University of California
Santa Barbara

Extracting Computational Representations of Place with Social Sensing

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Geography

by

Song Gao

Committee in charge:

Professor Krzysztof W. Janowicz, Chair
Professor Emerita Helen Couclelis
Professor Emeritus Michael F. Goodchild

June 2017

The Dissertation of Song Gao is approved.

Professor Emerita Helen Couclelis

Professor Emeritus Michael F. Goodchild

Professor Krzysztof W. Janowicz, Committee Chair

June 2017

Extracting Computational Representations of Place with Social Sensing

Acknowledgements

It has been my great pleasure for studying and working in the Spatio-Temporal Knowledge Observatory (STKO) Lab at the UCSB Geography Department during the past five years. I have got a lot of very valuable mentoring and support from my advisor, committee members, professors, colleagues, friends, and family members. First, I would like to express my special appreciation and thanks to my advisor Dr. Krzysztof Janowicz. I have been so fortunate to develop my knowledge base and research skills under his great guidance. Dr. Janowicz is a very knowledgeable, kind, and inspiring mentor. I have witnessed his creative approach to preparing our lab members for succeeding in our doctoral milestones, research projects, publications, scholarship applications, and career preparations. I would also greatly thank Dr. Michael F. Goodchild who encouraged me to apply for UCSB graduate school back to year 2008. My dream came true in 2012! I am so lucky to have him in my committee and conduct research with him as a graduate assistant researcher. Dr. Goodchild is a very knowledgeable scientist and always gives me a lot of encouragement and brilliant suggestions for my research direction. I would also like to thank Dr. Helen Couclelis, who is a very knowledgeable scholar and spent a lot of time on discussions with me and reviewing my manuscripts. Also, I appreciate the support from the UCSB Center for Spatial Studies for employing me as the GIS HelpDesk. In addition, I would like to thank Jack & Laura Dangermond who financially provide fellowships and travel grants to me for support my conference travels and presentations. I would like to greatly thank STKO Lab members and collaborators for great discussions and teamwork spirit in research: Ben, Grant, Yingjie, Jay, Bo, Yiting, Rui, Gengchen, Li, Kang, Blake, Wenwen, and Linna. I much appreciate great support from many other people at UCSB. Last but not the least, I would like to sincerely thank my parents and my wife Tiange for supporting me throughout my PhD journey.

Curriculum Vitæ

Song Gao

Education

2017	Ph.D. in Cartography and Geographic Information Science, University of California, Santa Barbara, United States.
2012	M.S. in Cartography and Geographic Information Systems, Peking University, China.
2009	B.S. Honors in Geography, Beijing Normal University, China.

Employment

2013,2016,2017	Teaching Assistant, Department of Geography, UCSB.
2016	Teaching Assistant, Bren School of Environmental Science & Management, UCSB.
2016 Summer	Course Lecturer, Department of Geography, UCSB.
2016 Spring	Course Lecturer, Department of Earth Science, UCSB.
2013-2016	Graduate Student Researcher, Department of Geography, UCSB.
2012-2016	GIS HelpDesk at the Center for Spatial Studies, UCSB.
2015 Summer	Software Developer Intern at Esri Inc. Application Prototype Lab.
2014 Summer	Software Engineering Intern at Apple Inc Maps Team.
2009-2011	Teaching and Research Assistant, Institute of Remote Sensing and Geographic Information Systems, Peking University.

Fellowships, Scholarships, and Awards

2017	Austrian Academy of Sciences ($\ddot{O}$AW) Young Researcher Award in GIScience
2017	International Cartographic Association (ICA) Scholarship
2017	UCSB Graduate Division Dissertation Fellowship
2016	National Award for Outstanding Chinese PhD Students Study Abroad
2016	Finalist Prize in AAG Geographic Information Science and Systems Specialty Group Honors Student Paper Competition
2015	Outstanding Student Scholarship of Chinese-American Engineers and Scientists Association of Southern California
2015	Third-place Prize in the Esri Intern Weekend of Innovation Hackathon
2014	The Jack & Laura Dangermond Graduate Fellowship
2014	UCSB Geography Excellence in Research Award

2014	The Cartography and Geographic Information Society Doctoral Scholarship Award
2014	Second Prize in Learning Analytics and Knowledge Data Challenge
2013	Finalist Prize in the AAG Robert Raskin Mashup Mapping Competition
2013	AAG International Geographic Information Fund Student Award
2009-2012	First-Class Academic Scholarships at Peking University
2012	The Youth Scientist Competition Award at Peking University
2011	First Prize in the Esri China GIS Web Application Development Contest
2011	Best Paper Award in the National RS&GIS Chinese Graduate Students Academic Forum
2009	Outstanding Graduate Award in Beijing Normal University
2006	Excellent Student Leadership Award in Beijing Normal University
2006-2008	National Fellowships for Outstanding Undergraduates

Travel Grants

2013-2017	Jack & Laura Dangermond Student Travel Grants
2017	NSF Travel Grant for Mobility Science Workshop at OSU
2016	NSF Travel Grant for Mobility Science Workshop at UT-Austin
2016	NSF Travel Grant for 2016 ACM SIGSPATIAL Conference
2015	Travel Grant for Twenty-Year Anniversary of the International Early Career Scholars Vespucci Institute in GIScience
2014	NSF Travel Grant for the 10th Reasoning Web Summer School
2014	NSF Travel Grant for the International Workshop on Big Data and Urban Informatics
2013	NSF Travel Grant for 2013 ACM SIGSPATIAL Conference
2013	NSF Travel Grant for 2013 CyberGIS All-Hands Meeting

Publications

- **Book Chapters**

1. **Song Gao**. (2017) Big Geo-Data. In Laurie A. Schintler and Connie L. McNeely (Eds): *Encyclopedia of Big Data*, Springer.
2. **Song Gao**. (2017) Spatial Scientometrics.. In Laurie A. Schintler and Connie L. McNeely (Eds): *Encyclopedia of Big Data*, Springer.

3. **Song Gao**, Gengchen Mai. (2017) Mobile GIS and Location-Based Services. In Bo Huang, Thomas J. Cova, and Ming-Hsiang Tsou et al.(Eds): *Comprehensive Geographic Information Systems*, Elsevier. Oxford, UK.
4. Max J. Egenhofer, Keith C. Clarke, **Song Gao**, Teriitutea Quesnot, W. Randolph Franklin, Yuan May, Coleman David. (2016) Contributions of GIScience over the Past Twenty Years. In Harlan Onsrud & Werner Kuhn (Eds), *Advancing Geographic Information Science: The Past and Next Twenty Years*, GSDI Association Press, 9-34.
5. **Song Gao**. (2016) Spatiotemporal Autocorrelation Analysis for Pattern Recognition on Geospatial Big Data. in Harlan Onsrud & Werner Kuhn (Eds), *Advancing Geographic Information Science: The Past and Next Twenty Years*, GSDI Association Press, 273-280.
6. **Song Gao**, Ying Long. (2015) Finding Public Transportation Community Structure based on Large-Scale Smart Card Records in Beijing. In Ying Long & Zhenjiang Shen (Eds), *Geospatial Analysis to Support Urban Planning in Beijing*, Springer GeoJournal Library (116), 155-167.

- **Peer-reviewed Journal Articles**

7. **Song Gao**, Krzysztof Janowicz, Helen Couclelis. (2017 in press) Extracting Urban Functional Regions from Points of Interest and Human Activities on Location-based Social Networks. *Transactions in GIS*, DOI:10.1111/tgis.12289.
8. **Song Gao**, Krzysztof Janowicz, Daniel R. Montello, Yingjie Hu, Jiue-an Yang, Grant McKenzie, Yiting Ju, Li Gong, Benjamin Adams, Bo Yan. (2017) A Data-Synthesis-Driven Method for Detecting and Extracting Vague Cognitive Regions. *International Journal of Geographical Information Science*, 31(6): 1245-1271.
9. **Song Gao**, Linna Li, Wenwen Li, Krzysztof Janowicz, Yue Zhang. (2017) Constructing Gazetteers from Volunteered Big Geo-Data based on Hadoop. *Computers, Environment and Urban Systems*, DOI:10.1016/j.compenvurbsys.2014.02.004.
10. **Song Gao**. (2015) Spatio-Temporal Analytics for Exploring Human Mobility Patterns and Urban Dynamics in the Mobile Age. *Spatial Cognition & Computation*, 15(2), 86-114.
11. **Song Gao**, Yu Liu, Yaoli Wang, Xiujun Ma. (2013) Discovering Spatial Interaction Communities from Mobile Phone Data. *Transactions in GIS*, 17(3):463-481.
12. **Song Gao**, Yaoli Wang, Yong Gao, Yu Liu. (2013) Understanding Urban Traffic Flow Characteristics: A Rethinking of Betweenness Centrality. *Environment and Planning B: Planning and Design*, 40(1):135-153.
13. Blake Regalia, Grant McKenzie, **Song Gao**, Krzysztof Janowicz. (2016) Crowdsensing Smart Ambient Environments and Services. *Transactions in GIS*, 20(3):382-398.

14. Zhifeng Wu, Yanwei Chai, Anrong Dang, Jianhua Gong, **Song Gao**, Yang Yue, Dong Li, Lin Liu, Xingjian Liu, Yu Liu, Ying Long, Feng Lu, Chengzhi Qin, Hui Wang, Peng Wang, Wei Wang, Feng Zhen. (2015) Geography Interact with Big Data: Dialogue and Reflection. *Geographical Research*, 34(12):2207-2221.
15. Grant McKenzie, Krzysztof Janowicz, **Song Gao**, Li Gong. (2015) How Where Is When? On the Regional Variability and Resolution of Geosocial Temporal Signatures for Points Of Interest.*Computers, Environment and Urban Systems*, 54:336-346.
16. Yingjie Hu, **Song Gao**, Krzysztof Janowicz, Bailang Yu, Wenwen Li, Sathya Prasad. (2015) Extracting and Understanding Urban Areas of Interest Using Geotagged Photos. *Computers, Environment and Urban Systems*, 54, 240-254.
17. Yu Liu, Xi Liu, **Song Gao**, Li Gong, Chaogui Kang, Ye Zhi, Guanghua Chi, Li Shi. (2015) Social sensing: A new approach to understanding our socio-economic environments. *Annals of the Association of American Geographers*, 105(3), 512-530.
18. Wenwen Li, Miaomiao Song, Bin Zhou, Kai Cao, **Song Gao**. (2015) Performance improvement techniques for geospatial web services in a cyberinfrastructure environment - A case study with a disaster management portal. *Computers, Environment and Urban Systems*, 2015.04.003.
19. Yingjie Hu, Krzysztof Janowicz, Sathya Prasad, **Song Gao**. (2015) Metadata Topic Harmonization and Semantic Search for Linked-Data-Driven Geoportals: A Case Study Using ArcGIS Online. *Transactions in GIS*, 19(3), 398-416.
20. Grant McKenzie, Krzysztof Janowicz, **Song Gao**, Jiue-An Yang, Yingjie Hu.(2015) POI Pulse: A Multi-Granular, Semantic Signatures-Based Approach For The Interactive Visualization Of Big Geosocial Data. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 50(2), 71-85.
21. Yaoli Wang, **Song Gao**, Yu Liu. (2013) Exploration into Urban Street Closeness Centrality and its Application Methods: A case study of Qingdao. *Geographic Research*, 32(3):452-464.
22. Jiansheng Wu, Li Huang, Yu Liu, Jian Peng, Weifeng Li, **Song Gao**, Chaogui Kang. (2013) Traffic Flow Simulation Based on Call Detail Records. *Acta Geographica Sinica*, 67(12):1657-1665.
23. Yu Liu, Chaogui Kang, **Song Gao**, Yu Xiao, Yuan Tian. (2012) Understanding Intra-Urban Trip Patterns from Taxi Trajectory Data. *Journal of Geographical Systems*, 14(4):463-483.
24. Yu Liu, Fahui Wang, Yu Xiao, **Song Gao**. (2012) Urban Land Uses and Traffic Source-Sink Areas: Evidence from GPS-Enabled Taxi Data in Shanghai. *Landscape and Urban Planning*, 106(1):73-87.
25. Yu Liu, Yu Xiao, **Song Gao**, Chaogui Kang, Yaoli Wang. (2011) A Review of Human Mobility Research Based on Location Aware Devices.*Geography and Geo-Information Science*, 27(4):8-13.

26. Yong Gao, **Song Gao**, Runqiang Li, Yu Liu. (2010) A Semantic Geographical Knowledge Wiki System Mashed up with Google Maps. *Science in China Series E: Technological Sciences*, 53:52-60.

- **Peer-reviewed Articles in Conferences and Workshops**

27. Yuchen Zhang, **Song Gao**, Yiting Ju. Identifying Geographic Features through Map Interpretation: A Case Study of Craters Using Controlling Variables. In *Proceedings of the 28th International Cartographic Conference (ICC 2017)*, Washington, DC, USA, July 2-7, 2017.
28. **Song Gao**, Sathya Prasad. Employing Spatial Analysis in Indoor Positioning and Tracking Using Wi-Fi Access Point. In *Proceedings of 8th ACM SIGSPATIAL Workshop on Indoor Spatial Awareness (ISA 2016)*, Oct 31-Nov 3, 2016, Burlingame, California, USA. DOI: 10.1145/3005422.3005425.
29. Gengchen Mai, Krzysztof Janowicz, Yingjie Hu, **Song Gao**. ADCN: An Anisotropic Density-Based Clustering Algorithm. In *Proceedings of the 24th International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL 2016)*, Oct 31-Nov 3, 2016, Burlingame, California, USA.
30. **Song Gao**, Krzysztof Janowicz, Dingwen Zhang. (2016) Designing a Map Legend Ontology For Searching Map Content. In conjunction with the 15th International Semantic Web Conference, In *the Proceedings of the 7th Workshop on Ontology and Semantic Web Patterns (WOP 2016)*, Oct.17-21, Kobe, Japan.
31. **Song Gao**, Rui Zhu, Gengchen Mai. (2016) Identifying Local Spatiotemporal Autocorrelation Patterns of Taxi Pick-ups and Drop-offs. In *the Ninth International Conference on Geographic Information Science (GIScience'16)*, Sep 27-30, Montreal, Canada.
32. Krzysztof Janowicz, Yingjie Hu, Grant McKenzie, **Song Gao**, Blake Regalia, Gengchen Mai, Rui Zhu, Benjamin Adams, Kerry Taylor. (2016) Moon Landing or Safari? A Study of Systematic Errors and their Causes in Geographic Linked Data. In *the Ninth International Conference on Geographic Information Science (GIScience'16)*, Sep 27-30, Montreal, Canada.
33. **Song Gao**, Yuan Zeng. (2016) Where to Meet: A Context-Based Geoprocessing Framework to Find Optimal Spatiotemporal Interaction Corridor for Multiple Moving Objects. In *the International Workshop on Analysis of Movement Data (Conjunct with GIScience'16)*, Sep 27, Montreal, Canada.
34. Blake Regalia, Krzysztof Janowicz, **Song Gao**. (2016) VOLT: A Provenance-Producing, Transparent SPARQL Proxy for the On-Demand Computation of Linked Data and its Application to Spatiotemporally Dependent Data. *In the Proceedings of the 13th Extended Semantic Web Conference(ESWC'16)*, May 29-June 2, 2016, Anissaras, Crete, Greece.

35. Jae Hyun Lee, **Song Gao**, Konstadinos G Goulias (2016) Comparing the Origin-Destination Matrices from Travel Demand Model and Social Media Data. In *the Transportation Research Board 95th Annual Meeting (TRB 2016)*, Jan 10-14, Washington, D.C., USA.

36. Yingjie Hu, Krzysztof Janowicz, Sathya Prasad, **Song Gao**. (2015) Enabling Semantic Search and Knowledge Discovery for ArcGIS Online: A Linked-Data-Driven Approach. *17th AGILE Conference on Geographic Information Science*.

37. **Song Gao**, Jiuean Yang, Bo Yan, Yingjie Hu, Krzysztof Janowicz, Grant McKenzie. (2014) Detecting Origin-Destination Mobility Flows From Geotagged Tweets in Greater Los Angeles Area. In *the Eighth International Conference on Geographic Information Science (GIScience'14)*, Sep 23-26, 2014, Vienna, Austria.

38. **Song Gao**, Jiuean Yang, Krzysztof Janowicz, Yingjie Hu, Bo Yan. (2014) TrajAnalyst: Matching Data to Trajectory Analysis Modules via a Conceptual Framework. In *GIScience'14 Workshop on Analysis of Movement Data*, Sep 23-26, 2014, Vienna, Austria.

39. **Song Gao**, Yingjie Hu, Krzysztof Janowicz, Grant McKenzie. (2013) A Spatiotemporal Scientometrics Framework for Exploring the Citation Impact of Publications and Scientists. In *ACM SIGSPATIAL GIS 2013*, Nov. 5-8, Orlando, FL, USA.

40. Yingjie Hu, Grant McKenzie, Jiue-An Yang, **Song Gao**, Krzysztof Janowicz. (2014) A Linked-Data-Driven Web Portal for Learning Analytics: Data Enrichment, Interactive Visualization, and Knowledge Discovery. In *Proceedings of the Workshops at the 2014 International Conference on Learning Analytics and Knowledge*, Mar. 24-28, Indianapolis, IN, USA.

41. **Song Gao**, Krzysztof Janowicz, Grant McKenzie, Linna Li. (2013) Towards Platial Joins and Buffers in Place-Based GIS. In *ACM SIGSPATIAL COMP'13*, Nov. 5, Orlando, FL, USA.

42. **Song Gao**, Hao Yu, Yong Gao, Yingle Sun. (2010) A Design of RESTful Style Digital Gazetteer Service in the Cloud Computing Environment. In *Proceedings of 18th International Conference on Geoinformatics*. June 18-20, 2010, Beijing, China.

43. Chaogui Kang, **Song Gao**, Xing Lin, Yu Xiao, Yihong Yuan, Yu Liu, Xiujun Ma. (2010) Analyzing and Geo-visualizing Individual Human Mobility Patterns using Mobile Call Records. In *Proceedings of 18th International Conference on Geoinformatics*. June 18-20, Beijing, China.

44. **Song Gao**, Yu Liu, Xing Lin. (2009) Geo-Wiki: A Semantic Geographical Knowledge Wiki System. In *Proceedings of 6th International Symposium on Digital Earth (ISDE6)*. Sept. 8-12, Beijing, China.

45. **Song Gao**, Tiechun Wang, Yue Zhao. (2008) Integrating SWAT and Xin'anjiang Model Based on ArcGIS Desktop. In Proceedings of 5th SWAT International Conference. Oct. 15-19, Beijing, China.

Services

- GIS HelpDesk at the UCSB Center for Spatial Studies (spatial@ucsb)
- Student Board of Directors for *American Association of Geographers CyberInfrastructure Specialty Group*
- Student Board of Directors for *International Association of Chinese Professionals in Geographic Information Sciences*
- Peer Reviewer for *Proceedings of the National Academy of Sciences*
- Peer Reviewer for *PLOS ONE*
- Peer Reviewer for *Physica A: Statistical Mechanics and its Applications*
- Peer Reviewer for *International Journal of Geographical Information Science*
- Peer Reviewer for *Transactions in GIS*
- Peer Reviewer for *Computers, Environment and Urban Systems*
- Peer Reviewer for *International Journal of Digital Earth*
- Peer Reviewer for *Journal of GIScience & Remote Sensing*
- Peer Reviewer for *International Journal of Urban Sciences*
- Peer Reviewer for *Journal of Spatial Science*
- Peer Reviewer for *ISPRS International Journal of Geo-Information*
- Peer Reviewer for *Urban Geography*
- Peer Reviewer for *Journal of Land Use Science*
- Peer Reviewer for *IET Intelligent Transport Systems*
- Peer Reviewer for *Public Transport*
- Co-organizer for the Workshop on *Spatial Data on the Web (SDW16)*, in the GIScience 2016 conference.
- Co-organizer for the Symposium on *Human Dynamics Research: Visualization and Analytical Tool Development*, in the AAG Annual Meeting.
- Co-organizer for the Spatiotemporal Symposium on *Emerging Topics in Data-driven Geography*, in the AAG Annual Meeting.
- Student Assistant for GeoVocampSB2013, GeoVocampSB2014 Workshops, ISWC 2013, Geoinformatics 2009 Conferences
- President of Chinese Students and Scholars Association in the U.S. Southwestern region (SWCSSA, 501(c) non-profit organization)
- Voluntary teaching for K-12 education during the Geography Awareness Weeks.

Abstract

Extracting Computational Representations of Place with Social Sensing

by

Song Gao

Place-based GIS are at the forefront of GIScience research and characterized by textual descriptions, human conceptualizations as well as the spatial-semantic relationships among places. The concepts of places are difficult to handle in geographic information science and systems because of their intrinsic vagueness. They arise from the complex interaction of individuals, society, and the environment. The exact delineation of vague regions is challenging as their borders are vague and the membership within a region varies non-monotonically and as a function of context. Consequently, vague regions are difficult to handle computationally, e.g., in spatial analysis, cartography, geographic information retrieval, and GIS workflows in general. The emergence of big data brings new opportunities for us to understand the place semantics from large-scale volunteered geographic information and data streams, such as geotags, texts, activity streams, and GPS trajectories. The term "social sensing" describes such individual-level big geospatial data and the associated analysis methods. In this dissertation, I present a generalizable, data-driven framework that complements classical top-down approaches by extracting the representations of vague cognitive regions and function regions from bottom-up approaches using spatial statistics and machine learning techniques with various social sensing sources. I demonstrate how to derive crisp boundaries for cognitive and functional regions from points of interest data, and show how natural language processing techniques can enrich our understanding of places and form a foundation for the semantic characterization of place types and the generalization of regions.

This work makes contributions to the development of computational methodologies for extracting vague cognitive regions and functional regions using data-driven approaches as well as the novel semantic generalization processing technique.

Contents

Chapter 1

Introduction and Motivation

Chapter 1 provides a general introduction to this dissertation. First, it introduces the background and the motivation of this research with a literature review on the roles of space and place, types of regions in Geography and GIScience, and explains why the computational representation of place is challenging with regard to its intrinsic vagueness and the formalization need in computerized information systems and databases. And then, two theoretical frameworks of *Semantic Signatures* and *Social Sensing* to study the geographical and socioeconomic environment in the Big Data age from the GIScience literature are summarized. Moreover, the chapter poses three research questions and associated research hypothesis. Finally, the organization structure of this dissertation is outlined.

1.1 Space and Place in GIScience

Space and *Place* are two fundamental concepts in geography [1, 2]. In the past decades' development of GIS and spatial analysis methods, there exist rich studies about the role of space but only a few about the role of place from the GIScience perspec-

tive. Space is more abstract and generic while the notion of place is more tangible to humans. Agnew (2011) [2] suggested that the definition of place includes three aspects: (1) *location* – where an activity or object is located; (2) *locale* – the environment where everyday human activities take place; and (3) *sense of place* – the experiences offered by a place or a community to a group of people and their shared perceptions and conceptualization of a place. Goodchild (2011) discussed the formalization of the concept of place and addressed the relationship between the informal world of human discourse and the formal world of digitally represented geography [3]. He also reviewed the role of place in GIS, in formal gazetteers, in volunteered geographic information (VGI) and on defining context. The typical spatial perspective in GIS is mainly based on geometric reference systems that include coordinates, distances, topologies, and directions; while the alternative "platial" perspective is usually characterized by place names and textual descriptions as well as semantic relationships between places [4]. Couclelis (1992) discussed the commonality and differences among several related terms including *location, space, place, and region* [5]. She also compared the different notions of mathematical space, physical space (absolute space and relative space), socioeconomic space, behavior space and experiential space. In our study, a place is characterized by its semantics and human conceptualizations instead of just an abstract geometry in space.

A place name is usually taken to differentiate one place from another or as a mental handle for communication. These names, however, are not unique identifiers and there is also ambiguity to which specific space they exactly refer to. In order to locate place names on a map with precise coordinates and to support geographic information retrieve (GIR), efforts have been taken to convert place to space. One major mechanism is the use of gazetteers, which conventionally contains three core elements: *place names, feature types, and footprints* [6, 7]. In GIR, place plays an important role for interlinking other information. [8] analyzed the Excite Web query logs to investigate the extent and

variation of Web queries containing geographic terms, and found that geographically related queries formed nearly one fifth of all queries submitted to Excite in which the terms occurring most frequently being place names. Researchers have made significant efforts toward georeferencing place names and linguistic descriptions to the surface of the Earth, such as using ontologies of place [9], using fuzzy objects [10], using a qualitative spatial reasoning framework [11], using probability models in combination with uncertainty [12, 13], using kernel-density estimation[14, 15] and using description logics [16]. The *granularity* of locations or the *scale* of maps to which place descriptions can refer has also attracted researchers' interests [17, 18, 19]. To answer the question "where it is" or locating place names, the granularity is very important because places are most likely hierarchically organized. [19] developed two algorithms to identify the finest level of location granularity of a paragraph of place descriptions with spatial relations or without that. They also proposed a categorization of spatial relationships which include *topological relations*, *relative orientation relations*, *absolute orientation*, *qualitative distance relations*, and *quantitative distance relations*. In summary, state-of-the-art developments in computational representations of place for georeferencing, mapping and information search applications mainly rely on "points of interest" but not "areas of interest". More generally speaking, searching for "place of interest" is still challenging since the intrinsic vagueness of place.

The concepts of place in general are difficult to handle in geographic information science and systems. One gap lies between the vagueness of place in human mind and the formalization need for place-based representations in computerized information systems. One of the most challenging questions is what are those core attributes of place that should be stored in GIS? How to derive them computationally? The emergence of big data brings new opportunities for us to understand the place semantics from large-scale volunteered geographic information and data streams, such as georeferenced tags, images,

activity streams, and GPS trajectories. This provides a great opportunity to extract the computational representations of place for GIS using data-driven approaches, which will be the focus of this dissertation.

1.2 Types of Regions in Geography and GIS

Geographic region usually describes a spatial extent at or (near) the Earth surface characterized by the similarity or invariance of a set of properties with respect to their magnitudes [20, 21, 22]. Note that *space* and *place* are two alternative views that complement each other but focus on different aspects in geography and GIS as described in the previous section, while the concept of *region* generally refers to the internal similarity and external dissimilarity of locational properties.

There exist various taxonomies of region types in the literature. A popular taxonomy was proposed by Hartshorne (1969) [20], including *formal, functional*, and *general* regions. *Formal* regions are "distinguished by a uniformity of one or more characteristics", such as soil regions, dialect region, and political regions. *Functional* regions are "defined by the particular set of activities or interactions that occur within it". *General* regions "stand out in one's mind", which often refer to perceptual regions that only exist as a human conceptualization. Furthermore, Montello (2003) [21] split the *formal region* into *administrative* region and *thematic* region, and proposed another taxonomy of regions which consists of four types: *administrative, thematic, functional, and cognitive* regions. *Administrative* regions are formed by political entities which reflect property control and geopolitical power, such as census blocks, tracts, county, state and country. *Thematic* regions are categorized and mapped by one or multiple characteristics such as population. *Functional* regions are formed with the consideration of patterns of spatial interactions among separate locations. And finally *Cognitive regions* are formed by people's informal

perceptions, attitudes, and memories.

In daily dialogues, people usually communicate via the administrative divisions in a hierarchy [23] (e.g., *Santa Barbara County, California State*), or vernacular place names and cognitive regions (e.g. *downtown, southern California, or northern England*) instead of geometric coordinates when location needs to be explored or specified. Thus, these types of regions have to be formalized and transformed from human mind into computational models with computerized binary code. The concepts of *vague cognitive regions* in this dissertation align with existing literature and are difficult to handle in geographic information science and systems. They arise from the complex interaction of individuals, society, and the environment. The exact delineation of vague cognitive regions is challenging as their borders are vague and the membership within a region varies non-monotonically and as a function of context. Consequently, vague cognitive regions are difficult to handle computationally, e.g., in spatial analysis, cartography, geographic information retrieval, and GIS workflows in general. In typical GI systems, regions have usually been represented using sharp boundaries but vague cognitive regions have intrinsic fuzziness (i.e., admitted ambiguity) and partial membership instead of full membership for those locations within the region boundary [24]. In this research, we will investigate the computational representations (membership, boundary, and thematic characteristics) for vague cognitive regions and how to sharpen the indeterminate boundary based on point-based observations and spatial clustering methods for mapping and geovisualization purpose or the point-in-region spatial analysis in GIS.

The concept of *functional region* will be used throughout this dissertation. It has different definitions in the literature of regional science and urban geography studies. One popular definition is that *functional regions* are usually characterized by connections or interactions between different areas and locational entities and those connections or interactions could be labor, capital, human movement, transportation, commodity, ser-

vices, and so on [25, 26, 21]. One classic approach to delineating functional regions is based on the journey-to-work commuting flows [27]. The objective of delineating boundaries for different regions is maximizing the interactions within the same region while minimizing the interactions between different regions [25, 28]. There are four general classes of functional regionalization procedures in the literature: (1) hierarchical clustering [25]; (2) multistage aggregation [27, 29]; (3) central place aggregation [30]; and (4) modularity-based network approaches [31, 32].

In this dissertation, we use part of the Hartshorne's definition on *functional region* with emphasis on "supporting the particular set of activities and depending on the structure of the area" [20]. The functions of certain (sub)regions are originally defined in urban planning and then reshaped by actual needs and usages of human activities [33]. The activities include living, working, shopping, eating, recreation, and so on. We will use place venues that support specific types of human activities on the ground as a proxy to delineate functional regions with various co-location patterns of place types. The same type of points of interest (POIs) can be located in different land use types and may also support different functions. For example, *restaurants* are found in residential areas and in commercial areas, as well as in industrial areas. The main function of the place-type *universities* is education, but they also support sports activities, music shows, and so on. The function of a region can be grounded by the co-location pattern of POIs that may be considered as a good proxy.

1.3 Semantic Signatures and Social Sensing

1.3.1 Semantic Signature

Semantic signatures proposed by Janowicz (2012) [34] is a concept that is used to characterize different types and instances of places from three perspectives, namely, *spatial bands, temporal bands, and thematic bands* in analogy to the spectral signatures in remote sensing. In the *POIPulse* project, researchers extracted the semantic signatures from large scale point of interest (POI) data and developed an interactive information observatory of places in Los Angeles.

A semantic generalization of space-based maps into place-based maps based on mining the semantic signatures of places may bridge the gap between geometries and human cognition, and help the understanding of complex interactions between human and places. [35] analyzed the spatial-semantic interaction of POI types (e.g., bar, cafe, restaurant, and post office) based on OpenStreetMap data. They argued that such an approach can assist volunteered geographic information (VGI) [36] contributors in suggesting the types of new features, cleaning up existing data, and integrating data from different sources. However, there are several problems about the feature representations in OpenStreetMap data which might affect the spatial-semantic interaction patterns. First, most of the polygonal feature instances in OpenStreetMap are represented as *single-polygon* instead of *multi-polygon*. For example, the POI type *university* has been suggested by OpenStreetMap community tagging as "amenity=university". However, the individual elements of the amenity such as the buildings of the university should not be tagged as "building=university" which were actually common to see in OpenStreetMap. Thus, the *university* type might be classified as spatially clustered group that was in fact resulted from the feature annotation problems, i.e., *university buildings* tend to be spatially clustered but might not be true for the POI type of *university*. Second, the fuzzy feature

representations (e.g., downtown, Southern California) are still missing from VGI sources, such as on the OpenStreetMap. But human frequently communicate with such fuzzy places. Third, the hierarchical relationships were missing and thus it didn't directly support spatial reasoning. In fact, some of the thematic tags associated with the *subClassOf* property of a POI might also be true for the parent entity. For instance, considering the spatial containment relationship, the characteristic tags attached to a BBQ facility inside a beach park should be included to this park for further semantic similarity calculation. If the spatial patterns of POI types could be studied, the multi-polygon regions and their hierarchical relationships might be derived, e.g, the spatial footprint of a university could be derived from the spatial distribution of university buildings, parking lots, dinning and residential halls, and so forth. Fourth, the topological relationships are still missing for most features except the nodes and ways for street networks. The study of geometric properties and spatial relations unaffected by the continuous change of shape or size of places could be beneficial to the semantic map generalization process.

When users navigate digital maps, different hierarchical levels of places and regions associated with their spatial footprints and thematic (or functional) topic annotations should be dynamically shown on the place-based maps. Such semantically meaningful maps could be customized and beneficial for different groups of maps users in variety of domains. For example, the tourist maps would highlight the prominent landmarks, attractions and transportation accessibility. The related POI types, entities and their geometries should have higher weights than others during the semantic generalization process for particular purposes. A user might be interested in finding a locally characteristic hotel near popular *bar* regions in a city core, while he wants to make sure that the distance or time cost from the bar region (which spatially contains the potential target hotel) to the airport should be within certain acceptable threshold. But current space-based maps doesn't support such functionality since it usually don't have the vague

cognitive regions nor the urban functional zones. To solve this problem, we aim to develop a data-synthesis driven framework for automatically extracting the vague cognitive regions and the urban functional zones, as well as the semantically generalized regions in this research.

1.3.2 Social Sensing

We have entered an era of Big Data. Emails, blogs, photos, videos, and geographic datasets have been generated every day by possibly anyone with access to digital devices with necessary technologies. These data are also widely disseminated over the Internet through websites such as Twitter, Flickr, Youtube and OpenStreetMap, commonly referred to user-generated content (UGC). For the first time in history, people are able to collect vast volumes of data on various aspects of human life, from conversations about daily life and fluctuation of emotions, to discussions on major scientific breakthroughs and debate on critical political decisions. Big Data is "big" not only because it involves a huge volume of data, but also because of the velocity, the variety, the high-dimensionality, and inter-linkage characteristics of those datasets [37, 38]. The Web has lowered the barrier to produce, share, publicize, and access voluminous and various information about all sorts of things, linked to places. VGI [36] has gradually been taking the lead as one of the largest sources of geographic data. In addition to features with explicit locational information stored in geodatabases, places are also mentioned and discussed in social media, blogs, and news, and forums. This type of unstructured geographic information is abundant, with a great potential to benefit scientific research and decision making.

With the rapid development of information and communications technology (ICT), location-based services (LBS), location-based social networks (LBSN), the emergence of big geo-data brings new opportunities for researchers to understand our physical and

socioeconomic environments. Several types of big geo-data (such as mobile phone call detailed records, geotagged social media posts, taxi trajectories, smart card transactions, and LBSN check-in behaviors) are available to capture the spatiotemporal patterns of human activities and thus provide an alternative approach to uncovering land uses and exploring how cities function in a fine temporal resolution [39, 40, 41, 42, 32, 43, 44]. The term *social sensing* was introduced by Liu et al. (2015) [45] for such individual-level geospatial big data and the associated analysis methods. The word "sensing" suggests two natures of data. First, they can be viewed as the analogue and complement of remote sensing, as big data can capture well socioeconomic features while conventional remote sensing data do not have such privilege. Second, in social sensing data, each individual plays the role of a sensor. Social sensing data contain rich information about spatial interactions and environmental semantics, which go beyond the scope of traditional remote sensing data.

In the big data era, researchers have been investigating issues on synthesizing multi-sources, including data representativeness, scale quality, and new analytics methods to deal with social sensing data [45, 46]. As shown in Figure 1.1, analyzing large-scale datasets collected at the individual-level about human movement and activities, social ties, emotion and perception could help reveal the aggregated patterns about geographic environment such as the spatial interaction flows, land uses, and the place semantics. In this dissertation, the semantics of place and computational models of place in GIScience are the focus.

1.4 Research Questions and Research Hypotheses

In this dissertation, three research questions that help set the main scope and the 'boundary' of our place-based GIS research in this work are described in the following.

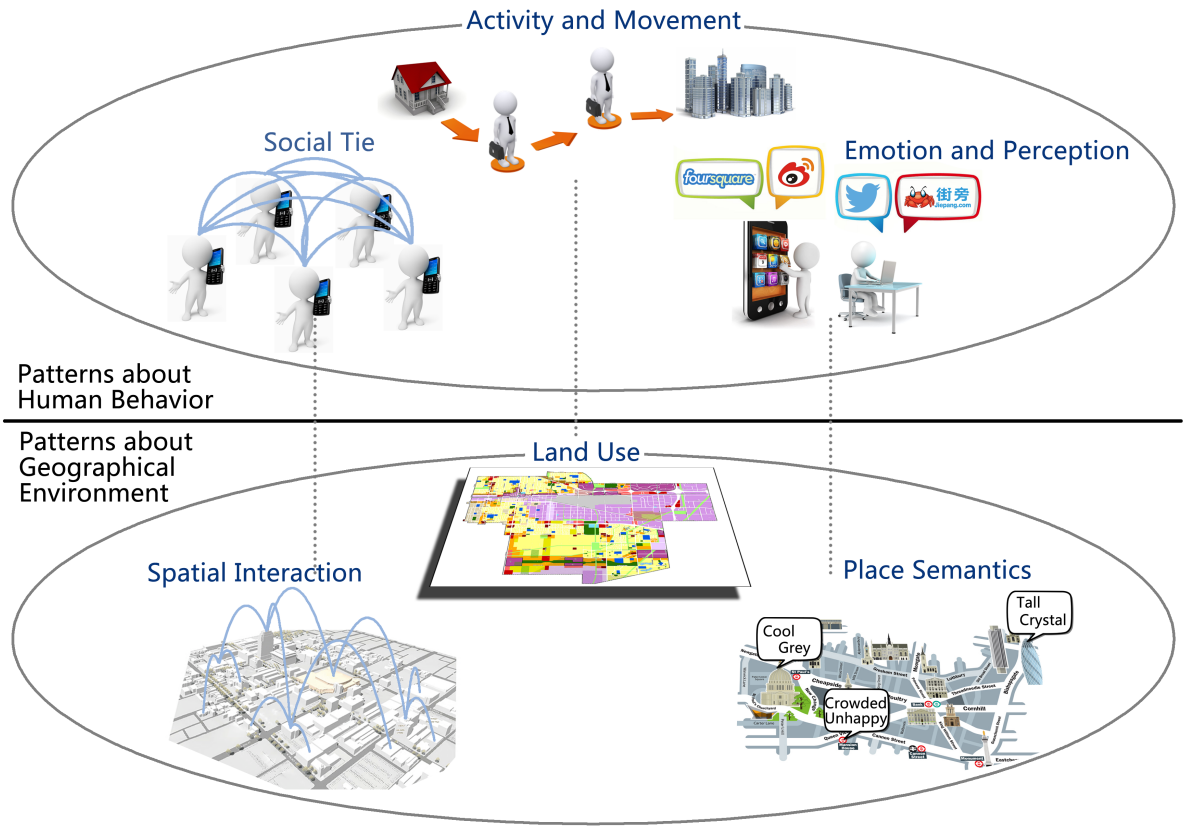


Figure 1.1: The social sensing diagram (Credit to Xi Liu).

RQ1: Is it possible to extract computational representations for vague cognitive regions from archival user-generated content such that they align with direct assessments of human participants using concordance measures?

RQ2: Is it possible to extract the co-location patterns of different place types and underlying characteristics that could be utilized to describe functional regions that support specific types of human activities?

RQ3: What kinds of analyses/operations on places can be employed for deriving semantically generalized regions in order to address the regional/cultural variability by broadening the thematic topics that form the functional regions?

In order to answer these research questions, the following stated research hypotheses

are the assumptions based on previous place-based studies or observations that require to be tested using well-designed scientific experiments.

Hypothesis 1: Labeled geo-referenced data extracted from user-generated content can be used to estimate the location and shape of vague cognitive regions as well as to reproduce their membership strength of individual locations to said regions as compared to direct assessments from human participants tests.

In order to test this hypothesis, we will conduct a data-synthesis-driven research compared with the traditional human participates survey in a vague cognitive region study with regard to the conceptual membership, the boundary, and the thematic characteristics of 'Southern California' and 'Northern California' (Chapter 2).

Hypothesis 2: The co-location distributions of different place types (e.g., restaurants, museums, parking lots) are indicative of functional regions that support specific human activities and those underlying patterns could be derived by applying topic modeling on POI data and corresponding user check-in behavior extracted from location-based social networks.

In order to test this hypothesis, we will study the spatial patterns of different POI types as well as their mixture patterns in ten most populated urban areas in the U.S., and then use the statistical topic modeling technique to learn their co-location patterns in order to derive urban functional topics and regions (Chapter 3).

Hypothesis 3: The thematic topics of regions can be generalized into broadened ones based on the semantic generalization operation on place-type hierarchy of points of interest and associated human activities on location-based social networks.

In order to test this hypothesis, we will develop a semantic generalization process based on topic modeling on LBSN points of interest data and their place-type hierarchal structure and apply in Los Angeles to search for semantically generalized regions with the target high-level place types. (Chapter 4).

1.5 Dissertation Structure

This dissertation is structured based on an accumulation of four individual but related articles including chapter 2, 3, and 4, and respectively, which will address the aforementioned research questions with hypothesis on the umbrella of place-based research in GIScience. The remainder of this dissertation is organized as follows.

Chapter 2 presents a data-synthesis-driven approach for studying vague cognitive regions. This work explains why vague cognitive regions and the concept of place in general are difficult to handle computationally, e.g., in spatial analysis, cartography, geographic information retrieval, and GIS workflows in general. In order to conquer this challenge, this work introduces a comprehensive framework that consists of deriving the membership score, hardening the crisp boundary, and extracting the thematic characteristics from natural language descriptions of places to computationally represent the vague cognitive regions. A case study has been conducted for the concepts of "Southern California" and "Northern California". The presented framework contains several core techniques for extracting computational representations (e.g., thematic topics, spatial footprints) of place, which will also be used in other chapters as well.

Chapter 3 presents a statistical framework to discover semantically meaningful topics of place types and further derive functional regions based on the co-occurrence patterns of place types and spatial clustering techniques in relation to urban planning. The framework applies latent Dirichlet allocation (LDA) topic modeling and incorporates user check-in activities on location-based social networks. A case study has been conducted using a large corpus of about 100,000 Foursquare venues and user check-in behavior in the ten most populated urban areas of the United States. This research shows that a region can support multiple functions but with different proportions, while the same type of functional region can span multiple geographically non-adjacent locations. Since each

region can be computationally modeled as a vector consisting of multinomial topic distributions, similar regions with regard to their thematic topic signatures can be identified. Compared to remote sensing images which mainly uncover the physical landscape of urban environments, the proposed popularity-based POI topic modeling approach can be seen as a complementary social sensing view on urban space based on human activities.

Chapter 4 presents a semantic generalization framework for converting point-based representation of place into region-based representations witch rich semantics. This research develops a new methodology that can take both spatial distributions of venues and the place-type hierarchical relationships into consideration to derive spatially and semantically coherent high-level generalized regions. While this research focuses on the theoretical contribution, a case study of extracting the semantic regions that relate to the *Beach*, *Shopping*, and *Asian Food* topics in Los Angeles has been conducted using the proposed semantic generalization methodology.

Finally, Chapter 5 summarizes this dissertation. Particularly, we answer the posed three research questions and state our research contributions. We also discuss the broader implications and limitations of this research, and present planned future work.

Chapter 2

A Data-Synthesis-Driven Method for Detecting and Extracting Vague Cognitive Regions

Chapter 2 presents a data-synthesis-driven approach for studying vague cognitive regions. This work explains why vague cognitive regions and the concept of place in general are difficult to handle computationally, e.g., in spatial analysis, cartography, geographic information retrieval, and GIS workflows in general. In order to conquer this challenge, this work introduces a comprehensive framework that consists of deriving the membership score, hardening the crisp boundary, and extracting the thematic characteristics from natural language descriptions of places to computationally represent the vague cognitive regions. A case study has been conducted for the concepts of "Southern California" and "Northern California". The presented framework contains several core techniques for extracting computational representations (e.g., thematic topics, spatial footprints) of place, which will also be used in the following chapters as well.

Peer-reviewed Publication	
Title	A Data-Synthesis-Driven Method for Detecting and Extracting Vague Cognitive Regions
Authors	Song Gao, Krzysztof Janowicz, Daniel R. Montello, Yingjie Hu, Jiue-an Yang, Grant McKenzie, Yiting Ju, Li Gong, Benjamin Adams, Bo Yan.
Venue	International Journal of Geographical Information Science
Editors	Yuan May
Publisher	Taylor and Francis
Pages	31(6): 1245-1271
Submit Date	15 August, 2016
Accepted Date	12 December, 2016
Publication Date	08 January, 2017
Copyright	Reprinted with permission from Taylor and Francis

Abstract: The concepts of cognitive regions and places are notoriously difficult to manage in geographic information science. They arise from the complex interaction of individuals, society, and the environment, their exact delineation is challenging as borders are vague, membership of places within a region varies non-monotonically, and homogeneity and regularity between raters cannot be assumed. Consequently, regions and places are difficult to handle computationally, e.g., in spatial analysis, cartography, geographic information retrieval, and GIS workflows in general. In a recent study, Montello and colleagues have devised a novel grid-based task in which participants rate the membership of individual cells to a given region and contrasted this new task to standard boundary-drawing. The authors used the regions of 'Northern' and 'Southern' California for their experiment on thematically influenced regions. They concluded that membership is about *attitude, not just latitude*. In this work, we reproduce their study by approaching it from a computational fourth paradigm perspective, i.e., by the synthesis of high volumes of heterogeneous data provided by various sources. We compare our results in identifying these regions to the results from Montello et al. and discuss differences and commonalities. At its core, however, this paper is not about 'Northern' and 'Southern' California but about the differences in study and task design, advantages and limitations of both approaches, and about the relation between conventional human

participants tests and the increasingly popular data-synthesis-driven research designs in GIScience.

2.1 Introduction

In its broadest sense, the concept of a *region* describes a bounded spatial extent characterized by the similarity or invariance of a set of properties. This includes the region defined by the property of *always facing away from the Earth*, i.e., the dark side of the moon, as well as regions defined by convention such as the thoracic anatomical region that encompasses the chest. Geographic information science is typically concerned with regions in geographic space that enable us to differentiate places inside of a region from those outside of it [21]. This includes *administrative regions* with fiat, institutional boundaries [47, 48] where the membership of places is exclusively determined by a binary containment relation [49], e.g., all counties in the state of California are completely and equally within California. Consequently, such regions do not have a graded structure; Santa Barbara County is not a lesser part of California than Los Angeles County. Interestingly, such administrative regions are generally the only type of regions that can accurately be described by the infinitely thin-line geometries that dominate GIS to date [50]. Instead, geographic regions typically have boundaries that are more or less vague.

Boundary vagueness occurs for one or more of a variety of specific reasons; [21] listed *measurement, temporal, multivariate, contested*, and *conceptual* vagueness. For example, the boundaries of the Kashmir region are disputed. Nonetheless, India, China, and Pakistan have their own national policies that exactly specify those boundaries [51]; this is contested vagueness. Other types of regions, such as *thematic regions*, are potentially multivariate. For example, the precise boundaries of ecological biomes can neither be acquired by measurementas this would require an infinity dense mesh of simultaneous

observations of all their properties, nor by theoretical considerationsas the concept of a biome is not specified to a degree that would enable the extraction of crisp boundaries [52, 21]. Consequently, thematic regions generally have two-dimensional boundaries and a graded structure. Places near the boundary may be less characteristic of the region than those in the center. In fact, the boundary zone between two regions is often of particular scientific interest, such as in studies of the upper timberline [48, 53]. As noted by [54], fiat boundaries are often projected onto physical space without a clear discontinuity of property values, e.g., in the case of valleys and their relation to mountains, or by introducing different kinds of barriers [55].

Another type of region arises from the complex interaction of individuals, society, and the environment. These *cognitive regions* [21] are informal regions that are also characterized by vague boundaries [52] and variable membership functions. Furthermore, the membership of places within a cognitive region may vary non-monotonically; membership strength does not necessarily decrease towards the boundaries, and in theory may vary up and down within the region. Cognitive regions can also vary in extent, shape, and location among groups and individuals, and can be highly specific to a local population; therefore neither homogeneity nor regularity can be assumed. Consequently, cognitive regions and places are difficult to handle computationally, e.g., in spatial analysis, cartography, geographic information retrieval, and GIS workflows in general. Interestingly, the spatial properties of cognitive regions are driven by individual and cultural beliefs about thematic properties to such a degree that metric, directional, or mereotopological [56] properties are relaxed or even ignored. For instance, as will be discussed later in this work, San Diego (SD) is perceived as less *Southern California* than is Los Angeles (LA), despite SD being more than 150 km to the south of LA. We call this a *platial effect* (rather than a *spatial effect*) in this paper, to highlight the fact that thematic and cultural aspects of the landscape can distort or relax spatial properties.

Understanding, assessing, and characterizing cognitive regions and their vague boundaries have been ongoing research activities for years. To give just a few examples, the *egg-yolk* theory proposes the use of concentric subregions to distinguish between an inner (certain) subregion, the *yolk*, and one or more outer, less certain region, called the *white*, those jointly form the *egg* [57]. In their *Where's downtown* paper, [58] reviewed three strategies to elicit an individual's representation of a region: by sketching the boundary, through a binary regular grid, and by selective binary trial-and-error sampling. Prior to this, [59] analyzed the cognition of neighborhood continuity and form by their residents. In their most recent work, [60] (MFP, for short) proposed a novel grid-based technique in which participants rated the membership of individual cells at a high resolution. This allows participants to express their beliefs about non-uniform region membership and vague boundaries in detail, and it puts few restraints on the spatial distribution of region membership patterns. For example, it allows membership variation to weaken and strengthen not non-linearly but even non-monotonically.

In the MFP study, 44 students from UCSB were presented with an outline map of California covered by a hexagonal tessellation of 90 cells (see Figure 2.1). The students were asked to rate each and every cell on a 1-7 scale, with 1 meaning very *Northern Californian*, 7 meaning very *Southern Californian*, and 4 meaning equally northern and southern Californian. The students were explicitly asked to base their judgment on not just cardinal directions but what people informally mean when they say *Northern California* and *Southern* California, i.e., to take feelings, lifestyles, and so forth into account. Those regions are widely known to locals and colloquially referred to often as *NorCal* and *SoCal*. Participants were asked to take their best guess for cells that they felt unsure about. Each of the 90 hexagons covers an area of approximately 4920 km^2. The tessellation was considered to be a (relatively) high-resolution grid by the study authors, considering that rating 90 cells for the entire state represents much higher

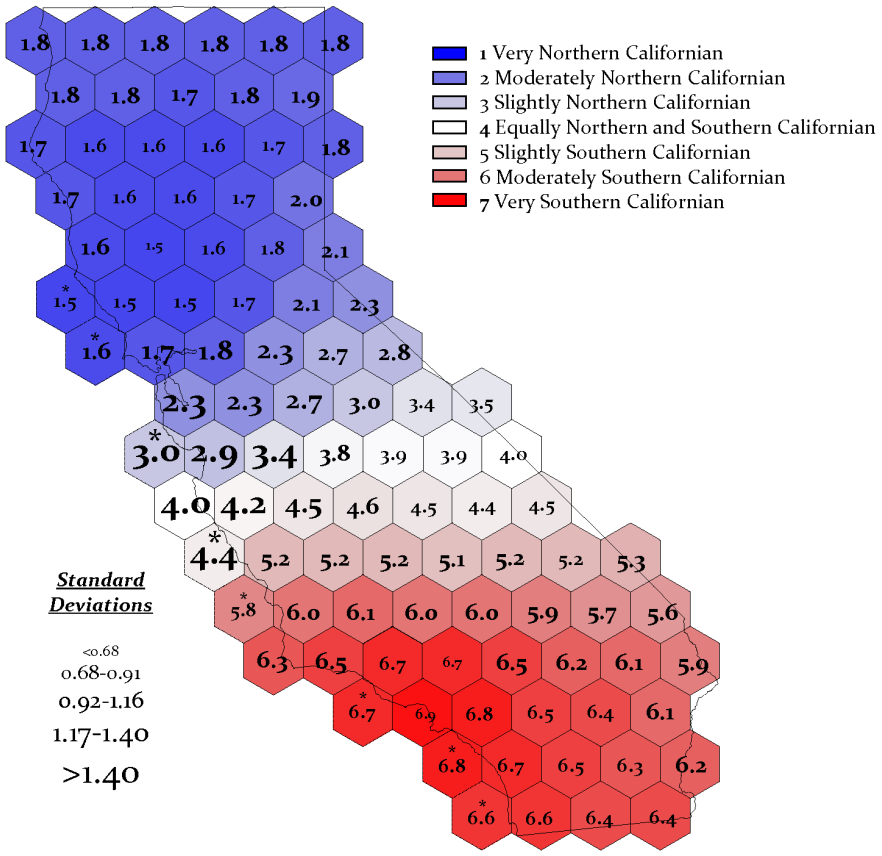


Figure 2.1: Means and standard deviations of ratings of *Northern* and *Southern* California based on [60]; dashed borders and asterisks indicate interpolated cells. Those cells marked with an asterisk were not part of the original study but have been added by us in order to fully cover the land area of California.

spatial resolution than is common when, e.g., participants divide the city into two regions of north and south, or three regions of north, south, and central. The statement also implies that rating 90 cells is close to the maximum that can be meaningfully asked of human participants. A detailed description of all studies, the Alberta control study, the study design, and the participants, can be found in the original MFP publication.

Figure 2.1 shows a slightly altered reproduction of the results of the MFP study. Cells with an asterisk are not part of the original study but have been added by us via linear interpolation in order to fully cover the land area of California and thereby collect

postings from these areas. Point-in-polygon analysis has been used to aggregate point observations and assign them to the hexagonal cells; more details are provided in Section 2.3.3. Interestingly, region membership is not monotonic, i.e., cells south of another cell may be less *Southern* California and cells to the north of another cell may be less *Northern* California. For instance, the hexagon containing the city of El Centro which borders Mexico, is considered to be less *Southern* California than the cell containing Santa Barbara which is to the northwest of Los Angeles. Similarly, on average, the cells in the San Francisco Bay Area are considered to be more *Northern* California than the most northern cells on the outline map. Furthermore, there is a clear coast-inland trend by which places at the coast are considered to be more *Northern* or *Southern* California than places to the east at the same latitude. This leads to a vague central boundary between *Northern* or *Southern* California that is of heterogeneous thickness, being thinner to the west and thicker to the east. Finally, the standard deviations across participant rankings are higher for northern than southern core areas of the respective regions. As we noted above, we call such phenomena *platial effects* to highlight the fact that thematic and cultural aspects of the environment can distort or relax spatial properties, including distance, direction, latitude & longitude, size, and so on.

The MFP research is a representative example of human participants studies carried out in cognitive and behavioral geography [61], spatial cognition, and geographic information science. It demonstrates a new methodology—grid-based interval-level rating—by applying it to an interesting geographic phenomenon. In this work, we reproduce their study, using the same California example and grid-based interval-level rating. However, we approach data acquisition and study design from a radically different angle, namely from a computational *fourth-paradigm* perspective [62], i.e., through the synthesis of high volumes of heterogeneous data provided by various online sources [63]. In this paper, we discuss the differences in study and task design between the two approaches, present the

results of this computational approach, compare them to the original human-participants study, and relate our results to the ongoing debate over the use of social media in GI-Science.

The research contributions of this work are as follows:

- We propose an automatable (and thus scalable) framework which can synthesize multiple heterogeneous datasets from different sources to study vague cognitive regions.
- We compare the results from our *data-synthesis-driven* approach with those from a human-participants experiment, and discuss the pros and cons of the two approaches.
- In addition to the grid-based membership study, we also approximate crisp boundaries for the cognitive regions and explore their underlying thematic topics.
- We explore the use of topic modeling to gain further insights into how vague cognitive regions can be represented and delineated.

To date, the literature on data-synthesis-driven approaches to quantitative geographic analysis is very sparse. Online social media records represent a form of secondary archival data[1] [64], which is not particularly novel in itself. However, the automated filtering and analysis of such data, particularly to analyze cognitive concepts such as cognitive regions, is novel. We introduce the term *data-synthesis-driven* here as an alternative to the popular notion of *data-intensive* science for two reasons. First, the term *data-intensive* could be misunderstood as implying that the MFP work (or any other work

[1]From a broad research-methods perspective, such as that in Montello and Sutton (2013), social media records are not data until they are coded for contentthey are sources of data. In this paper, however, we follow the convention of the data-synthesis (big data) research community and refer to the collected records as data.

along the same lines) is not heavily based on data merely on grounds of the amount of data used. Second, we believe that the real and radical novelty of the fourth paradigm lies in the way data are acquired and handled, and in the role they play in asking certain types of scientific questions [63].

The remainder of this chapter is structured as follows. In Section 2.2, we discuss existing studies related to the present work. Next, Section 2.3 presents the design of our study, the required data collection, changes that had to be made to the data from the MFP study for comparison to our work, as well as the processing workflow and methods employed. Section 2.4 presents our results, and compares them to the results of the original MFP study. Section 2.5 discusses the broader impact of this research, and finally, Section 2.6 summarizes this work and gives an outlook for future research and technology directions.

2.2 Related Work

Cognitive places are examples of vague places that are also referred to as vernacular places [65, 66], at least when they are concepts shared by groups of people and not idiosyncratic to one person. While typically not included in authoritative gazetteers, vague places are frequently used in our everyday dialogue, such as when describing locations and asking directions. The intrinsic nature of a vague place is its boundary vagueness, as seen in examples such as *downtown*. Fuzzy-set-based methods have been widely used to extract the indeterminate boundaries of vague places in GIScience and spatial cognition [24, 58]. Given their indispensable role in human thought and culture, researchers have conducted studies to acquire a better understanding of vague places. Based on a human-participants study, [67] discussed the user needs and implications for vague place modeling. [68] harvested Web pages related to particular vague places in the

UK, and identified their approximate boundaries based on the geo-referenced locations in the pages. [69] proposed a point-set-based region model to approximate vague areal objects and conducted a cognitive experiment to investigate the borders of *South China*. [70] collected geotagged *Flickr* data for studying vague places, and constructed spatial boundaries using kernel density estimations. Recently, [71] presented a computational framework which employed natural language processing and machine learning techniques to derive the geographic footprint of the cognitive region *historic center of Vienna* based on the TripAdvisor website and OpenStreetMap entries, and validated the results by comparing them with a historical map of the city.

Social media provides an alternative data source for studying the interactions between people and places. While often being criticized for concerns of representativeness [72, 73], social media data nevertheless reflect the behavior of millions of users throughout the world, and therefore have value [74, 75, 76]. The wide availability of social media has greatly enriched traditional volunteered geographic information (VGI) approaches, such as OpenStreetMap and Wikimapia [77, 78, 79]. Unlike these traditional VGI platforms which focus on online collaborative mapping, geotagged social media data reflect the spatial footprints of people in the real world, and therefore can be employed for studying human behavior. For example, [80] demonstrated a strong positive correlation between traffic flow in the greater Los Angeles area and geotagged Twitter data. Using geotagged Flickr data, [81] developed a bottom-up approach to construct place entities that can help enrich official gazetteers. Also based on Flickr data, [82] extracted urban areas of interest (AOI) for six different cities in the past ten years, and analyzed the spatiotemporal dynamics of the extracted AOI.

2.3 Study Design

In this section, we describe the datasets used, our workflow and methods, pre-processing steps, and the three analysis tasks we performed in order to reproduce the MFP study with a data-synthesis-driven approach.

2.3.1 Data Collection

In contrast to the MFP study, we did not collect data by interacting with selected participants but by automatically observing the use of terms in existing data. To do so, we filtered our data with two sets of keywords. The first grouped the keywords "SoCal", "South California", and "Southern California" into one set, which we call *SoCal*, and the keywords "NorCal", "North California", and "Northern California" into a second set, which we refer to as *NorCal*.

With these two sets of keywords, we collected data from five sources: *Flickr, Instagram, Twitter, Wikipedia,* and *TravelBlog.org*. *Flickr* is a photo sharing portal that stores millions of tagged and geo-referenced pictures. We believe that *Flickr* represents a more tourism-oriented view of California than the other social media sources. *Twitter* and *Instagram* are examples of online social media networks that are popular among both residents and visitors to California. These sources capture daily activities, news, visited points of interest, and so forth. We retrieved geo-referenceable entries from *TravelBlog* that provides trajectory-style data and capture outdoor locations well, including parks. While all these sources provide data and views from individuals, *Wikipedia* provides a consensus truth (broader agreement) about *NorCal* and *SoCal*, as articles containing these terms are the results of edits done by a larger community. As shown in Table 2.1, we collected 344,475 data entries/postings (203,713 for *SoCal* and 140,762 for *NorCal*) within the contiguous California State boundary (without islands). As for social media

Table 2.1: Data collection counts from five different sources

Source	SoCal Group	NorCal Group	Total
Flickr	22,132	19,706	41,838
Instagram	169,648	116,984	286,632
Twitter	10,376	3,294	13,670
Travel Blogs	107	78	185
Wikipedia	1,450	700	2,150
SUM	203,713	140,762	344,475

postings, the location mentions in the content might be different from where they were generated. We discuss the distinction between *the said place* and *the locale* further in Section 2.5. However, we only selected those georeferenced (Twitter and Instagram) postings which were generated from mobile devices and provided the users' GPS coordinates; therefore, we can be confident from where the postings were actually generated. More detailed information about each source is presented below.

(a) **Flickr**: We extracted 41,838 postings contributed by 1,338 unique users that contain the keywords (tags) mentioned above for the *SoCal* group and the *NorCal* group from 99.3 million Flickr photos taken from 2004 until 2014 and released by Yahoo Labs [83]. The photos are either geo-referenced manually or by the built-in positioning technologies in the mobile device or the camera. The location could either be the place where a photo was taken or the location of an object in the photo. Automatic recording by a GPS receiver is always the former case, while the human manually georeferenced photos could be either way. The photo metadata includes photo ID, title, description, textual tags, time when a photo was taken and uploaded, latitude and longitude etc.

(b) **Instagram**: Instagram is an online mobile photo (and video)-sharing social networking service. According to a Pew Research report [84], Instagram has grown in popularity with more than half (53%) of internet-using young adults (age 18 to 29) using

the service. The content shared on Instagram is georeferenced by built-in positioning technologies on mobile devices or by manually selecting the location from the preloaded Facebook gazetteer. We retrieved a total of 286,632 geo-referenced and *SoCal* & *NorCal* keyword-filtered postings by 79,371 unique users between 2011 and 2015. The metadata of the Instagram media includes the media ID, user ID, latitude, longitude and textual captions.

(c) **Twitter**: Combining the Twitter Streaming API and Search API, we retrieved a total of 13,670 geo-referenced and *SoCal* & *NorCal* keyword-filtered tweets posted by 8,482 unique users during the winter of 2014–2015. When posted from an Android or iOS application, the locations of the tweets were geo-referenced by the built-in positioning technologies if the user opted in to the location service. The metadata of our tweet collection includes the tweet ID, user ID, latitude, longitude, and the textual content of each tweet.

(d) **TravelBlogs.org**: Over 440,000 raw blog entries were downloaded. Each place name was matched to an entry in the GeoNames gazetteer, providing a latitude and longitude. More detailed information about the geoparsing procedure for these unstructured, natural language documents can be found in [75]. We extracted 185 travel blogs which mentioned at least one of the keywords from the *SoCal* or *NorCal* sets. Because this is such a small number of travel blogs extracted, we combined them with the Wikipedia articles discussed below for further analysis.

(e) **Wikipedia**: We extracted 2,150 articles which contained the *SoCal* or *NorCal* group of keywords, and inside the California State boundary. If the articles were not directly geo-referenced, information from DBpedia was applied for geo-referencing [85, 75].

Among the five selected sources, the number of data entries vary substantially, due to API access restrictions, limited geo-referenced content, and so forth. We discuss the

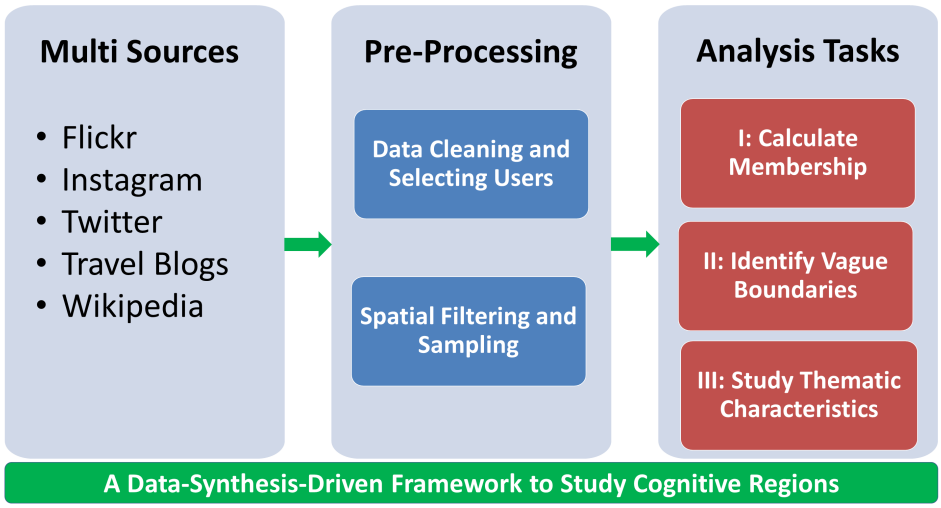


Figure 2.2: The processing framework for studying cognitive regions using a data-synthesis-driven approach.

differences among these data sources in the discussion Section.

2.3.2 Workflow & Methods

The overall analysis procedure for our data-synthesis-driven approach involves (a) extracting data that are frequently tagged with *SoCal* and *NorCal* in social media postings, (b) examining the spatial patterns of these data, and (c) defining the variability and boundary vagueness of *SoCal* and *NorCal*. The most challenging part of this approach is to select a large number of good quality data that meet our criteria from the raw data. We design a standard processing workflow (see Figure 2.2) to calculate the membership scores for the hexagon-cell-based representation of cognitive regions (**Task I**), to identify and characterize the vague boundaries (**Task II**), and to extract prominent thematic topics tied to cognitive regions from the natural language descriptions (**Task III**).

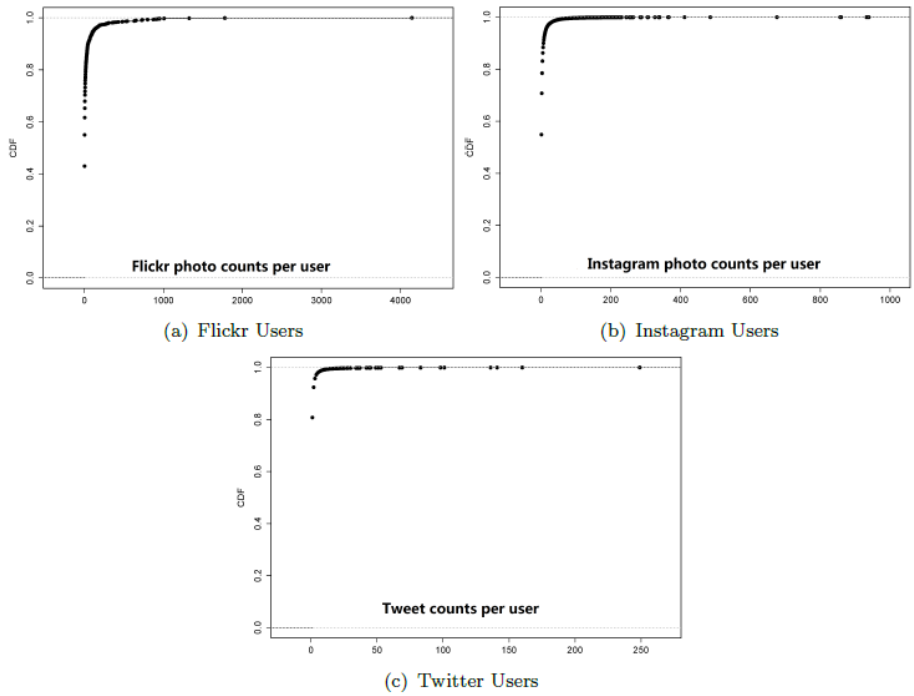
Pre-processing Step 1. Cleaning Data and Selecting Appropriate Users and Contributed Entries

The information shared on social media and online crowdsourcing platforms usually follows a power-law distribution [86, 87], which means most of the postings are contributed by a few users. In our case, we do not want the resulting patterns to be dominated by the most active users. In order to reduce such effects, we limited the number of entries contributed by each user. First, we calculated a cumulative probability distribution function (CDF) for the posting counts per user (Figure 2.3) to decide on an appropriate threshold. From there, we find the number of entries at which the majority (i.e., 90% was chosen) of the users posted.

Taking Flickr photo postings as an example, the 90th percentile threshold value is 41 photos for the *SoCal* group and 40 for the *NorCal* group. This means that about 90% of the users posted no more than 41 photos for *SoCal* and 40 photos for *NorCal*. For users who contributed less than or equal to the percentile threshold p, all photos are kept. For users who contributed more photos, we randomly selected photos up to the threshold.

Pre-processing Step 2. Spatial Clustering of Entries and Sampling for Each User

Second, we limited the number of posts by the same users, to avoid having a few users dominate the overall patterns of a specified local region. For a given local region (within a certain search radius), we value contributions from multiple users because they represent a consensus among the general public for this region, which is similar to a human-participants test. Therefore, we spatially filtered out repeated postings from a single user within a search radius of 100 meters so that we retained only one post per user in this local region.



(a) Flickr Users (b) Instagram Users

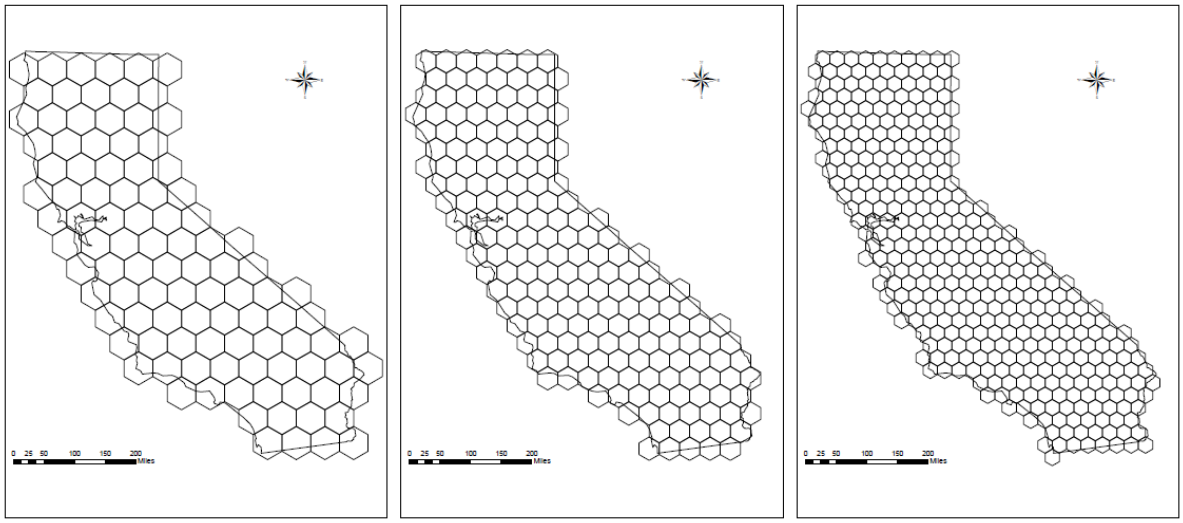
(c) Twitter Users

Figure 2.3: The cumulative distribution functions (CDF) of entries contributed per user in Flickr, Instagram and Twitter.

2.3.3 Analyzing Selected Data

Task I. Calculating Membership Scores

After the data filtering and clustering performed during the pre-processing steps, we applied point-in-polygon analysis to aggregate point observations to three different hexagonal tessellations at three different resolutions (Figure 2.4). The first level of hexagonal tessellation has the same spatial resolution as used in the MFP study, with each hexagon covering about 4920 km^2. The second-level and the third-level hexagons are at higher resolution, covering a half (2460 km^2) and a quarter (1230 km^2) of the first-level area in each cell, respectively. Varying the spatial resolution in a data-synthesis-driven approach is easy to do, while increasing the resolution is difficult in a traditional human-participants survey, since participants can be overwhelmed by a large number of cells to rate.



(a) Resolution 1 (b) Resolution 2 (c) Resolution 3

Figure 2.4: The hexagon-based tessellations at different spatial resolutions.

After spatially joining the point observations associated with the *SoCal* and *NorCal* group keywords to the hexagon grids, we obtained two occurrence counts in each cell for a given data source. Let S_i^j denote the occurrence counts of *SoCal* mentions and N_i^j as the *NorCal* mentions, where i is the hexagon ID at one of three resolution levels of tessellation grids, and j represents the data sources: Flickr, Instagram, Twitter or Travel Blogs & Wikipedia. Cells with a sum count of N_i^j and S_i^j less than 10 were considered as providing insufficient observations, and were therefore filtered out before the quantitative computation and comparison steps. We created two simple measures to derive the membership value of cells (Equation 2.1 and 2.2)

$$M1_i = S_i^j - N_i^j \tag{2.1}$$

$$M2_i = S_i^j / S_{max}^j - N_i^j / N_{max}^j \tag{2.2}$$

where S^j_{max} represents the maximum occurrence counts of *SoCal* mentions per cell across the whole study area for a given data source j; and N^j_{max} is the maximum for *NorCal* mentions. The purpose of $M1$ is to quantify the absolute occurrence differences per cell while $M2$ measures a normalized ratio difference. Next, the cells are classified and rated from 1 to 7 for each data source based on ranking percentiles. From these, the spatial distribution maps of cell memberships for each data source were derived.

We also computed the mean values $M1^{mean}_i$ and $M2^{mean}_i$, as well as standard deviations $M1^{sd}_i$ and $M2^{sd}_i$ for each cell for both measures across all data sources. For both $M1$ and $M2$, the higher the value of the means, the more likely the cell is rated as being a *SoCal* (or *NorCal*) cell.

In order to determine the inter-source agreement of different data sources among the cells, we took each data source as one *rater layer* and index the cells that had sufficient observation counts in all layers with the ranks (1∼k) sorted by their occurrence counts. This results in 4 sets of tuples [cell-id, rank (1∼k)]. For instance, a cell with the ID 19 may have a value of 2 in Twitter (*moderately NorCal*) but a value of 1 in Instagram (*strongly NorCal*). We use *Kendalls Coefficient of Concordance (W)* [88] to assess the agreement among these different rater layers.

To do this, assume there are m sources rating n subjects in rank order from 1 to k. Let $r_{i,j}$ represents the rating a source j gives to a subject i. Let R_i be the total ranks given to the subject i (i.e., $\sum_{j=1}^{m} r_{i,j}$) and $\bar{R}$ be the mean of R_i, the sum of the squared deviations S can be calculated as by Equation 2.3. Then the Kendall's W is defined as given by Equation 2.4.

$$S = \sum_{i=1}^{n} (R_i - \bar{R})^2 \tag{2.3}$$

$$W = \frac{12S}{m^2(n^3 - n)} \tag{2.4}$$

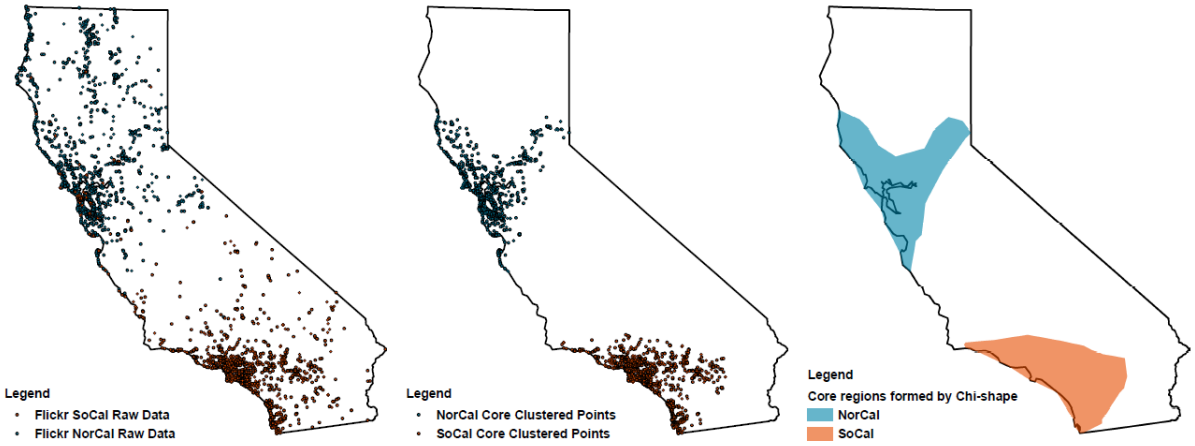
Task II. Extracting Continuous Boundaries of Cognitive Regions

Task I employed a discrete approach based on a hexagonal grid to calculate the membership score of each individual cell. In the second task, we aimed at determining the core regions of *NorCal* and *SoCal* using a continuous approach by approximating the boundaries of these two cognitive regions. While perceived borders of vague regions often vary among individuals [21], our goal here is to extract the core regions which are agreed upon by most people.

We use three social media sources, namely Flickr, Twitter, and Instagram, to identify the core regions. Using multiple sources helps ensure that the identified regions are not artifacts of one particular data source. In addition, it also reduces the potential bias introduced by the different user demographics of different social media platforms.

We applied a two-step workflow to extract the approximate regional boundaries for *NorCal* and *SoCal*. In step 1, we performed spatial clustering and identified point clusters based on geo-referenced social media data. This step considers each mention, e.g., a tweet about *NorCal* or *SoCal*, as a *vote* for the corresponding region and identify as those core areas that are agreed upon by a significant number of people. In step 2, we constructed polygons from the identified point clusters. While such polygons may not be completely consistent with the understanding of each individual, they can provide intuitive delineations of the general areas. In addition, these constructed polygons can be used to support spatial queries, e.g., *show me all the hotels in SoCal*. Figure 2.5 illustrates this workflow, where subfigures 2.5a and 2.5b show the clustering process, and subfigures 2.5b and 2.5c demonstrate the polygon construction.

To identify point clusters from geo-referenced social media postings, we use DBSCAN



(a) Raw data (b) Clustered points (c) Constructed polygons

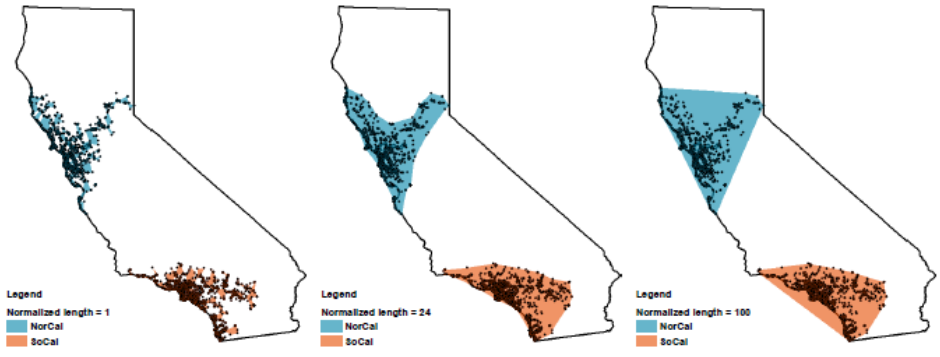
Figure 2.5: The workflow for extracting continuous boundaries for the cognitive regions of *NorCal* and *SoCal* (the visualized dataset is based on Flickr).

which is a density-based spatial clustering algorithm [89]. Compared with distance-based clustering methods such as K-means or K-medoid, DBSCAN has two advantages which make it more suitable for our task. First, DBSCAN can identify clusters with any arbitrary shape. In this research, the shapes of the potential cognitive regions are unknown, and DBSCAN can help discover their perceived boundaries. Second, DBSCAN is robust to noise which commonly exist in social media data. Clustering methods, such as K-means, will classify noise observations into clusters and therefore can distort the derived regions.

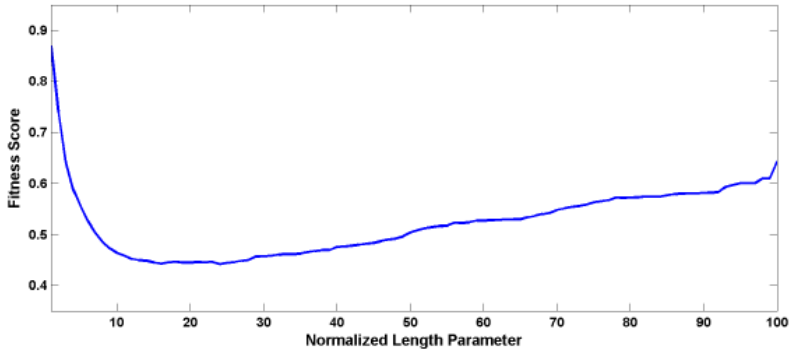
DBSCAN requires two parameters, namely ε and *MinPts*. ε defines the search radius while *MinPts* specifies the minimum number of data points within the said search radius. The two parameters together define a density threshold; clusters are identified at the locations whose density values are higher than the defined threshold. To find a proper ε value, we performed a nearest neighbor analysis on the three social media datasets as suggested by [89]. We assumed 1% of the data were noise, and found the 99th percentile of the nearest neighbor distance (NND) in each dataset. Accordingly, the 1% of data points

which were further away from the vast majority of observations, were considered as noise. We calculated the average of the 99th NND percentiles for the three data sources and used the averaged value for ε. For *MinPts*, we cannot use a single absolute value (e.g., 4) as in traditional DBSCAN applications, since the number of data entries from different sources varied significantly. For example, the number of Instagram postings was much larger than those of the other sources (see Table 2.1). Consequently, it would have been much easier for Instagram observations to form clusters than for the other two sources, if a single value were used for *MinPts*. To address this issue, we used percentages instead of absolute counts for *MinPts*, namely 1%, 2%, and 3% of the total number of postings per data source, to model the vague nature of the cognitive regions. Other settings could be explored in future work, with larger values shrinking the core region.

With point clusters identified, the second step was to construct polygons to approximate the boundaries of the cognitive regions. A *convex hull* approach has been used in many studies to represent the minimum bounding shape for a group of points [90]. Such a hull, however, is unable to accurately delineate for the shapes of point clusters. The chi-shape algorithm, proposed by [91], computes a *concave hull* for a set of points. The chi-shape algorithm requires a normalized length parameter λ_P, which ranges from 1 to 100. A value of 1 creates polygons which are closest to the original point set, but may generate spiky edges (Figure 2.6a). A larger value of λ_P will create smoother boundaries but also generates more empty space within the polygon. When λ_P is set to 100, the constructed polygon is equivalent to a convex hull (Figure 2.6c). Recent work by [92] proposes a fitness function which balances the complexity and the emptiness of the constructed polygon. Based on their work, we iterated λ_P from 1 to 100, and identified the optimal λ_P value (which is 24 in our experiment) that achieves the minimum value for the fitness function. Figure 2.6d shows the resulting curve plot. The polygon generated with $\lambda_P = 24$ is shown in Figure 2.6b respectively.



(a) $\lambda_P = 1$ (b) $\lambda_P = 24$ (c) $\lambda_P = 100$



(d) Curve plot for fitness and λ_P

Figure 2.6: Constructed polygon representations for the cognitive regions of *NorCal* and *SoCal* using different λ_P values.

Task III. Inferring Thematic Characteristics via Topic Modeling

Having identified and delineated regions, we explored what these regions have in common with each other and how they differ. To do this, we use *topic modeling* over social media. We selected the *Resolution 3* (Figure 2.4c) spatial data layer as our basis for topic modeling given that it offers the most detailed depiction of California that we assessed in our experiments, allowing for nuanced changes in topics to have an impact. Each of the social layers (Flickr, Twitter, Instagram) was spatially intersected with the Resolution 3 hexagons, and the unstructured textual data were grouped and aggregated to the individual hexagon level. Next, the data were cleaned to remove standard English

stop-words, non-alphabetic characters and words consisting of less than three characters. The words for each hexagon were then stemmed[1] and place names were removed via DBPedia Spotlight[2] and manual extraction.

We applied *Latent Dirichlet Allocation (LDA)* [93] for topic modeling using the MALLET toolkit [94]. *LDA* is a generative, unsupervised model that takes a bag-of-words approach to constructing topics. In this case, the corpus consists of all hexagons in California while the textual references within each hexagon make up a single document. The topics are constructed by exploring the co-occurrence of words in each document. Provided these topics, each hexagon could then be thematically defined as a distribution across all topics. For this research, and in line with previous work [95, 96, 97], we used 60 topics. The resulting topic distributions were then assigned back to the hexagons allowing for visual and statistical representation through thematic layers.

2.4 Results and Discussions

2.4.1 Membership Variability and Comparisons with Survey

Figure 2.7 depicts the spatial distributions and membership values of the *SoCal* and *NorCal* cognitive regions from the aforementioned data sources at three different resolutions. The cells rated as most *NorCal* are color-coded as blue, whereas *SoCal* cells are colored red. Darker colors represent a higher degree of membership. The core region of *NorCal* is around San Francisco (the Bay Area) which is roughly 300 miles south of the northern border of California. On the other side, cells that are most highly rated as being *SoCal* are around the greater Los Angeles area, which is more than 100 miles north of the southern border of the state. The boundary between *NorCal* and *SoCal* is

[1]Using the Snowball stemming method http://snowball.tartarus.org
[2]http://spotlight.dbpedia.org

vague and quite similar for the different data sources with respect to its shape, width, and location. However, the data do show different boundary transition patterns due to varying number of postings across data sources. The Instagram dataset has adequate observations in the transition zone between *NorCal* and *SoCal*, while other sources don't have sufficient data. All data sources reveal intensity values that are higher for both regions at the coasts and decrease towards the east, mostly failing to cross our minimum threshold to be considered part of either vague region. These results confirm the existence of a *platial effect* that distorts and relaxes spatial properties such as cardinal directions, substantially altering its monotonic variation across the landscape.

The average pattern across the social media sources is shown in Figure 2.8; each cell contains the mean of classified ranking percentiles (on a 1-7 scale) across all data sources. Figure 2.8 makes it evident that the cognitive regions we derive from our data-synthesis approach are highly similar to those from the original human-participants survey by [60], although not identical. Both empirical approaches show that the *NorCal-SoCal* distinction is mostly relevant to the west coast, including the coast ranges, beach communities, and metropolitan areas of San Francisco-Oakland-San Jose, Los Angeles, and San Diego. Indeed, the data-synthesis approach leaves several cells unclassified as either *NorCal* or *SoCal*, especially cells in the middle and eastern parts of the state (more below). Thus, both approaches show clearly that the boundary between the two cognitive regions is not homogeneous but wedge-shaped, being much narrower toward the west coast and broader toward the east; the data-synthesis boundary area is even a trapezoid or truncated wedge. Both approaches show that the locations of core intensity for the cognitive regions of *NorCal* and *SoCal* are not at the northern and southern state borders, respectively, but considerably south of the northern border and north of the southern border. The two approaches identify a Southern California core that is virtually identically locatedencompassing downtown Los Angeles and the west side, including the coast. The

core of *NorCal*, however, is identified by our data-synthesis approach as quite a bit further south than it is by the human-participants approach of MFP. It is essentially the San Francisco Bay area for the data-synthesis approach, while it is north of that for the MFP approach, around the confluence of the counties of Lake, Colusa, Yolo, Napa, and Sonoma. This difference aside, the data-synthesis approach agrees with the human-participants approach that the concepts of *NorCal* and *SoCal* are not merely latitudinal but attitudinal (i.e., both reveal platial effects).

To compare our results with the MFP results quantitatively, we computed Spearmans rank correlation ρ between the four layers from social media sources and the single layer from the human-participants survey, for the 69 cells which had sufficient social media data. As Table 2.2 shows, the correlations are uniformly very high for each of the four sources with the human-participants data; averaging across all four sources, the correlation with the human-participants data is 0.870 for scoring function $M1$, based on absolute occurrence differences, and 0.882 for scoring function $M2$, based on normalized ratio differences. All these high correlations are significant at p-value < 0.001 (df $= 67$), indicating that our automated approach generated membership results for these cognitive regions that closely approximate those of direct human raters. Moreover, Kendalls rank correlation τ is 0.712 for $M1$ and 0.721 for $M2$ respectively, which also implies a positive ordinal association between our approach and the human-participants approach.

As shown in Table 2.3, the value for Kendall's W (0.953, p-value < 0.001) shows a high agreement among our four data sources with respect to the membership rankings of all cells. Kendall's W remains very high (0.929, p-value < 0.001) even after adding the survey ranks from the MFP study as the fifth source, demonstrating a consistency between our data-synthesis-driven approach and the human-participants survey. In other words, the effects we see are not merely artifacts of a specific data source (and its user community).

Table 2.2: Correlation between the data-synthesis-driven results and the human-participants results from the original MFP study

Source	ρ (M1)	ρ (M2)	τ (M1)	τ (M2)
Flickr	0.881	0.880	0.721	0.719
Instagram	0.867	0.856	0.711	0.701
Twitter	0.874	0.838	0.714	0.673
TravelBlogs & Wikipedia	0.897	0.878	0.747	0.718
Means	0.870	0.882	0.712	0.721

Table 2.3: Kendall's coefficient of concordance W

Source	Four Raters	Five Raters
Kendall's W	0.953	0.929
Chi-sq	259	316
p-value	< 0.001	< 0.001

Figure 2.8 also shows the standard deviations (SDs) for each cell, as were presented by MFP (Figure 2.1). Our pattern of SDs is starkly different than that for MFPs results. MFP found the least variabilitythe greatest consensusfor cells at and near the core of *NorCal* and *SoCal*. The boundary cells between the cores show the greatest variability. This would perfectly fit a pattern of statistical range restriction near the extremes of the scale (i.e., floor and ceiling effects), except that the MFP participants agreed a great deal that the eastern cells making up the boundary were neither *NorCal* nor *SoCal*. Our data-synthesis SDs show a complex pattern. They clearly do not reveal any statistical range restriction, perhaps understandable given the cell values are not based on a direct numerical rating scale. References to *NorCal* are highly variable for cells making up the core of that region, while they are very consistent for cells making up the core of *SoCal*. Apparently users of the selected social media sources agree more strongly about the spatial reference of *SoCal* than they do about *NorCal*. In general, we find high variance for cells in the northern half of the state and low variance for cells in the southern half.

2.4.2 Sharpening the Boundaries

Like the MFP results, we generated boundaries for *NorCal* and *SoCal* that are vague or approximate. In our results, that is because social media references to *NorCal* and *SoCal* terms do change abruptly at or for some precise location on the landscape. For the MFP results, people do not express the belief that there is a precise transition location for these regions. In other words, whether considered a cultural phenomenon or a mental phenomenon (or both), these cognitive regions are *conceptually* vague [21].

However, aside from the basic-research motivation of understanding the nature of vague cognitive regions, we can apply our understanding to improving the functionality of various geographic information technologies. In several contexts, such as geographic information retrieval, this functionality will be increased by sharpening (also called hardening) the vague boundaries. The data-synthesis approach can be used to do this, even though we recognize that the cognitive boundary as such remains conceptually vague.

Here we discuss our results from applying DBSCAN clustering and the chi-shape algorithm to Flickr, Twitter, and Instagram results to "precisify" the vague cognitive regions by computing crisp boundaries for their core areas.

We do this by varying the threshold of reference density we require to include a cell as being in one of the two regions (Figure 2.9). For each of the subfigures, graduated colors (from light to dark) represent the extracted polygons based on minimum density thresholds of 1%, 2%, and 3% of the total number of data observations respectively. Naturally, the region hulls shrink as we increase the threshold density. This is due to the fact that the 3% threshold puts a higher DBSCAN requirement for point clusters to be formed than the 2% threshold. However, the boundaries formed by the 3% threshold are also more reliable since they are derived from more observations.

In Figure 2.9 (d), we overlap the results of all three data sources to identify the

common core regions for *NorCal* and *SoCal*. These identified common cores can be combined in different ways to fit specific applications. For example, a GIS which requires high *precision* for its spatial query results can employ the overlapped core region (i.e., those in the darkest color). In contrast, an application that needs high *recall* for its retrieved result can use a spatial union of the 1% polygons derived from the three sources. As can be seen in Figure 2.9 and Figure 2.7, there is a substantial overlap between the regions derived from the three different datasets. This consistency indicates that these regions are not mere artifacts of a particular dataset, but reflects a broader and shared understanding.

2.4.3 Thematic Characteristics

For Task III, we modeled topics associated with *NorCal* and *SoCal* social media postings using latent Dirichlet allocation. This topic modeling approach considers the co-occurrence of words in a document and constructs topics from those words often occurring together. Upon examination, one can see that these topics are often thematically related and coalesce around properties such as those related to *Nature*, *Food*, or *Hiking*. Figure 2.10 shows three examples of the total of 60 topics generated via the topic model. Each topic is shown as a map of California with the color of each hexagon determined by the probability value of that topic appearing in that cell. The word cloud associated with each map shows the top terms contributing to that topic.

Figures 2.10a and 2.10b both depict topics related to physical features in the environment and the outdoors. Words such as *Mountain*, *Park* and *Tree* contribute highly to both topics. There is a clear geospatial difference in the topics, however, with Figure 2.10a showing high density in the southern interior, and Figure 2.10b presenting higher probability values in the center and northern parts of the state. These are examples of

topics that are clearly influenced by the linguistic characteristics of individuals contributing data from either *NorCal* and *SoCal*. In contrast, Figure 2.10c presents a topic that is split east and west rather than north and south. This topic lists the highest probabilities along the coast, consisting of words such as *Beach*, *Ocean* and *Surf*. Both *NorCal* and *SoCal* are equally represented in this map showing that social data contributors mention words related to this topic regardless of the norther/southern California split.

From a purely visual representation in Figure 2.10, one could assume that there is no clear topic-wise distinction between the two cognitive regions of *SoCal* and *NorCal*. However, this is not the case. To demonstrate this, we selected ten prototypical *SoCal* and ten *NorCal* hexagons based on the membership intensity values reported in the original MFP paper. We extracted topic distributions for these hexagons and calculated the *Kullback-Leibler divergence (KLD)* [98] for hexagons within *NorCal*, *SoCal* and between both. *KLD* is a measure of the difference between two probability distributions. Low values indicate similar distributions while higher values suggest dissimilar distributions. Figure 2.11 shows these *KLD* values plotted as a smoothed histogram.

The core hexagons for *SoCal* are highly similar in terms of their distribution of topics. Core hexagons in *NorCal* are also quite similar to each other though slightly less than those for *SoCal*. This reflects the less cohesive data for *NorCal* we have discussed previously. When comparing inter-region hexagons, we find a peak *KLD* value that is much larger, indicating substantially greater topic dissimilarity between *NorCal* and *SoCal* cells than between cells within each region separately. In short, the intra-region topic similarities are substantially higher than the inter-region similarities. This means that while no single topic on its own is sufficient to distinguish the two cognitive regions from social media posts, the 60 topics can jointly distinguish between *SoCal* and *NorCal*. This is an important finding, as it suggests that everyday conversation is "geo-indicative" [99] to a degree where it can likely be used to discriminate regions and other geographic

properties and entities [100].

2.5 Broader Implications

This study revisits the work of [60] using a very different data-synthesis-driven approach to obtaining data, instead of a human-participants survey. We demonstrated that a data-synthesis-driven approach can be successfully used to reproduce cognitive regions and membership like those established with a direct study of human research participants. We have also demonstrated how the used data and methods can be applied to go beyond previous work by extracting hardened hulls to represent these regions and how to study their thematic topics via topic modeling. Using the example of the informal cognitive regions of *SoCal* and *NorCal*, our work proposes an approach to deriving human conceptions of places, including regions, from social media data sources. The approach potentially captures not only spatial patterns but also the semantics of cognitive regions.

Our results suggest it is possible to reproduce the results of a direct human-participants study by mining existing social media postings from the Web. While we do not argue that such a data-synthesis driven approach can or should entirely replace human-participants testing, the data-synthesis approach has the clear advantage that it can be repeated for a wide set of cognitive regions at flexible spatial scales without running into the limitations of participants testing, e.g., the limited number of participants, limited attention span, variable knowledge of local geography, and so forth.

This research raises some further issues about the data-synthesis-driven approach. First is the difference between *the said place* which a person tags or mentions in a social-media entry and *the locale* where the person is located when posting the entry. *The said place* is not necessarily the same as *the locale*, since people can post any message about any place no matter where they are. In fact, we assume this might happen fairly often.

In our data, for instance, the tag *SoCal* was sometimes mentioned in a small number (about 2%) of entries posted from the Bay Area, a core part of *Northern California*; while about 1% of the tag *NorCal* mentions were posted from the core part of *Southern California*. This is the nature of crowdsourced data. Researchers must pay attention to this issue when interpreting the experiment results. Different types of location inferences and insights can be extracted from the social Web [101, 102]. For this reason, we used two membership measures *M1* and *M2* to focus on relative differences and proportions instead of raw counts of place mentions. The results validated our proposed metrics. In future work, natural language processing techniques (e.g., place name disambiguation, preposition and contextual analysis) can be employed in analyzing social media entries to better differentiate the said place and the locale.

There are also some arguments (e.g., sampling bias) with regard to the data-intensive paradigm in scientific research. The results of this study, however, suggest that user-generated social media data at least partially do reflect people's experiences, focus, opinions and interests in places. Thus, these rich datasets can be synthesized as *social sensors* to support the study of vague cognitive regions in geography and GIScience.

An advantage of this data-synthesis-driven processing and geocomputation framework is the flexibility with which one can change the spatial resolution of hexagons or any other polygonal tessellation used to discretize the landscape. This includes not only finer resolutions but coarser, more aggregated resolutions. If there were a theoretical argument to do so, one could even create a tessellation with multiple scales in a single layer. For example, the cognitive regions of *NorCal* and *SoCal* appear to apply much more to coastal California than the Central Valley, the Sierra Nevada, or the eastern deserts; thus, one might want to tessellate the state with a higher resolution in the coastal areas. Besides the potential for resolution of nearly unlimited fineness, we recognize the general appropriateness of matching the scale of ones analysis to the scale of the phenomenon

one studies [103]. More generally, we recognize that the analytic possibilities of the data-intensive approach may create phenomena that are not psychologically plausible and can thus be misleading. We were able to analytically harden (sharpen) the boundaries of our cognitive regions, but individual people typically do not have this ability and their conceptions of informal regions likely do not have such precise boundaries.

The human-participants approach asks individual people to directly express the degree to which they believe a particular place should be considered Northern or Southern California. This means that data relevant to the concept or feature of interest (*NorCal* and *SoCal*) are generated for all locations within the study framework (California). A limitation of the data-synthesis-driven approach is that cells lacking sufficient observations have to be filtered out, which means that comprehensive spatial coverage is lost, unlike a human-participants survey. These missing-data cells are places with small numbers of residents and visitors, including areas within national forests, large water bodies, or mountain ridges. Alternatively, another way to look at this is that when people make the *NorCal-SoCal* distinction (as cognitive regions), they are referring only to western, coastal California, maybe mostly to just the San Francisco Bay Area versus the Los Angeles Area. In that case, the human-participants approach might be misleading because it required people to apply a distinction to every location within the state, even if the person never thinks of that distinction as being relevant to places like the Sierra Nevada, the northeastern Modoc Plateau, or the southern deserts. Alternatively, one could allow human participants to rate only cells that relate to the regional distinction as they understand it.

The human-participants approach asks directly for expressions of ones beliefs about informal regions, including both their spatial properties and their thematic associations. The data-intensive approach is indirect, collecting communications that include a verbal reference to *NorCal* or *SoCal* but not asking anyone explicitly what they actually think

about these regions. As a case in point, modeling the topical references in the social media postings showed us that they can statistically segregate the two regions, but it told us nothing about the thematic content of themes related to *NorCal* and *SoCal*. That is, it did not tell us what thematic associations come to mind when people use one of the two region terms rather than the other; a human-participants study could presumably do this directly. The same can arguably be done with topic modeling in a future study but may require additional data sources. The data-synthesis approach will often tap into cultural conventions that may or may not correspond closely to the beliefs of individuals. Presumably, such reference occur in social media on many occasions when the creator of the message is not thinking at all about the characteristics of places or the regions of California. Considering all of these issues though, we find it even more impressive how much agreement we find between our approach and that of MFP.

Going back to geographic information retrieval as one of the application areas of research on vague cognitive regions, there is one interesting question that we have not addressed so far. Although highly problematic for large areal features, the vast majority of geographic features, be it museums or mountains, is still represented by point coordinates. Google Maps, for instance, includes such point features for both *NorCal* and *SoCal*[3]. How representative are these locations with respect to the identified regions in both the original study and our replication? Interestingly, the *SoCal* point coordinates from Google Maps are located in the middle of the desert between interstates I-15 and I-40, more precisely at about (34.96, -116.42). This puts Google's *SoCal* marker about 180km to the northeast of the centroid (33.81, -117.68) computed for the 3% common core region (near Anaheim, CA). Google's *NorCal* marker (38.84, -120.9) is placed near Garden Valley, CA northeast of Sacramento, CA . This is about 160km to the northeast

[3]The interface will accept both of these terms and map them to 'Northern California' and 'Southern California', respectively.

of the centroid identified by our work (37.96, -122.21) which is located in the broader Bay Area. In other words, the map markers for both regions differ substantially from the result obtained by MFP and our work. They also do not follow the west-east trend, where membership intensity values to both regions are higher at the coasts.

2.6 Conclusion

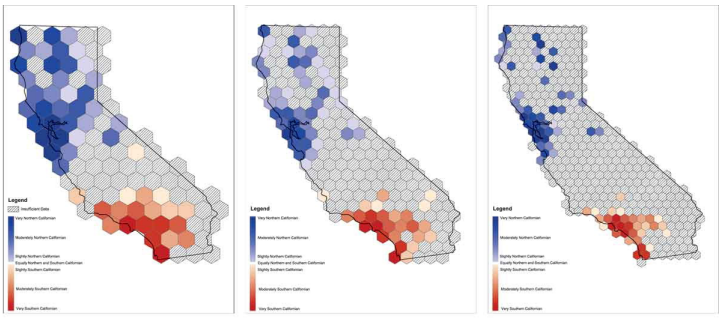
In this research, we investigated using a data-intensive approach to determining vague cognitive regions. We compared them to the corresponding MFP study based on human participants which validated our proposed approach. Using data sourced from social media including Flickr, Instagram, Twitter, Travel Blogs, and Wikipedia pages, we derived region membership scores for cells within the state of California that correlated significantly to those in the original study, both in terms of Spearmans as well as Kendalls rank correlation statistics. Overall, the shapes of *NorCal* and *SoCal* were quite similar for the two empirical approaches, including the non-monotonicity of the two regions and the heterogeneity of their vague boundaries. Most importantly, our work showed the same *platial effects* observed in the original study. Furthermore, our work examined the implications of increasing the spatial resolution of the tessellations on the cognitive regions that result.

In addition to assessing membership scores within the hexagons, we further explored the continuous boundaries and the core regions for *NorCal* and *SoCal*. A two-step workflow based on the DBSCAN clustering method and the chi-shape algorithm was designed to generate approximate boundaries for the cognitive regions. Experiments were conducted to select optimal parameters for the workflow, and we observe consistency among the polygon representations that are derived from the different datasets.

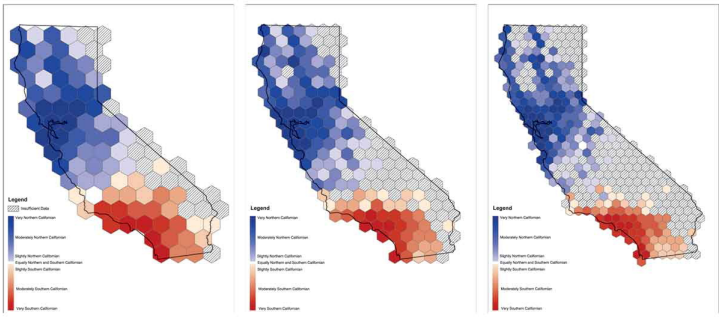
We also explored thematic associations for *NorCal* and *SoCal* with the help of topic

modeling. This generated various topics most often associated with different regions of California on our social media sources. Comparing the topic distributions of prototypical *NorCal* and *SoCal* hexagons shows high similarity within each region and a lower similarity between the two regions.

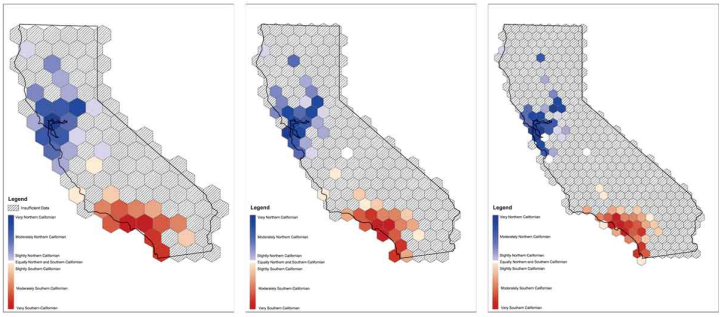
In sum, our paper is about the prospects for utilizing multiple social media sources to apply a data-synthesis approach to extracting and characterizing informal geographic concepts and features, such as the cognitive regions of *NorCal* and *SoCal*. Our study sheds light on differences in the methodology of traditional human-participants approach and the increasingly popular data-synthesis approach, suggests advantages and limitations of both approaches, and points to future avenues for research and system design in GIScience.



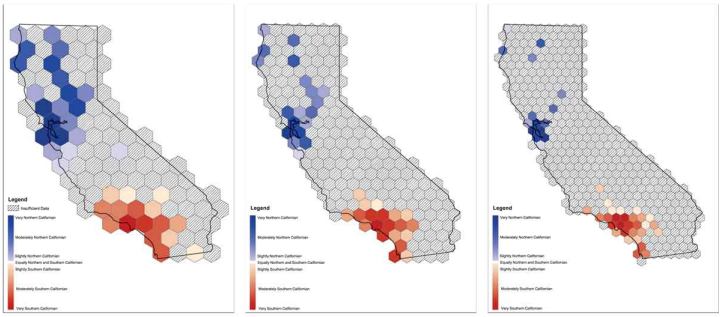
(a) Flickr Res. 1 (b) Flickr Res. 2 (c) Flickr Res. 3



(d) Instagram Res. 1 (e) Instagram Res. 2 (f) Instagram Res. 3



(g) Twitter Res. 1 (h) Twitter Res. 2 (i) Twitter Res. 3



(j) Travel & Wiki 1 (k) Travel & Wiki 2 (l) Travel & Wiki 3

Figure 2.7: The spatial distribution of membership measures derived from different data sources at three different resolutions.

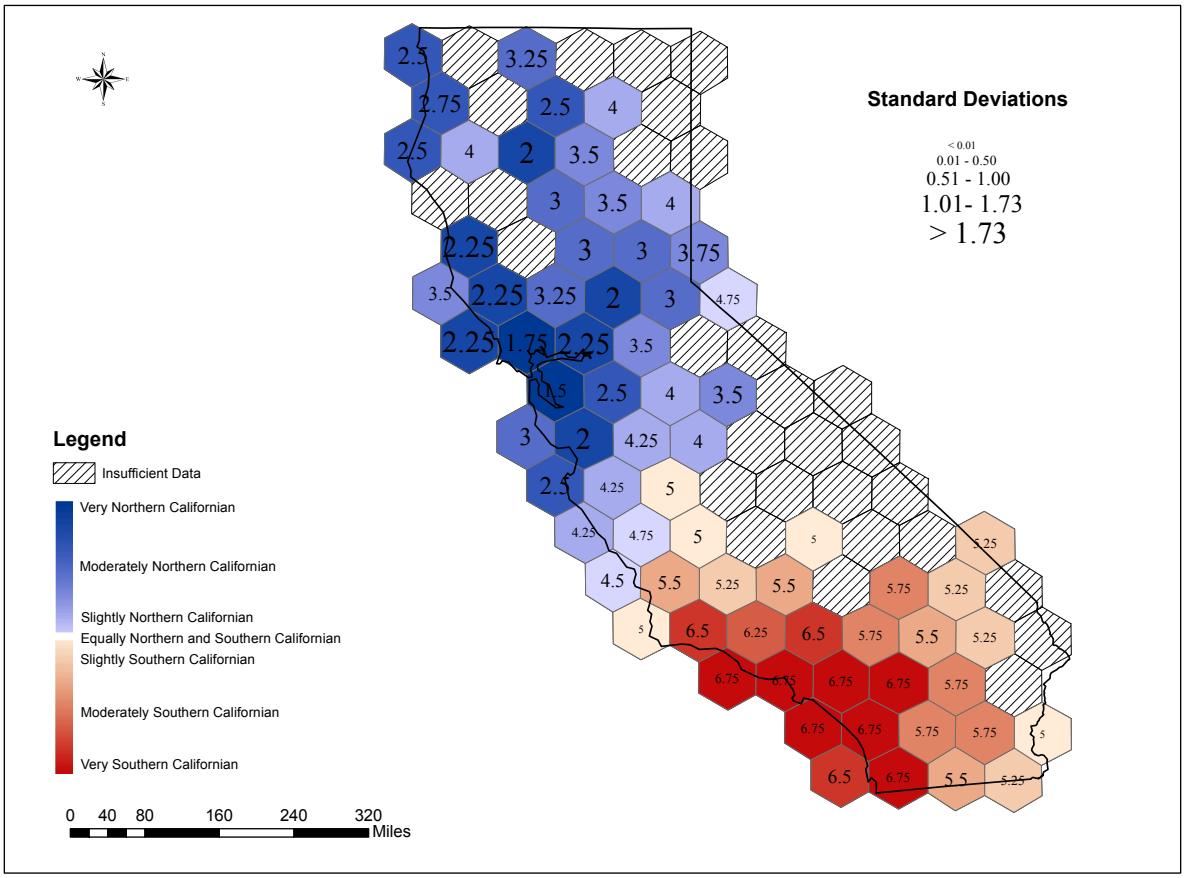


Figure 2.8: The results of identifying *SoCal* and *NorCal* cognitive regions using the data-synthesis-driven ranking percentiles.

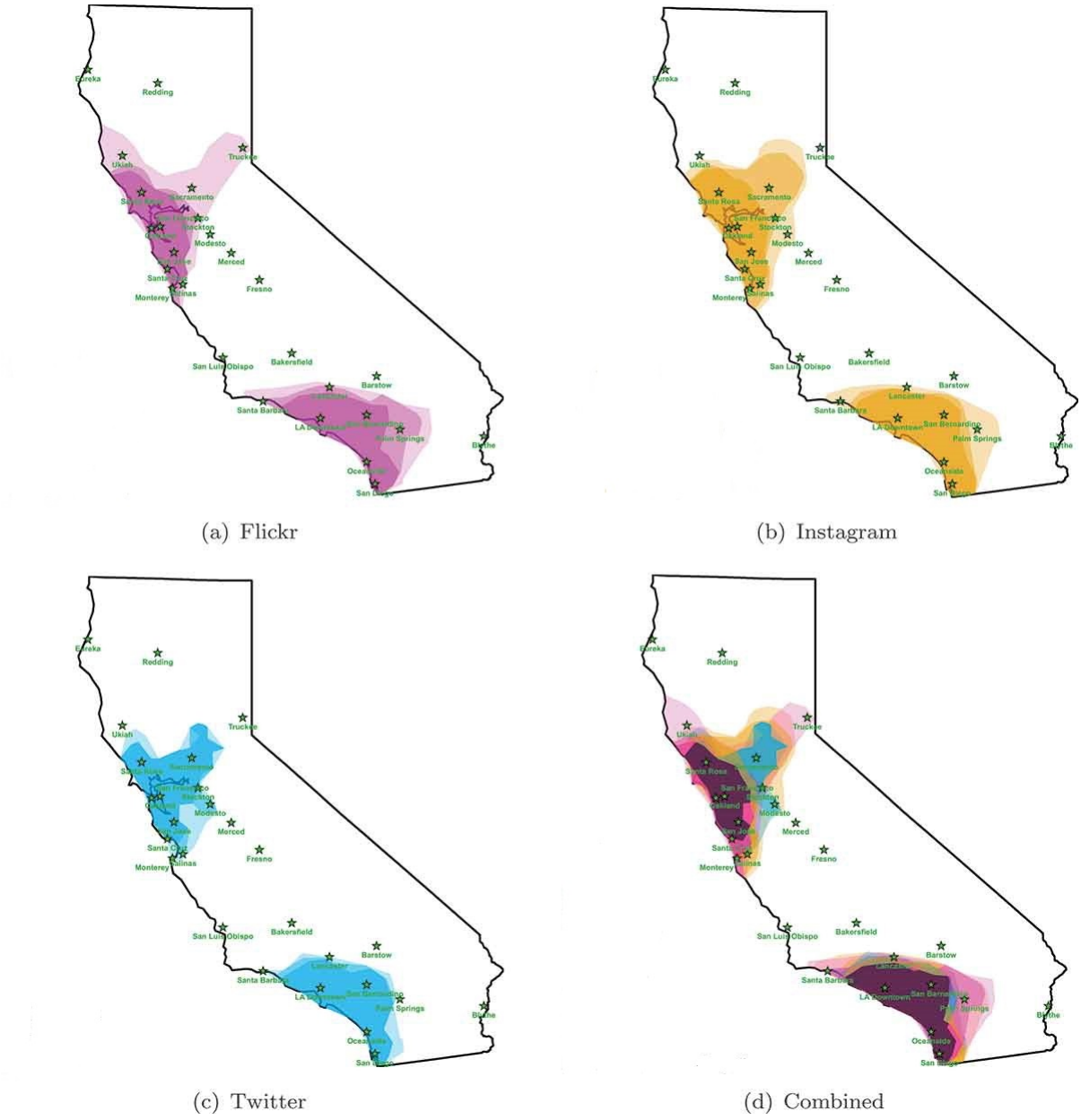


Figure 2.9: Core regions of *NorCal* and *SoCal* extracted using different datasets.

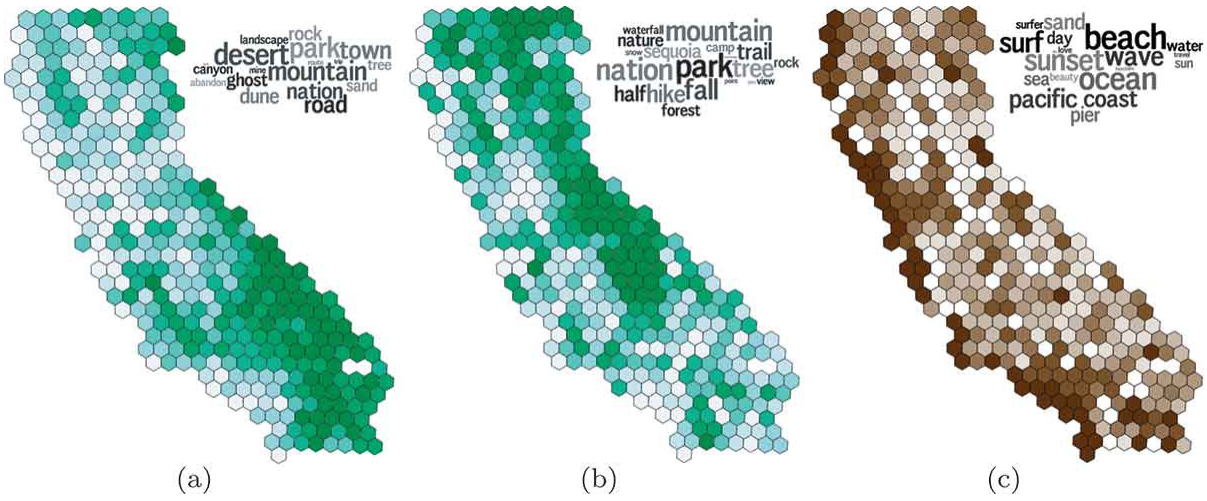


Figure 2.10: Three topics mapped to California along with their related word clouds. The darker the chromatic hue, the more prominent are the topics of terms in the postings from a particular cell.

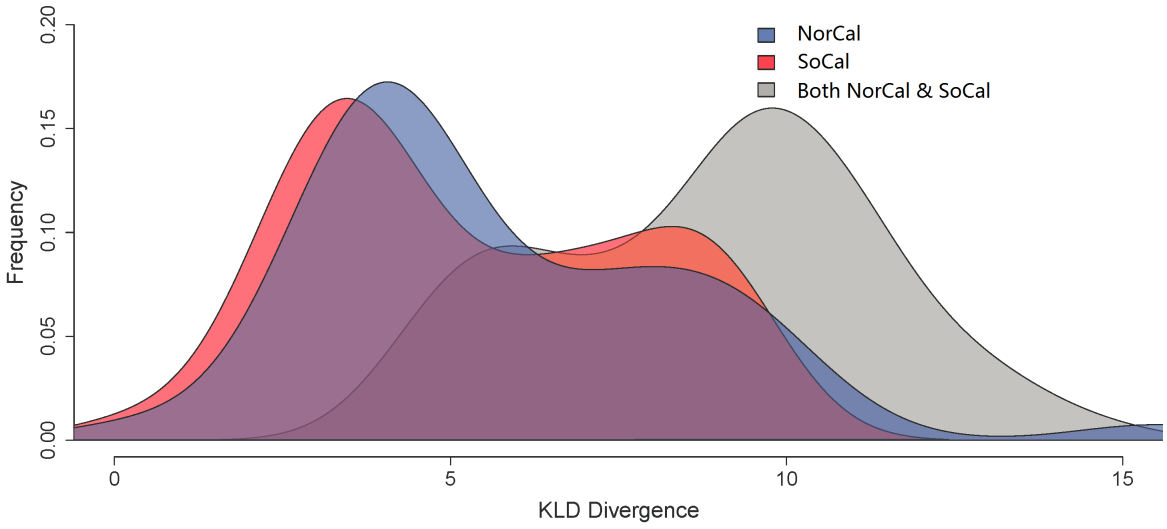


Figure 2.11: Kullback-Leibler divergence showing similarity of topics within *SoCal* hexagons and similarity of topics within *NorCal* hexagons, and dissimilarity of topics between both.

Chapter 3

Extracting Urban Functional Regions from Points of Interest and Human Activities on Location-based Social Networks

Chapter 3 presents a statistical framework to discover semantically meaningful topics of place types and further derive functional regions based on the co-occurrence patterns of place types and spatial clustering techniques in relation to urban planning. The framework applies latent Dirichlet allocation (LDA) topic modeling and incorporates user check-in activities on location-based social networks. A case study has been conducted using a large corpus of about 100,000 Foursquare venues and user check-in behavior in the ten most populated urban areas of the United States. Compared to remote sensing images which mainly uncover the physical landscape of urban environments, the proposed popularity-based POI topic modeling approach can be seen as a complementary social sensing view on urban space based on human activities.

Peer-reviewed Publication	
Title	Extracting Urban Functional Regions from Points of Interest and Human Activities on Location-based Social Networks
Authors	Song Gao, Krzysztof Janowicz, Helen Couclelis.
Venue	Transactions in GIS
Editors	John P. Wilson
Publisher	John Wiley & Sons
Pages	TBD
Submit Date	13 February, 2017
Accepted Date	28 April, 2017
Publication Date	In Press
Copyright	Reprinted with permission from John Wiley & Sons

Abstract: Data about points of interest (POI) have been widely used in studying urban land use types and for sensing human behaviors. However, it is difficult to quantify the right mix or the spatial relations among different POI types indicative of specific urban functions. In this research, we develop a statistical framework to help discover semantically meaningful topics and functional regions based on the co-occurrence patterns of POI types. The framework applies the latent Dirichlet allocation (LDA) topic modeling technique and incorporates user check-in activities on location-based social networks. Using a large corpus of about 100,000 Foursquare venues and user check-in behaviors in the ten most populated urban areas of the United States, we demonstrate the effectiveness of our proposed methodology by identifying distinctive types of latent topics and further, by extracting urban functional regions using K-means clustering and Delaunay triangulation spatial constraints clustering. We show that a region can support multiple functions but with different probabilities, while the same type of functional region can span multiple geographically non-adjacent locations. Since each region can be modeled as a vector consisting of multinomial topic distributions, similar regions with regard to their thematic topic signatures can be identified. Compared to remote sensing images which mainly uncover the physical landscape of urban environments, our popularity-based POI

topic modeling approach can be seen as a complementary social sensing view on urban space based on human activities.

3.1 Introduction

Cities support a variety of functions that relate to land use types, including residential, commercial, industrial, transportation, and business regions and infrastructure, while affording different types of human activities, such as living, working, commuting, shopping, eating, and recreation. Rapid urbanization and new construction have caused land use changes and urban expansions in many areas. Remote sensing images together with spatial metrics have been widely used to classify urban land use and monitor change at different spatial scales [104, 105, 106]. However, human activities usually take place in different types of points of interest (POIs). Remote sensing techniques perform well in extracting physical characteristics, such as land surface reflectivity and texture of urban space but are not good in identifying functional interaction patterns or in helping understand socioeconomic environments [107, 45]. Compared to other datasets and methods in remote sensing and field mapping, using POI data, social media, and their associated methods can lead to a better understanding of individual-level and group-level utilization of urban space at a fine-grained spatial and temporal resolution. Rich *social sensing* techniques can help bridge the semantic gap between land use classification and urban functional regions. The function of a place is determined by what type of activities can occur there [34, 33]. The same types of POIs can be located in different land use types and may also support different functions. For example, *restaurants* are found in residential areas and in commercial areas, as well as in industrial areas. The main function of *universities* is education, but they also support sports activities, music shows, and so on. Previous studies have demonstrated that different POI types have distinctive

semantic signatures [34] (i.e., spatial, temporal, and thematic distributions) based on crowd-sourced location-based social media data analysis, in analogy to spectral bands in remote sensing [108]. There is a growing trend of using location-awareness sensing data (e.g., trajectories from mobile phones), POI data, and social media feeds to study the spatial and social structure of urban environments [107, 109, 110, 45, 108, 111, 112]. However, few studies have investigated the latent relationships among different types of POIs and how they spatially interact with each other to support urban functions, such as education, business, and shopping. In this research, we aim to develop a data-driven framework to discover urban functional regions from POIs and associated human activities on location-based social networks (LBSN).

We argue that geographic knowledge and measures of spatial distribution over POI types (categories) can be employed to derive latent classification features for these said types, which will then enable the detection and the abstraction of higher-level functional regions (i.e., semantically coherent areas of interest) such as shopping areas, business districts, educational areas, and tourist zones. To test this claim, we will study the co-occurrence patterns of different POI categories as well as the associated human activities (i.e., mobility, check-ins, reviews, and comments), and thus employ analytical measures to quantify their differences and conduct classifications of functional regions.

The contributions of this research are as follows:

- We propose a novel framework to study urban functional regions by employing data about Points Of Interest and human activities derived from social media.
- We incorporate location-based social network user check-ins into a probabilistic topic modeling technique to discover functional co-occurrence patterns of different POI types.
- The proposed method can support functional inferences for specific type of re-

gions and thus serve as a new heuristic to enable the search for similar urban places/regions, based on their POI-type distributions and corresponding human activities, and using natural language processing and machine learning techniques.

The remainder of this article is structured as follows. Section 3.2 introduces background material and related work. Next, Section 3.3, discusses the datasets used and the selection of study areas. Section 3.4 introduces the methods used in our framework and specifically LDA. In Section 3.5, we present the results of topic modeling to characterize, cluster, and compare functional regions. Next, we conclude our work and point out directions of future work in Section 3.6.

3.2 Related Work

With the increasing popularity of travel blogs, volunteered geographic information (VGI), location-based social networks (LBSN), and so forth, researchers have developed a variety of place-based studies that employ datasets from these various sources. For instance, [113] presented a topic modeling methodology to estimate geographic regions from unstructured, non geo-referenced text on Wikipedia and travel blogs by computing a density surface of geo-indicative topics over the Earth's surface. The proposed framework combined natural language processing techniques, geostatistics methods, and data-driven bottom-up semantics. In order to evaluate the use of topic modeling techniques on the extraction of thematic characteristics of places, [114] applied that approach on a set of travel blog entries to identify the themes that are most closely associated with specific places around the world. Their proposed method is capable of measuring the degree to which certain themes are local or global, as well as analyzing thematic changes over time. POI data play an important role for human activity-based land use, transportation, and environmental models. [110] utilized the Yahoo online POI data together with publicly

available aggregated employment data from census at the block group level to derive fine resolution of disaggregated land use estimates (i.e., employment by category) at the city block level. For the evaluation, they first used a variety of machine learning algorithms to match and cluster POI types into a labeled business establishment taxonomy, and then compared it with ground-truth data from commercial business data vendors. The results demonstrated that their proposed method got a better goodness of fit with a lower relative mean squared error for the estimated employment population across all city blocks than that from the traditional uniform-distribution disaggregation approach. As for LBSN applications, [115] proposed a method to classify the geographical areas and LBSN users based on place types and the users' check-in statistics in Foursquare venues. The experiments were conducted in the metropolitan cities of London and New York and identified similar regions and user groups in each city. However, they didn't consider the temporal pattern of user activities. Later on, [116] employed both POI type information and the temporal patterns of taxi pick-ups/drop-offs in segmented map regions, utilized a topic modeling method based on latent Dirichlet allocation and Dirichlet multinomial regression techniques, and discovered various urban functional regions in the city of Beijing. The extracted region clusters were annotated as nine different groups: *diplomatic and embassy areas, education and science areas, developed residential areas, emerging residential areas, developed commercial/entertainment areas, developing commercial/entertainment areas, regions under construction, areas of historic interests, and nature & parks*. However, such rich multiple datasets that complement each other in the same city and especially high-precision mobility data are usually hard to fully access. One challenging issue is how to semantically classify and label the regions that are found given only one data source, and how to find similar places and regions across different cities. [117] proposed a novel observation-to-generalization place model and employed natural language processing techniques to derive place attributes. The proposed meth-

ods can support similar-place-search functions and the case studies were conducted using over 600,000 place articles on Wikipedia as a proof of concept. Later, [118] presented a novel method to enrich the place information on linked knowledge graphs using thematic signatures that are derived from unstructured text through topic modeling. This method can also be used to clean miscategorized places on the linked data cloud. In another study, [119] developed a semantic region growing algorithm based on the density of POIs on OpenStreetMap to extract places that afford certain type of human activities, e.g., shopping areas. In their model, four features including the number of *banks & ATM, restaurants, tourist facilities, and subcategories of shops* were used to identify the shopping areas/settings. They then compared the similarity of shopping areas/settings based on the four features in two European capital cities: Vienna and London. By incorporating human spatio-temporal activity data from social media, [120] extracted the spatial distribution hotspots of six types of urban functions (i.e., *Travel & Transport, Education Resource, Shop & Service, Nightlife Spot, Outdoor & Recreation, Food & Restaurant*) in the cities of Boston and Chicago. [121] introduced a low-rank-approximation-based model to detect functional regions based on 15 million social media check-in records in the city of Shanghai, China. This method discovered latent spatio-temporal human activity patterns and linked these with different functional regions. Researchers are also interested in the regional differences on discovering thematic characteristics of different POI types. [122] identified the most and least spatially varying place types and compared their thematic signatures internationally. The ongoing trend in this research direction lies in data-synthesis-driven approaches to study places and vague cognitive regions as well as the semantic generalizations of urban settings [119, 123, 124].

In summary, there is a variety of research studying places and place types from human data traces, including spatio-temporal human mobility patterns that can reveal the functions of regions. However, only a few studies have simultaneously considered both

POI information and human activities on location-based social network to derive urban functional regions. Moreover, to the best of our knowledge, there is no thorough discussion of the robustness of discovered urban and regional functional areas using different numbers of topics and clusters. There has also been no attempt to develop an urban function ontology based on the structure of POIs using a bottom-up approach.

3.3 Study Area and Datasets

3.3.1 Study Area

Urban areas – cities for short – are the highly populated places on the planet and include metropolitan regions, urban districts, towns, and suburbs. In order to explore the thematic characteristics and semantic clusters of urban areas in connection with urban functions, the ten most populated U.S. cities based on the 2015 population census: *New York, Los Angeles, Chicago, Houston, Philadelphia, Phoenix, San Antonio, San Diego, Dallas, and San Jose* and their surrounding metropolitan regions were selected as our study areas. The cartographic boundaries of those ten metropolitan areas are downloaded from the U.S. Census Bureau's TIGER geographic database[1].

3.3.2 Points of Interest Dataset

People usually go to different POIs for different kinds of activities, e.g., studying, working, dining, shopping, and relaxing. We assume that the spatial distributions and interactions of different types of POIs reflect particular urban functions. Location-based social networks such as Foursquare have created traces of social interactions based on the physical location of users. In these LBSN systems, users can check-in to a venue (i.e., a

[1] `https://www.census.gov/geo/maps-data/data/cbf/cbf_msa.html`

POI), rate it, and share their comments or tips. As shown in Figure 3.1, we first randomly generated 200 points as search locations in each urban area and then identified the surrounding Foursquare venues with their attribute information including *name, location coordinates, place category, number of check−ins, number of checked users, number of tips,and the rating score* in each search locations. Note that because of the Foursquare developer API limits, we only retrieved at most 50 nearby venues given a random search point. The POI data were collected in December 2016 and the attribute information for all venues is a historic snapshot at that time. There is a total of 480 different POI types in our data. Figure 3.2 shows the empirical cumulative density function (CDF) of the distance distribution between each Foursquare venue and the corresponding search location. Steeper curves (with larger slope values) before reaching the relatively steady state (about 95% cumulative probability) show that more POIs are closer to the search locations given the same number of Foursquare venues. In order to generate most 'nearby' POIs around each search location, we further spatially filtered out those venues outside the 95% inverse CDF distance threshold; i.e., we only selected those venues within a relatively small search distance. The distance thresholds differ among cities as shown in Table 3.1.

3.4 Methods

3.4.1 Popularity-based Probabilistic Topic Model

Probabilistic topic models have been widely used to discover latent thematic characteristics and their structure when analyzing large sources of textual documents [93, 125, 126]. The *latent Dirichlet allocation (LDA)* is among the most popular topic modeling methods. LDA is an unsupervised generative probabilistic model that takes a bag-of-

Table 3.1: The 95% inverse CDF distance thresholds for all the ten urban areas

City Name	95% Inverse-CDF Distance Threshold (km)
Chicago	7.389
Dallas	8.994
Houston	8.647
Los Angeles	4.474
New York	7.123
Philadelphia	8.894
Phoenix	7.969
San Antonio	7.475
San Diego	7.837
San Jose	4.822
Average	7.363

words approach (which implies that the order of words in the document does not matter) to constructing topics. The key idea of LDA is that documents can be represented as a joint probability distribution over latent topics and each topic is characterized by a distribution over words [93]. Assume that there are total K number of topics associated with N words in the document corpus D, and α and η represent the prior parameters for the Dirichlet document-topic and topic-word distribution respectively. The mathematical relationship between the latent variables and the observed variables is described below:

$$p(\beta_{1:K}, \theta_{1:D}, Z_{1:D}, W_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{i=1}^{D} p(\theta_d) (\prod_{n=1}^{N} p(Z_{(d,n)}|\theta_d) p(W_{(d,n)}|\beta_{1:K}, Z_{(d,n)})) \quad (3.1)$$

As shown in Figure 3.3, the generative process can be described as follows:

I. Let β_k denote a probabilistic distribution over the word vocabulary for a given topic k, and draw $\beta_k \sim \text{Dir}(\eta)$;

II. Let θ_d represent the topic proportions for the d_{th} document, and draw $\theta_d \sim \text{Dir}(\alpha)$;

III. Let $Z_{(d,n)}$ denote the topic assignment for the n_{th} word in document d and $W(d,n)$ represent the n_{th} word in document d from a fixed vocabulary, and draw the multinomial distributions: $Z_{(d,n)} \sim \text{Multinomial}(\theta_d)$ and $W_{(d,n)} \sim \text{Multinomial}(\beta_{z_{(d,n)}})$.

In order to compute the conditional distribution of the topic structure given the word observations in documents, the expectation−maximization (EM) algorithm and *Gibbs sampling* are most the commonly used methods. After finishing the computation, two matrices θ and β associated with topic proportions and assignments are generated. More detailed notations, calculations, and explanations can be found in [93].

In analogy with LDA's use of textual materials, we take the type (e.g., school, park, restaurant) of each POI as a word, the search region that contains those POIs as a document, and an urban function or a land use as a topic that represents thematic characteristics and the semantics of places. By running the LDA topic modeling technique, we can find the posterior probabilistic distribution of each POI type in a certain type of region conditioned on the search region's topic assignments. The LDA model running on POI types generates summaries of thematic place topics (e.g., beach promenades, art zones, shopping areas) with a discrete probability distribution over POI types for each topic, and infers per-search-region probability distributions over topics. For example, one would assume that a *beach promenade* topic should contain venues such as *beach*, *seafood restaurants*, and *surfing spots*; while a *shopping area* would more likely contain *clothing stores*, *cosmetics shops*, and *shoe stores*.

Another important concept in our method is the *popularity* of a POI as captured by its LBSN user check-in behaviors. For example, the neighborhood of a football stadium usually has only one instance of *stadium* surrounded by dozens of *sports bars*, *restaurants*, and *parking lots*. However, a stadium usually attracts thousands of visitors and

is the dominant feature of its neighborhood. Thus this particular POI type makes said neighborhood distinct from other neighborhoods (e.g., nightlife zone), which also contain *cocktail bars* and *restaurants*. We need to address such an human activity effect during the generation of the document-word frequency matrix. More specifically, we will rescale the POI type occurrences according to their associated POI instance check-in counts. The rescaling process can be represented as follows:

$$Freq_{(d,t)} = \lceil \sum_{i} Log(V_{(d,t,i)}) \rceil \tag{3.2}$$

where $Freq_{(d,t)}$ represents the rescaled occurrence for a POI type t given a search region d; and $V_{(d,t,i)}$ is the number of unique users who have contributed their check-ins for a venue i that belongs to the same POI type t in the same search region d. The ceiling function $\lceil \sum_i Log(V_{(d,t,i)}) \rceil$ gets the least integer greater than or equal to the value of $\sum_i Log(V_{(d,t,i)})$.

We then test whether such an unsupervised popularity-based LDA topic modeling technique can support the discovery of characteristic semantic regions across different U.S. cities with a similar structure of POI type mix distribution.

Finding the appropriate number for latent topics is important but difficult given a dataset using the LDA topic model. Several metrics and methods have been developed to address this issue. [95] used the Gibbs sampling algorithm to obtain samples from the posterior distribution over topic assignments Z at different choices of the total K number of topics, and then calculated a log-likelihood $P(W|Z,k)$. The value of K at which the log-likelihood gets the maximum and stabilizes after hundreds of iterations will be taken as the right number of topics for a specific document corpus. With the consideration of one issue that sometimes words have too many overlaps across those generated latent topics, [127] proposed a density-based method for adaptive LDA model selection. The key

idea of this algorithm is to maximize the intra-topic similarity while minimize the inter-topic similarity. They calculated the average cosine distance between pairwise topics with their word assignments and then used a heuristic to find the most stable topic structure given the best K value based on the topic density measure. [128] viewed the LDA topic model as a matrix factorization mechanism and applied the symmetric Kullback−Leibler (KL) divergence [98] on the distributions generated from topic-word and document-topic matrices for finding the right number of topics. The best K value at which the symmetric KL-divergence is the minimum would derive the most discriminative topics and their distributions become orthogonal. In several empirical geographic information studies [118, 108, 124], different K numbers (e.g., 60, 100, 300) of topics have been deployed to investigate place characteristics. An optimal value of K may vary between different datasets and has influence on the thematic similarity of POI types [122].

3.4.2 Functional Region Aggregation

After deriving those latent thematic topics by running the proposed popularity-based LDA model, each region can be represented as a vector of the K-dimensional POI type topics. Those regions that are semantically similar in the topic space might contribute to the same urban function and can be aggregated into the same cluster as a functional region. Two clustering approaches are applied in this work: *K-means clustering* and *the Delaunay triangulation spatial constraints clustering methods*.

K-means clustering only takes the thematic characteristics of multivariate topic distributions of places into consideration without any spatial constraints [129]. It is an unsupervised clustering approach in which the number K needs to be predefined. The *silhouette* criterion has been widely used for determining an appropriate value of K [130]. The *silhouette* value $s(i)$ quantifies how well an object i is appropriately clustered. The

range of *silhouette* value is between -1 and 1. A high $s(i)$ value (close to 1) indicates that an object is appropriately clustered and is very dissimilar from other clusters. In the region clustering process for our POI datasets, we tried different K values ranging from 1 to 30 and identified the maximum average silhouette value across all clusters and chose that K as the optimal K-means clustering parameter for reporting the corresponding results.

Delaunay triangulation spatial constraints clustering has been introduced by [131]. This approach consists of three steps: (1) building a connectivity graph to capture the adjacency relations between points based on Delaunay triangulation spatial constraints; (2) creating a minimum spanning tree (MST) [132] from the neighboring connectivity graph with minimizing the sum of the dissimilarities over all the edges of the tree; (3) partitioning the derived MST into different subtrees as spatial clusters using a hierarchical division strategy to minimize the intra-cluster square deviations. More implementation details about this clustering algorithm can be found at [131].

3.5 Analysis and Results

3.5.1 Topic Modeling Results

As proposed in Section 3.4, before running the LDA model, we first incorporated the number of visitors for each venue as a popularity score in the rescaling process to generate a new document-word matrix (i.e., a search region−POI type occurrence matrix) across all search regions in the ten urban areas. Next, we evaluated the performance of different choices of K as the total number of topics for the LDA topic model using three introduced measures. As shown in Figure 3.4, by choosing the value of K from 5 to 200 and then running LDA topic models on our POI data, we derived different topic assignment

results. The measure proposed by [95] aims to maximize the log-likelihood of word-topic probability in the documents, while other two measures [127, 128] aim to minimize the proposed criteria. In the ideal case, one would expect those three measures converge at the same value of K. Unfortunately, in empirical studies, they do not necessarily present such perfect convergence patterns. In our parameter tuning experiments, the optimal K value for the "CaoJuan2009" and "Arun2010" measures is in the range of $90-160$, while the "Griffiths2004" metric gets relatively stable when K reaches 130 topics.

Therefore, we set $K = 130$ as the total number of topics and ran 2000 iterations of the Gibbs sampling process to derive the posterior probabilistic distribution over topic assignments. In Figure 3.5, we show nine of those interesting topics related to urban functions. Note that the probability assignments for those POI types are weighted and ranked by their *term frequency−inverse document frequency* (i.e., POI type frequency−inverse region frequency) so that each topic can display more distinctive and meaningful POI types that are directly proportional to the frequency in documents while inversely proportional to the region frequency at which a POI type occurs in the whole corpus. For instance, *coffee shop* has a very high frequency but also widely exists in most of the regions in our POI data so it plays a less important role than other categories (e.g., *theme park*) in distinguishing the function of a region. Some of those meaningful topics are illustrated as follows.

Topic 67 is a shopping-plaza topic that consists of various frequent occurring POI types including *shopping mall*, *accessories store*, *chocolate shop*, and a few restaurants. It is one of the most prominent topics across all cities. *Topic 109* is a beach-related topic that consists of *beach, surf spot, island, beach bar, pier, and so on.* In terms of spatial distribution, one would expect such topics should be located in coastal or lake-side cities only. Both *Topic 21* and *Topic 25* contain *history museum* and *art museum*, but *Topic 21* is more related to college/university regions since it also contains *pool, college rec*

center, tourist information center, and several other educational facilities; while *Topic 25* is more likely an art museum district since it also consists of *art gallery, antique shop, scenic lookout, and so on*. *Topics 6, 117, and 119* relate to outdoor sports and leisure activity places, such as *national park, ski lodge, gym, golf course, tennis court*, and various restaurants and studios. *Topic 36* and *Topic 74* describe mix distribution patterns of *bar, restaurant, government building, residential apartment, and business service*, which may suggest central-city areas.

In order to explore the variability of the above discovered nine topics while changing the total number of topics, we further investigate whether we can find exactly matching or most similar topics with different values for K. Two evaluation criteria, namely *Cosine Similarity* and *Jaccard Index*, were applied for this purpose [133]. Assume that each topic vector is a sequence of probabilistic values between 0 and 1 for all the 480 POI categories. Considering each pair of one target topic (e.g., Topic 6 when K=130) and another one comparing topic (e.g., Topic 1 when K=10), the cosine similarity measures the cosine of the angle between two non-zero vectors defined using an inner product. It is well suited to evaluate sparse vectors such as document-word matrices and the topic-POI matrices in our experiments. Unlike the cosine similarity that is frequently used for numeric vectors, the Jaccard index is a popular similarity measure for binary and categorical data, which is defined as the cardinality of the intersection divided by the cardinality of the union of two sets. We use the Jaccard index to quantify the topic structure similarity for their top-fifteen probabilistically ranked POI types. The larger the value, the more similar two topics are, where 1 equals a perfect match while 0 indicates no overlapping top-terms (i.e., POI types) in the comparison of two topics. The comparison batch processing was conducted from K=10 to K=150 with a step of 10. During each run with a given K, the maximum similarity values to each of the nine topics were computed. As shown in Figure 3.6, the maximum cosine similarities for *Topics 74 and 117* reach almost 1 and remained

stable when the total number of topics exceeded 30. As for *Topics 6, 36, 67, and 109*, we can also identify most semantically similar ($\geq$0.9) topics to them with K value equals to 150, 120, 150 or 140 respectively. This indicates the stability of identifying those prominent urban functional topics related to frequently co-occurrent physical facilities and services, a variety of bars and restaurants, and leisure activity places. However, we cannot find very similar ($\geq$0.8) topics to *Topics 21, 25, and 119* when choosing different K values, which implies that these topics may be more characteristic of a specific K value. In a similar manner, we analyze the topic structure similarity using the Jaccard index. As shown in Figure 3.7, those low similarity values illustrate the large composition variability existed in the top-fifteen probabilistically ranked POI types for all discovered topics with different K values.

In short, rather than a traditional top-down approach for describing urban functions based on familiar compositions of POI types, we demonstrated a bottom-up statistical topic-learning approach for finding underlying co-occurrence relationships among different types of POIs based on data on human activities extracted from location-based social networks.

3.5.2 Searching for Similar Places

Searching for similar places is an important task in geographic information retrieval and also valuable in many applications, such as tourism, real estate, and immigration. People may consider many factors such as job market, affordability, natural environment, and quality of life. When people consider moving into new cities, they may also want to know how they will like these new places and whether they can find similar neighborhoods to the ones they will be leaving. Such places typically contain a mix of types of POIs that people would like to visit. Fortunately, such information can be retrieved from popular

location-based social network platforms that have been used as a lens of social sensing to capture human-place interactions. In the following, we illustrate this idea with two scenarios:

(1) Search for similar regions given a dominant theme. We selected the city of Denver as our target city, which was ranked as the best metropolitan area to live in the U.S. according to a survey[2] from the U.S. News in 2016. It has a variety of local attractions and support many activities. Here we aim to find regions that are similar to those represented by *Topic 25*, which is related to art districts. We collected the Foursquare POIs and user check-in data for Denver by randomly sampling search locations and then searching for 50 nearby POIs given each sample location. Based on the aforementioned data processing procedures and the LDA topic model by incorporating the popularity score based on unique Foursquare check-in users, we can infer the probabilistic combination of different topics for a search region given its POI type co-occurrence pattern. As shown in Figure 3.8, within this search neighborhood, we discovered a high probabilistic topic distribution for *Topic 25*, which consists of a variety of prominent POI types such as *art museum, art gallery, history museum, concert hall and American restaurant*. Such a place may serve multiple functions. The second largest probabilistic topic in this search neighborhood is *Topic 121* that contains a large percentage composition of *brewery places*. By looking up other geographic background information and Web pages, we realize that local people actually also identify this region near the *"Santa Fe Dr."* as an "Art District" in Denver, which attracts many local residents, artists and tourists[3]. This example illustrates the inference capability of our method to identify similar neighborhoods given certain thematic characteristics.

(2) Search for similar places considering all themes. After running the LDA topic

[2]`http://realestate.usnews.com/places/rankings/best-places-to-live`
[3]`http://www.denver.org/about-denver/neighborhood-guides/artdistrict-on-santa-fe`

model, each place can be represented as a multinomial distribution of K-dimensional POI type topics, denoted as a probability vector *[p_1,p_2,...,p_k]*, where all the probability values sum to one. Thus we can apply a variety of probabilistic distance or similarity measures (e.g., Hellinger distance, cosine similarity, and Jensen-Shannon divergence (JSD) [134] to quantify the pairwise similarity among all search regions in our POI data with regard to their POI type mix distributions. JSD is a symmetric distance measure derived from the Kullback-Leibler divergence (KLD) asymmetric distance measure between two probability distributions P and Q [98].

$$KLD(P|Q) = \sum_i P(i) log \frac{P(i)}{Q(i)} \tag{3.3}$$

$$M = \frac{P+Q}{2} \tag{3.4}$$

$$JSD(P|Q) = JSD(Q|P) = \frac{KLD(P|M)}{2} + \frac{KLD(Q|M)}{2} \tag{3.5}$$

The JSD is bounded by 0 and 1 if using the base 2 logarithm for the two KLD relative entropy calculation. And thus we can define a JSD-similarity metric ($S_{(JSD)}$) as follows:

$$S_{(JSD)} = 1 - JSD(Q|P) \tag{3.6}$$

where the base 2 logarithm is used in the KLD and JSD calculations.

Therefore, according to the proposed similarity measure $S_{(JSD)}$, we can analyze the pairwise similarity among our randomly selected search places that contains those POIs. Figure 3.9 shows a JSD-similarity matrix for 200 randomly selected places in Los Angeles, derived from one part of our whole dataset in Section 3.3, and each place is represented as a vector of 130-dimensional thematic topics. The similarity score in each grid is between

0 and 1. The higher the value, the more similar the two places are with regard to their topic distributions. The values in the diagonal are all 1. By visualizing this similarity matrix, one can easily identify two anomalous red stripes (i.e., the labeled R_{th} row and the C_{th} column) with relatively low similarity values across the grid cells. Interestingly, as shown in Figure 3.10, further investigation reveals that this place was sampled at a location inside the *Disneyland Resort* in the Los Angeles metropolitan area, which is very different from all other randomly sampled places and the dominant topic (*Topic 56*) has unusual POI types such as *theme park, theme park ride/attraction, and gift shop*. The frequent co-occurrence of those distinctive types of POIs in this region causes the very low similarity to all other places. Thus, given any place, we can find the most similar or dissimilar places in another geographic region based on this similarity matrix.

3.5.3 Discovering Functional Regions

Another goal for us is to discover urban functional regions where semantically similar places group together. As described in Section 3.4, two clustering methods are applied for aggregating similar places into functional regions. Figure 3.11 shows the K-means clustering result for 200 randomly sampled places in Los Angeles. The *silhouette* value for determining the optimal number of clusters in Los Angeles is 15, and thus we group those places into fifteen clusters. The circles with the same color on the map belong to the same cluster within which POI structures are more similar in their topic space of types. The numeric label on the top of each circle displays the top ranked topic that has the largest probability over the 130-dimensional thematic topic vector in this location. Note that purely K-means clustering doesn't consider any spatial constraints, and thus distant places sharing similar functions or thematic characteristics can also be grouped into the same cluster. For example, several places are dominated by food-related *Topic 30*, which

contains frequent distributions of various restaurants such as *Korean restaurant, Mongolian restaurant, Portuguese restaurant, and Polish restaurant* are grouped into *Cluster 8*, although those places are spatially separated. However, if we take spatial constraints into considerations, only places that are semantically similar in the topic space and also located near each other can be aggregated into the same cluster. Figure 3.12 shows the Delaunay-triangulation-spatial-constraints clustering result for those 200 sampled places in Los Angeles. Note that we keep the same color scheme for the visualization of two clustering results, but those clusters in the same color from two maps are not identical. In the West Coast area, we can see that several places are dominated by the beach *Topic 109* and related leisure activity categories are spatially clustered together into *Cluster 7*. Although *Cluster 7* and *Cluster 3* are spatially close and share beach characteristic, *Cluster 3* tends to have another dominant POI type (*shopping plaza*) in this region, which distinguishes it from *Cluster 7 and Cluster 10*.

By analyzing the spatial distribution of similar places and clusters, researchers can have a better understanding of how certain types of POIs co-locate in order to serve different urban functions from the bottom-up perspective. In addition, urban planners or managers are able to further investigate the needs for complementary physical facilities and services related to the thematic characteristics derived from the human activities on the location-based social networks. This keeps in line with the human-centered and community-oriented perspectives in traditional top-down urban planning and design. Furthermore, we create the bounded functional regions as *convex polygons* derived from those points in the same cluster (Figure 3.12). This can help geographic information service providers develop topic-related POI search services within certain functional regions. Because the POI type assignments for all topics are semantically interpretable, we can also select multiple dimensions of topics in geographic information queries such as the *beach + shopping plaza* topics. *Cluster 3* (in Figure 3.12) would be a good candidate

since it has a mix of the dominant beach topic and the shopping topic.

In addition, in order to test the robustness of discovered urban functional areas with different probabilistic topics, we perform a series of clustering result comparisons by choosing different numbers of topics ranging from 10 to 150. We use their corresponding probabilistic POI type compositions as clustering features and run both K-means clustering and the Delaunay-triangulation-spatial-constraints clustering. Two popular metrics for comparing clustering results are applied in our tests: the *Rand index* [135] and *normalized mutual information (NMI)* [136]. The Rand index measures the percentage of decision agreements between two clustering results X and Y. It contains two types of decision agreements: (1) the number of pairs of search locations within the same clustering region in X that are also in the same clustering region in Y; (2) the number of pairs of search locations that are in different clustering regions in X and also in different clustering regions in Y. The NMI quantifies the mutual dependence/similarity between two clustering results using information theory. The detailed formula descriptions can be found in the original article [136]. For both the Rand index and NMI, their value range is between 0 and 1 and larger values indicate higher similarity between two clustering results. Figure 3.13 shows the K-means clustering comparisons using the Rand index and the NMI metric between the target scenario (130 topics and 15 clusters) and other scenarios with different number of topics but with the same total number of clusters. Figure 3.14 shows the comparison results for the Delaunay-triangulation-spatial-constraints clustering in a similar manner. We find that the Rand index keeps a high value around 0.85 for both clustering methods, which indicates a large percentage of agreements on the clustering membership of those search locations and derived functional areas. But the NMI values show a fluttering pattern that indicates the existence of cluster membership variability. Furthermore, as for both evaluation metrics, the Delaunay-triangulation-spatial-constraints clustering has a higher similarity value in most comparison scenarios

and seems to be more stable than the K-means clustering results. It may imply that the spatial constraints play a role in deriving the functional regions.

3.6 Conclusion

In this work, we develop a statistical framework that applies the LDA topic modeling technique and incorporates user check-ins on LBSN in order to help discover semantically meaningful topics and functional regions based on co-occurrence patterns of POI types. The "functions" derived from probabilistic topic modeling techniques can reveal the latent structure of POI mixtures and the semantics of places. Based on a large corpus of about 100,000 Foursquare venues and check-in behavior in the ten most populated urban areas in the U.S., we demonstrate the effectiveness of proposed methodology by identifying distinctive types of latent topics and further, by extracting urban functional regions using the K-means clustering and the Delaunay triangulation spatial constraints clustering methods. A region can have multiple functions but with different probabilities, while the same type of functional region can span multiple geographically non-adjacent locations. Compared with the remote sensing images that mainly uncover the physical landscape of urban environments, results derived from the popularity-based POI topic model can be seen as a complementary social sensing view of urban space based on human activities and the place settings of urban functions. However, there may exist gaps between the real-world business establishments and the online available POI information. Data-fusion and cross-validation relying on multiple sources may help reduce such gaps.

Although we have successfully identified several types of semantically meaningful urban functional topics, LDA topic modeling is an unsupervised approach that has certain limitation with respect to discovering plausible urban functions. In the future, we plan to investigate additional semantic signatures such as those incorporating the spatial pat-

terns of POI distributions and using supervised-versions of probabilistic topic models to compare the performance of two families of topic models (unsupervised or supervised) in discovering urban functional regions. Last but not least, we also aim at developing a functional region ontology by combing the data-driven approach as outlined in this work with the top-down knowledge engineering approach based on our understanding of urban functional regions from human geography and urban planning.

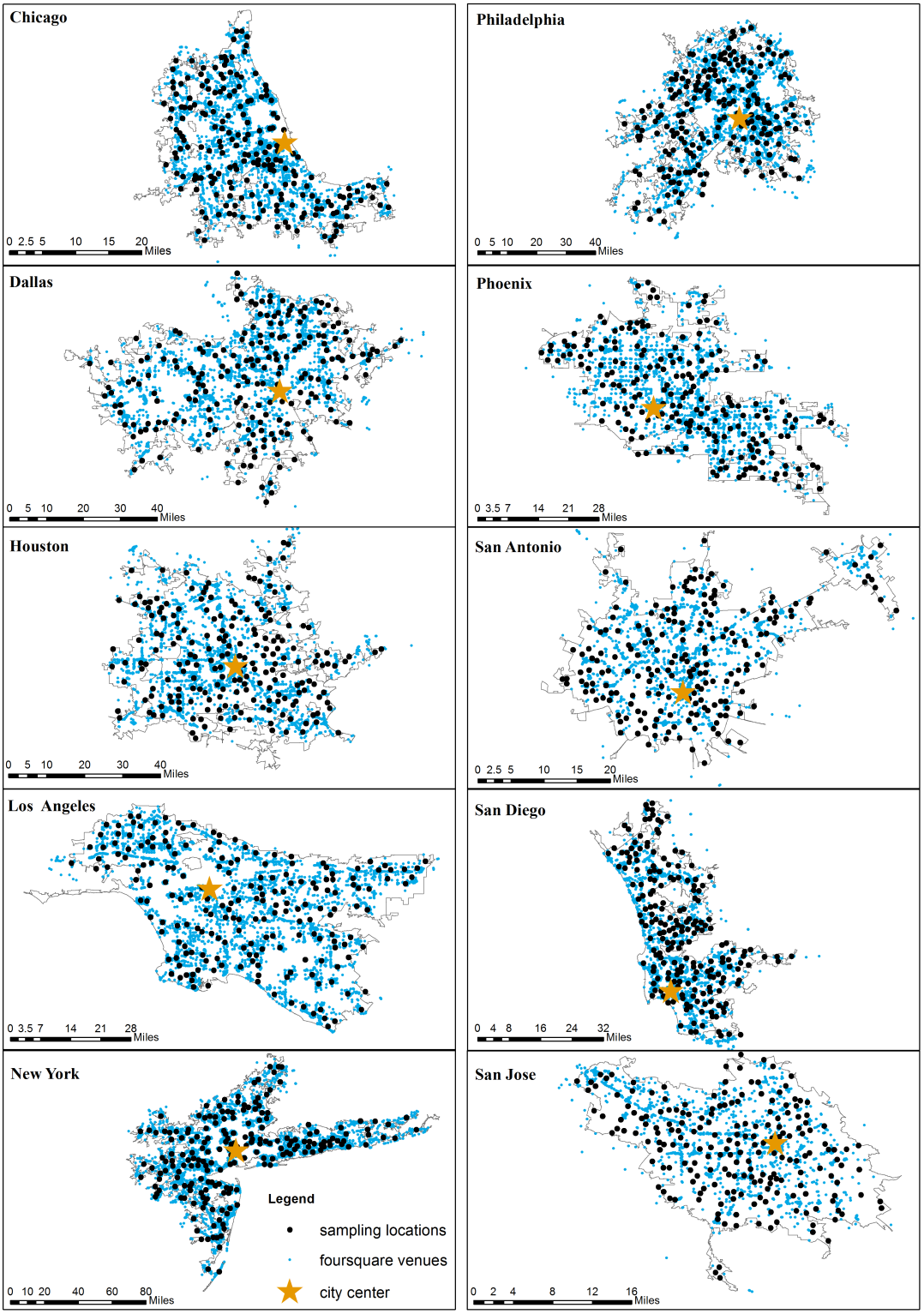


Figure 3.1: The spatial distributions of sampling locations and the collected Foursquare venues (POIs) in ten urban areas.

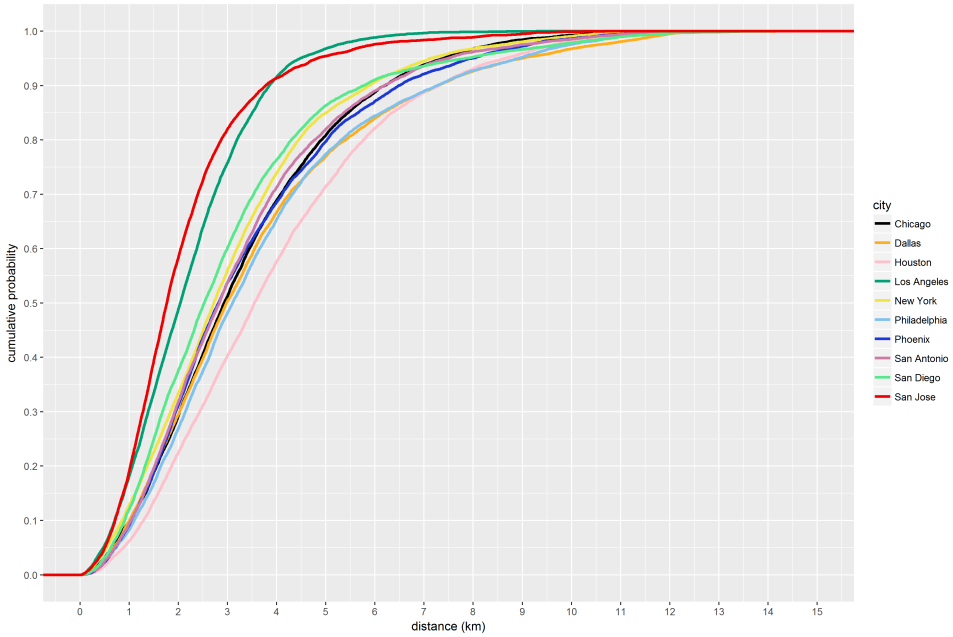


Figure 3.2: The cumulative density function of the distance distribution between each Foursquare venue and the corresponding search location.

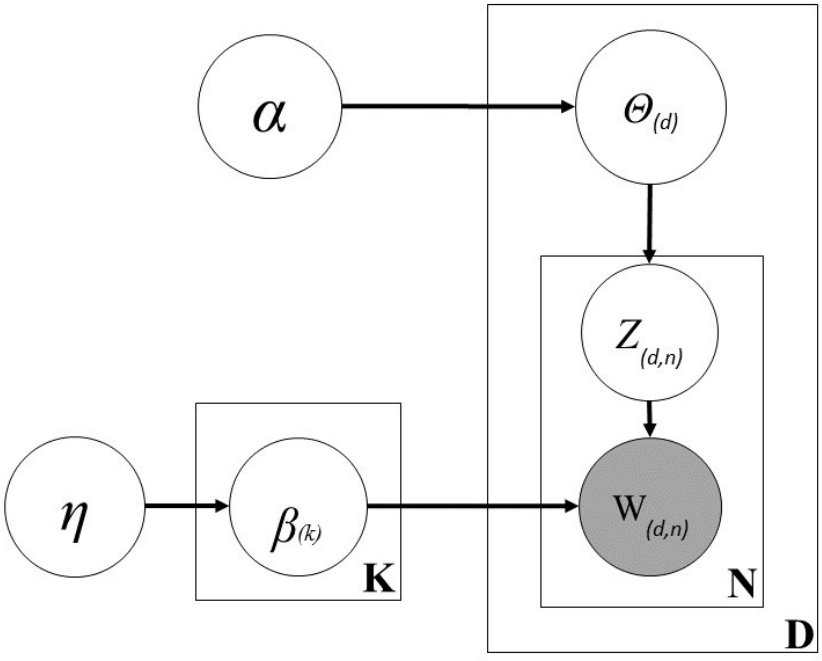


Figure 3.3: The graphical representation for latent Dirichlet allocation. The topic-related random variables in the generative process are the unshaded nodes while the observed words in documents are represented as a shaded node. The rectangles are "plate" notation that denotes replication.

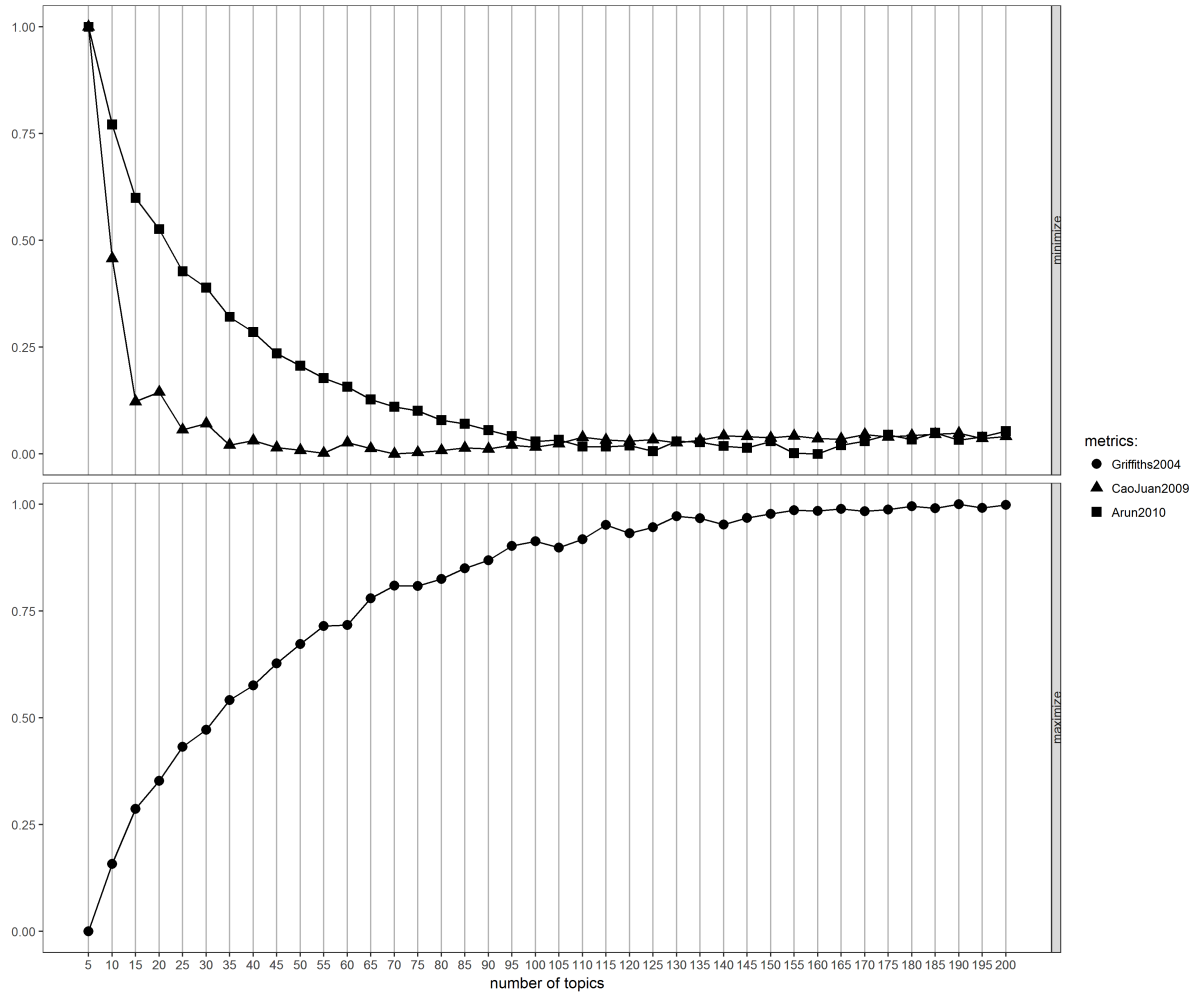


Figure 3.4: Find an appropriate K value for the total number of topics using three metrics.

Topic 6	
Category	Prob.
gas station	0.309651
italian restaurant	0.028574
flower shop	0.013235
national park	0.002077
ski lodge	0.000968
jewish restaurant	0.000899
auditorium	0.000847
southern food	0.000226
ice cream shop	0.000123
farmers market	0.000109
club house	0.000105
bbq joint	0.000100
pizza place	0.000100
winery	0.000072
grocery store	0.000059

Topic 21	
Category	Prob.
pool	0.122166
history museum	0.047285
historic site	0.043474
college basketball court	0.015107
concert hall	0.012485
art museum	0.012412
college rec center	0.010488
park	0.007814
sculpture garden	0.007216
outdoor sculpture	0.005112
college soccer field	0.004028
college cafeteria	0.003933
tourist information center	0.003731
molecular gastronomy	0.003120
stables	0.002947

Topic 25	
Category	Prob.
museum	0.065636
art museum	0.047585
art gallery	0.046691
american restaurant	0.038677
record shop	0.025063
antique shop	0.024183
building	0.002540
cycle studio	0.001103
health food store	0.000462
history museum	0.000430
soup place	0.000350
concert hall	0.000281
scenic lookout	0.000264
animal shelter	0.000179
burger joint	0.000179

Topic 36	
Category	Prob.
yoga studio	0.105001
science museum	0.065819
boutique	0.029987
gay bar	0.015371
sculpture garden	0.012688
government building	0.008197
israeli restaurant	0.004401
apartment / condo	0.003005
pakistani restaurant	0.002829
street food gathering	0.001212
track stadium	0.000872
college baseball diamond	0.000602
mexican restaurant	0.000542
gym / fitness center	0.000526
cheese shop	0.000481

Topic 67	
Category	Prob.
shopping mall	0.207709
accessories store	0.056738
chocolate shop	0.013896
shoe store	0.000288
breakfast spot	0.000282
gaming cafe	0.000196
optical shop	0.000180
post office	0.000114
bistro	0.000105
dumpling restaurant	0.000096
korean restaurant	0.000090
german restaurant	0.000080
herbs & spices store	0.000079
airport terminal	0.000078
outlet store	0.000076

Topic 74	
Category	Prob.
bar	0.511221
board shop	0.000046
asian restaurant	0.000046
brewery	0.000036
ice cream shop	0.000030
parking	0.000025
buffet	0.000025
business service	0.000019
apartment / condo	0.000016
fried chicken joint	0.000012
resort	0.000007
gourmet shop	0.000007
lighthouse	0.000006
indian restaurant	0.000006
train	0.000006

Topic 109	
Category	Prob.
beach	0.285864
surf spot	0.028952
italian restaurant	0.015458
island	0.005920
beach bar	0.004078
board shop	0.003793
bridge	0.001484
indie theater	0.001235
pier	0.001187
outdoor sculpture	0.001121
sri lankan restaurant	0.001040
bistro	0.000891
nature preserve	0.000851
arepa restaurant	0.000751
neighborhood	0.000726

Topic 117	
Category	Prob.
italian restaurant	0.082055
fast food restaurant	0.000131
gym	0.000064
golf course	0.000056
sushi restaurant	0.000044
salon / barbershop	0.000043
boutique	0.000034
café	0.000030
szechuan restaurant	0.000030
japanese restaurant	0.000030
paella restaurant	0.000027
men's store	0.000023
caribbean restaurant	0.000017
deli / bodega	0.000016
massage studio	0.000014

Topic 119	
Category	Prob.
french restaurant	0.090092
cocktail bar	0.072534
lounge	0.035774
tennis court	0.005389
whisky bar	0.003636
american restaurant	0.000697
dry cleaner	0.000168
pizza place	0.000118
café	0.000117
art museum	0.000110
bakery	0.000106
jazz club	0.000082
chinese restaurant	0.000074
neighborhood	0.000068
cycle studio	0.000040

Figure 3.5: Nine interesting topics with their top-15 ranked POI types related to urban functions.

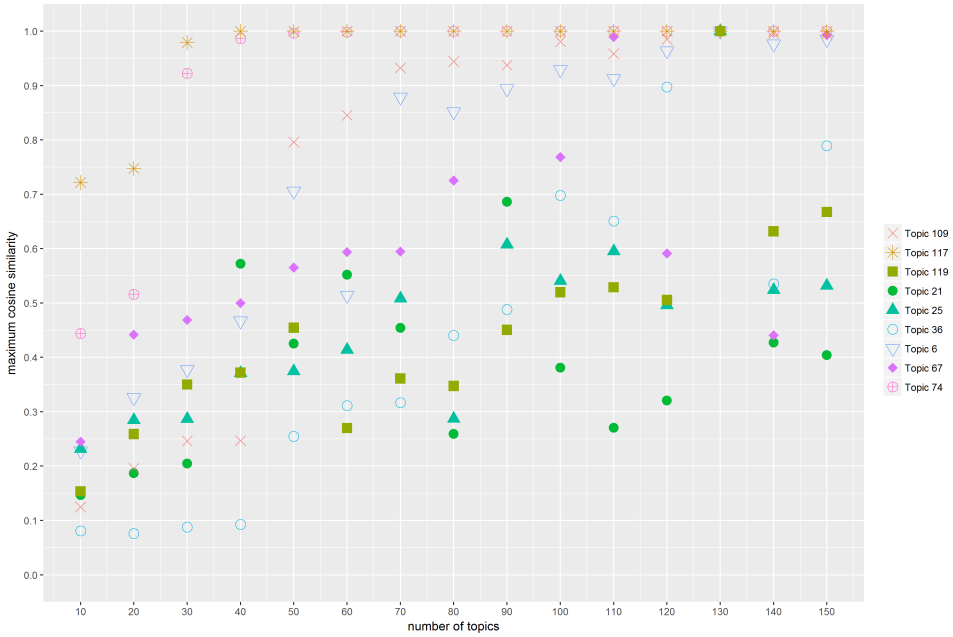


Figure 3.6: Maximum cosine similarity between the selected nine topics and the resulting topic models by choosing different total number of topics.

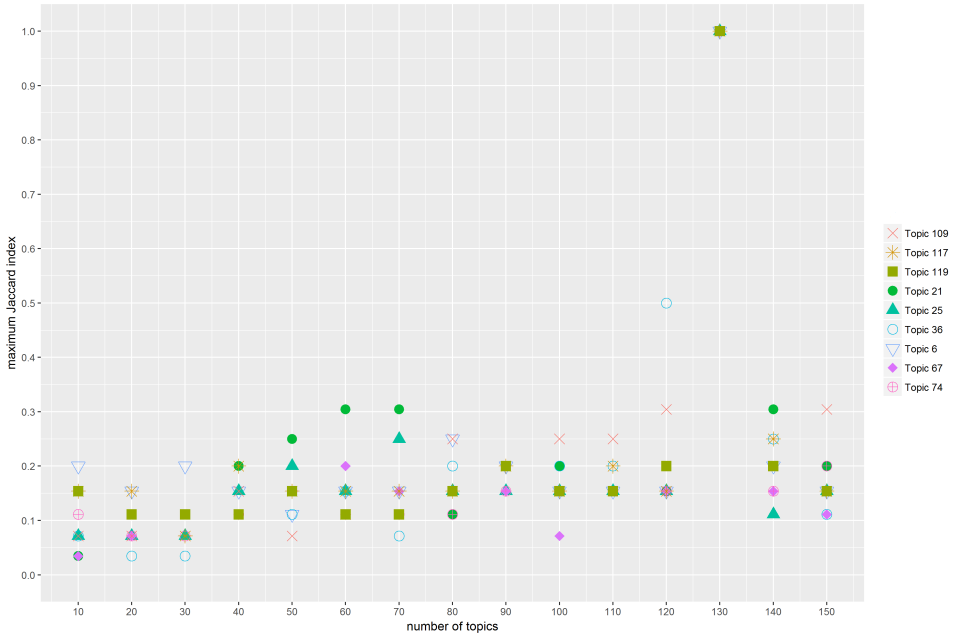


Figure 3.7: Maximum Jaccard similarity between the top-15 POI types of the selected nine topics and that from the resulting topic models by choosing different total number of topics.

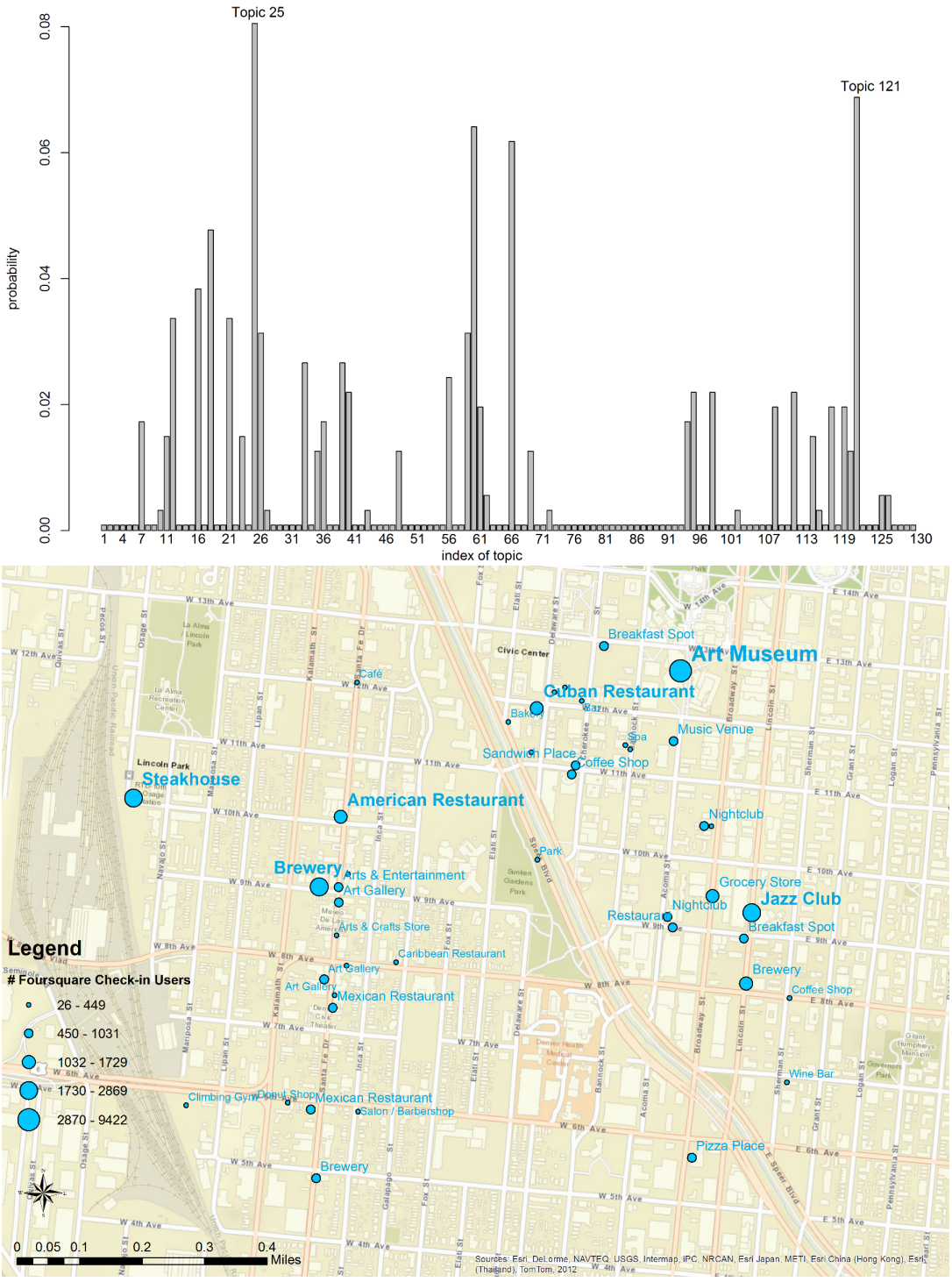


Figure 3.8: The topic probability distribution and the spatial distribution of Foursquare POIs around the Denver art district and museums.

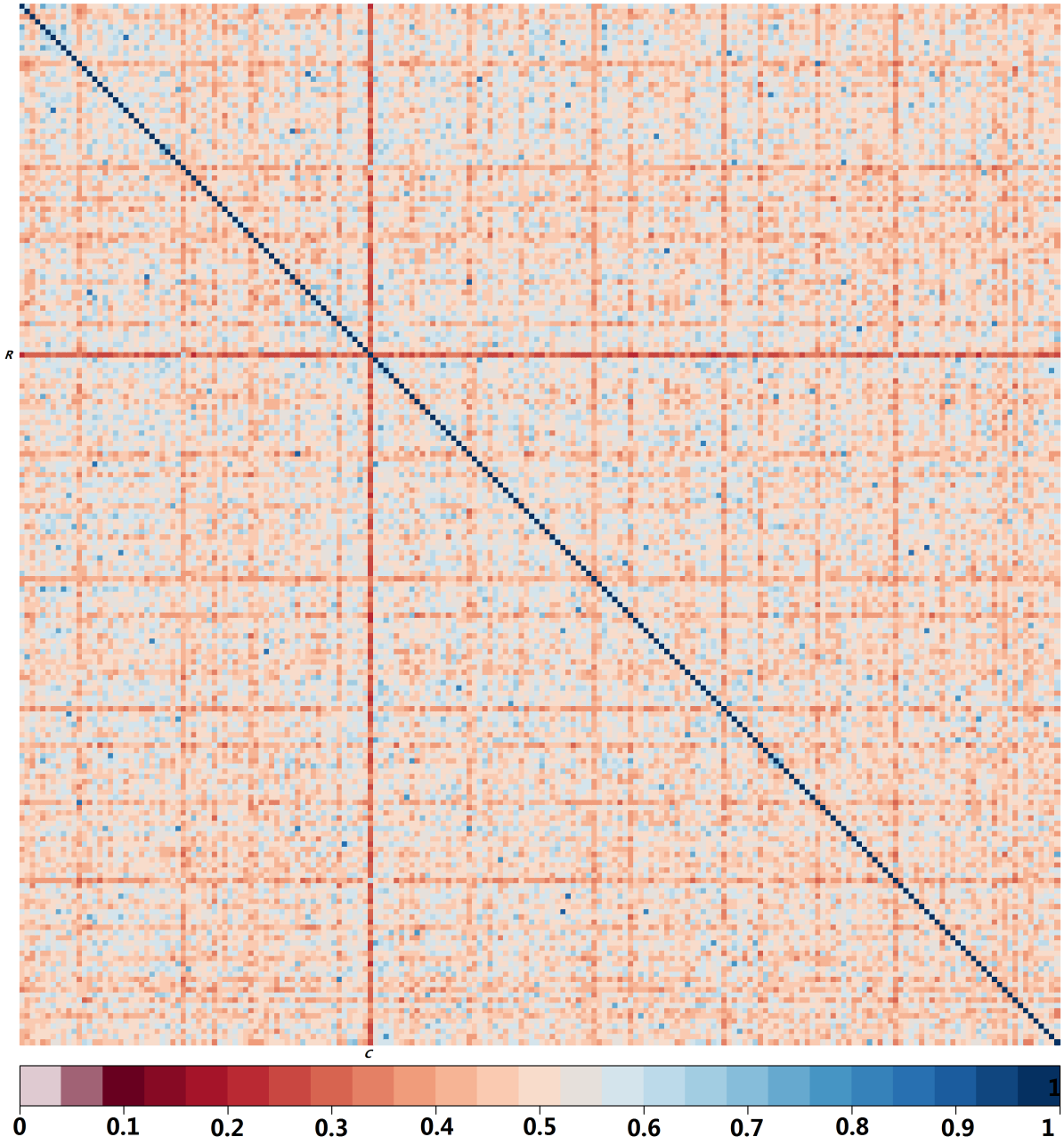


Figure 3.9: The JSD-similarity matrix for 200 randomly selected places in Los Angeles, where each place consists of 130 dimensional thematic topics.

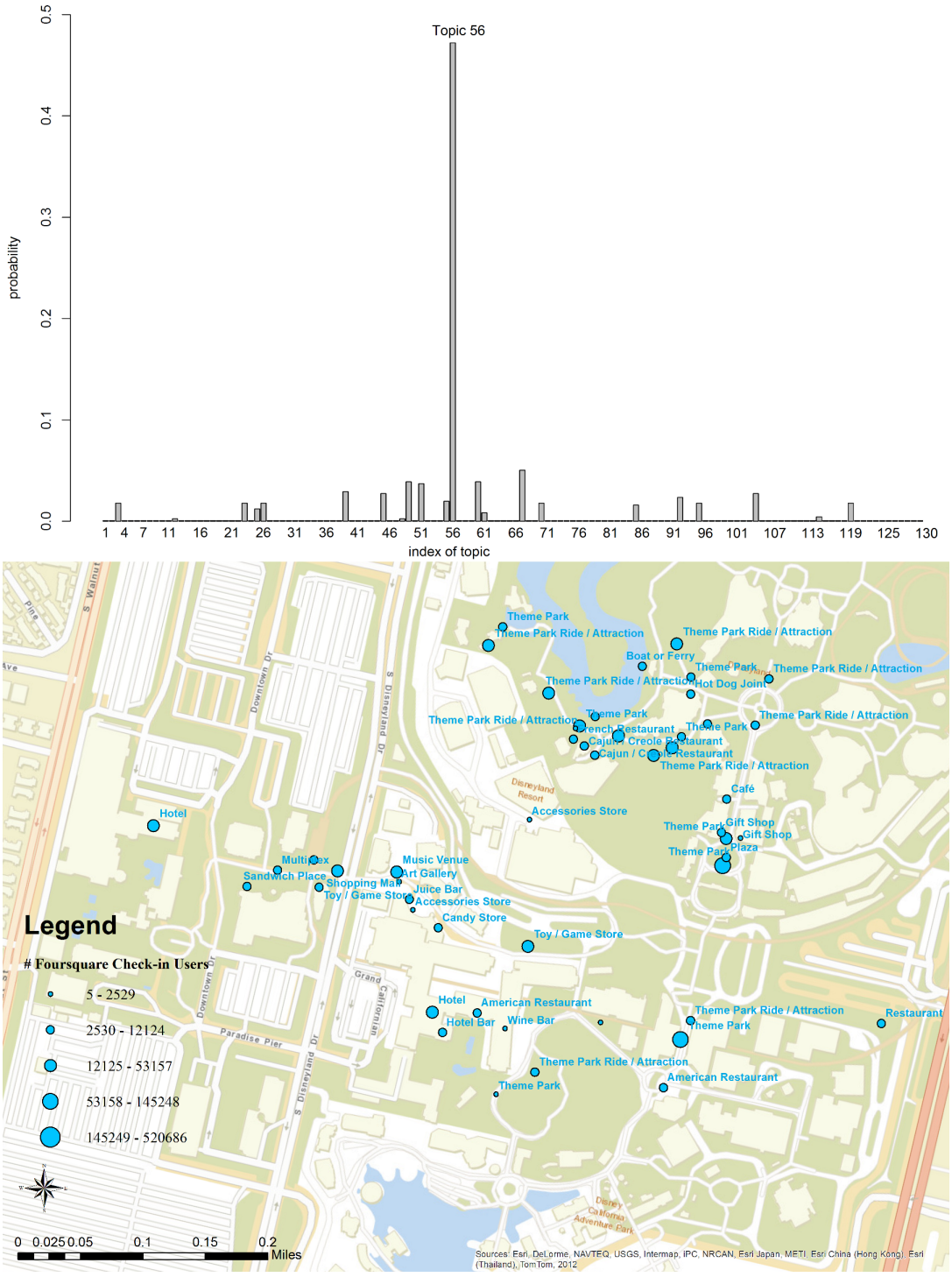


Figure 3.10: The topic probability distribution and the spatial distribution of Foursquare POIs in the Los Angeles Disneyland Resort.

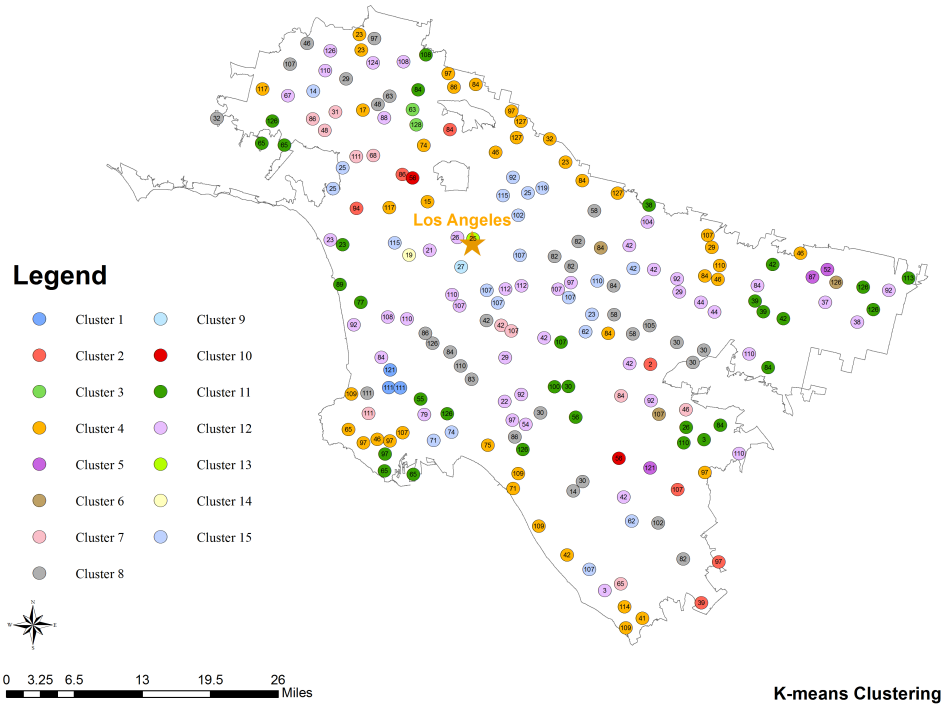


Figure 3.11: The K-means clustering result for 200 sampled places in Los Angeles.

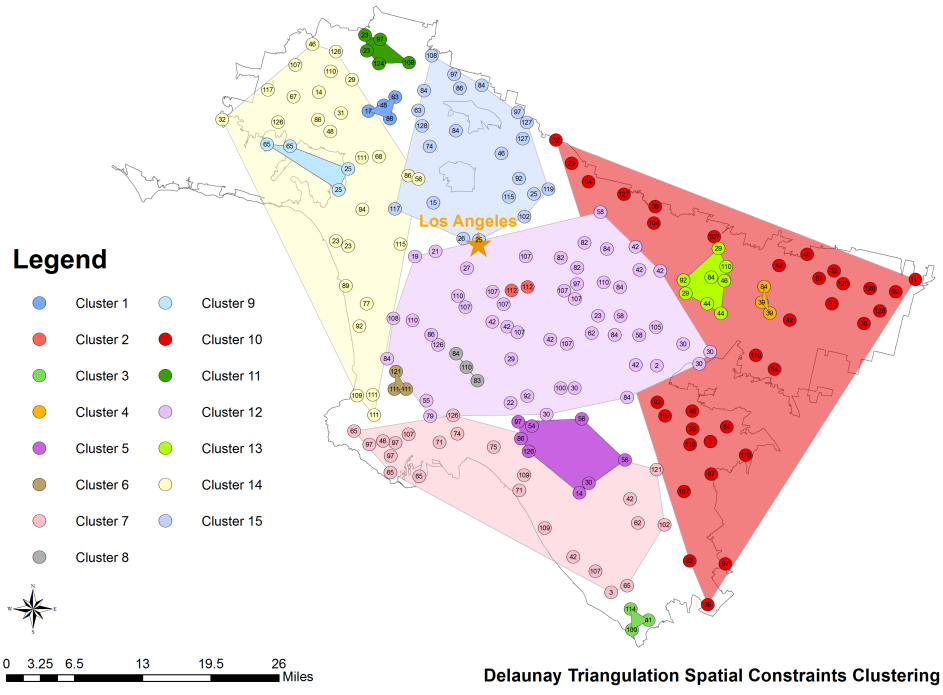


Figure 3.12: The Delaunay triangulation spatial constraints clustering and convex polygon generation result for 200 sampled places in Los Angeles.

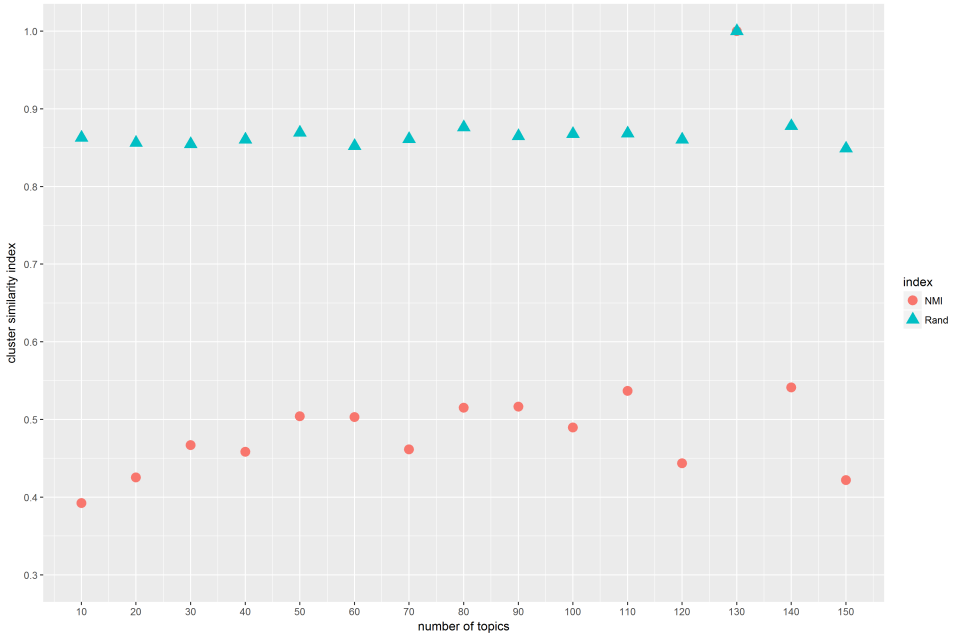


Figure 3.13: K-means clustering similarity evaluation using the NMI and Rand metrics with different number of topics.

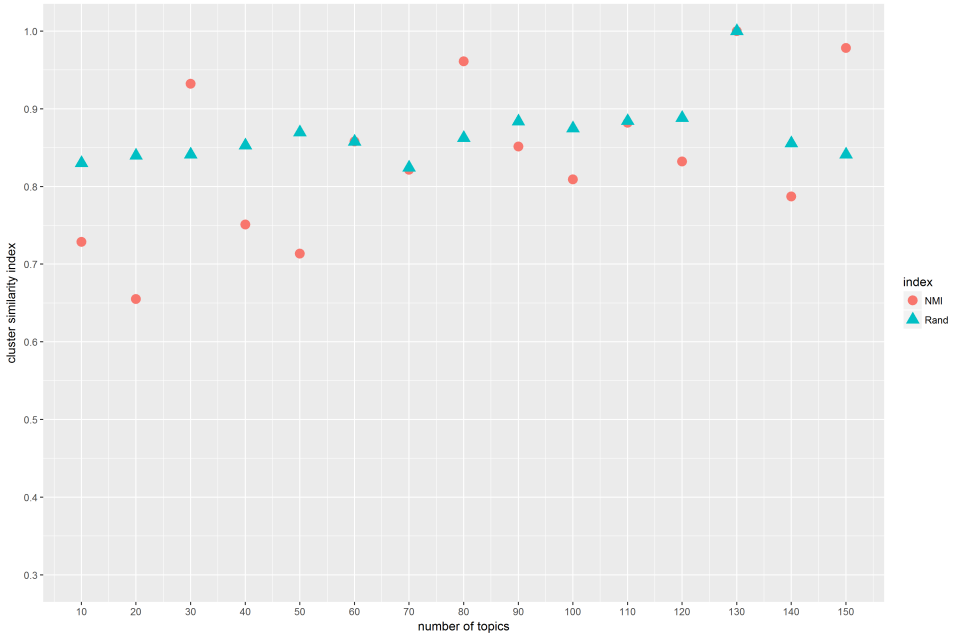


Figure 3.14: Delaunay triangulation spatial constraints clustering similarity evaluation using the NMI and Rand metrics with different number of topics.

Chapter 4

Semantic Generalization of Regions

Chapter 4 presents a semantic generalization framework for converting point-based representation of place into region-based representations witch rich semantics. This research develops a new methodology that can take both spatial distributions of venues and the place-type hierarchical relationships into consideration to derive spatially and semantically coherent high-level generalized regions in order to address the regional/cultural variability by broadening the thematic topics that form the functional regions. While this research focuses on the theoretical contribution, a case study of extracting the semantically generalized regions that relate to the *Beach*, *Shopping*, and *Asian Food* topics in Los Angeles has been conducted using the proposed semantic generalization methodology.

Abstract: Map generalization is the process whereby information is selected and represented on a map for adapting to a specific map scale with not necessarily preserving all intricate cartographic details. Most of existing generalization methods mainly rely on geometry-based operations while the semantics of places and regions may not be well preserved. In this study, we introduce a novel semantic generalization framework operated on points of interest data to generate semantically coherent regions that bridge the gap between the abstract geometries and the human conceptualization of places. The proposed generalization methodology takes both spatial distributions of venues and place-type hierarchical relationships into consideration to derive spatially and semantically coherent high-level generalized regions. While this research focuses on the theoretical contribution, a case study of extracting the semantic regions that relate to the *Beach*, *Shopping*, and *Asian Food* topics in Los Angeles has been conducted with a spatial sampling technique using large-scale Foursquare venue data. The results demonstrate the effectiveness of our proposed semantic generalization framework for converting point-based representation of places into region-based representations with coherent semantics in order to address the regional/cultural variability by broadening the thematic topics that form the functional regions.

4.1 Introduction

Map generalization, or cartographic generalization, is the process whereby information is selected and represented on a map in a way that adapts to the scale of the display medium of the map, not necessarily preserving all intricate geographical or other cartographic details. It bridges the gap between the *scientific* and *artistic* perspective of the cartography [137].

As described in [138],"map makers are human". Every map is a reflection of both

objective realities and subjective elements. It emphasizes the importance of human mind by stating that the map is drawn by human hands but controlled by operations in a human mind. A series of versions of the textbook led by Arthur H. Robinson [139] summarized the cartographic generalization process as having three significant components: (1) to select the objects and features to be shown on the map; (2) to simplify their forms; (3) to evaluate the relative significance of the items to make sure that the appearance of the important items are more prominent. The principle of selection was proposed in [140], which is expressed as an equation relating the number of occurrences of a particular feature at the source map scale and at derived map scale. Later, it has been further developed into the formal structure of generalization and become a standard reference: *simplification, classification, symbolization, induction*. Robinson also speculated on the significance of subjectivity towards the generalization process and distinguished two types of cartographic generalization: one is the *intellectual generalization* which focuses on the selection and organizing the map items and features that need to be portrayed; and the other is the *visual generalization* which addresses the symbolization and visualization effectiveness.

With the invention of computer-assisted digital environment, the digital map generalization continuously attracts cartographers' attention. The computer-assisted generalization and the spatial modeling process can be simulated by strategies based on human understanding and not by a mere sequence of operational processing steps. Thus, a conceptual framework for automated map generalization has been proposed by [141], which consists of five steps: (1) structure recognition; (2) process recognition, (3) process modeling, (4) process execution, and (5) medium display. With the fast development of the geographic information systems, the map display of both vector and raster spatial data needs more holistic investigation regarding the complex generalization conditions, measures, controls and the needs. [142] gave a comprehensive research on a logic framework

of the digital map generalization processing which consists of three main components: (1) the intrinsic objectives of *why* to generalize, (2) the assessment of situation which indicates *when* to generalize, and (3) an understanding of *how* to generalize using spatial and attribute transformation. The goal of the generalization is to reduce the data volume of storing while keeping the key features. However, the map generalization also needs to address the importance of geography which indicates preserving the recognizability of geographical features and their positional accuracy. [143] introduced the conceptual basis for geographic line generalization (e.g., contours, streams, shorelines, roads, railways) and stated that "Geographical generalization must incorporate information about the geometric structure of geographic phenomena."

The past decades of research on digital map generalization mainly rely on the **geometric** properties of spatial data but not considering the **platial effects**. A semantic generalization of space-based maps into place-based maps may bridge the gap between the abstract geometries and human cognition, and help a better understanding of the interaction between human concepts and places.

In the following, we will present a novel methodology for the semantic region generalization by integrating spatial sampling, topic modeling, and platial buffer techniques.

The remainder of this chapter is structured as follows. Section 4.2 reviews the related work. Next, Section 4.3 introduces the proposed semantic generalization framework. In Section 4.4, we present a case study in the urban area of Los Angeles. Finally, we conclude our work and point out directions of future work in Section 4.5.

4.2 Related Work

Due to the proliferation of location-based social network data, researchers started considering the integration of spatial patterns, temporal rhythms, and thematic attributes of

POIs to derive more semantically meaningful regions or zones. In the *Livehoods Project*, [144] introduced a clustering-based research methodology for studying the structure and composition of a city and deriving the neighborhood boundaries based on large-scale Foursquare POI data. [113] presented a topic modeling methodology to estimate geographic regions from unstructured, non geo-referenced text on Wikipedia and travel blogs by computing a density surface of geo-indicative topics over the Earth's surface. The proposed framework combined natural language processing techniques, geostatistics methods, and data-driven bottom-up semantics. [109] proposed a density-based clustering framework for extracting and understanding urban AOI from Flickr geotagged photos and then constructing concave-hull bounding polygons from point clusters. [145] proposed a geo-clustering approach for the detection of area-of-interest (AOI) and their underlying semantics. Based on the Flickr photo tagging data in Greece, they divided the study region into grid cells and merging adjacent similar tagging cells. After extracting those regions, they applied the textual analysis to rank the geo-cluster tags and selected top-ranking ones as representational tags. The results were evaluated based on both quantitative and qualitative metrics from human subject survey.

The harvesting of rich place-based data and associated human activities from social media or VGI is a novel research exploration in the map generalization field. [119] developed a semantic region growing algorithm based on the density of POIs on OpenStreetMap to extract places that afford certain type of human activities, e.g., shopping areas. In their model, four features including the number of *banks & ATM, restaurants, tourist facilities, and subcategories of shops* were used to identify the shopping areas/settings. The growth of regions relies on spatial adjacency and place type similarity.

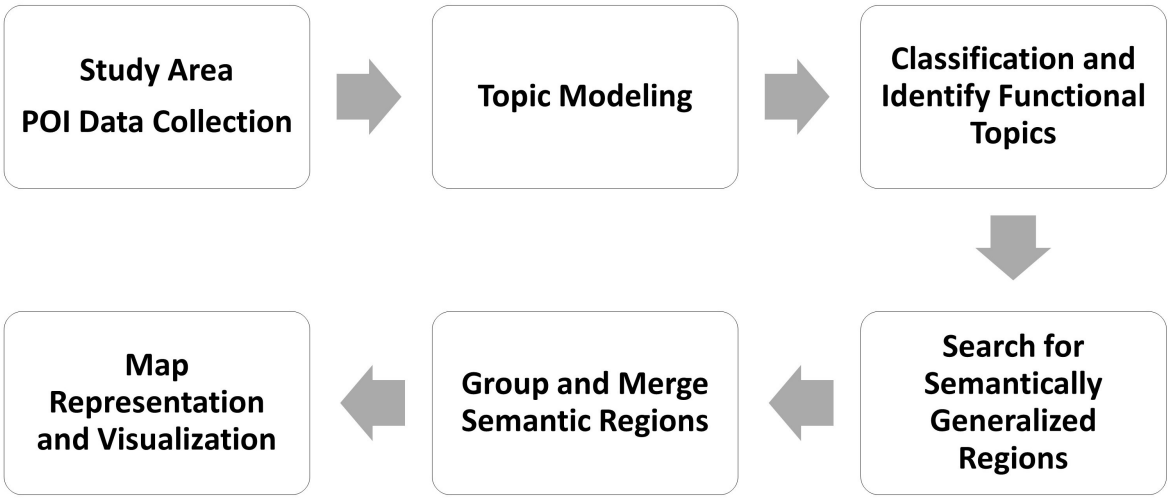


Figure 4.1: The workflow of semantic generalization processing.

4.3 Methodology

In this research, we aim to develop a new methodology that can take both spatial distributions of POIs and their hierarchical relationships into consideration to derive spatially and semantically coherent high-level generalized regions.

To this end, we propose a novel semantic generalization processing framework that can produce semantically coherent high-level generalized regions and their polygonal representations on maps from points of interest (POI) data. The workflow consists of the following six steps (in Figure 4.1):

(1) **POI Data Collection:** Most of popular location-based service providers and mapping companies (such as Yelp, Foursquare, and Google Maps) provide their application programming interfaces (API) for allowing programmers to access millions of POI data but limited to their intrinsic data coverage and API request limits. For a given POI (a venue), we could extract its placename, address, coordinates, categories/types, users' visit statistics, reviews, ratings, and tips, etc. Note that a POI could have one or multiple

place types. For example, a venue in Santa Barbara named as *The Neighborhood Bar & Grill* has three user-assigned place types on Foursquare: *Bar*, *Arcade*, and *General Entertainment* (Figure 4.2). Users think that it provides not only drinks but also pool tables and ping-pong tables for entertainment activities. Human activities could happen at one type or multiple types of POIs. Such crowdsouced POI data reflect what kind of places that people visit but only those available with the digital format of POI databases.

In the sampling process, generally speaking, there are two approaches for collecting POI data. The first approach divides the study area into regular grid cells with the same size (e.g., 100-meter side length) and utilizes the centroid of each cell as the search location for requesting the nearby available POI from data providers. The other approach randomly generates a large number of points with coordinate information spatially bounded by the study area and then uses them as search locations for nearby POI queries. Note that because of the POI data provider API access limits, we can only retrieve certain number (e.g., 50 in Foursquare) of nearby venues given a random search point. Thus, we need to generate large number of random search points to fully cover the study area. The retrieved raw data can be further uniquely filtered and selected based on the POI unique identifier and the spatial boundaries of urban areas. A unique set of POI data with attribute information is generated after the preprocessing for the study area.

(2) **Topic Modeling:** As introduced in Chapter 3, in analogy to topic modeling of textual documents in natural language processing, we take the place type (e.g., restaurant, pub, park) of each POI as a word, a region that contains those POIs as a document, and an urban function or a land use as a topic that represents thematic characteristics and the semantics of place. By running the LDA topic modeling technique, we can find the posterior probabilistic distribution of each POI type in a given region conditioned on the search region's topic assignments. The LDA model running on POI types generates

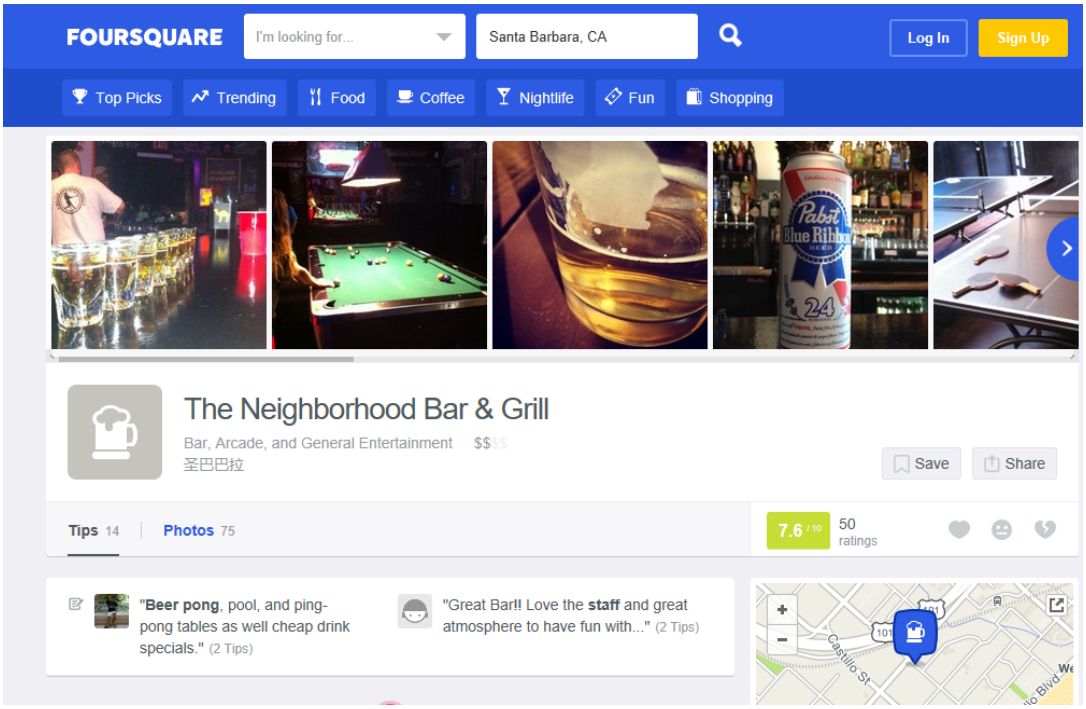


Figure 4.2: The screenshot of a POI named as *The Neighborhood Bar & Grill* on Foursquare.

summaries of thematic place topics (e.g., beach promenades, art zones, shopping areas) with a discrete probability distribution over POI types for each topic, and infers per-region multi-probability distributions over topics. For example, one would assume that a *beach promenade* topic should contain venues such as *beach*, *seafood restaurants*, and *surfing spots*; while a *shopping area* would more likely contain *clothing stores*, *cosmetics shops*, and *shoe stores*.

Another important concept in our method is the *popularity* of a POI as captured by its LBSN user check-in behaviors. For example, the neighborhood of a football stadium usually has only one instance of *stadium* surrounded by dozens of *sports bars*, *restaurants*, and *parking lots*. However, a stadium usually attracts thousands of visitors and is the dominant feature of its neighborhood. Thus this particular POI type makes said neighborhood distinct from other neighborhoods (e.g., nightlife zone), which also contain *cocktail bars* and *restaurants*. We need to address such a human activity effect during the

generation of the document-word frequency matrix. More specifically, we will rescale the POI type occurrences according to their associated POI instance check-in counts. The rescaling process can be represented as follows:

$$Freq_{(d,t)} = \lceil \sum_{i} Log(V_{(d,t,i)}) \rceil \tag{4.1}$$

where $Freq_{(d,t)}$ represents the rescaled occurrence for a POI type t given a search region d; and $V_{(d,t,i)}$ is the number of unique users who have contributed their check-ins for a venue i that belongs to the same POI type t in the same search region d. The ceiling function $\lceil \sum_i Log(V_{(d,t,i)}) \rceil$ gets the least integer greater than or equal to the value of $\sum_i Log(V_{(d,t,i)})$.

Furthermore, one advantage of using topic modeling techniques is that researchers can define their own vocabulary for model training such that derived topics only reveal a pre-defined vocabulary (i.e., place types in our case).

(3) **Identify Functional Topics and Clustering:** After running the LDA topic modeling on place documents that contain a variety of POI type information, we can identify those semantically meaningful topics over a set of different POI types. Note that the probability assignments for those POI types are weighted and ranked by their *term frequency−inverse document frequency* (i.e., POI type frequency−inverse region frequency) so that each topic can display more distinctive and meaningful POI types that are directly proportional to the frequency in documents while inversely proportional to the region frequency at which a POI type occurs in the whole corpus. For instance, *coffee shop* has a very high frequency but also widely exists in most of the regions in our POI data so it plays a less important role than other categories (e.g., *theme park*) in distinguishing the function of a region.

After deriving those latent thematic topics by running the proposed popularity-based

LDA model, each region document can be represented as a vector of multi-dimensional POI type topics. Those regions that are semantically similar in the topic space might contribute to the same urban function and can be aggregated into the same cluster as a functional region. Their mean topic probabilistic distributions can be taken as the representative vector $(\overline{P}_{topic1}, \overline{P}_{topic2}, ..., \overline{P}_{topicK})$ of sampled topics for a theme (e.g., shopping plaza), which will also be used in the semantic generalization process.

(4) **Search for Semantically Generalized Regions:** Searching for similar places is an important task in geographic information retrieval and also valuable in many applications, such as in tourist map exploration. After running the LDA topic model, each place can be represented as a multinomial distribution of K-dimensional POI type topics, denoted as a probability vector *[p_1,p_2,...,p_k]*, where all the probability values sum to one. Thus we can apply a variety of probabilistic distance or similarity measures (e.g., Hellinger distance, cosine similarity, and Jensen-Shannon divergence (JSD) [134] to quantify the pairwise similarity among all search regions in our POI data with regard to their POI type mix distributions.

One challenge is finding the semantically generalized regions with similar topics. For example, because of the natural weather conditions, one can find a winery bars & art gallery neighborhood in California such as *the Funk Zone*[1] in the city of Santa Barbara but may not find exactly the same type of neighborhood in the city of Columbus, Ohio. However, if replacing 'winery' by 'brewery', we can find a very similar neighborhood: *the Franklinton* that also consists of a dozen of brewery venues, bars, and artist places. One common characteristic of the two regions is that both have a variety of drinking spots for human social interactions and make people relaxed and immersed in such an art district with inspiration. These place types, including *Wine Bar, Brewery, Cocktail Bar etc.*, all

[1] `https://funkzone.net/funk-zone-directory/`.

Table 4.1: The ten top-level venue categories in Foursquare and the number of subcategories

Type	Subcategories	Type	Subcategories
Arts & Entertainment	64	Outdoors	104
College & University	39	Professional & Other Places	105
Event	11	Residence	6
Food	341	Shop & Service	170
Nightlife Spot	25	Travel & Transport	54

belong to the generalized type: *Nightlife Spot* in the Foursquare category schema[2].

The hierarchical structure exists in all the POI database schemata, which allows us to build the semantic relationships between place types such as the parent-child relation and the sibling relation in a hierarchical tree of place types. In Foursquare, there are over 900 POI types and can be generalized into ten top-level (more abstract) categories accessed in January 2017 as shown in Table 4.1. The *Food* type has the most subcategories (341). Its hierarchical diagram is visualized in Figure 4.3. We find that the tree structure of place types is also heterogeneously organized in which different subtrees have varying levels (depths) of children nodes. For example, *the Latin American Restaurant* has four levels of subcategories while *the Asian Restaurant* has three levels of subcategories.

In this work, we propose a semantic generalization approach based on the place type hierarchy and topic modeling techniques. In [146], a novel place-based operator: *platial buffer* was introduced. It is based on the topological distance (connectivity or hierarchy) and semantic relations between places. Rada et al. (1989) introduced a conceptual distance between two entities by counting the number of links in the shortest path on the semantic net [147]. Given such a distance, if someone is interested in higher-level interlinked place types, we can derive this information based on the *platial buffer* operation for identifying n-degree connected place types in a place graph. The n-degree represents the number of links that connect the two target entities using the shortest

[2]`https://developer.foursquare.com/categorytree`.

path on a semantic graph. Here, we define a different version of the n-degree buffering based on the POI type hierarchy. The degree n is the steps for a target node moving from its current level to the higher n level in a tree [148]. As shown in Figure 4.4, for a set of nodes [A, B, C, D, E, F, G] in four hierarchical levels (0,1,2,3), 0 represents the root level; the smaller the numeric value is, the higher the hierarchy level is for all nodes on this level. The root-node A represents the most abstract place type while node B and node C are subcategories of A, so on and so forth. The path for node D and node E to reach B is just 1-step upper move. Therefore, if we apply the first-degree generalization buffer on node D or node E, we should get the place type of node B since they are directly connected parent-child nodes. The generalization process 'buffer' is always towards the more abstract place type. Accordingly, we should reach the root-node A by applying the second-degree generalization buffer on node D or node E from level 2 to level 0.

The *scale* is an important concept cartographic generalization. In analogy to it, during the training phase for topic modeling, we can pre-define the level of detail on the POI type hierarchy and *buffer* those lower-level place types to the high-level types that are within the required generalized levels. Formally, we can represent this processing as follows:

$$\left\{ \begin{array}{r} B \sqsubseteq A, \\ D \sqsubseteq B, \\ D(i), \end{array} \right\} \vdash B(i), A(i). \tag{4.2}$$

It means that if we know that a POI instance i ("Sichuan Spicy Food") belongs to a place type D (Chinese Restaurant); place type D is a subcategory of place type B (Asian Restaurant); and place type B is a subcategory of place type A (Food). We can infer that the POI instance i also belongs to higher-level place type B in the hierarchy level $L1$ and place type A in the hierarchy level $L0$ (Figure 4.4). After the preprocessing, all

the POI types are generalized to specific levels in the hierarchical tree and then trained using LDA topic modeling.

In the phase of searching for generalized regions, we first select a referencing region topic vector $\overrightarrow{\mathbf{R}} : (\overline{P}_{topic1}, \overline{P}_{topic2}, ..., \overline{P}_{topicK})$ with its representative topic distributions for a theme (e.g., shopping plaza). One search challenge is that those semantic regions might have varying sizes and thus the classic spatial buffer operation with a fixed search distance might not be appropriate for our purpose. After investigation, we take the Openshaw's spatial sampling technique [149] to generates a large number of search circles with random sizes, and places them (throws them) randomly over the study area. Those POIs that spatially fall into a search circle will be the candidates. Then their associated place types will be used as words for the search-region document. Based on the trained LDA topic model generated in the previous step, the topic distribution $\overrightarrow{\mathbf{S}}$ for a search region can be derived. Before comparing the topic similarity between the search region topic vector $\overrightarrow{\mathbf{S}}$ and the referencing region topic vector $\overrightarrow{\mathbf{R}}$, we also take the area size of the search circle as an adjustment factor into consideration. The key idea is that we hope to generate the regions to be as homogeneous as possible in the topic space. Larger regions that usually contain richer place types should be more restrictive comparisons. Only those large-size regions that have a higher similarity to the referencing topic vector than that of small-size regions could stay. Mathematically, we define the similarity function to control the sampling process:

$$Similarity(\overrightarrow{\mathbf{R}}, \overrightarrow{\mathbf{S}}) = 1 - JSD(\overrightarrow{\mathbf{R}}, \overrightarrow{\mathbf{S}}) - f(r) \tag{4.3}$$

$$KLD(\overrightarrow{\mathbf{R}} | \overrightarrow{\mathbf{S}}) = \sum_i \overrightarrow{\mathbf{R}}(i) log_2 \frac{\overrightarrow{\mathbf{R}}(i)}{\overrightarrow{\mathbf{S}}(i)} \tag{4.4}$$

$$\overrightarrow{\mathbf{M}} = \frac{\overrightarrow{\mathbf{R}} + \overrightarrow{\mathbf{S}}}{2} \tag{4.5}$$

$$JSD(\overrightarrow{\mathbf{R}}|\overrightarrow{\mathbf{S}}) = JSD(\overrightarrow{\mathbf{S}}|\overrightarrow{\mathbf{R}}) = \frac{KLD(\overrightarrow{\mathbf{R}}|\overrightarrow{\mathbf{M}})}{2} + \frac{KLD(\overrightarrow{\mathbf{S}}|\overrightarrow{\mathbf{M}})}{2} \tag{4.6}$$

$$f(r) = C - e^{-r}, (0 \leqslant C \leqslant 1) \tag{4.7}$$

The Jensen-Shannon divergence (JSD) is a symmetric distance measure derived from the Kullback-Leibler divergence (KLD) asymmetric distance measure between two probability distributions [98].The JSD is bounded by 0 and 1 if using the base 2 logarithm for the two KLD relative entropy calculation. The original similarity without considering the area adjustment factor can be defined as (1-JSD). The area adjustment factor $f(r)$ for a circle can be defined using the radius r and a constant parameter C with a range of [0,1]. The exponential component e^{-r} makes sure that the derived value can be only within the range of [0,1]. In applications, the constant parameter C can be tuning using labeled ground truth data. After the processing, only those search regions that meet the requirement of similarity threshold δ could keep.

(5) **Group and Merge Semantically Interlinked Regions:** After running the search step, a set of regions that meet the topic similarity criteria might spatially overlap, we can dissolve those regions that are similar to the same topic theme (e.g., beach promenades, shopping plaza). In addition, we can also merge multiple semantically interlinked topics into one, such as the $ThemePark$ and $Hotel$ topics as tourist regions.

(6) **Map Representation and Visualization:** In order to generate the final semantic generalization map, the transparency setting is applied for each map layer and the geographic background information as a basemap is also added for a context-aware

geovisualization.

4.4 Case Study

In this section, we present a case study using the proposed semantic generalization method and apply it on the Foursquare POI data in Los Angeles to generate a place-based map relevant to the *Beach*, *Shopping*, and *Asian Food* topics. As shown in Figure 4.5, over 85,000 Foursquare POI data in the urban area of Los Angeles were collected for the testing and the POI Data in the ten most populated urban areas used in the previous Chapter 3 will be utilized as our training data for topic modeling. Each POI has an attribute information including *name, location coordinates, place type, number of check−ins, number of checked users, number of tips,and the rating score.*

Before running the LDA topic model, we first buffered the place type of each POI to its corresponding Level-2 category. For example, as shown in Figure 4.3 and Figure 4.6 for the *Food* subtree, all the subcategories of *Asian Restaurant* including *Chinese Restaurant* and its children types, *Korean Restaurant* and its children types, *Thai Restaurant* and so on will be generalized to the same place type *Asian Restaurant* in the topic modeling training phase. In addition, we still incorporated the number of checked-in visitors for each POI as a popularity score in the rescaling process to generate a new document-word matrix (i.e., a search region−POI type occurrence matrix) across all sampling locations in the ten urban areas.

Next, we evaluated the performance of different choices of K as the total number of topics for the LDA topic model using three introduced measures. As shown in Figure 4.7, by choosing the value of K from 5 to 200 and then running LDA topic models on our POI data for the second-level place types, we derived different topic assignment results. The measure proposed by [95] aims to maximize the log-likelihood of word-

topic probability in the documents, while other two measures [127, 128] aim to minimize the proposed criteria. In our parameter tuning experiments, the optimal K value for the "CaoJuan2009" and "Arun2010" measures is in the range of $90 - 110$, while the "Griffiths2004" metric gets relatively stable when K reaches 100 topics. Therefore, we set $K = 100$ as the total number of topics and ran 2000 iterations of the Gibbs sampling process to derive the posterior probabilistic distribution over topic assignments for place types within the second-level hierarchy.

In Figure 4.8, we show six of those topics related to *Beach, Shopping, and Asian Restaurant*. Note that the probability assignments for those POI types are weighted and ranked by their *term frequency−inverse document frequency* (i.e., POI type frequency−inverse region frequency) so that each topic can display more distinctive and meaningful POI types that are directly proportional to the frequency in documents while inversely proportional to the region frequency at which a POI type occurs in the whole corpus.

Two topics *20* and *34* are shopping-plaza topic that consists of various frequent occurring POI types (e.g., men's clothing store, women's clothing store, and shoes store) that were generalized into the more abstract place types: *shopping mall* and *clothing store*. Interestingly, if we search for the region that has the top-ranked shopping *Topic 67* (in previous chapter) before semantic generalization, only the *Westfield Topanga Shopping Plaza* in Los Angeles can be found (Figure 4.9). However, after the place-type-buffer operation, nine shopping regions can be successfully identified and have top-ranked *Topic 20* or *Topic 34* in their multi-dimensional probabilistic topic distributions. As shown in Figure 4.10, after overlaying the OpenStreetMap data layers, we can clearly find that all of nine sampling locations contain at least one big shopping plaza. The histogram of JSD-based similarity values among the nine identified shopping regions is shown in Figure 4.11. The mean of the similarity distribution is about 0.59 with a standard deviation of 0.08. Therefore, we can use the similarity value (mean + standard deviation) = 0.67

as the threshold in the searching phase for finding shopping regions.

For the new beach-related topic 87, the top-ranked place-type that has the largest probability assignment for this topic is the second-level category *beach* with a probabilistic value 0.46 and has increased by 58.6% from its original probability 0.29 before semantic generalization. The mean of similarity values among those discovered beach regions is high around 0.8 and will be taken as the threshold for searching beach regions in the testing data.

Three topics 3, 18, and 26 are related to the *Asian Food* theme that consists of variety of place types including *asian restaurant, tea room, bubble tea shop, dumpling restaurant, and so on.* The *asian restaurant* was generalized from a lot of aforementioned subcategories such as *Chinese Restaurant*, *Korean Restaurant* and *Thai Restaurant.* We would assume that the regions that are dominated by the same higher-level place type should be more similar after the semantic buffer. For instance, we selected three sampling regions (with unique document identifiers: LA-Doc-92, LA-Doc-93, and LA-Doc-146) that belong to the *Asian Food* topic after the semantic generalization topic modeling process. Their original topic distributions and the spatial distributions are shown in Figure 4.12, 4.13, and 4.14 respectively. We can find that the region (LA-Doc-92) consists of frequent place types such as *Korean restaurant, Ramen restaurant (Janpanese), Grocery Store, Sandwich Place etc.*; the region (LA-Doc-93) is dominated by the *Thai Restaurant* topic; and the region (LA-Doc-146) is a typical *Chinese Restaurant* region. The topic similarity between (LA-Doc-146) and (LA-Doc-92) is 0.557 and increased to 0.674 after generalization; the topic similarity between (LA-Doc-146) and (LA-Doc-93) is 0.442 and increased to 0.508 after generalization. The Shannon information entropy [150] (with a logarithmic-base 2) for their topic distributions are also slightly reduced accordingly, region (LA-Doc-92) reduces from 5.70 to 5.42; region (LA-Doc-92) reduces from 5.57 to 5.28; and region (LA-Doc-92) reduces from 5.29 to 4.8. Generally speaking, the entropy

refers to disorder or uncertainty. The reduced topic entropy indicates a more predictable structure after the semantic generalization, which meets our expectation as well. The mean of the similarity distribution is about 0.64 with a standard deviation of 0.06. Therefore, we can use the similarity value (mean + standard deviation) = 0.70 as the threshold in the searching phase for finding *Asian Food* regions.

After we set the search similarity thresholds for the three groups of topics respectively, the experiments of randomly throw-circles with varying sizes for the spatial sampling are conducted as shown in Figure 4.15. The POIs within each search circle are selected and their generalized place-types are used for the topic prediction using the previous trained LDA topic model. Based on the criteria introduced in Section 4.3, we further derived the semantic generalized regions that meet the topic similarity requirement for *Beach, Shopping, and Asian Food* (in Figure 4.16). We find that the beach-topic regions are not necessarily continuously distributed along the beach side although most of them are still close to the Ocean. It might because that some regions lack sufficient venues or they are dominated by other thematic topics. Note that some of those discovered regions with varying topic themes are spatially overlapping together, which indicates that those regions have co-existing multiple prominent topics. For example, as shown in Figure 4.17, the identified region (LA-Doc-192) could be a good candidate region for both shopping and eating Asian Food.

4.5 Conclusion

In summary, in this work, we demonstrate the effectiveness of the proposed semantic generalization methodology on extracting the semantic regions that relate to the *Beach*, *Shopping*, and *Asian Food* topics in Los Angeles. The proposed theoretical framework can also be applied in other thematic topics and the extraction of associated semantic

generalized regions. The throwing-circle spatial sampling technique is also just one implementation in the search phase. Other techniques based on the place-type hierarchy can be developed for the semantic generalization process as well. The spatial footprints of the vernacular places and cognitive regions might also be constructed based on the semantic generalization framework.

Note that the testing process for large-scale region documents is computationally expensive. Thus the presented results might only reflect the scope of the collected data. This problem also raises the need for high-performance computation and scalable geoprocessing framework in future work.

Figure 4.3: The hierarchical structure of the POI type 'Food' and its subcategories Foursquare. Only some types are labeled because of a large visual load.

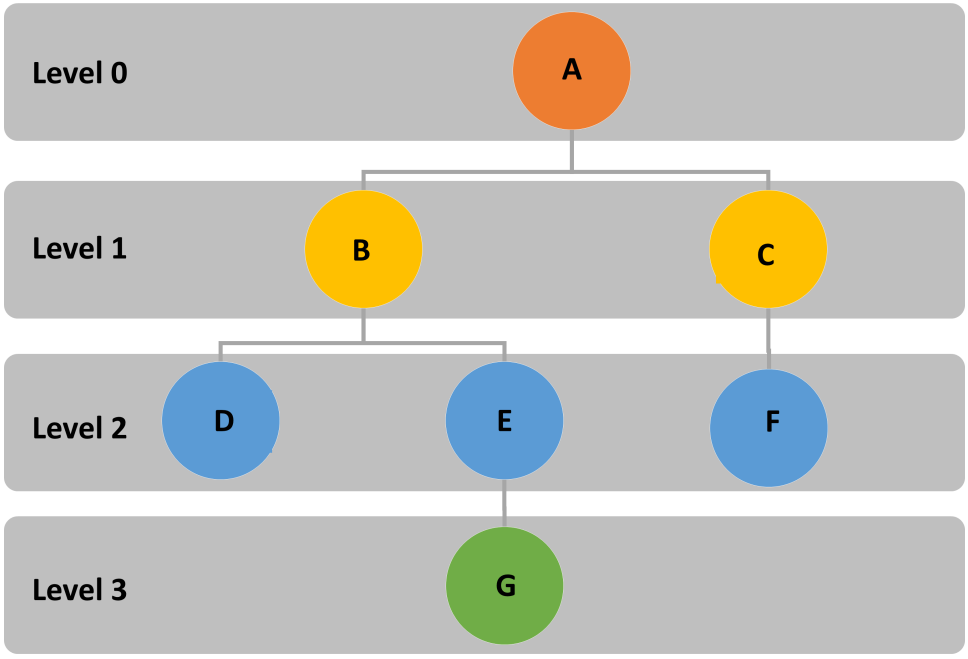


Figure 4.4: The hierarchical tree structure of nodes.

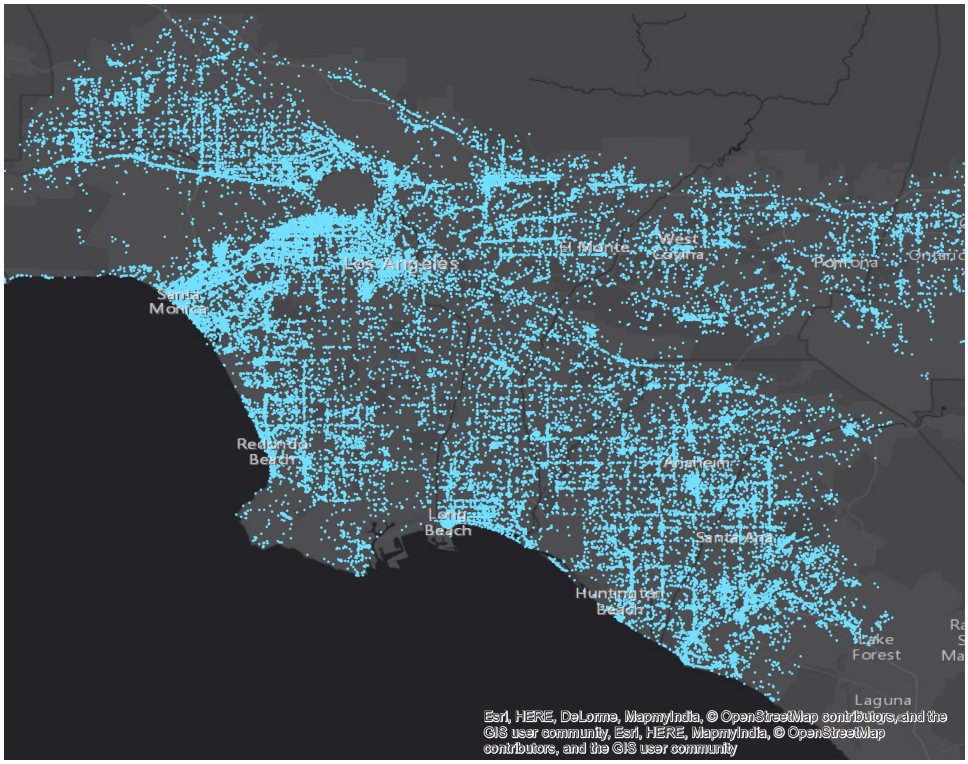


Figure 4.5: The collected Foursquare POI data in Los Angeles.

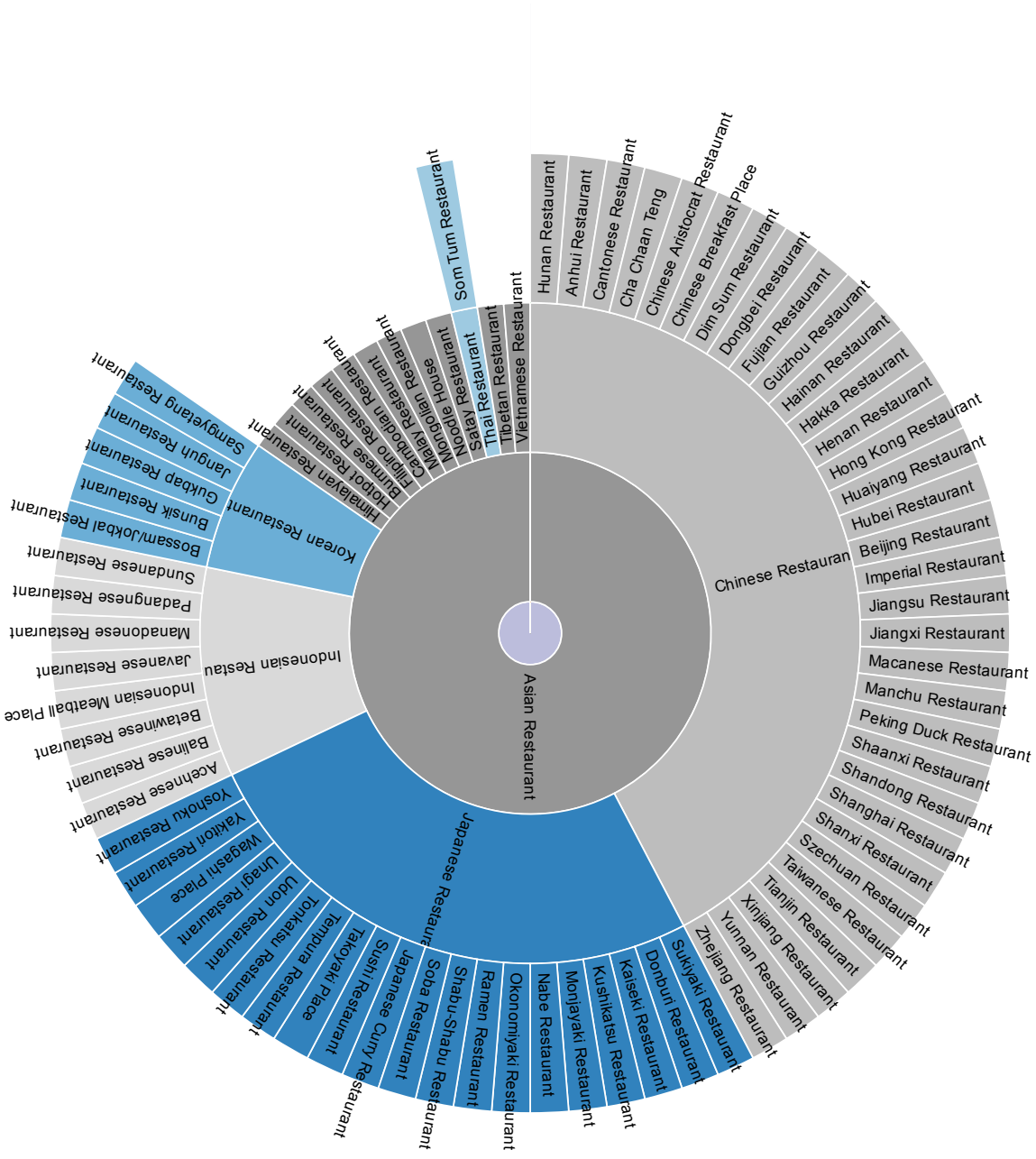


Figure 4.6: The hierarchical structure of the POI type 'Asian Restaurant' and its subcategories Foursquare.

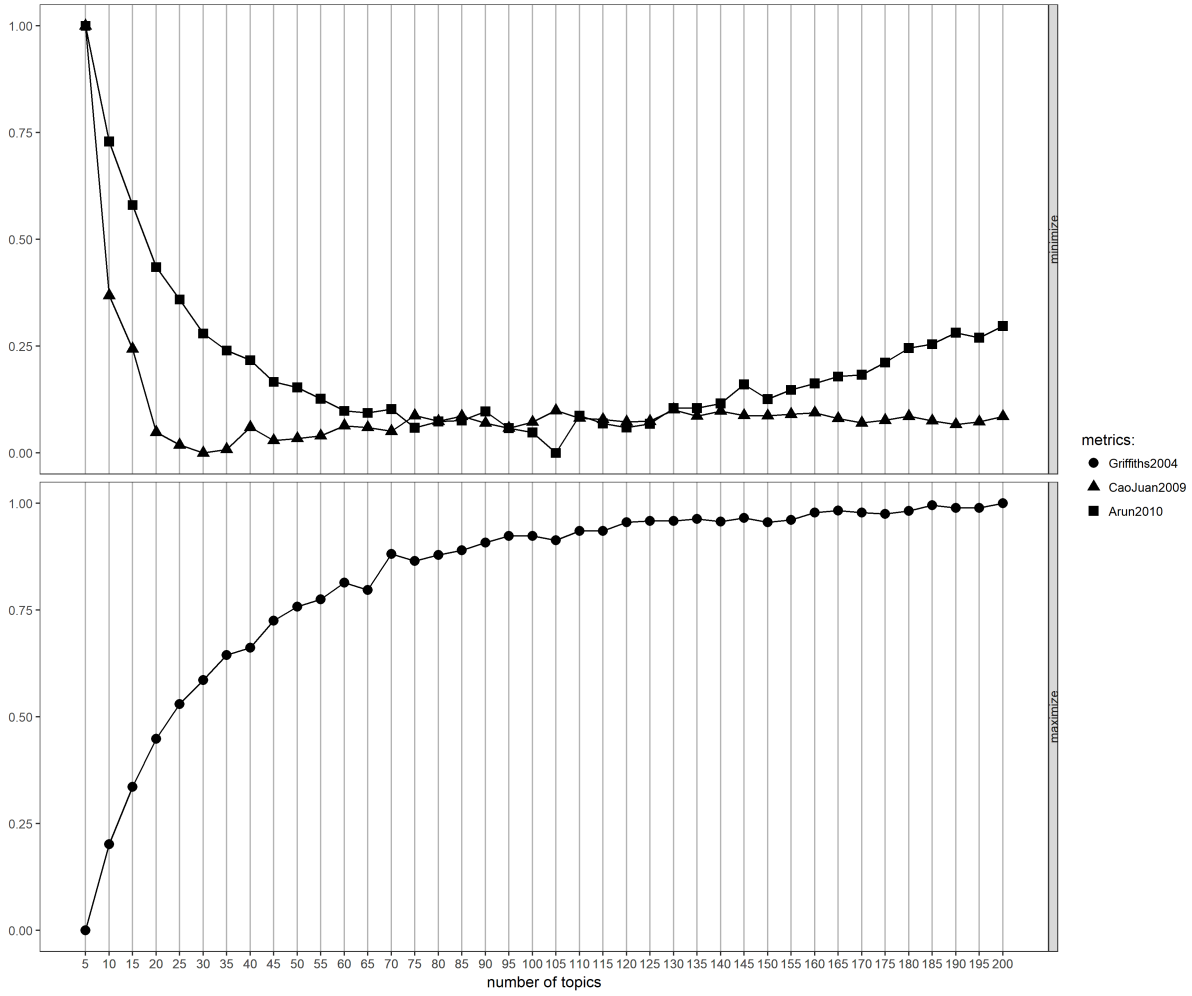


Figure 4.7: Find an appropriate K value for the total number of topics on the Level 2 categories using three metrics.

Topic 3	
Category	**Prob.**
asian restaurant	0.341470
tea room	0.050206
dessert shop	0.001295
coffee shop	0.000781
dumpling restaurant	0.000657
sandwich place	0.000356
burger joint	0.000286
diner	0.000270
seafood restaurant	0.000166
italian restaurant	0.000158
steakhouse	0.000152
college bookstore	0.000146
athletics & sports	0.000096
spiritual center	0.000040
community center	0.000022

Topic 18	
Category	**Prob.**
asian restaurant	0.185822
vegan restaurant	0.111357
bubble tea shop	0.062268
hawaiian restaurant	0.009922
animal shelter	0.000756
hobby shop	0.000668
coffee shop	0.000212
juice bar	0.000182
breakfast spot	0.000150
gunrange	0.000028
nature preserve	0.000022
clothing store	0.000017
fastfood restaurant	0.000017
convenience store	0.000016
brewery	0.000015

Topic 26	
Category	**Prob.**
asian restaurant	0.281990
coffee shop	0.050223
smoothie shop	0.028113
falafel restaurant	0.014617
bakery	0.000260
hardware store	0.000222
clothing store	0.000203
sandwich place	0.000132
garden center	0.000032
fastfood restaurant	0.000018
convenience store	0.000017
brewery	0.000016
hotel	0.000016
seafood restaurant	0.000016
bar	0.000016

Topic 20	
Category	**Prob.**
clothing store	0.982139
chocolate shop	0.000466
outlet store	0.000220
wings joint	0.000138
pet store	0.000071
food & drink shop	0.000055
music venue	0.000052
miscellaneous shop	0.000035
fastfood restaurant	0.000011
laser tag	0.000011
convenience store	0.000011
brewery	0.000010
hotel	0.000010
seafood restaurant	0.000010
bar	0.000010

Topic 34	
Category	**Prob.**
shopping mall	0.433009
coffee shop	0.073150
athletics & sports	0.006028
outdoor supply store	0.003895
american restaurant	0.002169
hunting supply	0.001597
dessert shop	0.001488
brewery	0.000248
bar	0.000237
park	0.000233
italian restaurant	0.000226
museum	0.000172
farm	0.000133
gastropub	0.000123
hobby shop	0.000040

Topic 87	
Category	**Prob.**
beach	0.459725
italian restaurant	0.088272
municipalities	0.015231
american restaurant	0.010109
bar	0.007362
harbor / marina	0.006024
boat or ferry	0.004842
asian restaurant	0.001714
island	0.000859
coffee shop	0.000636
board shop	0.000589
lighthouse	0.000305
athletics & sports	0.000272
food & drink shop	0.000244
steakhouse	0.000226

Figure 4.8: The six selected topics relevant to beach, shopping, and Asian restaurant.

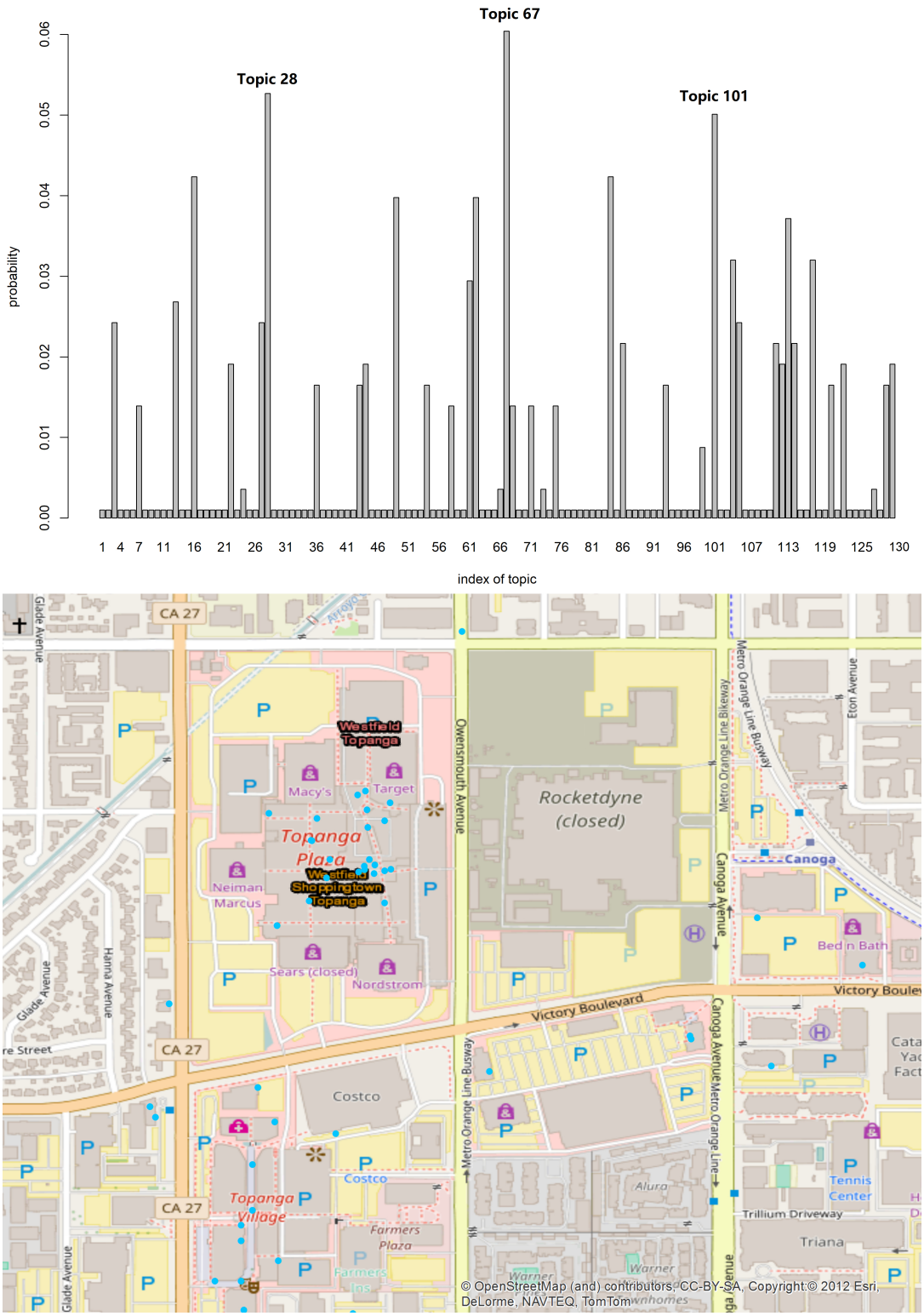


Figure 4.9: The topic probability distribution and the spatial distribution of Foursquare POIs around the Westfield Topanga Shopping Plaza in Los Angeles.

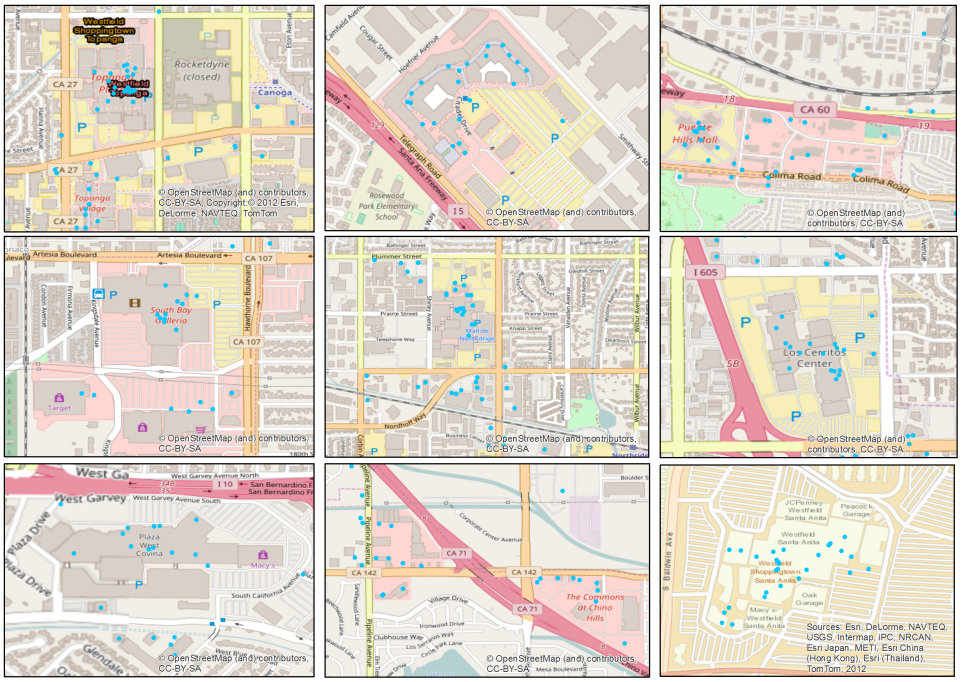


Figure 4.10: The spatial distribution of Foursquare POIs (blue dots) around nine prominent shopping regions in Los Angeles.

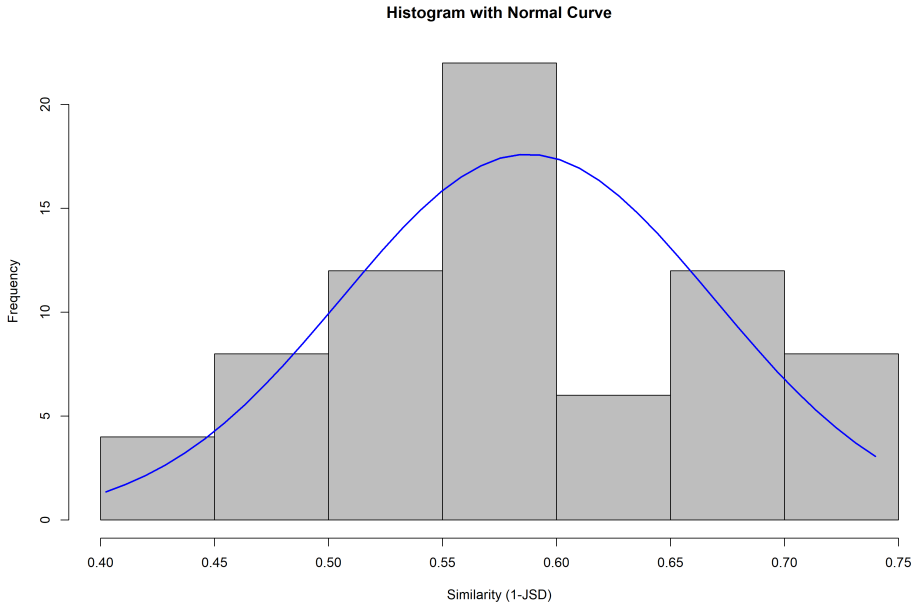


Figure 4.11: The histogram of JSD-based similarity values among the nine identified shopping regions in Los Angeles.

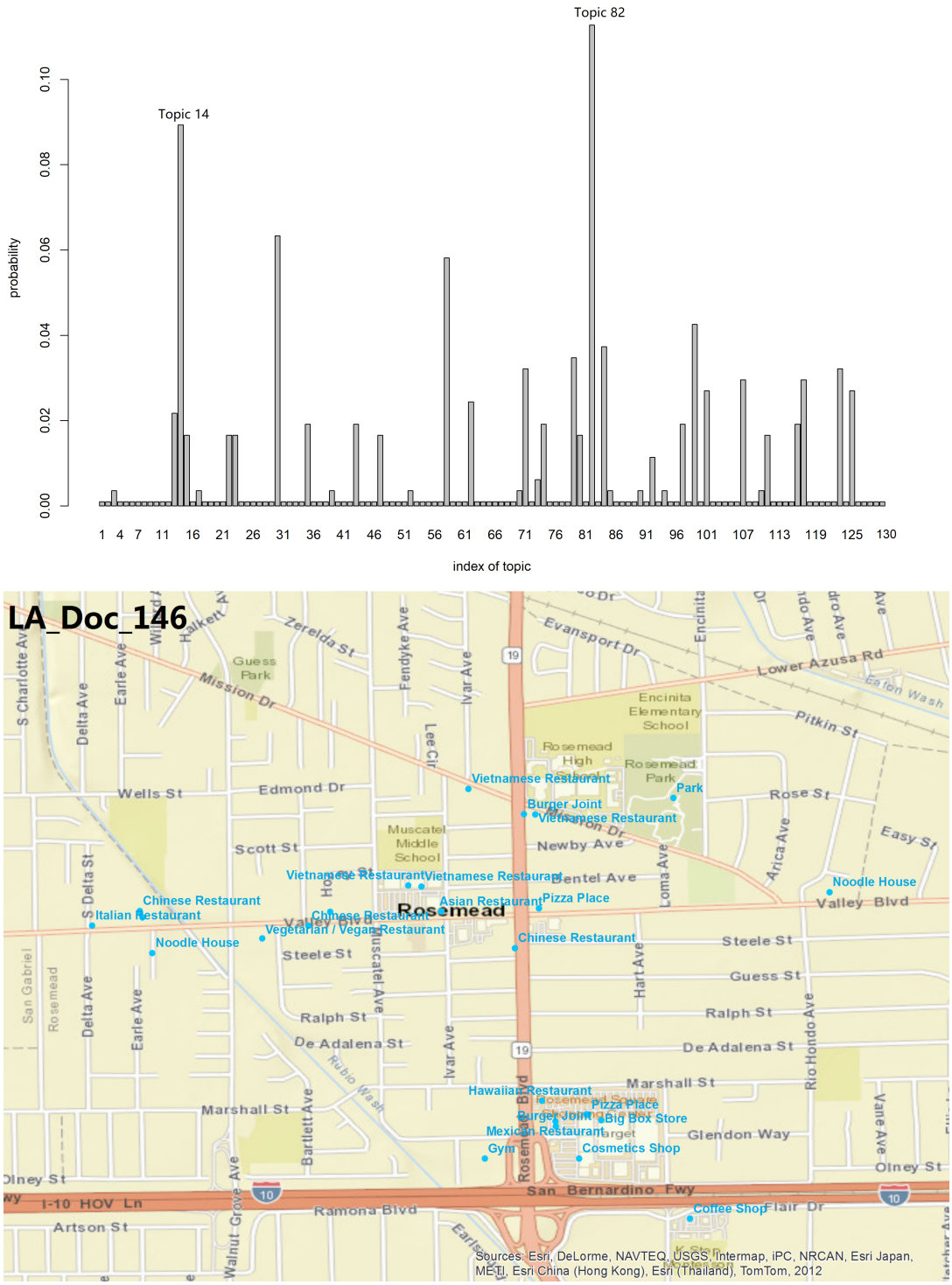


Figure 4.12: The topic distribution and the spatial distribution of Foursquare POIs in the search region (LA-Doc-146).

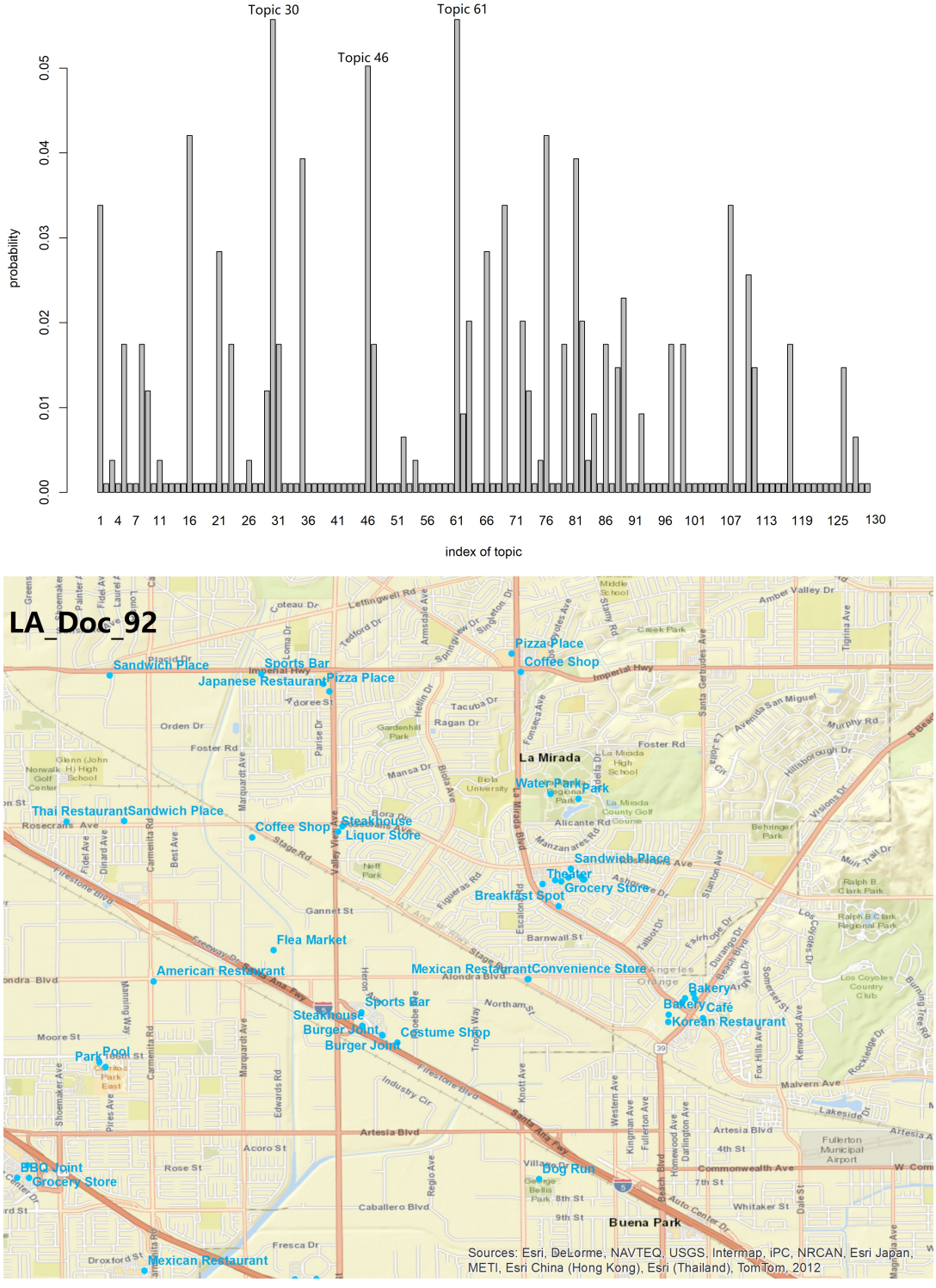
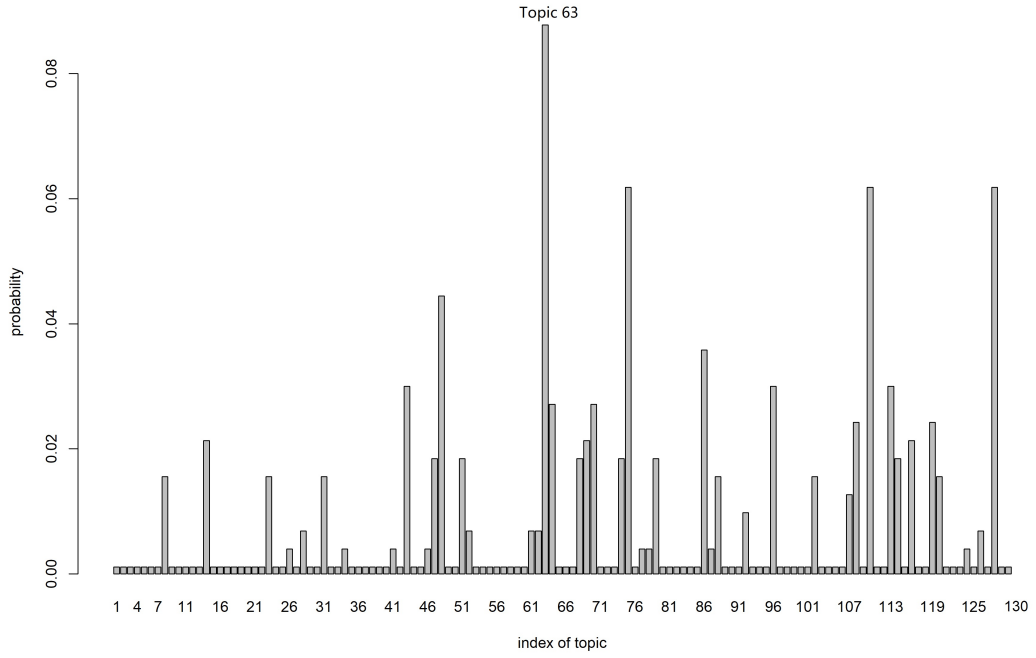


Figure 4.13: The topic distribution and the spatial distribution of Foursquare POIs in the search region (LA-Doc-92).



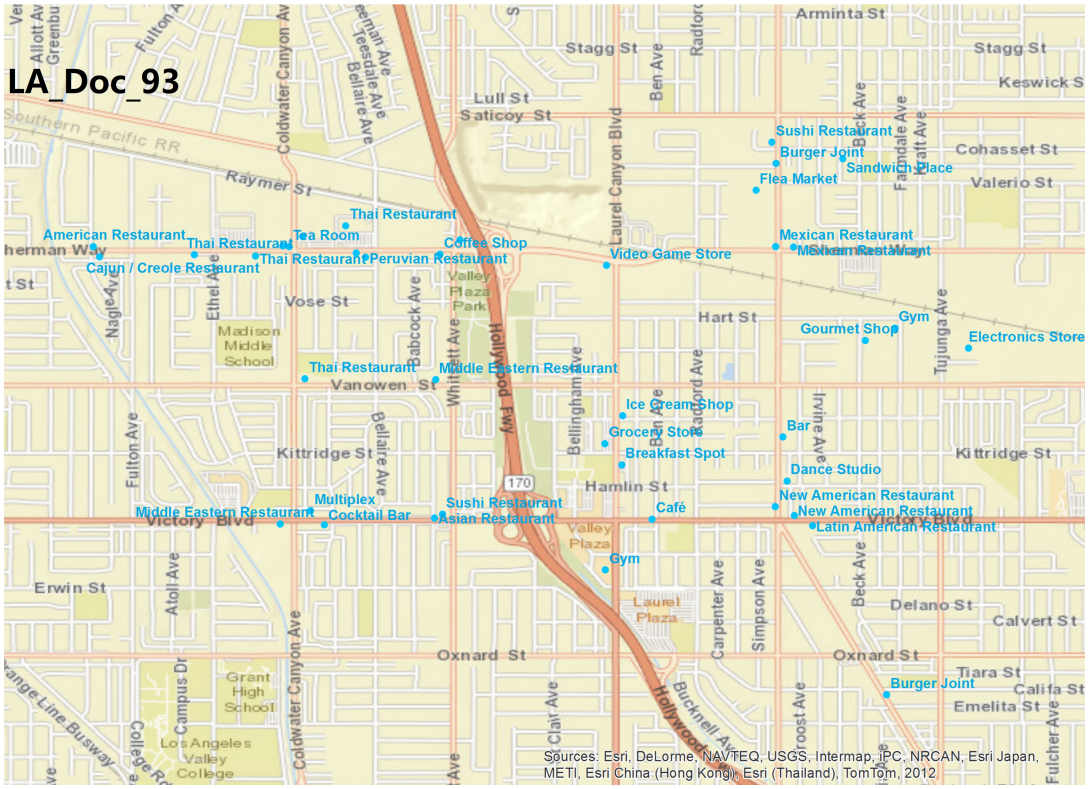


Figure 4.14: The topic distribution and the spatial distribution of Foursquare POIs in the search region (LA-Doc-93).

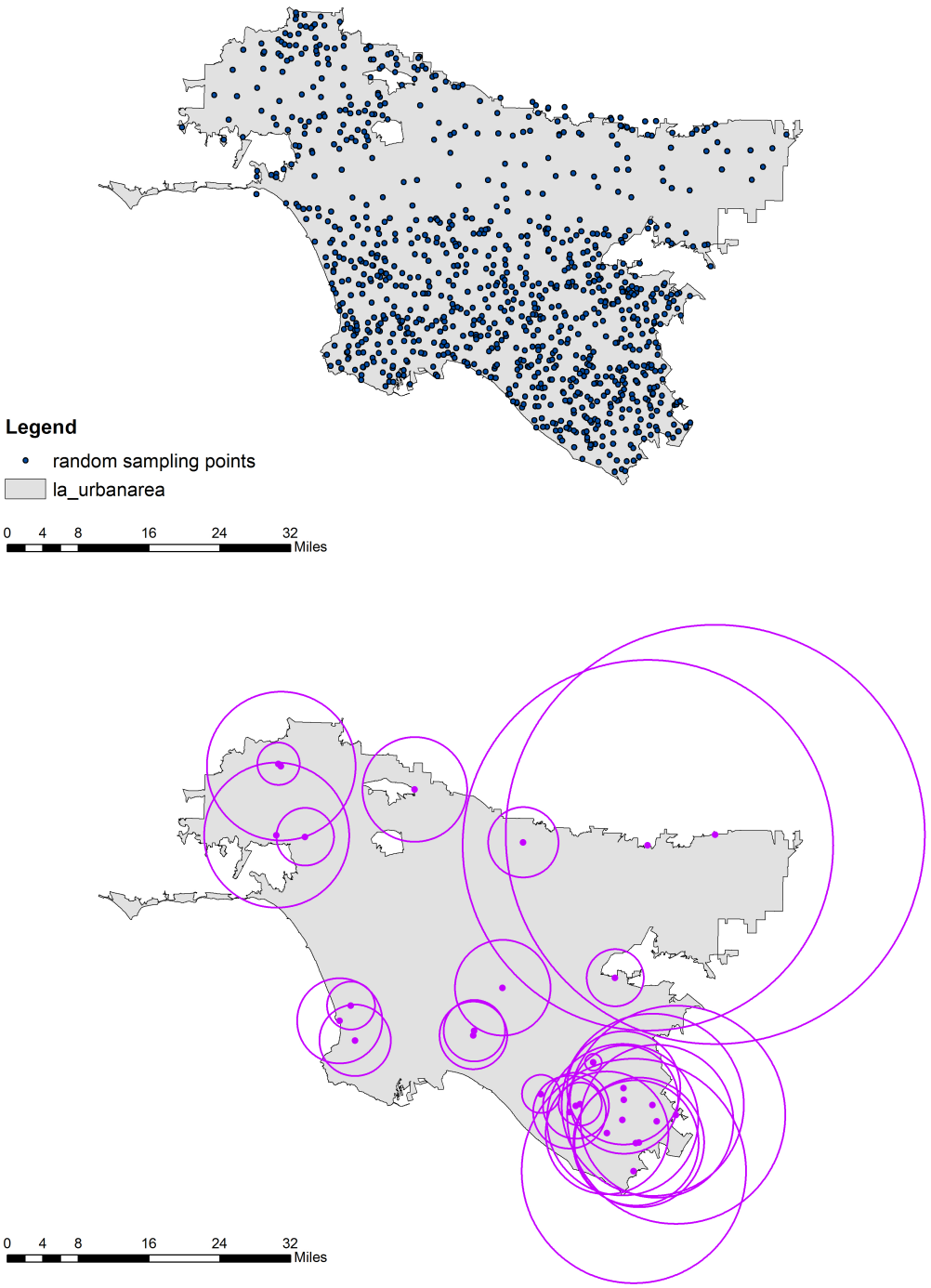


Figure 4.15: The spatial locations of those randomly search regions and an illustration for throwing varying-size circles in the study area.

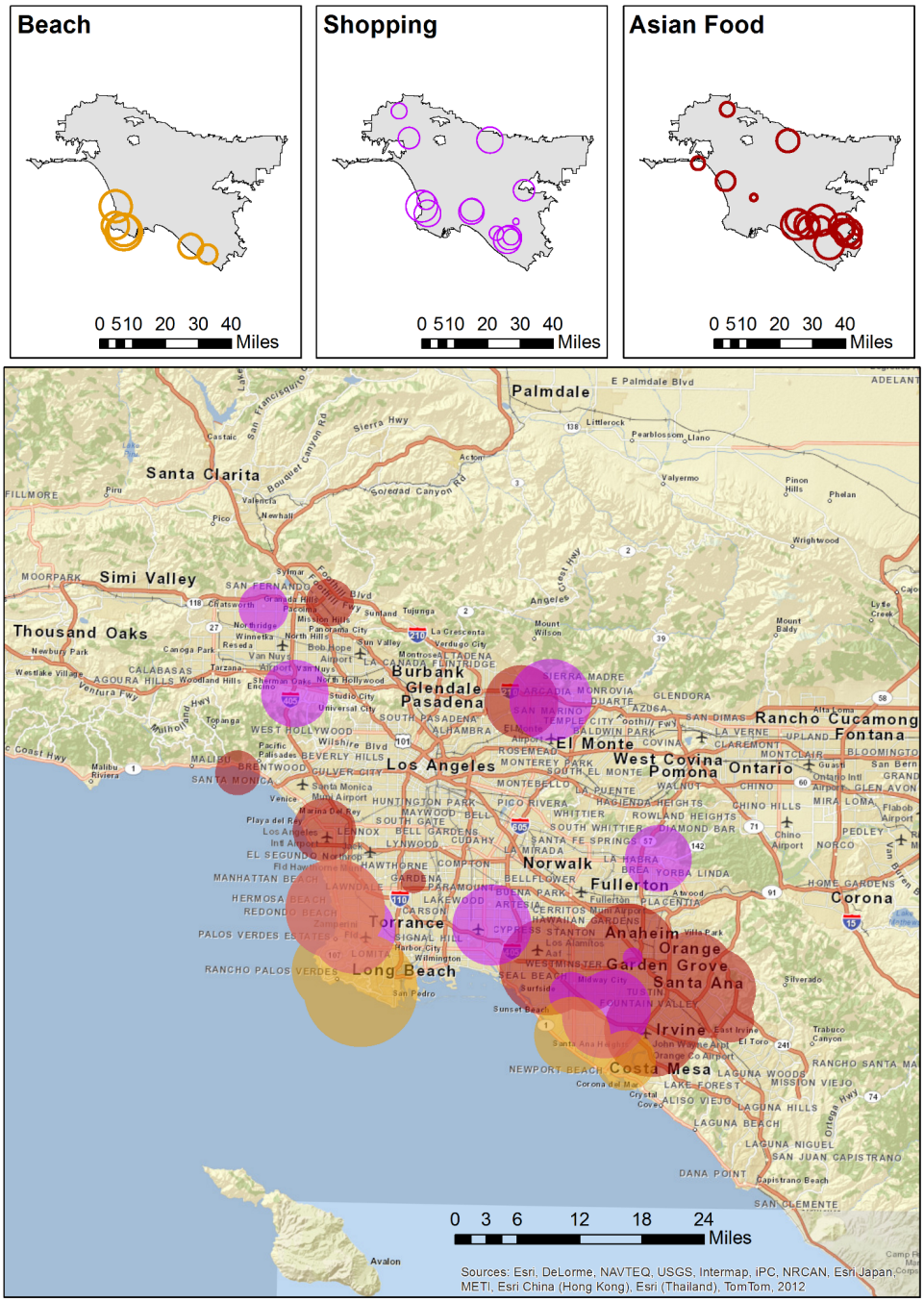


Figure 4.16: The resulting semantic generalization of regions for the three topics: Beach, Shopping, and Asian Food in Los Angeles.

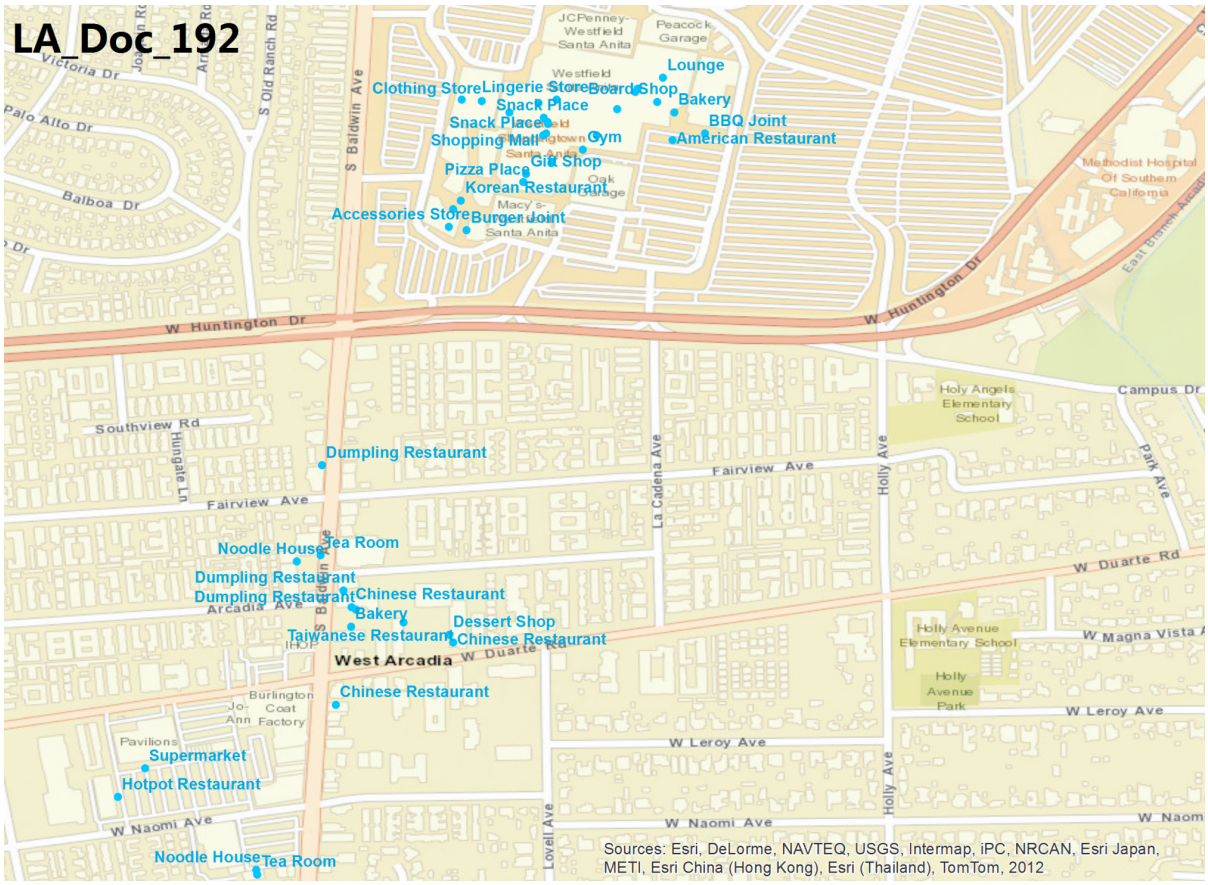


Figure 4.17: The spatial distribution of Foursquare POIs in the search region (LA-Doc-192).

Chapter 5

Conclusions and Future Work

Chapter 5 summarizes this dissertation. Particularly, we answer the posed three research questions and state our research contributions. We also discuss the broader implications and limitations of this research, and present planned future work.

5.1 Conclusions

In this section, we first conclude this dissertation by answering the aforementioned three research questions.

RQ1: *Is it possible to extract computational representations for vague cognitive regions from archival user-generated content such that they align with direct assessments of human participants using concordance measures?*

In Chapter 2, we investigated using a data-intensive approach to determining vague cognitive regions. We compared them to the corresponding MFP study based on human participants which validated our proposed approach. Using data sourced from social media including Flickr, Instagram, Twitter, Travel Blogs, and Wikipedia pages, we derived region membership scores for cells within the state of California that correlated signifi-

cantly to those in the original study, both in terms of Spearmans as well as Kendalls rank correlation statistics. The results also show a high agreement/concordance among our user-generated data sources with respect to the membership rankings of all cells, even after adding the survey ranks from the MFP study as the fifth source, demonstrating a consistency between our data-synthesis-driven approach and the human-participants survey. In other words, the effects we see are not merely artifacts of a specific data source. Overall, the shapes of *NorCal* and *SoCal* were quite similar for the two empirical approaches, including the non-monotonicity of the two regions and the heterogeneity of their vague boundaries. Most importantly, our work showed the same *platial effects* observed in the original study. Furthermore, our work examined the implications of increasing the spatial resolution of the tessellations on the cognitive regions that result.

In addition to assessing membership scores within the hexagons, we further explored the continuous boundaries and the core regions for *NorCal* and *SoCal*. A two-step workflow based on the DBSCAN clustering method and the chi-shape algorithm was designed to generate approximate boundaries for the cognitive regions. Experiments were conducted to select optimal parameters for the workflow, and we observe consistency among the polygon representations that are derived from the different datasets.

We also explored thematic associations for *NorCal* and *SoCal* with the help of topic modeling. This generated various topics most often associated with different regions of California on our social media sources. Comparing the topic distributions of prototypical *NorCal* and *SoCal* hexagons shows high similarity within each region and a lower similarity between the two regions. The temporal characteristics of regions were not studied in this work and related works will be conducted with regard to different granularity of temporal bands in future work.

RQ2: *Is it possible to extract the co-location patterns of different place types and underlying characteristics that could be utilized to describe functional regions that support*

specific types of human activities?

In Chapter 3, we developed a statistical framework that applied the LDA topic modeling technique and incorporated user check-ins on LBSN in order to help discover semantically meaningful topics and functional regions based on the co-location patterns of POI types. The "functions" derived from probabilistic topic modeling techniques can reveal the latent structure of POI mixtures and the characteristics of places. Based on a large corpus of about 100,000 Foursquare venues and check-in behavior in the ten most populated urban areas in the U.S., we demonstrate the effectiveness of proposed methodology by identifying distinctive types of latent topics and further, by extracting urban functional regions using the K-means clustering and the Delaunay triangulation spatial constraints clustering methods. A region can have multiple functions but with different probabilities, while the same type of functional region can span multiple geographically non-adjacent locations. Compared with the remote sensing images that mainly uncover the physical landscape of urban environments, results derived from the popularity-based POI topic model can be seen as a complementary social sensing view of urban space based on human activities and the place settings of urban functions. However, there may exist gaps between the real-world business establishments and the online available POI information. Data-fusion and cross-validation relying on multiple sources may help reduce such gaps.

Although we have successfully identified several types of semantically meaningful urban functional topics, LDA topic modeling is an unsupervised approach that has certain limitations with respect to discovering plausible urban functions. One limitation of this particular research is that we cannot systematically evaluate the accuracy of those derived functional regions without labeled ground truth data or the detailed urban land-use GIS data. But we have tested the intrinsic robustness of identifying functional topics with different parameter settings. The variability analyses were carried out at two levels: the

topic-level and the cluster-level. At the topic level, we found the stability in identifying prominent urban functional topics related to frequently co-occurrent physical facilities and services, a variety of bars and restaurants, and leisure activity places regardless of the total number of topics. But the topic composition of top-ranked POI categories varies in different scenarios. It implies the variability of the semantic structure of functional topics. Although choosing an optimal K in topic modeling can either maximize the log-likelihood of the term-topic probability in the training document corpus, or minimize the inter-topic similarity, we may miss the opportunity for discovering some interesting topic composition structures that can only be identified with a different K value or with other model parameter settings. At the cluster level, a series of clustering result comparisons by choosing different numbers of topics were evaluated using the Rand index and the NMI metric. We found a large percentage of agreements on the clustering membership of those search locations with their surrounding POIs and derived functional areas that can be supported by the mix types of POIs.

RQ3: *What kinds of analyses/operations on places can be employed for deriving semantically generalized regions in order to address the regional/cultural variability by broadening the thematic topics that form the functional regions?*

In Chapter 4, presented a novel methodology for the semantic region generalization by integrating spatial sampling, topic modeling, and platial buffer techniques. We proposed a novel semantic generalization processing framework that can produce semantically coherent high-level generalized region representations on maps from POI data. The workflow consists of six steps: (1) POI Data Collection; (2) Topic Modeling; (3) Identify Functional Topics and Clustering; (4) Search for Semantically Generalized Regions (based on platial buffer in the hierarchical structure of POI types); (5) Group and Merge Semantically Interlinked Regions; and (6) Map Representation and Visualization.

Based on the collected over 85,000 Foursquare POI data and associated human

check-in behaviors, we successfully extracted semantically coherent regions that relate to the *Beach*, *Shopping*, and *Asian Food* topics in Los Angeles and demonstrated the effectiveness of the proposed semantic generalization methodology to address the regional/cultural variability by broadening the thematic topics that form the functional regions. The proposed theoretical framework can also be applied in other thematic topics and the extraction of associated semantic generalized regions. The throwing-circle spatial sampling technique is also just one implementation in the search phase. Other techniques based on the place-type hierarchy can be developed for the semantic generalization process as well. The spatial footprints of the vernacular places and cognitive regions might also be constructed based on the semantic generalization framework.

In sum, this dissertation presents a comprehensive computational framework for extracting the representations of place (using vague cognitive regions, functional regions, and semantically generalized regions as examples) and those representations could be further coded in geographic information systems and applied in spatial search or other applications. This research sheds light on differences in the methodology of traditional human-participants approach and the increasingly popular data-synthesis driven approaches, suggests advantages and limitations of both approaches, and points to future avenues for research and system design in GIScience.

Last but not least, advanced places-based studies can engage citizens in knowledge production and sharing experiences on places for sustainable development of our society and environment.

5.2 Research Contributions

The main contribution of this dissertation is an advancement to the computational representations and models of place in GIScience, which is a challenging issue in geographic information retrieval, mapping and information visualization, and GIS processing workflows in general. The broad umbrella term for place-centered analyses and the formalization of place in GIScience has been informally defined as "Place-based GIS" [3, 4]. Central to all research branches concerned with place-based GIS is the computational modeling of place with regard to human mind or activities in order to assess human-place interactions. This dissertation has made a number of contributions to the place-based GIS theories, methodologies and applications, which are summarized as follows:

- **A Theoretical Contribution for Computational Models of Place in GIScience.** As is known from the computer modeling and programming implementation perspective, *coordinates* should be included in the "Space Class Constructor" to represent a location on the Earth Surface. In order to model the concepts of place from informal human conversations and natural language descriptions into formalized computerized informations systems, we suggest that *thematic topics* and *location* are the key characteristics that need to be included in the "Place Class Constructor". In addition, a place could also have *placename*, *time constraint*, and *semantic linkages to other places and entities*. The *thematic topics* of regions contain one prominent or multiple co-located place-types as well as the characteristics described in natural language. The *location* of a place maybe specified by its *membership* or by a *geometry*.

- **A Formalized Computational Representation for Vague Cognitive Regions.** Most vague cognitive regions only exist in human mind but are not sufficiently represented in GIS or Web mapping systems. In order to let computers and

GIS handle vague cognitive regions, they have to be formalized and transformed from human mind into computer binary code. This research develops a framework for representing vague cognitive regions with their fuzzy membership scores, geometric boundaries, and textual thematic characteristics.

- **A Data-Synthesis-Driven Method for Extracting Vague Cognitive Regions and Functional Regions.** Although vague cognitive regions and urban functional regions have been extensively studied in spatial cognition and urban studies, most existing works were conducted in a top-down research design manner and relied on expert knowledge. In this dissertation, we present an automatable framework that can synthesize multiple heterogeneous datasets from different sources to study vague cognitive regions and urban functional regions, which can provide a human-centered social sensing view of those concepts and bottom-up derived computational representations.

- **A Popularity-based Statistical Topic Modeling Technique for Characterizing the Semantics of Place.** Probabilistic topic models have been widely used to discover latent thematic characteristics and their structure when analyzing large sources of textual documents. In analogy to the use of textual documents in topic modeling, this research takes the place type of each POI as a word, the search region that contains those POIs as a document, and an urban function or a land use as a topic that represents thematic characteristics and the semantics of places. In order to address the human activity effect, the popularity of places based on their user check-in statistics has been taken into consideration during the sampling process. The case studies in this dissertation demonstrate that the proposed popularity-based LDA topic modeling technique can learn low-dimensional thematic representations to differentiate different vague cognitive regions or dis-

tinctive types of urban functions.

- **A Novel Methodology for Semantic Generalization of Regions.** The existing research on digital map generalization mainly rely on the geometric properties of spatial data but not considering the platial effects. A semantic generalization of space-based maps into place-based maps may bridge the gap between the abstract geometries and human cognition, and help a better understanding of the interaction between human concepts and places. This research develops a new methodology that can take both spatial distributions of POIs and the place-type hierarchical relationships into consideration to derive spatially and semantically coherent high-level generalized regions in order to address the regional/cultural variability by broadening the thematic topics. The presented throwing-circle spatial sampling technique is also just one implementation. Other techniques based on the place-type hierarchy can be developed for the semantic generalization processing as well.

5.3 Broader Implications and Limitations

This research has its own limitations on the data sampling, methods, and derived results that could be improved in the future. In the following, we summarize these limitations and discuss possible improvements.

- **Bias and Noise in Social Media Data.** In order to capture the human effect and take human activities into consideration for the study of place in this research, large-scale multi-source social media data have been collected and processed. However, the datasets might still be bias to only specific groups of people's opinions instead of that from the whole population, although we have applied several filtering and resampling strategies to best reduce such bias. It also raises some further issues

about the data-synthesis-driven approach. There is a difference between *the said place* which a person tags or mentions in a social-media entry and *the locale* where the person is located when posting the entry. *The said place* is not necessarily the same as *the locale*, since people can post any message about any place no matter where they are. Such mismatch is common to see in our crowdsourced data. In future work, natural language processing techniques (e.g., place name disambiguation, preposition and contextual analysis) can be employed in analyzing social media entries to better differentiate the said place and the locale.

- **Ground-truth Testing.** Regardless of the extracted urban functional regions or the semantically generalized regions, human participant experiments or other labeled ground-truth data are still necessary to further test the performance of the proposed methods in this dissertation. Depending on the ground-truth data availability, we may also collect and process the location-based social network data in other study areas in future work.

- **The Temporal Dimension of Place.** Although we didn't explicitly include the role of "time" in characterizing different places and regions, we do recognize the importance of the temporal dimension of place. Existing researches show that place-types can be uniquely identified by the temporal patterns of their visitors. For instance, people are more likely to visit restaurants during typical lunch and dinner hours than at midnight, while visiting bars at night than at daytime [151, 97]. Since we took POIs as a proxy for delineating functional regions that support specific human activities, those extracted regions may only be tangible at specific time periods. In addition, certain types of places (e.g., *football stadium*) are more prone to regional and international differences with respect to the temporal check-in behavior than others (e.g., *drug store & pharmacy* because of the socioeconomic

and culture differences) [152]. Moreover, some places and cognitive regions only exist at specific historical time, and thus those entries should be only available for place-based queries with certain time-constraints in the databases or file systems.

- **Non-spatial Topic Modeling Technique.** In this dissertation, we applied the non-spatial version of LDA topic modeling technique to discover the hidden co-occurrence patterns of POIs to support specific urban functions. However, the original methods didn't consider the spatial distributions of "place-type terms" in "region documents". Therefore, the distance metrics or spatial relations among those "terms" were still not the first-citizen during model training and prediction processes. Thus the results didn't explicitly show the spatial patterns (with distance measure) among place types. We will discuss potential solutions for this in the future work section.

5.4 Future Work

5.4.1 A Combination of Bottom-up with Top-down Ontology Design for Urban Functional Regions

Based on the analysis results in Chapter 3, we showed that several latent topics of POI types are spatially and semantically related to certain urban functions. For example, the college/university topic that consists of *buildings, pool, sports fields, and apartments*, is also co-located with several *restaurant* and *bar* like topics; the *shopping plaza* topic is often also co-located with the *parking* and *resort* like topics. This reveals the underlying relations of how POI categories function in urban settings. We have also discovered various urban functional regions as clusters of multinomial topic distributions over POI categories. However, one limitation is that we cannot systematically evaluate

the accuracy of those derived functional regions without labeled ground truth data or the detailed urban land-use GIS data. Future work need to include the labeled groundtruth data as testing cases for comparisons.

One broader question of this research is whether we can automatically identify those topological and hierarchical relations in order to support the development of an ontology for urban functional regions. As shown in Figure 5.1, we applied the Ward hierarchical clustering method [153] on those 130-topics derived from the aforementioned LDA topic model. Each topic is a 480-dimensional probabilistic vector over all POI types in our datasets. Those semantically related topics are grouped together in each step by minimizing the increment of within-cluster variance after merging. This process repeats until all topic vectors merge into the same group. This tree diagram is derived from a bottom-up approach and can be used as a starting point with regard to constructing the urban functional region ontology. However, it does not yet include the spatial relationships nor the dichotomous relationship among POIs. It may be more promising to combine this bottom-up approach with the top-down approach of the expert urban geographers or planners to develop a more holistic ontology in the future.

5.4.2 Spatial-LDA Topic Modeling Methods

Currently, topic modeling techniques that originated from natural language processing domain have been widely applied in many other domains including geography. Although, the non-spatial version of topic modeling could also generate the co-occurrence patterns of places. However, the original methods didn't consider the spatial distributions of "place-type terms" in "region documents". Therefore, the distance metrics or spatial relations among those "terms" were still not the first-citizen during model training and prediction processes.

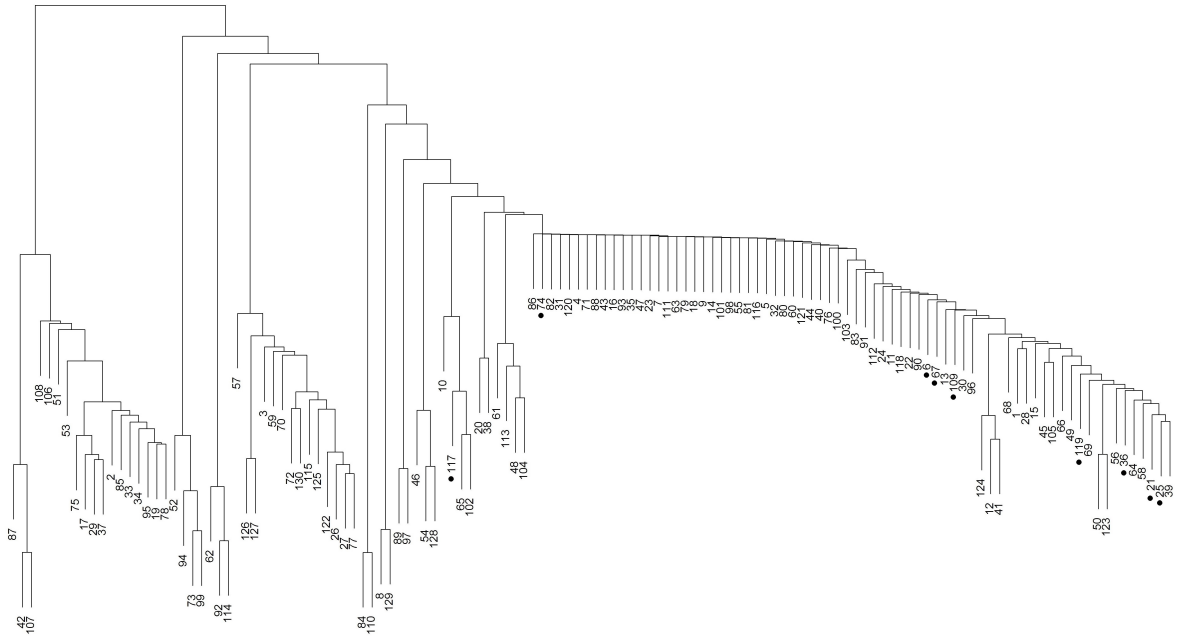


Figure 5.1: The dendrogram for the hierarchal clustering result on the 130-LDA topics using the Ward clustering method. The topics highlighted with a black filled-in circle are those mentioned in chapter 3.

Generally speaking, there exists at least two groups of approaches to including the distance metric into consideration. The first group of methods rely on a deep understanding of the statistical topic modeling methods and can directly change the probabilistic formulas for developing new family of spatial-metric-oriented topic modeling methods.

The second groups of methods rely on the oversampling of places or rescaling process when generating region documents. For example, we could encode the distance bins into the "term" as an interaction format of "place A - distance bin (or a topology) -place B". And then running the LDA topic modeling technique on those interaction terms, we could generate the spatial-interaction patterns of POI types, such as "bar-close-bar" and "fire station-intermediate distance-hospital". One challenge is how to decode those interaction terms from discovered topics. Basically, each place-type can interact with all other types, which will generate a ginormous vocabulary for the topic modeling and the

results will be very hard to interpret.

Another idea is to unitize the spatial indexing tree during the POI spatial sampling process, we can use the hierarchal level of the spatial index for all those POIs to generate different region documents. We assume that those generated documents intrinsically carry the spatial interaction information and thus the discovered topics could reveal be spatial co-location patterns of places.

5.4.3 Automatic Semantic Generalization of Regions in Multi-Scales

The presented semantic generalization framework in Chapter 4 only considers the scenario with a fixed spatial scale, but didn't consider the dynamic exploration process across multiple spatial scales.

One future study could be exploring the *conditions* under which the semantic generalization should be invoked, and what *operators* that the semantic generalization procedure should employ, as well as the *measures* on which the generalization evaluation should rely.

To this end, future study will explore the automatic semantic generalization theories and techniques using OpenStreetMap Data and POI databases. Case studies of identifying and abstracting different scales of vague cognitive regions will be employed to evaluate the proposed operators, measures, controls, and so on for enabling automatic semantic generalization procedures. To develop the automated semantic generalized maps, the generalization operators should be appropriate to multiple scales and the quantitative characteristics of the selection should be consistent. The evaluation criteria should be developed to assess the usability of the seamlessly automated semantic map generalization system.

Future research work will eventually integrate all aforementioned aspects into the

semantically enabled place-based digital map generalization process. An implementation of place-based mapping system will support searching, browsing and exploring the semantically coherent *neighborhoods, downtowns, functional regions, and vernacular places* in multi-scales.

5.4.4 Crowdsourcing Place Graphs for Supporting Knowledge Discovery

Based on the experiments conducted in the dissertation, we find that many places and place types are frequently co-located in physical space or co-existed in social media posts. The linkages among those places and place types can be constructed as a graph. We plan to evaluate the effectiveness of data-driven approach for constructing place graphs from user generated content and to quantify the place relatedness in human mind in future.

Two types of place graphs will be developed. One is the place graph based on natural language descriptions reflecting human perception and the other is based on human behavior in social media data. To measure place relatedness from natural language descriptions, the co-occurrence model will be employed to quantify the frequency that two places occur together given a smoothing window. The co-occurrence frequency will then be normalized to give a quantitative measure between places. To measure the place relatedness from human behavior, the human check-in pattern from social media data will be examined. The number of co-checkins from the same users in different places will be used to quantify the relatedness between these places. While only two types of place graphs will be constructed, each type will contain multiple graphs depending on the specific data source. For example, the natural language based place graphs can be generated through data from News Articles, Wikipedia, Travel Blogs, and even government textual documentation. Similarly, the human behavior-based place graphs can be

developed using data from Foursquare, Twitter, or Flickr.

The developed place graphs can be used to facilitate spatial search and knowledge discovery. In addition, for users who have visited one place, a Web service can recommend closely related places based on the graph. This recommendation can also be applied to Spatial Data Infrastructures (SDI). A user who have browsed the geospatial data of one city may also be recommended with the data in closely related cities.

The quality and usefulness of the constructed place graphs will be evaluated based on their capability to meet the application requirements. Thus, two evaluation experiments will be conducted for the two applications respectively. To evaluate the improvement of information search based on place graph, a human participant experiment will be conducted and the baseline will be the purely space-based search. 50 human participants will be recruited, and each need to evaluate 20 queries and the quality of returned results from platial search and spatial search. The human participants will give a score from 0 to 5, and all the scores will be summarized to see if the platial approach received a statistically significant larger score than the pure spatial approach.

To evaluate the recommendation system, another human participant experiment will be conducted. This experiment will compare the place-based recommendation with space-based recommendation (i.e., recommending places only based on the geographic distance). Similarly, 50 participants will be recruited, and each participant will need to give a score from 0 to 5 for the recommended places. It is expected that the recommendation based on place graphs will be scored higher than the space-based recommendation.

5.4.5 Investigating Place-based Analysis Functionality

The most challenging part in the place-based research should be the formalization of platial analysis functionalities to their spatial counterparts.

As a start, we have developed the platial-join and platial-buffer operations [4]. In analogous to the spatial join, the purpose of the platial join is to attach the properties or characteristics from the join entities to the target place using semantics references. In other words, the platial join operation involves the aggregation of properties (attributes) from one or multiple place entities to the target place entity based on merge rules (such as sum, average, first, last) and their topological predicates, including the "part-whole" relation, the "part-whole", the "parent-child" relation, and other topological relationships, e.g., touch, overlap, equals, contains, disjoint, and intersects [154]. A merge rule is applied when more than one entity are matched to a target place.

The platial buffer operation involves identifying neighboring places (first-degree buffer) or other n-degree connected places for a target based on the semantic relations (upper-level class and lower-level class). The n-degree represents the number of shortest-path steps that connect the places under consideration on a semantic net.

As presented by Goodchild (2015)[155], if someone asks a spatial question: "Is A near B?" In a spatial approach, the answer is yes if B is within a buffer around A, and the buffered distance defines "near". In a platial analysis approach, the answer is yes if A and B are within some higher-level place (or place-type) C, and the difference in level or the graph path-length defines "near". Another question: "Is B nearer to A than is C?" In a spatial approach, it is required to compute the distance of (A, B) and the distance of (A, C). In a platial approach, B is nearer to A if A and B are in the same higher-level place hierarchy and C is not.

A number of research questions are worth of investigation in future work. Example include but not limit to: What should be the platial association function? What about place-based map algebra? Does "platial hot-spots" exist?

Bibliography

[1] Y.-F. Tuan, *Space and place: The perspective of experience*. U of Minnesota Press, 1977.

[2] J. Agnew, *Space and place*, in *The SAGE handbook of geographical knowledge* (J. Agnew and D. Livingstone, eds.), pp. 316–330. Thousand Oaks: Sage, (Chapter 23), 2011.

[3] M. F. Goodchild, *Formalizing place in geographic information systems*, in *Communities, Neighborhoods, and Health* (L. Burton, S. Kemp, M.-C. Leung, S. Matthews, and D. Takeuchi, eds.), pp. 21–33. New York: Springer, 2011.

[4] S. Gao, K. Janowicz, G. McKenzie, and L. Li, *Towards platial joins and buffers in place-based gis*, in *Proceedings of the 1st ACM SIGSPATIAL international workshop on computational models of place (COMP2013)*, pp. 42–49, 2013.

[5] H. Couclelis, *Location, place, region, and space*, *Geography's inner worlds* **2** (1992) 15–233.

[6] L. L. Hill, *Core elements of digital gazetteers: placenames, categories, and footprints*, in *Research and advanced technology for digital libraries*, pp. 280–290. Springer, 2000.

[7] M. F. Goodchild, *The alexandria digital library: review, assessment, and prospects*, *D-Lib Magazine* **10** (2004), no. 5.

[8] M. Sanderson and J. Kohler, *Analyzing geographic queries*, in *SIGIR Workshop on Geographic Information Retrieval*, vol. 2, 2004.

[9] C. B. Jones, H. Alani, and D. Tudhope, *Geographical information retrieval with ontologies of place*, in *Spatial information theory*, pp. 322–335. Springer, 2001.

[10] D. R. Montello, M. F. Goodchild, J. Gottsegen, and P. Fohl, *Where's downtown?: Behavioral methods for determining referents of vague spatial queries*, *Spatial Cognition & Computation* **3** (2003), no. 2-3 185–204.

[11] X. Yao and J.-C. Thill, *Spatial queries with qualitative locations in spatial information systems*, *Computers, environment and urban systems* **30** (2006), no. 4 485–502.

[12] Q. Guo, Y. Liu, and J. Wieczorek, *Georeferencing locality descriptions and computing associated uncertainty using a probabilistic approach*, *International Journal of Geographical Information Science* **22** (2008), no. 10 1067–1090.

[13] Y. Liu, Q. Guo, J. Wieczorek, and M. F. Goodchild, *Positioning localities based on spatial assertions*, *International Journal of Geographical Information Science* **23** (2009), no. 11 1471–1501.

[14] C. B. Jones, R. S. Purves, P. D. Clough, and H. Joho, *Modelling vague places with knowledge from the web*, *International Journal of Geographical Information Science* **22** (2008), no. 10 1045–1065.

[15] L. Li and M. F. Goodchild, *Constructing places from spatial footprints*, in *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information*, pp. 15–21, ACM, 2012.

[16] J. Bernad, C. Bobed, E. Mena, and S. Ilarri, *A formalization for semantic location granules*, *International Journal of Geographical Information Science* **27** (2013), no. 6 1090–1108.

[17] D. R. Montello, *Scale and multiple psychologies of space*, in *Spatial information theory a theoretical basis for gis*, pp. 312–321. Springer, 1993.

[18] T. Tenbrink and S. Winter, *Variable granularity in route directions*, *Spatial Cognition & Computation* **9** (2009), no. 1 64–93.

[19] D. Richter, S. Winter, K.-F. Richter, and L. Stirling, *Granularity of locations referred to by place descriptions*, *Computers, Environment and Urban Systems* **41** (2013) 88–99.

[20] R. Hartshorne, *Perspective on the Nature of Geography.* Rand McNally, 1959.

[21] D. R. Montello, *Regions in geography: Process and content*, *Foundations of geographic information science* (2003) 173–189.

[22] P. L. Knox and S. A. Marston, *Human geography: Places and regions in global context (7th Edition).* Pearson, 2015.

[23] R. G. Golledge, *The nature of geographic knowledge*, *Annals of the Association of American Geographers* **92** (2002), no. 1 1–14.

[24] P. A. Burrough and A. Frank, *Geographic objects with indeterminate boundaries*, vol. 2. CRC Press, 1996.

[25] L. A. Brown and J. Holmes, *The delimitation of functional regions, nodal regions, and hierarchies by functional distance approaches*, *Journal of Regional Science* **11** (1971), no. 1 57–72.

[26] M. W. Smart, *Labour market areas: uses and definition*, *Progress in planning* **2** (1974) 239–353.

[27] M. G. Coombes, A. E. Green, and S. Openshaw, *An efficient algorithm to generate official statistical reporting areas: the case of the 1984 travel-to-work areas revision in britain*, *Journal of the operational research society* (1986) 943–953.

[28] V. T. Noronha and M. F. Goodchild, *Modeling interregional interaction: Implications for defining functional regions*, *Annals of the Association of American Geographers* **82** (1992), no. 1 86–102.

[29] M. Konjar, A. Lisec, and S. Drobne, *Methods for delineation of functional regions using data on commuters*, in *Proceedings of the 13-th AGILE International Conference on Geographic Information Science, Portugal*, 2010.

[30] C. Karlsson and M. Olsson, *The identification of functional regions: theory, methods, and applications*, *The Annals of Regional Science* **40** (2006), no. 1 1–18.

[31] C. J. Farmer and A. S. Fotheringham, *Network-based functional regions*, *Environment and Planning A* **43** (2011), no. 11 2723–2741.

[32] S. Gao, Y. Liu, Y. Wang, and X. Ma, *Discovering spatial interaction communities from mobile phone data*, *Transactions in GIS* **17** (2013), no. 3 463–481.

[33] C. Zhong, X. Huang, S. M. Arisona, G. Schmitt, and M. Batty, *Inferring building functions from a probabilistic model using public transportation data*, *Computers, Environment and Urban Systems* **48** (2014) 124–137.

[34] K. Janowicz, *Observation-driven geo-ontology engineering*, *Transactions in GIS* **16** (2012), no. 3 351–374.

[35] C. Mülligann, K. Janowicz, M. Ye, and W.-C. Lee, *Analyzing the spatial-semantic interaction of points of interest in volunteered geographic information*, in *Spatial information theory*, pp. 350–370. Springer, 2011.

[36] M. F. Goodchild, *Citizens as sensors: the world of volunteered geography*, *GeoJournal* **69** (2007), no. 4 211–221.

[37] S. Madden, *From databases to big data*, *Internet Computing, IEEE* **16** (2012), no. 3 4–6.

[38] K. Janowicz, S. Scheider, T. Pehle, and G. Hart, *Geospatial semantics and linked spatiotemporal data–past, present, and future*, *Semantic Web* **3** (2012), no. 4 321–332.

[39] C. Kang, S. Gao, X. Lin, Y. Xiao, Y. Yuan, Y. Liu, and X. Ma, *Analyzing and geo-visualizing individual human mobility patterns using mobile call records*, in *Geoinformatics, 2010 18th International Conference on*, pp. 1–7, IEEE, 2010.

[40] M. Ye, D. Shou, W.-C. Lee, P. Yin, and K. Janowicz, *On the semantic annotation of places in location-based social networks*, in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 520–528, ACM, 2011.

[41] Y. Liu, C. Kang, S. Gao, Y. Xiao, and Y. Tian, *Understanding intra-urban trip patterns from taxi trajectory data*, *Journal of geographical systems* **14** (2012), no. 4 463–483.

[42] Y. Liu, F. Wang, Y. Xiao, and S. Gao, *Urban land uses and traffic source-sink areas: Evidence from gps-enabled taxi data in shanghai*, *Landscape and Urban Planning* **106** (2012), no. 1 73–87.

[43] Y. Liu, Z. Sui, C. Kang, and Y. Gao, *Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data*, *PloS one* **9** (2014), no. 1 e86026.

[44] S. Gao, Y. Wang, Y. Gao, and Y. Liu, *Understanding urban traffic-flow characteristics: a rethinking of betweenness centrality*, *Environment and Planning B: Planning and Design* **40** (2013), no. 1 135–153.

[45] Y. Liu, X. Liu, S. Gao, L. Gong, C. Kang, Y. Zhi, G. Chi, and L. Shi, *Social sensing: A new approach to understanding our socioeconomic environments*, *Annals of the Association of American Geographers* **105** (2015), no. 3 512–530.

[46] S. Gao, *Spatio-temporal analytics for exploring human mobility patterns and urban dynamics in the mobile age*, *Spatial Cognition & Computation* **15** (2015), no. 2 86–114.

[47] B. Smith and A. C. Varzi, *Fiat and bona fide boundaries*, *Philosophical and Phenomenological Research* (2000) 401–420.

[48] A. Galton, *On the ontological status of geographical boundaries*, *Foundations of geographic information science* (2003) 151–171.

[49] A. U. Frank, *The prevalence of objects with sharp boundaries in gis*, *Geographic objects with indeterminate boundaries* (1996) 29–40.

[50] H. Couclelis, *People manipulate objects (but cultivate fields): beyond the raster-vector debate in gis*, in *Theories and methods of spatio-temporal reasoning in geographic space* (A. U. Frank, I. Campari, and U. Formentini, eds.), pp. 65–77. Springer, 1992.

[51] M. F. Goodchild, *Looking forward: Five thoughts on the future of gis*, *Esri ArcWatch* (2011).

[52] B. Bennett, *What is a forest? on the vagueness of certain geographic concepts*, *Topoi* **20** (2001), no. 2 189–201.

[53] F.-K. Holtmeier, *Mountain timberlines: ecology, patchiness, and dynamics*, vol. 36. Springer Science & Business Media, 2009.

[54] D. M. Mark, B. Smith, and B. Tversky, *Ontology and geographic objects: An empirical study of cognitive categorization*, in *Spatial Information Theory. Cognitive and Computational Foundations of Geographic Information Science* (C. Freksa and D. M. Mark, eds.), pp. 283–298. Springer, 1999.

[55] E. White and K. Stewart, *Barrier dynamics for gis: a design pattern for geospatial barriers*, *International Journal of Geographical Information Science* **29** (2015), no. 6 1–16.

[56] R. Casati and A. C. Varzi, *Parts and places: The structures of spatial representation*. MIT Press, 1999.

[57] A. G. Cohn and N. M. Gotts, *The egg-yolkrepresentation of regions with indeterminate boundaries*, *Geographic objects with indeterminate boundaries* **2** (1996) 171–187.

[58] D. R. Montello, M. F. Goodchild, J. Gottsegen, and P. Fohl, *Where's downtown?: Behavioral methods for determining referents of vague spatial queries*, *Spatial Cognition & Computation* **3** (2003), no. 2-3 185–204.

[59] S. C. Aitken and R. Prosser, *Residents' spatial knowledge of neighborhood continuity and form*, *Geographical Analysis* **22** (1990), no. 4 301–325.

[60] D. R. Montello, A. Friedman, and D. W. Phillips, *Vague cognitive regions in geography and geographic information science*, *International Journal of Geographical Information Science* **28** (2014), no. 9 1802–1820.

[61] D. R. Montello, *Cognitive geography*, *International encyclopedia of human geography* **2** (2009) 160–166.

[62] A. J. Hey, S. Tansley, K. M. Tolle, *et. al.*, *The fourth paradigm: data-intensive scientific discovery*, vol. 1. Microsoft Research Redmond, WA, 2009.

[63] K. Janowicz, F. van Harmelen, J. A. Hendler, and P. Hitzler, *Why the data train needs semantic rails.*, *AI Magazine* **36** (2015), no. 1 5–14.

[64] D. R. Montello and P. Sutton, *An introduction to scientific research methods in geography and environmental studies (2nd edition)*. Sage Publications, 2013.

[65] L. Hollenstein and R. Purves, *Exploring place through user-generated content: Using flickr tags to describe city cores*, *Journal of Spatial Information Science* (2010), no. 1 21–48.

[66] R. Purves, A. Edwardes, and J. Wood, *Describing place through user generated content*, *First Monday* **16** (2011), no. 9.

[67] C. Davies, I. Holt, J. Green, J. Harding, and L. Diamond, *User needs and implications for modelling vague named places*, *Spatial Cognition & Computation* **9** (2009), no. 3 174–194.

[68] C. B. Jones, R. S. Purves, P. D. Clough, and H. Joho, *Modelling vague places with knowledge from the web*, *International Journal of Geographical Information Science* **22** (2008), no. 10 1045–1065.

[69] Y. Liu, Y. Yuan, D. Xiao, Y. Zhang, and J. Hu, *A point-set-based approximation for areal objects: A case study of representing localities*, *Computers, Environment and Urban Systems* **34** (2010), no. 1 28–39.

[70] L. Li and M. F. Goodchild, *Constructing places from spatial footprints*, in *Proceedings of the 1st ACM SIGSPATIAL international workshop on crowdsourced and volunteered geographic information*, pp. 15–21, ACM, 2012.

[71] H. Hobel, P. Fogliaroni, and A. U. Frank, *Deriving the geographic footprint of cognitive regions*, in *Geospatial Data in a Changing World*, pp. 67–84. Springer, 2016.

[72] L. Li, M. F. Goodchild, and B. Xu, *Spatial, temporal, and socioeconomic patterns in the use of twitter and flickr*, *Cartography and Geographic Information Science* **40** (2013), no. 2 61–77.

[73] Z. Tufekci, *Big questions for social media big data: Representativeness, validity and other methodological pitfalls*, *arXiv preprint arXiv:1403.7400* (2014).

[74] M.-H. Tsou, J.-A. Yang, D. Lusher, S. Han, B. Spitzberg, J. M. Gawron, D. Gupta, and L. An, *Mapping social activities and concepts with social media (twitter) and web search engines (yahoo and bing): a case study in 2012 us presidential election*, *Cartography and Geographic Information Science* **40** (2013), no. 4 337–348.

[75] B. Adams, G. McKenzie, and M. Gahegan, *Frankenplace: Interactive thematic mapping for ad hoc exploratory search*, in *24th International World Wide Web Conference*. IW3C2, 2015.

[76] E. Steiger, R. Westerholt, and A. Zipf, *Research on social media feeds–a giscience perspective*, in *European Handbook of Crowdsourced Geographic Information* (C. Capineri, M. Haklay, H. Huang, V. Antoniou, J. Kettunen, F. Ostermann, and R. Purves, eds.), pp. 237–254. London: Ubiquity Press, 2016.

[77] M. F. Goodchild, *Citizens as sensors: the world of volunteered geography*, *GeoJournal* **69** (2007), no. 4 211–221.

[78] M. Haklay and P. Weber, *Openstreetmap: User-generated street maps*, *Pervasive Computing, IEEE* **7** (2008), no. 4 12–18.

[79] L. N. Mummidi and J. Krumm, *Discovering points of interest from users map annotations*, *GeoJournal* **72** (2008), no. 3-4 215–227.

[80] S. Gao, J.-A. Yang, B. Yan, Y. Hu, K. Janowicz, and G. McKenzie, *Detecting origin-destination mobility flows from geotagged tweets in greater los angeles area*, in *Proceedings of the Eighth International Conference on Geographic Information Science*, pp. 1–4, 2014.

[81] C. Keßler, P. Maué, J. T. Heuer, and T. Bartoschek, *Bottom-up gazetteers: Learning from the implicit semantics of geotags*, in *GeoSpatial semantics* (K. Janowicz, M. Raubal, and S. Levashkin, eds.), pp. 83–102. Springer, 2009.

[82] Y. Hu, S. Gao, K. Janowicz, B. Yu, W. Li, and S. Prasad, *Extracting and understanding urban areas of interest using geotagged photos*, *Computers, Environment and Urban Systems* **54** (2015) 240 – 254.

[83] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, *The new data and new challenges in multimedia research*, *arXiv preprint arXiv:1503.01817* (2015).

[84] M. Duggan, N. B. Ellison, C. Lampe, A. Lenhart, and M. Madden, *Social media update 2014*, tech. rep., Pew Research Center, January, 2015.

[85] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann, *Dbpedia-a crystallization point for the web of data*, *Web Semantics: science, services and agents on the world wide web* **7** (2009), no. 3 154–165.

[86] H. Kwak, C. Lee, H. Park, and S. Moon, *What is twitter, a social network or a news media?*, in *Proceedings of the 19th international conference on World wide web*, pp. 591–600, ACM, 2010.

[87] S. Gao, L. Li, W. Li, K. Janowicz, and Y. Zhang, *Constructing gazetteers from volunteered big geo-data based on hadoop*, *Computers, Environment and Urban Systems* (2017).

[88] M. G. Kendall and B. B. Smith, *The problem of m rankings*, *The annals of mathematical statistics* **10** (1939), no. 3 275–287.

[89] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, *A density-based algorithm for discovering clusters in large spatial databases with noise.*, in *Kdd*, pp. 226–231, AAAI Press, 1996.

[90] F. P. Preparata and S. J. Hong, *Convex hulls of finite sets of points in two and three dimensions*, *Communications of the ACM* **20** (1977), no. 2 87–93.

[91] M. Duckham, L. Kulik, M. Worboys, and A. Galton, *Efficient generation of simple polygons for characterizing the shape of a set of points in the plane*, *Pattern Recognition* **41** (2008), no. 10 3224–3236.

[92] F. Akdag, C. F. Eick, and G. Chen, *Creating polygon models for spatial clusters*, in *Foundations of Intelligent Systems* (T. Andreasen, H. Christiansen, J.-C. Cubero, and Z. W. Ra, eds.), pp. 493–499. Springer, 2014.

[93] D. M. Blei, A. Y. Ng, and M. I. Jordan, *Latent dirichlet allocation*, *Journal of Machine Learning Research* **3** (2003), no. Jan 993–1022.

[94] A. K. McCallum, *Mallet: A machine learning for language toolkit*, tech. rep., University of Massachusetts Amherst, 2002.

[95] T. L. Griffiths and M. Steyvers, *Finding scientific topics*, *Proceedings of the National Academy of Sciences* **101** (2004), no. suppl 1 5228–5235.

[96] B. Adams and K. Janowicz, *Thematic signatures for cleansing and enriching place-related linked data*, *International Journal of Geographical Information Science* **29** (2015), no. 4 556–579.

[97] G. McKenzie, K. Janowicz, S. Gao, J.-A. Yang, and Y. Hu, *Poi pulse: A multi-granular, semantic signatures-based approach for the interactive visualization of big geosocial data*, *Cartographica: The International Journal for Geographic Information and Geovisualization, The University of Toronto Press* **50** (2015), no. 2 71–85.

[98] S. Kullback and R. A. Leibler, *On information and sufficiency*, *The Annals of Mathematical Statistics* **22** (1951), no. 1 79–86.

[99] B. Adams and K. Janowicz, *On the geo-indicativeness of non-georeferenced text.*, in *ICWSM*, pp. 375–378, 2012.

[100] M. M. Louwerse and N. Benesh, *Representing spatial structure through maps and language: Lord of the rings encodes the spatial structure of middle earth*, *Cognitive science* **36** (2012), no. 8 1556–1569.

[101] Y. Ikawa, M. Vukovic, J. Rogstadius, and A. Murakami, *Location-based insights from the social web*, in *Proceedings of the 22nd International Conference on World Wide Web*, pp. 1013–1016, ACM, 2013.

[102] O. Ajao, J. Hong, and W. Liu, *A survey of location inference techniques on twitter*, *Journal of Information Science* **41** (2015), no. 6 855–864.

[103] D. R. Montello, *Scale in geography*, in *The international encyclopedia of social and behavioral sciences (2nd ed.)* (J. Wright, ed.). Oxford: Elsevier, 2013.

[104] M. J. Barnsley and S. L. Barr, *Inferring urban land use from satellite sensor images using kernel-based spatial reclassification*, *Photogrammetric Engineering and Remote Sensing* **62** (1996), no. 8 949–958.

[105] M. Herold, H. Couclelis, and K. C. Clarke, *The role of spatial metrics in the analysis and modeling of urban land use change*, *Computers, Environment and Urban Systems* **29** (2005), no. 4 369–399.

[106] E. Banzhaf and M. Netzband, *Monitoring urban land use changes with remote sensing techniques*, *Applied Urban Ecology: A Global Framework* (2012) 18–32.

[107] T. Pei, S. Sobolevsky, C. Ratti, S.-L. Shaw, T. Li, and C. Zhou, *A new insight into land use classification based on aggregated mobile phone data*, *International Journal of Geographical Information Science* **28** (2014), no. 9 1988–2007.

[108] G. McKenzie, K. Janowicz, S. Gao, J.-A. Yang, and Y. Hu, *Poi pulse: A multi-granular, semantic signature–based information observatory for the interactive visualization of big geosocial data*, *Cartographica: The International Journal for Geographic Information and Geovisualization* **50** (2015), no. 2 71–85.

[109] Y. Hu, S. Gao, K. Janowicz, B. Yu, W. Li, and S. Prasad, *Extracting and understanding urban areas of interest using geotagged photos*, *Computers, Environment and Urban Systems* **54** (2015) 240–254.

[110] S. Jiang, A. Alves, F. Rodrigues, J. Ferreira, and F. C. Pereira, *Mining point-of-interest data from social networks for urban land use classification and disaggregation*, *Computers, Environment and Urban Systems* **53** (2015) 36–46.

[111] E. Steiger, R. Westerholt, and A. Zipf, *Research on social media feeds–a giscience perspective*, *European Handbook of Crowdsourced Geographic Information* (2016) 237.

[112] Y. Yao, X. Li, X. Liu, P. Liu, Z. Liang, J. Zhang, and K. Mai, *Sensing spatial distribution of urban land use by integrating points-of-interest and google word2vec model*, *International Journal of Geographical Information Science* **31** (2017), no. 4 825–848.

[113] B. Adams and K. Janowicz, *On the geo-indicativeness of non-georeferenced text*, in *ICWSM*, pp. 375–378, 2012.

[114] B. Adams and G. McKenzie, *Inferring thematic places from spatially referenced natural language descriptions*, in *Crowdsourcing Geographic Knowledge*, pp. 201–221. Springer, 2013.

[115] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil, *Exploiting semantic annotations for clustering geographic areas and users in location-based social networks.*, *The Social Mobile Web* **11** (2011) 02.

[116] J. Yuan, Y. Zheng, and X. Xie, *Discovering regions of different functions in a city using human mobility and pois*, in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 186–194, ACM, 2012.

[117] B. Adams, *Finding similar places using the observation-to-generalization place model*, *Journal of Geographical Systems* **17** (2015), no. 2 137–156.

[118] B. Adams and K. Janowicz, *Thematic signatures for cleansing and enriching place-related linked data*, *International Journal of Geographical Information Science* **29** (2015), no. 4 556–579.

[119] H. Hobel, A. Abdalla, P. Fogliaroni, and A. U. Frank, *A semantic region growing algorithm: extraction of urban settings*, in *AGILE 2015*, pp. 19–33. Springer, 2015.

[120] X. Zhou and L. Zhang, *Crowdsourcing functions of the living city from twitter and foursquare data*, *Cartography and Geographic Information Science* **43** (2016), no. 5 393–404.

[121] Y. Zhi, H. Li, D. Wang, M. Deng, S. Wang, J. Gao, Z. Duan, and Y. Liu, *Latent spatio-temporal activity structures: a new approach to inferring intra-urban functional regions via social media check-in data*, *Geo-spatial Information Science* **19** (2016), no. 2 94–105.

[122] G. McKenzie and K. Janowicz, *The effect of regional variation and resolution on geosocial thematic signatures for points of interest*, in *Proceedings of the 2017 AGILE Conference*, pp. 237–256, Springer, 2017.

[123] H. Hobel, P. Fogliaroni, and A. U. Frank, *Deriving the geographic footprint of cognitive regions*, in *Geospatial Data in a Changing World*, pp. 67–84. Springer, 2016.

[124] S. Gao, K. Janowicz, D. R. Montello, Y. Hu, J.-A. Yang, G. McKenzie, Y. Ju, L. Gong, B. Adams, and B. Yan, *A data-synthesis-driven method for detecting and extracting vague cognitive regions*, *International Journal of Geographical Information Science* (2017) 1245–1271.

[125] M. Steyvers and T. Griffiths, *Probabilistic topic models*, *Handbook of Latent Semantic Analysis* **427** (2007), no. 7 424–440.

[126] D. M. Blei, *Probabilistic topic models*, *Communications of the ACM* **55** (2012), no. 4 77–84.

[127] J. Cao, T. Xia, J. Li, Y. Zhang, and S. Tang, *A density-based method for adaptive lda model selection*, *Neurocomputing* **72** (2009), no. 7 1775–1781.

[128] R. Arun, V. Suresh, C. V. Madhavan, and M. N. Murthy, *On finding the natural number of topics with latent dirichlet allocation: Some observations*, in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 391–402, Springer, 2010.

[129] J. MacQueen *et. al.*, *Some methods for classification and analysis of multivariate observations*, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 281–297, Oakland, CA, USA., 1967.

[130] P. J. Rousseeuw, *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis*, *Journal of Computational and Applied Mathematics* **20** (1987) 53–65.

[131] R. M. Assunção, M. C. Neves, G. Câmara, and C. da Costa Freitas, *Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees*, *International Journal of Geographical Information Science* **20** (2006), no. 7 797–811.

[132] J. C. Gower and G. Ross, *Minimum spanning trees and single linkage cluster analysis*, *Applied Statistics* (1969) 54–64.

[133] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques (Third Edition)*. Elsevier, Waltham, MA, 2011.

[134] J. Lin, *Divergence measures based on the shannon entropy*, *IEEE Transactions on Information Theory* **37** (1991), no. 1 145–151.

[135] W. M. Rand, *Objective criteria for the evaluation of clustering methods*, *Journal of the American Statistical association* **66** (1971), no. 336 846–850.

[136] A. Strehl and J. Ghosh, *Cluster ensembles—a knowledge reuse framework for combining multiple partitions*, *Journal of Machine Learning Research* **3** (2002), no. Dec 583–617.

[137] R. B. McMaster and K. S. Shea, *Generalization in digital cartography*, Association of American Geographers Washington, DC, 1992.

[138] J. K. Wright, *Map makers are human: Comments on the subjective in maps*, *Geographical Review* (1942) 527–544.

[139] H. Arthur Robinson, *Elements of cartography*. John Wiley And Sons, Inc; New York, 1958.

[140] F. Töpfer and W. Pillewizer, *The principles of selection*, *The Cartographic Journal* **3** (1966), no. 1 10–16.

[141] K. E. Brassel and R. Weibel, *A review and conceptual framework of automated map generalization*, *International Journal of Geographical Information System* **2** (1988), no. 3 229–244.

[142] K. S. Shea and R. B. McMaster, *Cartographic generalization in a digital environment: When and how to generalize*, in *Proceedings of AutoCarto*, vol. 9, pp. 56–67, 1989.

[143] D. M. Mark, *Conceptual basis for geographic line generalization*, in *Proc. 9th International Symposium on Computer-Assisted Cartography*, pp. 68–77, 1989.

[144] S. Scellato, A. Noulas, and C. Mascolo, *Exploiting place features in link prediction on location-based social networks*, in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1046–1054, ACM, 2011.

[145] E. Spyrou, M. Korakakis, V. Charalampidis, A. Psallas, and P. Mylonas, *A geo-clustering approach for the detection of areas-of-interest and their underlying semantics*, *Algorithms* **10** (2017), no. 1 35.

[146] S. Gao, K. Janowicz, G. McKenzie, and L. Li, *Towards platial joins and buffers in place-based gis.*, in *COMP@ SIGSPATIAL*, pp. 42–49, 2013.

[147] R. Rada, H. Mili, E. Bicknell, and M. Blettner, *Development and application of a metric on semantic nets*, *Systems, Man and Cybernetics, IEEE Transactions on* **19** (1989), no. 1 17–30.

[148] D. Harel and R. E. Tarjan, *Fast algorithms for finding nearest common ancestors*, *SIAM Journal on Computing* **13** (1984), no. 2 338–355.

[149] S. Openshaw, M. Charlton, C. Wymer, and A. Craft, *A mark 1 geographical analysis machine for the automated analysis of point data sets*, *International Journal of Geographical Information System* **1** (1987), no. 4 335–358.

[150] C. E. Shannon, *A mathematical theory of communication*, *ACM SIGMOBILE Mobile Computing and Communications Review* **5** (2001), no. 1 3–55.

[151] M. Ye, K. Janowicz, C. Mülligann, and W.-C. Lee, *What you are is when you are: the temporal dimension of feature types in location-based social networks*, in *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 102–111, ACM, 2011.

[152] G. McKenzie, K. Janowicz, S. Gao, and L. Gong, *How where is when? on the regional variability and resolution of geosocial temporal signatures for points of interest*, *Computers, Environment and Urban Systems* **54** (2015) 336–346.

[153] J. H. Ward Jr, *Hierarchical grouping to optimize an objective function*, *Journal of the American statistical association* **58** (1963), no. 301 236–244.

[154] D. A. Randell, Z. Cui, and A. G. Cohn, *A spatial logic based on regions and connection.*, *KR* **92** (1992) 165–176.

[155] M. F. Goodchild, *Space, place and health*, *Annals of GIS* **21** (2015), no. 2 97–100.