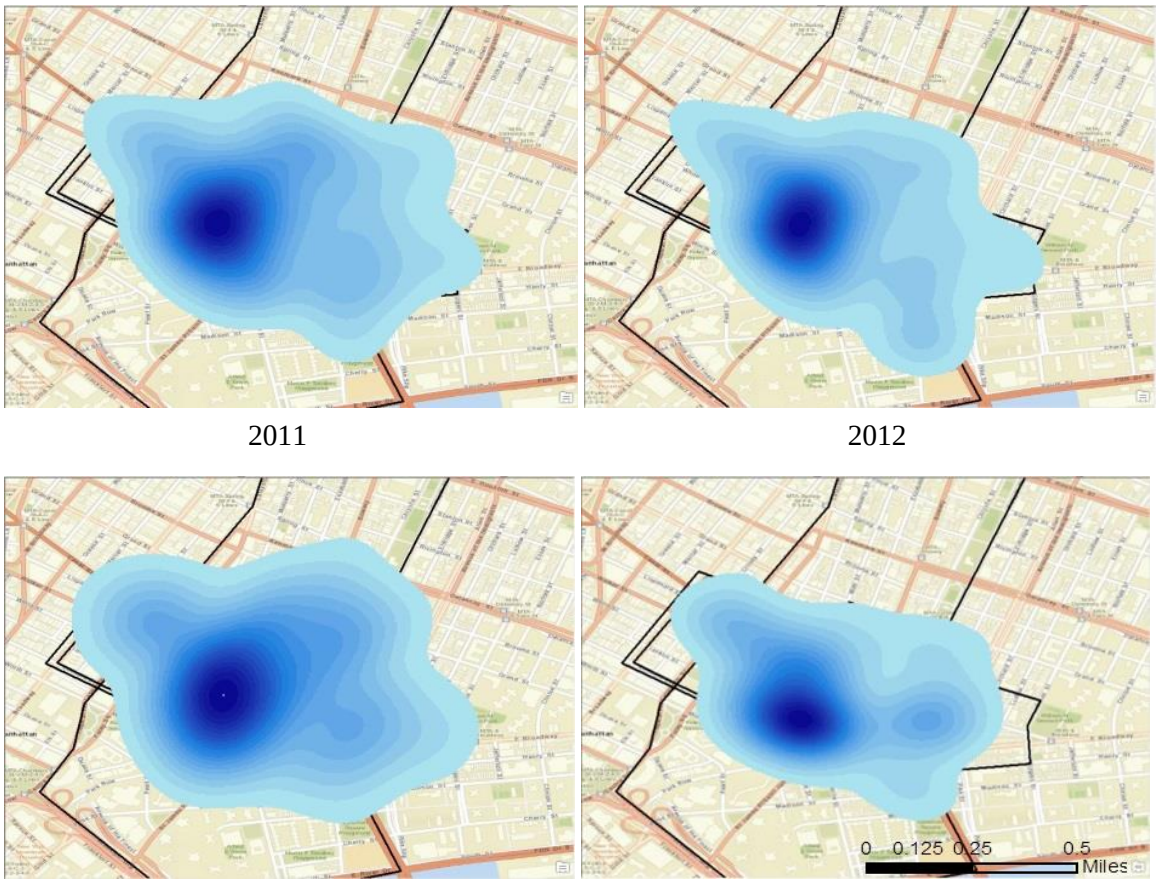
2004 - 2005 2006



2007 2008

2009 2010

Figure 26. Continued



2011 2012

2013 2014

Figure 26. Continued

According to the time series of geographic extents, the core area of Chinatown with the highest probability level has been consistently located in Canal Street, the Bowery, Worth Street, and Baxter Street with only minor differences through the years. However, Chinatown's south boundary shrank by one or two blocks in the first few years. The geographic extent of 2004 to 2005 reached south of the intersection of Saint James Place and Madison Street, but shrank about 200 meters in 2006 and 2007. Then the south edge moved a little north and located around the intersection of Saint James Place and James Street during 2008 to 2012. Even though the geographic extent expanded slightly south again across Madison Street in 2013, the whole trend of the south boundary was shrinking northward.

On the contrary, the north and east sides of the geographic extent have trended toward expansion. The west boundary was one or two blocks away from Essex Street from 2004 to 2010, and reached it in 2006. However, the geographic extents from 2011 to 2013 went beyond Essex Street. Similarly, the north side of the geographic extents expanded from 2004 to 2010, but gradually shrank from 2011 to 2012. Its north boundary even crossed Kenmare Street in 2010. The expansion peaked in 2013 and reached Sprint Street.

Chapter 5

Conclusions

5.1 Conclusions

The term *place*, having various functions in different settings, is most often used by geographers to define the context of a geographic area within which people conduct certain activities such as residing, shopping, and entertaining. The sense and identity of a place depends more on human and social attributes than on geometry. Thus, *place* usually has an ambiguous boundary. Online social network platforms have been developed and refined in recent decades, and have generated large volumes of geotagged data associated with time, location, and sometimes, users' perceptions. Indicating the spatio-temporal footprints of their contributors, these geotagged OSN data are a valuable source of knowledge about perceived understanding of places.

This empirical study proposes a probabilistic method to map the perceived geographic extent of colloquial place names associated with a place: select images tagged with the colloquial place names, conduct density estimation on the image set, and map the perceived geographic extent for a series of probability levels based on the appropriate density surface. For this study, a kernel density estimation tool with Gaussian kernel function in Geospatial Modeling Environment software was used to generate density surfaces. All six commonly used KDE bandwidth selecting algorithms supported in GME were tested and evaluated. These algorithms were plug-in estimator, smoothed cross validation, likelihood cross validation, two algorithms of biased cross validation, and least squares cross validation. Values of grid cells in the density surface were used as a measurement of probability to delimit the geographic extent of the case study places

based on a series of probability levels. Resulting density surfaces were reclassified with an interval of 5% of the maximum density value, and each new class was considered as a probability level. Using this method, the author generated perceived geographic extents of colloquial place names for three case studies using geotagged Flickr images.

When processing the data, preliminary boundaries for each case study were selected according to data distribution, physical barriers, and referral locations of each place in order to reduce computational cost and save time. Aiming at generating reliable and representative geographic extents, the author considered as many peoples' opinions as possible. Thus, if several images uploaded by the same person with the same coordinates and tags were uploaded within six hours, they were considered redundant and only one was retained as effective data. To filter images that indicate colloquial names of a place, image tags with complete names, aliases, name abbreviations, and name variations of each place were selected. The image subset tagged with colloquial place names of the Great Smoky Mountains National Park contains 24,225 images; the subset of Manhattan Chinatown contains 294,959 images; and there are 4,725 images for Taipei Shilin Night Market.

To select an appropriate method, different methods were used to validate resulting geographic extents according to the characteristics of selected case-study places. The Great Smoky Mountains National Park has an official name and boundary. Two indices, coverage rate and out rate, are defined to measure the consistency between perceived geographic extents and the official boundary. Higher coverage rates generally accompany higher out rates. Due to the natural characteristics of the national park, most of the perceived geographic extent is located at areas that are highly accessible, both physically

and digitally. The BCV/BCV2 result that shows a low probability level generally covers all areas that should be within the national park, but it may lose some details and also contain extra areas when compared to the results from SCV, plug-in, and CVh. However, it has a slightly larger coverage rate than LSCV while sharing similar out rates. Manhattan Chinatown and Taipei Shilin Night Market do not have an official boundary; instead, they have several well defined referral boundaries or location points. Manhattan Chinatown is home to the largest enclave of Chinese people in the Western Hemisphere, and has an intuitive characteristic from the surrounding areas, which is a concentration of Chinese stores and buildings with Chinese signs. Similarly, Taipei Shilin Night Market is composed of a traditional market and several surrounding streets full of roadside stands and small storefronts. Streets in the night market area usually do not have roadsides occupied by automobile or bicycle parking. Thus, the Street View service of Google Maps is used to visualize the sample areas and validate whether resulting geographic extents of these two case study places are appropriate. The geographic extent of Manhattan Chinatown from LSCV agrees with the verification results of Google Maps Street View at five out of six sample areas, similar to BCV/BCV2, and performs especially well at the southern border. The geographic extent result of Shilin Night Market using plug-in bandwidth selector is consistent with Google Maps Street View at four out of six sample areas, which is the best among the six bandwidth selector methods. Consequently, the density estimation surface using bandwidth selector LSCV was selected for Manhattan Chinatown, while the relatively appropriate result of Shilin Night Market uses plug-in bandwidth selector. The perceived geographic extents of the three places at different probability levels are mapped respectively.

Considering the fuzziness of colloquial terms used in daily life to refer to places, the distribution of images with some related tags were analyzed to see their relationships with the perceived geographic extent of places. Tags for “trail,” “cabin,” “cherokee” and “cadescove” were analyzed for the Great Smoky Mountains National Park while tags for Taiwanese snacks were explored for Shilin Night Market. The tags analyzed in this research were inspiring, but insufficient. Further studies are needed to more comprehensively understand colloquial terms to delimit colloquial place extents.

Spatio-temporal characteristics, patterns, and trends were analyzed for perceived geographic extents of Manhattan Chinatown and Taipei Shilin Night Market, which are formed by aggregation of man-made structures. Time-series of geographic extent snapshots taken at different times were generated using the aforementioned methods and parameters. According to the analyzed results, the south part of Shilin Night Market in the block bounded by Chengde Road, Jihe Road, Jiantan Road, and Wenlin Road, was torn down around 2011 and under construction after that, which agrees with Google Maps Street Views taken in 2012. By 2013 and 2014, this area could no longer be considered part of Shilin Night Market with high probability. From 2004 to 2010, Manhattan Chinatown seemed to be shrinking on the south side but expanding on the north and west sides. However, this trend seemed to stop in 2011. This study demonstrates some interesting findings of spatio-temporal analysis of the case studies, and brings a temporal perspective to the research of perceived geographic extents of places.

5.2 Limitations and Further Research

The concept of place is fundamentally vague in terms of identity and exact location. For this reason, perfect results are difficult to achieve even though appropriate methods and large amounts of data from a widely used data source are applied. The data used in this study were from Flickr, a popular online social network platform with a vast amount of image data. However, many studies indicate that the socioeconomic characteristics of OSN data contributors may affect the validity and accuracy of sociological research results [22]. This is an inherent weakness of geotagged OSN data with current social and economic conditions and limitations of technology. Race, age, education, income level, and employment of data contributors may impair how representative the perceived geographic extent of a colloquial place name is when it is used to serve a broader user community.

Furthermore, even though the total number of geotagged images within the preliminary study boundaries of each place is large, effective images tagged with colloquial place names are not abundant. The data subsets of certain time periods contain even fewer images. Insufficient image records used for density estimation may increase uncertainty and decrease reliability of the probability maps of perceived geographic extent of places.

To avoid the problems of image recognition and comprehensive semantic analysis, this research was simplified by assuming that the tags of an image reveal the location where the image was geotagged. In fact, though it is often the case, there are some images tagged with either the major content of the trip or content of the image, sometimes even something not at all related to the place. For example, an image taken on

the way to Shilin Night Market may be tagged by the person uploading the image as "shilinnightmarket" for personal image classification; people may take a geotagged picture of the beautiful scenic view when the Great Smoky Mountains National Park appears in horizon and tag it as "greatsmokymountains" despite it not being in the extent of the national park. Insufficient knowledge of the place may lead to errors when generating geographic extents. For example, it is unclear whether "Shilin Traditional Market" is an alias of Shilin Night Market or a different place. Other than image tags, titles, and descriptions, viewers' comments of an image may also supply information about the place. Such information should be taken into consideration in future research.

Even though this research examined many names, aliases, and variations of the places names being considered, some colloquial names may have been left out. Exploring additional related colloquial terms of a place, as well as the relationships of their distribution with the geographic extent of a place, is an initial attempt to delimit perceived geographic extent of colloquial place names, which inspires future studies.

LIST OF REFERENCES

1. Goodchild, Michael F, *Formalizing place in geographic information systems*, in *Communities, Neighborhoods, and Health*, Linda M. Burton, Stephen A. Matthews, ManChui Leung, Susan P. Kemp, and David T. Takeuchi, Editors. 2011, Springer. p. 21-33.
2. *Kernel Density (Spatial Analyst)*. Esri Kernel Density (Spatial Analyst) 2014 [cited 2014 April 5]; Available from: http://resources.arcgis.com/en/help/main/10.1/index.html#//009z0000000s000000 .
3. Jeffries, Adrianne. *The man behind Flickr on making the service 'awesome again'*. The Verge The man behind Flickr on making the service 'awesome again' 2013 [cited 2014 March 16]; Available from: http://www.theverge.com/2013/3/20/4121574/flickr-chief-markus-spiering-talks-photos-and-marissa-mayer.
4. *Flickr*. Wikipedia Flickr 2014 [cited 2014 June 6]; Available from: http://en.wikipedia.org/wiki/Flickr.
5. Wu, Suzanne and Michelle Boston. *Twitter and Privacy: Nearly one-in-five Tweets divulge user location through geotagging or metadata*. Twitter and Privacy: Nearly one-in-five Tweets divulge user location through geotagging or metadata 2013 [cited 2014 September]; Available from: https://pressroom.usc.edu/twitter-and-privacy-nearly-one-in-five-tweets-divulge-user-location-through-geotagging-or-metadata/.
6. Zickuhr, Kathryn. *Location-Based Services*. Location-Based Services 2013 [cited 2014; Available from: http://www.pewinternet.org/2013/09/12/location-based-services/.
7. Lee, Ryong, Shoko Wakamiya, and Kazutoshi Sumiya, *Discovery of unusual regional social activities using geo-tagged microblogs.* World Wide Web, 2011. **14**(4): p. 321-349.
8. Rattenbury, Tye, Nathaniel Good, and Mor Naaman. *Towards automatic extraction of event and place semantics from flickr tags*. in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. 2007. ACM.
9. Kent, Joshua D and Henry T Capello Jr, *Spatial patterns and demographic indicators of effective social media content during theHorsethief Canyon fire of 2012.* Cartography and Geographic Information Science, 2013. **40**(2): p. 78-89.
10. Kennedy, Lyndon, Mor Naaman, Shane Ahern, Rahul Nair, and Tye Rattenbury. *How flickr helps us make sense of the world: context and content in community-contributed media collections*. in *Proceedings of the 15th international conference on Multimedia*. 2007. ACM.
11. Jankowski, Piotr, Natalia Andrienko, Gennady Andrienko, and Slava Kisilevich, *Discovering landmark preferences and movement patterns from photo postings.* Transactions in GIS, 2010. **14**(6): p. 833-852.
12. Majid, Abdul, Ling Chen, Gencai Chen, Hamid Turab Mirza, Ibrar Hussain, and John Woodward, *A context-aware personalized travel recommendation system based on geotagged social media data mining.* International Journal of Geographical Information Science, 2013. **27**(4): p. 662-684.
13. Egenhofer, Max J and David M Mark, *Naive geography*. 1995: Springer.

14. Mark, David M and Max J Egenhofer. *Common-sense geography: foundations for intuitive geographic information systems*. in *GIS LIS-INTERNATIONAL CONFERENCE-*. 1996.
15. Mark, DM, MJ Egenhofer, and K Hornsby, *Formal models of commonsense geographic worlds: Report on the specialist meeting of research initiative 21*. Santa Barbara, California: National Center for Geographic Information and Analysis, 1997. **21**(97-2).
16. Gao, Song, Krzysztof Janowicz, Grant McKenzie, and Linna Li. *Towards Platial Joins and Buffers in Place-Based GIS*. in *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Computational Models of Place (COMP'2013)*. 2013.
17. Yao, Xiaobai and Bin Jiang, *Visualization of qualitative locations in geographic information systems.* Cartography and Geographic Information Science, 2005. **32**(4): p. 219-229.
18. Pasley, Robert C, Paul D Clough, and Mark Sanderson. *Geo-tagging for imprecise regions of different sizes*. in *Proceedings of the 4th ACM workshop on Geographical information retrieval*. 2007. ACM.
19. Jones, C. B., R. S. Purves, P. D. Clough, and H. Joho, *Modelling vague places with knowledge from the Web.* International Journal of Geographical Information Science, 2008. **22**(10): p. 1045-1065.
20. Goldberg, Daniel W, John P Wilson, and Craig A Knoblock, *Extracting geographic features from the internet to automatically build detailed regional gazetteers.* International Journal of Geographical Information Science, 2009. **23**(1): p. 93-128.
21. Goodchild, Michael F and Linda L Hill, *Introduction to digital gazetteer research.* International Journal of Geographical Information Science, 2008. **22**(10): p. 1039-1044.
22. Li, Linna, Michael F Goodchild, and Bo Xu, *Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr.* Cartography and Geographic Information Science, 2013. **40**(2): p. 61-77.
23. Ahern, Shane, Mor Naaman, Rahul Nair, and Jeannie Hui-I Yang. *World explorer: visualizing aggregate data from unstructured text in geo-referenced collections*. in *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*. 2007. ACM.
24. Keßler, Carsten, Patrick Maué, Jan Torben Heuer, and Thomas Bartoschek, *Bottom-up gazetteers: Learning from the implicit semantics of geotags,* in *GeoSpatial semantics*. 2009, Springer. p. 83-102.
25. Thomee, Bart and Adam Rae. *Uncovering locally characterizing regions within geotagged data*. in *Proceedings of the 22nd international conference on World Wide Web*. 2013. International World Wide Web Conferences Steering Committee.
26. Hollenstein, Livia and Ross Purves, *Exploring place through user-generated content: Using Flickr tags to describe city cores.* Journal of Spatial Information Science, 2010(1): p. 21-48.

27. Li, Linna and Michael F Goodchild. *Constructing places from spatial footprints*. in *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information*. 2012. ACM.
28. Gao, Song, Linna Li, Wenwen Li, Krzysztof Janowicz, and Yue Zhang, *Constructing gazetteers from volunteered Big Geo-Data based on Hadoop.* Computers, Environment and Urban Systems, 2014.
29. Montello, Daniel R., Michael F. Goodchild, Jonathon Gottsegen, and Peter Fohl, *Where's Downtown?: Behavioral Methods for Determining Referents of Vague Spatial Queries.* Spatial Cognition & Computation, 2003. **3**(2-3): p. 185-204.
30. Robinson, V. B., *A perspective on the fundamentals of fuzzy sets and their use in geographic information systems.* Transactions in GIS, 2003. **7**(1): p. 3-30.
31. Fisher, Peter, *Sorites paradox and vague geographies.* Fuzzy Sets and Systems, 2000. **113**(1): p. 7-18.
32. Liu, Yu, Yihong Yuan, Danqing Xiao, Yi Zhang, and Jiangquan Hu, *A point-set-based approximation for areal objects: A case study of representing localities.* Computers, Environment and Urban Systems, 2010. **34**(1): p. 28-39.
33. Grothe, Christian and Jochen Schaab, *Automated footprint generation from geotags with kernel density estimation and support vector machines.* Spatial Cognition & Computation, 2009. **9**(3): p. 195-211.
34. *Density Estimation*. Wikipedia Density Estimation 2014 2014.9]; Available from: http://en.wikipedia.org/wiki/Density_estimation.
35. Silverman, Bernard W and M Christopher Jones, *E. Fix and JL Hodges (1951): An important contribution to nonparametric discriminant analysis and density estimation: Commentary on Fix and Hodges (1951).* International Statistical Review/Revue Internationale de Statistique, 1989: p. 233-238.
36. Silverman, Bernard W, *Density estimation for statistics and data analysis*. Vol. 26. 1986: CRC press.
37. Danese, Maria, Maurizio Lazzari, and Beniamino Murgante, *Kernel density estimation methods for a geostatistical approach in seismic risk analysis: The case study of potenza hilltop town (Southern italy)*, in *Computational Science and Its Applications–ICCSA 2008*. 2008, Springer. p. 415-429.
38. Worswick, Christopher, *Adaptation and inequality: children of immigrants in Canadian schools.* Canadian Journal of Economics/Revue canadienne d'économique, 2004. **37**(1): p. 53-77.
39. Wolf, Christopher A and Daniel A Sumner, *Are farm size distributions bimodal? Evidence from kernel density estimates of dairy farm size distributions.* American Journal of Agricultural Economics, 2001. **83**(1): p. 77-88.
40. Portier, Kenneth, J Keith Tolson, and Stephen M Roberts, *Body weight distributions for risk assessment.* Risk analysis, 2007. **27**(1): p. 11-26.
41. Nakaya, Tomoki and Keiji Yano, *Visualising Crime Clusters in a Space‐time Cube: An Exploratory Data‐analysis Approach Using Space‐time Kernel Density Estimation and Scan Statistics.* Transactions in GIS, 2010. **14**(3): p. 223-239.
42. Shi, Xun, *Selection of bandwidth type and adjustment side in kernel density estimation over inhomogeneous backgrounds.* International Journal of Geographical Information Science, 2010. **24**(5): p. 643-660.

43. Zeng, Wei, Chi‐Wing Fu, Stefan Müller Arisona, and Huamin Qu. *Visualizing interchange patterns in massive movement data*. in *Computer Graphics Forum*. 2013. Wiley Online Library.
44. Sacks, Jerome and Donald Ylvisaker, *Asymptotically optimum kernels for density estimation at a point.* The Annals of Statistics, 1981: p. 334-346.
45. Jones, M Chris, James S Marron, and Simon J Sheather, *A brief survey of bandwidth selection for density estimation.* Journal of the American Statistical Association, 1996. **91**(433): p. 401-407.
46. Chiu, Shean-Tsong, *A comparative review of bandwidth selection for kernel density estimation.* Statistica Sinica, 1996. **6**(1): p. 129-145.
47. Hall, Peter, Simon J Sheather, MC Jones, and JS Marron, *On optimal data-based bandwidth selection in kernel density estimation.* Biometrika, 1991. **78**(2): p. 263-269.
48. Bowman, Adrian W, *An alternative method of cross-validation for the smoothing of density estimates.* Biometrika, 1984. **71**(2): p. 353-360.
49. Rudemo, Mats, *Empirical choice of histograms and kernel density estimators.* Scandinavian Journal of Statistics, 1982: p. 65-78.
50. Zambom, Adriano Zanin and Ronaldo Dias, *A review of kernel density estimation with applications to econometrics.* arXiv preprint arXiv:1212.2812, 2012.
51. Sain, Stephan R, Keith A Baggerly, and David W Scott, *Cross-validation of multivariate densities.* Journal of the American Statistical Association, 1994. **89**(427): p. 807-817.
52. Parzen, Emanuel, *On estimation of a probability density function and mode.* The annals of mathematical statistics, 1962: p. 1065-1076.
53. Jones, MC, James Stephen Marron, and Byeong U Park, *A simple root n bandwidth selector*. The Annals of Statistics, 1991: p. 1919-1932.
54. Duong, Tarn and Martin L Hazelton, *Cross‐validation Bandwidth Matrices for Multivariate Kernel Density Estimation.* Scandinavian Journal of Statistics, 2005. **32**(3): p. 485-506.
55. Wand, MP and MC Jones, *Multivariate plug-in bandwidth selection.* Computational Statistics, 1994. **9**(2): p. 97-116.
56. Gitzen, Robert A, Joshua J Millspaugh, and Brian J Kernohan, *Bandwidth selection for fixed-kernel analysis of animal utilization distributions.* Journal of Wildlife Management, 2006. **70**(5): p. 1334-1344.
57. Hägerstraand, Torsten, *What about people in regional science?* Papers in regional science, 1970. **24**(1): p. 7-24.
58. Nadi, Saeed and Mahmoud Reza Delavar. *Spatio-Temporal Modeling of Dynamic Phenomena in GIS*. in *ScanGIS*. 2003.
59. Langran, Gail. *Accessing spatiotemporal data in a temporal GIS*. in *Proc. Autocarto*. 1989.
60. Cheylan, Jean-Paul and Sylvie Lardon, *Towards a conceptual data model for the analysis of spatio-temporal processes: the example of the search for optimal grazing strategies*, in *Spatial Information Theory A Theoretical Basis for GIS*. 1993, Springer. p. 158-176.
61. Wolfson, Ouri, Bo Xu, Sam Chamberlain, and Liqin Jiang. *Moving objects databases: Issues and solutions*. in *Scientific and Statistical Database*

Management, 1998. Proceedings. Tenth International Conference on. 1998. IEEE.
62. Laube, Patrick, Stephan Imfeld, and Robert Weibel, *Discovering relative motion patterns in groups of moving point objects.* International Journal of Geographical Information Science, 2005. **19**(6): p. 639-668.
63. Erwig, Martin, Ralf Hartmut Gu, Markus Schneider, and Michalis Vazirgiannis, *Spatio-temporal data types: An approach to modeling and querying moving objects in databases.* GeoInformatica, 1999. **3**(3): p. 269-296.
64. Yu, Hongbo and Shih‐Lung Shaw, *Exploring potential human activities in physical and virtual spaces: a spatio‐temporal GIS approach.* International Journal of Geographical Information Science, 2008. **22**(4): p. 409-430.
65. Calabrese, Francesco, Francisco C Pereira, Giusy Di Lorenzo, Liang Liu, and Carlo Ratti, *The geography of taste: analyzing cell-phone mobility and social events*, in *Pervasive computing*. 2010, Springer. p. 22-37.
66. Versichele, Mathias, Tijs Neutens, Matthias Delafontaine, and Nico Van de Weghe, *The use of Bluetooth for analysing spatiotemporal dynamics of human movement at mass events: A case study of the Ghent Festivities.* Applied Geography, 2012. **32**(2): p. 208-220.
67. Yin, J., Z. E. Yin, H. D. Zhong, S. Y. Xu, X. M. Hu, J. Wang, and J. P. Wu, *Monitoring urban expansion and land use/land cover changes of Shanghai metropolitan area during the transitional economy (1979-2009) in China.* Environmental Monitoring and Assessment, 2011. **177**(1-4): p. 609-621.
68. Herold, Martin, Noah C Goldstein, and Keith C Clarke, *The spatiotemporal form of urban growth: measurement, analysis and modeling.* Remote sensing of Environment, 2003. **86**(3): p. 286-302.
69. Mallinis, G., D. Emmanoloudis, V. Giannakopoulos, F. Maris, and N. Koutsias, *Mapping and interpreting historical land cover/land use changes in a Natura 2000 site using earth observational data The case of Nestos delta, Greece.* Applied Geography, 2011. **31**(1): p. 312-320.
70. Xiubin, Li, *A review of the international researches on land use/land cover change [J]*. Acta Geographica Sinica, 1996. **6**.
71. Perez, Sarah. *Snapchat Is Now The #3 Social App Among Millennials*. Snapchat Is Now The #3 Social App Among Millennials 2014 Aug 11, 2014 Aug 15, 2014]; Available from: http://techcrunch.com/2014/08/11/snapchat-is-now-the-3-social-app-among-millennials/.
72. Liu, Sumang, *Online Social Network Friends and Spatio-temporal Proximity of Their Geotagged Photos – A Case Study of Flickr Data*. 2012, University of Tennessee.
73. Catt, Rev Dan. *100,000,000 geotagged photos (plus)*. 100,000,000 geotagged photos (plus) 2009 [cited 2014 June 10]; Available from: http://code.flickr.net/2009/02/04/100000000-geotagged-photos-plus/.
74. *Exchangeable image file format*. Wikipedia Exchangeable image file format 2014 June 25, 2014]; Available from: http://en.wikipedia.org/wiki/Exchangeable_image_file_format.
75. *Explore/Tag*. Flickr Explore/Tag 2014 Aug 15, 2014]; Available from: https://www.flickr.com/photos/tags/.

76. *Cherokee*. Wikipedia Cherokee 2014 [cited 2014/09/25; Available from: http://en.wikipedia.org/wiki/Cherokee.
77. Beyer, H.L. (2012). *Geospatial Modelling Environment* (Version 0.7.2.1). Available from: http://www.spatialecology.com/gme
78. Duong, Tarn. (2014). *Kernel smoothing* (Version 1.9.3). Available from: http://www.mvstat.net/tduong
79. *Great Smoky Mountains National Park*. Wikipedia Great Smoky Mountains National Park 2014 [cited 2014; Available from: http://en.wikipedia.org/wiki/Great_Smoky_Mountains_National_Park.
80. *Manhattan/Chinatown*. Wikitravel Manhattan/Chinatown 2014 [cited 2014 Apr. 10]; Available from: http://wikitravel.org/en/Manhattan/Chinatown.
81. *Cherokee - Great Smoky Mountains*. National Park Service Cherokee - Great Smoky Mountains 2014 [cited 2014/9/25; Available from: http://www.nps.gov/grsm/historyculture/cherokee.htm.
82. *Night markets in Taiwan*. Wikipedia Night markets in Taiwan 2014 [cited 2014/09/25; Available from: http://en.wikipedia.org/wiki/Night_markets_in_Taiwan.

VITA

Yuan Liu was born in Tangshan, China. She received her Bachelor of Engineering degree in Geographic Information Science and Technology from Tongji University in 2012, and then entered the Geography Department of the University of Tennessee, Knoxville the same year to pursue a Master's Degree. She has been studying space-time GIS and GIS-Transportation under the guidance of Dr. Shih-Lung Shaw. She will graduate in the Spring of 2015.