UNIVERSITY OF CALIFORNIA

Santa Barbara

What Questions Do Users Ask about GIS: an Analysis of Posts on GIS Stack Exchange

A Thesis submitted in partial satisfaction of the

requirements for the degree Master of Arts

in Geography

by

Jingyi Xiao

Committee in charge:

Professor Werner Kuhn, Chair

Professor Krzysztof Janowicz

Professor Karen Kemp, University of Southern California

December 2019

The thesis of Jingyi Xiao is approved.

Krzysztof Janowicz

Karen Kemp

Werner Kuhn, Committee Chair

December 2019

What Questions Do Users Ask about GIS: an Analysis of Posts on GIS Stack Exchange

by

Jingyi Xiao

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my committee chair Professor Werner Kuhn for his constructive suggestions, patient guidance, encouragement and generous devotion of the time throughout the process of working on and writing up the thesis. I would also like to express my great appreciation to my committee members, Professor Krzysztof Janowicz and Professor Karen Kemp for their valuable critiques, suggestions and generous support in completing this work.

I would also like to extend my thanks to the colleagues at the Center of Spatial Studies and the Department of Geography for their helpful suggestions.

I would like to thank my dear friends who keep me company.

Finally, my great thanks and appreciations go to my parents for their encouragement and support throughout my graduate study.

ABSTRACT

What Questions Do Users Ask about GIS: an Analysis of Posts on GIS Stack Exchange

by

Jingyi Xiao

In recent years, the easy access to geospatial data has increased the opportunities and demand for geospatial analysis across disciplines. Yet, it is often not obvious for non-geographers how to use Geographic Information Systems (GIS) or other geospatial analysis tools. A popular place for users to seek answers to their questions is online forums, such as GIS Stack Exchange. Studying questions on these forums may help uncover what users usually ask about GIS in order to answer their questions about the world. In this paper, a Latent Dirichlet Allocation (LDA) topic model is used to explore the topics users ask about on GIS Stack Exchange. The 15 topics identified from over 40,000 posts on GIS Stack Exchange cover a broad spectrum, ranging from geospatial information representation (e.g., coordinate systems), geospatial information retrieval (e.g., geospatial querying) to geospatial computing (e.g., raster and vector computations) and geospatial data visualization (e.g., mapping). The number of posts, answer rates and answer times vary by topic clusters, implying their relative popularity and difficulty. To complement the analysis, I also compared the top 20 tags used in GIS and Cross Validated (Statistics) Stack Exchange. The domination of software-related tags in GIS contrasts with the prevalence of tags standing for software-independent fundamental concepts (e.g., hypothesis test, probability and distributions) for statistics questions. This supports my observation, from the LDA topic model, that the questions users ask about GIS are mainly about data models and software procedures. It indicates that the conceptual basis for GIS has not yet reached the clarity and consensus found in statistics. This study contributes empirical evidence on gaps in the conceptual basis underlying GIS design, education, and training. It also offers insights to GIS educators and software developers who aim at making GIS more accessible and easier to use across disciplines, inspired by the success of statistics along these lines.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

I. Introduction

More and more geospatial data are becoming publicly available and used in various disciplines outside geography, where geospatial thinking and computing often provide novel perspectives on scientific and practical questions. For example, Type 2 diabetes mellitus has been found to be associated with physical and social environments of patients via Geographic Information Systems (GIS) (Christine et al., 2015); economist Paul Krugman (1991) regards economic geography as one of the most “striking features of real-world economics” that is still too often overlooked in data collection and analysis (p. 483); mapping intergenerational mobility by commuting zones helps in assessing the roles and relative importance of income, education, and other factors in social mobility (Chetty, Hendren, Kline, & Saez, 2014).

While geospatial computing and analysis could thus be beneficially used to aid research and problem-solving in many domains, it is often not obvious for non-geographers (and even geographers) how to adequately and effectively use GIS or other geospatial analysis tools (Sipe & Dale, 2003). GIS researchers and educators have been trying to make GIS more available across domains by establishing a software-independent conceptual basis for GIS. Such attempts include the dichotomy of field-based and object-based models (Peuquet, 1988; Goodchild, 1991; Couclelis, 1992), the unifying geo-atom model (Goodchild, Yuan, & Cova, 2007), the core concepts of spatial information (Kuhn, 2012), as well as taxonomies of geospatial analysis questions (Cappelli, 2013) and GIS functionalities (Dangermond, 1983; Rhind & Green, 1988; Albrecht, 1998; Gao & Goodchild, 2013). However, more often than not, these attempts are made in a top-down manner, often mainly based on the researchers’ own expertise and experience or small numbers of test subjects.

The ubiquity of web access has changed the way people communicate ideas and exchange knowledge and has made rich data sources available to researchers and practitioners. Question-answering (Q&A) sites such as Stack Exchange, Quara, Yahoo! Answers and others have become popular because they enable knowledge sharing by users all over the world in a timely manner. GIS users with various backgrounds nowadays are seeking and often obtaining answers to their questions through these online forums, in particular through GIS Stack Exchange, because they typically cannot afford the time to learn about the specifics of data structures and algorithms. Mining their questions promises valuable information on what they ask about GIS and what topics are interesting and/or difficult to them. Therefore, the research question of this thesis is what do users frequently ask about GIS in order to answer their questions about the world.

In this research, I have applied Latent Dirichlet Allocation (LDA) to explore over 40,000 posts on GIS Stack Exchange regarding what questions users ask about GIS. The posts were categorized into 15 topics, each topic being characterized by a list of terms with the highest occurrence probabilities and illustrated with actual posts as examples. I find that the 15 topics cover a broad spectrum of GIS notions, ranging from geospatial information representation (coordinate systems), geospatial information retrieval (e.g., geospatial querying) to geospatial computing (e.g., raster and vector computations) and geospatial data visualization (e.g., mapping). Clusters were then computed by the cosine similarity between the vectors representing the topics in the hyper-dimensional space of posts associated with them. The number of posts, view counts, answer rates and answer times per cluster reflect the popularity and/or difficulty of the cluster themes. In particular, the number of questions on raster and vector data and computations are higher than other topic clusters, and questions on (geospatial) querying and coordinate systems are usually answered in a shorter time. To complement my analysis from a different perspective, I compared the top 20 tags used in GIS and statistics questions on Stack Exchange. I find that almost all (19 out of 20) GIS tags are about software products (e.g. QGIS, ArcGIS-Desktop), data formats (e.g. raster, shapefile), and programming languages (e.g. Python, R), while the great majority of tags in statistics (16 out of 20) are about software-independent statistical concepts (e.g. probability, distribution, testing). This suggests that the conceptual basis for GIS has not yet reached the level of generality, clarity, and consensus found in statistics, making it harder for GIS users to seek answers to geospatial questions, as they are busy thinking about software procedures and their sequencing.

My work complements top-down approaches to organize the GIS domain with a bottom-up method mining the Q&A archives of GIS Stack Exchange. The outcome reveals the topics that current GIS users discuss. The findings offer facts and insights to GIS educators and developers who aim at making GIS more accessible and easier to use for a wider range of disciplines and practices. To the best of my knowledge, no analysis on GIS Stack Exchange has been conducted so far with similar goals.

The thesis is organized as follows. In Section 2, I discuss related work on GIS concepts and operations. Section 3 introduces the dataset used in this study, followed by the analysis and results in Section 4. Discussion of the results and limitations along with conclusions are presented in Section 5.

II. Related Work

Attempts to make GIS easier to learn and use loosely fall into two categories: (1) define conceptual models that could be perceived and understood by users without detailed technical knowledge or training; (2) classify GIS operations at a semantic level rather than at the level of data formats.

An early example of the first category is Nystuen (1963), who identified direction, distance and connection or relative position as fundamental spatial concepts. Spurred by the National Center for Geographic Information and Analysis (NCGIA), researchers took the initiative to identify the cognitive models and develop formal mathematical models (e.g. binary topological relationships (Egenhofer, 1989), cardinal directions reasoning (Frank, 1996)) of geospatial concepts along with their relations to natural language in the late 1980s. Peuquet (1988) suggested a framework with a dual conceptual model, i.e. location-based and object-based representations, as the basis of geographical phenomena. This idea was further developed by Goodchild (1991) and Couclelis (1992). Later, Marsh, Golledge and Battersby (2007) developed five levels of geospatial concepts for teaching and learning geography in the US, namely primitive, simple, difficult, complicated, and complex (levels of geospatial concepts). Their work was followed by the taxonomy of geospatial thinking proposed by Jo and Bednarz (2009) and the foundation concepts in geospatial thinking suggested by Janelle and Goodchild (2011).

However, these concepts were introduced for the purpose of facilitating and organizing geospatial thinking; they are not closely tied to the use of GIS or to questions these systems should be able to answer. Cappelli (2013) presented a taxonomy of geospatial analysis questions with six high-level categories of questions, including understanding where, measuring size, shape, and distribution, determining how places are related, finding the best locations and paths, detecting and quantifying patterns, and making predictions. Core concepts of spatial information were proposed as a high-level description of what geographic information is about (Kuhn & Ballatore, 2015). To the best of my knowledge, none of these proposals falling in my first category has been systematically validated for completeness or cognitive adequacy.

Among the approaches in my second category, targeting some form of GIS operations taxonomies, map algebra (Tomlin, 1983) stands out in many ways. It was an early and rare attempt to organize geospatial questions and the computations to answer them around a mathematically well-defined set of operations. In its commonly known practical form, map algebra provides a concise organization of operations on gridded geospatial data. These operations (i.e., local, focal, zonal, and global) were derived from the fundamental structures in the early computer programs SYMAP (Fisher, 1968) and GRID (Sinton & Steinitz, 1969), commonly seen as GIS ancestors. Later, Rhind and Green (1988), based on previous work and their own experience, presented a classification of GIS functions into data input and encoding, data manipulation, data retrieval, data analysis, data display, and database management for a heterogeneous scientific community. Dangermond (1983) suggested a classification of GIS functionality into map automation and database creation, analytic manipulation techniques, database manipulation techniques and graphic manipulation techniques. Albrecht (1998) developed a classification of domain- and data model-independent GIS operations, i.e. search/(re)classification, location analysis, terrain analysis, distribution/neighborhood, geospatial analysis/statistics, and measurements, from a series of user interviews. Gao and Goodchild (2013) identified a list of typical questions and computations, i.e. search/location/extent, data basics/processing/conversion, distributions/patterns/neighborhood, relations/associations, terrain/surface and time, and proposed this as a semantic framework for designing new GIS user interfaces. While the approaches in the second category are by their nature more empirical (classifying actual or sometimes desired GIS procedures), they are often based on the authors' experience, introspections and understanding, or on results from interviews with a manageably small group of GIS practitioners.

III. Dataset

Stack Exchange is a network that consists of over 170 question-and-answer (Q&A) communities on a wide variety of fields, ranging from science and technology to life, arts and culture. It offers rich information and has been broadly studied for a variety of purposes. For instance, Xia, Lo, Wang, and Zhou (2013) developed a tool *TagCombine* that can recommend tags for question-answering sites like Stack Overflow. Posts on Stack Overflow have also been explored to understand the most confusing programming concepts (Allamanis, 2013) and types of security-related questions (Yang, Lo, Xia, Wan, & Sun, 2016).

GIS Stack Exchange, as a Q&A community for cartographers, geographers, and GIS professionals, has had over 100,000 questions and 126,000 answers posted by around 100,000 users, providing abundant data about the use and understanding of GIS. As stated earlier, the objective of this study is to gain a better understanding of what questions users ask about GIS in order to answer questions about the world. Compared to alternative approaches such as user surveys, using the data on GIS Stack Exchange provides several advantages:

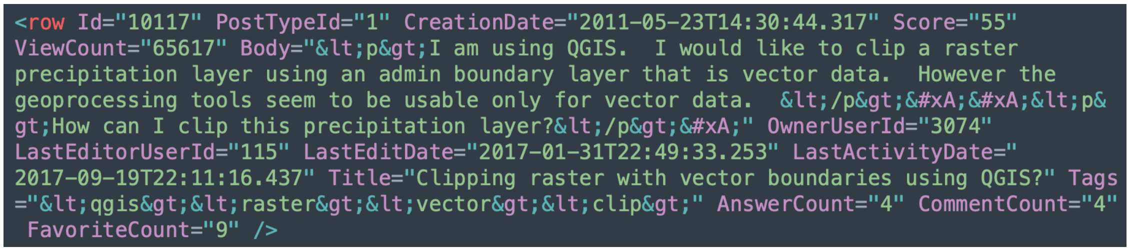
- It reveals what GIS users actually ask (in such an online community), as opposed to what they think they do not know, possibly prompted by researcher-suggested topics;
- The contents have been contributed by over 100,000 online users worldwide (though mainly from North America and Europe, biased toward English speakers), with backgrounds and expertise in many different domains;
- The data have been accumulated and timestamped since 2010, making an analysis over time feasible.

In short, this dataset is much more comprehensive and representative in quantity, geography, timespan and disciplines than data resulting from traditional information gathering methods like surveys.

Stack Exchange data is publicly available and free to download from *Internet Archive*[1]. The contents of the GIS site (as well as those of any other site) are available as a separate archive of XML files. The downloaded GIS Stack Exchange archive[2] contains Posts, Users, Votes, Comments, Post History and Post Links files. The file "Posts.xml" is the data used in this study. It consists of 106,149 posts between January 2010 and December 2018. An example of a post in "Posts.xml" is shown in Figure 1. Each post has an ID, post type ID (PostTypeId = 1 means the post is a question while 2 means it is an answer), owner user ID, last editor user ID, post body, tags, view counts, answer counts, comment counts, scores, favorite counts, create date, last edit date, and last active date.

[1] Internet Archive: https://archive.org/download/stackexchange.

[2] Data last updated on December 2nd, 2018.



```
<row Id="10117" PostTypeId="1" CreationDate="2011-05-23T14:30:44.317" Score="55"
ViewCount="65617" Body="&lt;p&gt;I am using QGIS.  I would like to clip a raster
precipitation layer using an admin boundary layer that is vector data.  However the
geoprocessing tools seem to be usable only for vector data.  &lt;/p&gt;&#xA;&#xA;&lt;p&
gt;How can I clip this precipitation layer?&lt;/p&gt;&#xA;" OwnerUserId="3074"
LastEditorUserId="115" LastEditDate="2017-01-31T22:49:33.253" LastActivityDate="
2017-09-19T22:11:16.437" Title="Clipping raster with vector boundaries using QGIS?" Tags
="&lt;qgis&gt;&lt;raster&gt;&lt;vector&gt;&lt;clip&gt;" AnswerCount="4" CommentCount="4"
 FavoriteCount="9" />
```

Figure 1. An example post in the file “Post.xml”.

All posts have tags, i.e., a collection of words or phrases chosen by the poster to describe the content of the post. Posters can create new tags if they are not already in the current tag collection. Tags comprise an unstructured variety of subject matters. Some refer to software platforms, others to programming or scripting languages, and yet others to tasks users want to solve (e.g. calculate a distance). Each tag has its own editable tag wiki produced by users to describe its meaning. It also has a tag frequency, indicating how many posts use that tag. For instance, the tag *vector*, which is defined in the wiki as “a coordinate-based data model that represents geographic features as points, lines and polygons”, has been used in 1,299 posts. Tags are mainly used in three ways:

- To identify or narrow down the topics that interest users. Users can browse posts by tags.
- To label a post in multiple ways. A post typically has multiple tags, referring to various aspects, in order to increase the discoverability of the post.
- To point “experts” to the questions they think they are able to answer.

The total number of distinct tags used in the dataset is 2,473. The ranking of the tags by frequency follows a power law distribution with a “long tail”, with R-squared (the goodness-of-fit criterion) of 0.892 shown in Figure 2, indicating the vast majority of posts uses only a small set of tags.

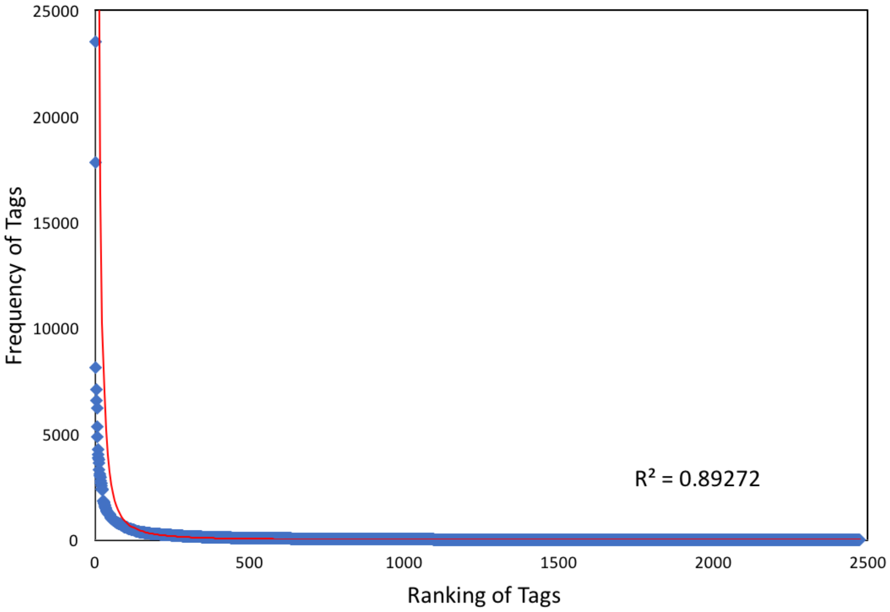


Figure 2. The ranking of the tags by frequency and their frequencies follow a power law distribution with R-squared of 0.892.

IV. Analysis and Results

In this section, I conducted three analyses on the data obtained from GIS Stack Exchange, namely topic identification, topic clustering, and tag comparison. An LDA topic model was used to identify the topics frequently discussed by GIS users. The clusters of topics were explored and visualized through a dendrogram, with some statistics revealing the thematic and temporal patterns formed by clusters. Furthermore, the 20 most frequently used tags on GIS Stack Exchange were compared with those on Cross Validated (statistics) Stack Exchange.

A. Topic Identification

The downloaded posts are largely concerned with six thematic areas (derived by manual inspection, not necessarily comprehensive), illustrated with actual posts shown in Table 1.

Table 1. Six thematic areas of the posts

Thematic Area	Example Posts
Data import, edit, and export	- Adding GPX files into ArcMap? (58,562 views; Tags: arcgis-desktop, arcmap, gpx) - How to export attribute table to Excel from QGIS? (66,428 views; Tags: qgis, qgis-plugins, attribute-table, export, excel)
Programming languages and libraries	- Installing GDAL with Python on windows? (94,886 views; Tags: python, gdal, installation, windows) - Programmatic authentication to ArcGIS Server secured layers via RESTful API (15,293 views; Tags: arcgis-10.1, arcgis-rest-api, security, authentication, arcgis-javascript-api)
Cartography	- Seeking examples of beautiful maps? (90,598 views; Tags: cartography) - Styling road maps with QGIS? (15,543 views; Tags: qgis, cartography, visualisation, references)
Data sources and data quality	- Listing available online WMS services (Weather, Land Data, Place Names)? (105,321 views; Tags: data, wms) - Seeking administrative boundaries for various countries? (77,601 views; Tags: data, global) - Where to get 2010 Census Block data? (40,172 views; Tags: data, census, united-states)
Learning materials	- Seeking QGIS tutorials and web resources? (17,679 views; Tags: qgis, pyqgis, qgis-3.0, references) - How do I develop my GIS programming skills? (15,521 views; Tags: python, c++, references)
Geospatial tools, operations and analyses	- Calculating polygon areas in QGIS? (125,951 views; Tags: qgis, shapefile, field-calculator, area) - Create a new layer from overlap between two layers? (35,810 views; Tags: qgis, intersection)

The last thematic area is of interest because it groups the questions that are about using GIS to directly answer questions about the world. However, it is impractical to go over the 106,149 posts manually to find the ones that fall into the last thematic area. As stated earlier, each post has at least one tag. A tag filtering scheme was therefore used to filter out the posts not belonging to the last thematic area. To do so, all 2,473 tags with tag wikis were evaluated by the author manually and filtered out if not related to the last thematic area. Finally, 257

(10.39%) tags were retained[3]. Posts were preserved if they have any of the 257 tags. The total number of posts remaining was 42,731 (40.26% of the original 106,149 posts).

Topic models can provide insights into large collections of texts. Latent Dirichlet Allocation (LDA, Blei, Ng & Jordan, 2003), as a generative probabilistic model for collections of data such as text corpora, helps find hidden topics in a collection of documents. As a well-known topic modeling method, LDA has been used widely for information extraction in various fields. Compared to other topic models such as Latent Semantic Analysis (LSA), probabilistic latent semantic analysis (PLSA) and hierarchical LDA (hLDA), LDA is more popular (with over 28,000 citations). Labeled LDA (Ramage, Hall, Nallapati, & Manning, 2009) requires topic labels in advance and cannot generate new topic labels for the texts, which means it cannot serve the purpose of disclosing the topics of the posts. Therefore, I chose LDA as the topic model generator.

I first applied text cleaning to the titles, text bodies, and accepted answers (if any) of the 42,731 posts. This step included the removal of code snippets, HTML tags, URLs, numbers, punctuation, stop words such as “a”, “the” and non-alphabetic characters. Then, words were lemmatized, i.e. transformed to their base form found in a dictionary.

After tag filtering and text cleaning, I computed the unigram, bigram and trigram frequencies of each post. A n-gram is a contiguous sequence of n words from a given text. Term sequences like 2-grams “coordinate system” and “raster calculator” are more meaningful than their isolated components “coordinate”, “system”, “raster” and “calculator” and need to be captured as wholes. Finally, all resulting terms (words or n-grams) were used

[3] See appendix for the full list of the 257 tags.

to construct a post-term matrix $\boldsymbol{P}$ for the LDA topic model, where $\boldsymbol{P}(i, j)$ represents the frequency of the *j*-th term in the *i*-th post.

Applying LDA to the data (using Python's machine learning library *scikit-learn*[4]) generated a probability distribution for each post over a set of topics (e.g., post 1 is 80% probability about topic 1 and 20% probability about topic 2), where each topic consisted of a set of terms with descending occurrence probability in that topic (e.g., term 1 has 60% occurrence probability in topic 1). The LDA method itself does not suggest a number of topics for given texts. To determine the number of topics, I tried different numbers of topics and chose a number that produced topics not too general or too repetitive. Actually, the topics were quite stable when topic number was chosen as 15 plus or minus 2. Therefore, the number of topics was set to 15, and the result is shown in Table 2. Each topic is represented by the 12 terms with the highest probabilities of occurring in the documents (posts) about that topic. I also supplied an actual post (rendered in italics) as an example for each topic. Since LDA does not label topics, the characterizations given in the *topic* column in Table 2 are only indicative of the posts associated with each topic (not just 12 terms but also the following terms that are not shown) for easier reference. Despite the risk of misrepresenting the topics coming out of LDA, this labeling is a commonly used approach (Yang, Lo, Xia, Wan, & Sun, 2016).

Table 2. Fifteen topics, the twelve terms with the highest probabilities, and example posts

Topic	**12 terms and example posts**
database computations	field, value, calculator, expression, field calculator, contour, string, date, number, calculate, raster calculator

[4] https://scikit-learn.org/stable/

	Writing conditional (if/then) statements into ArcGIS Field Calculator using Python Parser? (32,711 views)
attribute queries	table, attribute, join, column, attribute table, layer, field, id, row, data, add, select Selecting multiple values with Select by Attributes in ArcGIS Desktop? (56,339 views)
vector computations	polygon, geometry, postgis, st, polygons, area, intersect, overlap, boundary, result, intersection, function Selecting features within Polygon from another layer using QGIS? (59,625 views)
geometry	point, line, distance, buffer, create, points, qgis, layer, segment, polyline, end, tool Creating point features with exact coordinates in QGIS? (102,770 views)
ArcGIS	feature, class, arcgis, feature class, select, tool, layer, desktop, arcgis desktop, create, features, arcmap Merge intersecting polygons into one which are part of the same feature (53,653 views)
raster data	raster, image, value, pixel, dem, cell, elevation, band, data, resolution, tool, create Getting boundary of raster image as polygon in ArcGIS Desktop? (62,072 views)
raster computations	area, calculate, grid, distance, value, km, slope, meter, result, cell, length, surface Measuring area of raster classes? (29,593 views)
networks	network, route, time, tool, path, building, create, arcgis, way, problem, color, group Seeking open source route planning software? (27,598 views)
datasets	data, road, geospatial, object, land, filter, dataset, example, way, look, method, classification Full list of ISO ALPHA-2 and ISO ALPHA-3 country codes (79,323 views)
coordinate systems	projection, coordinate, epsg, crs, wgs, project, data, utm, zone, reference, transformation, datum Difference between projection and datum? (108,834 views)
coordinates	coordinate, point, lat, latitude, longitude, long, gps, lon, lat long, plot, location, convert Convert X,Y State Plane coordinates to decimal degrees (36,857 views)
data formats	file, qgis, shapefile, data, import, format, csv, convert, shapefiles, export, open, save Converting between KML and shapefile (SHP) format? (108,581 views)
programming	error, python, code, script, run, output, tool, arcpy, input, clip, function, model arcpy.MakeFeatureLayer in-memory layer still exists when subsequent step fails during testing (14,177 views)
map layers	layer, map, add, display, qgis, google, vector, click, change, image, openlayers, set, zoom, create, scale, geoserver What ratio scales do Google Maps zoom levels correspond to? (186,030 views)
web services	query, service, address, server, city, street, arcgis, data, api, database, county, state How to Geocode 300,000 addresses on the fly? (67,550 views)

Table 2 shows that the topics identified by LDA cover mainly GIS data models and computations. Yet, there are some interesting discoveries.

- *Vector* and *raster* related topics (*vector computations, geometry, networks, raster data, raster computations*) account for a third of all topics (i.e., 5 out of 15). The concepts of field and object do not show up in the topics, as the topics are cast entirely as data models, not as views of the world.
- *Distance* and *area* appear in both *raster* and *vector*-related topics, indicating that these operations are generic to both kinds of data models.
- *Coordinates* and *coordinate systems* appear as the only two topics that are above the level of data models.
- Geostatistics notions do not occur in any of the topics. This may be a reflection of the (unfortunate) separation of software usage between general geospatial analysis software (e.g., ArcGIS and QGIS) and geostatistics software (e.g., GeoDa), which has their own Q&A forums.
- Some topics are more related than others. For example, *raster computations* are more related to *raster data* than *programming*.

B. Topic Clustering

Having identified 15 topics of conversation in the posts considered relevant, I found that some of them are more related than others. Therefore, I tried to group them into clusters, in order to see if there are any patterns. In linguistics, the "distributional hypothesis" states that similar words tend to be used in similar contexts (Harris, 1954). Likewise, related topics are more likely to be mentioned in similar posts. Thus, I measured the similarity between topics and structured them hierarchically.

The LDA model generated a 42,731-by-15 post-topic matrix $\boldsymbol{T}$ where $\boldsymbol{T}(i,j)$ is the probability of the i-th post being about the j-th topic. Therefore, each topic can be represented by a vector of length 42,731 (i.e., the total number of the posts) in a hyper-dimensional space of posts. Similar topics tend to have similar vector representations. Hence, I define the similarity of two topics to be the cosine similarity between two vectors:

$$S(t_x, t_y) = \frac{t_x \cdot t_y}{||t_x||\ ||t_y||}$$

where t_x represents the vector of topic x in the post-topic matrix $\boldsymbol{T}$, and $S(t_x, t_y)$ represents the cosine similarity between the vector of topic x and topic y.

After the cosine similarity was computed for all pairs of topics, an agglomerative hierarchical clustering procedure was performed with Ward's method (Ward, 1963). It supplies the criterion for choosing the pair of topics (and/or clusters) to group based on minimum variance at each step. The result of the hierarchical structure between the 15 topics is shown in a dendrogram (Figure 3). Similar topics are connected by the same colored lines. The clusters turned out to be meaningful: *map layers* are closely related to *web services* while *vector computations* are closely related to *geometry* and *networks*. For easy reference, I label these clusters as follows:

- *Vector*: geometry, vector computations, networks
- *Raster*: raster data, raster computations, datasets
- *Querying*: attribute queries, database computations
- *Mapping*: web services, map layers
- *Coordinate systems*: coordinates, coordinate systems
- *Software platforms*: ArcGIS, programming, data formats

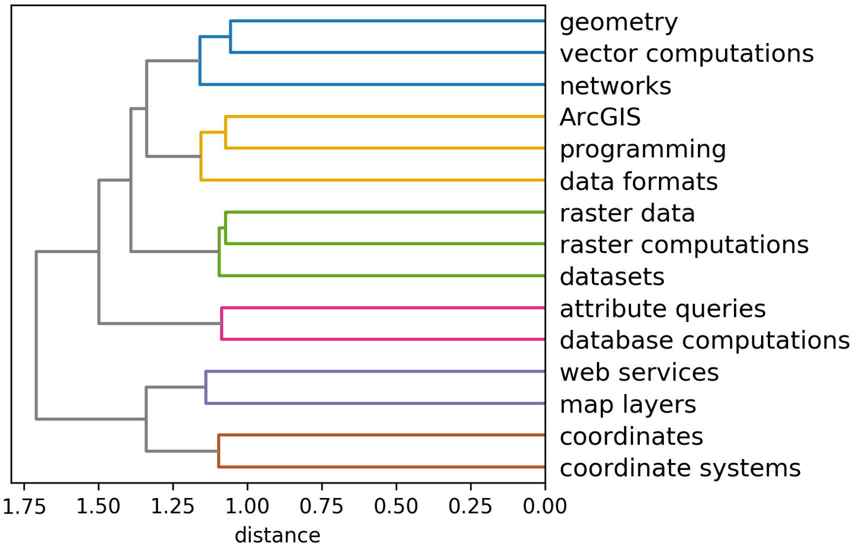


Figure 3. The dendrogram of the 15 topics.

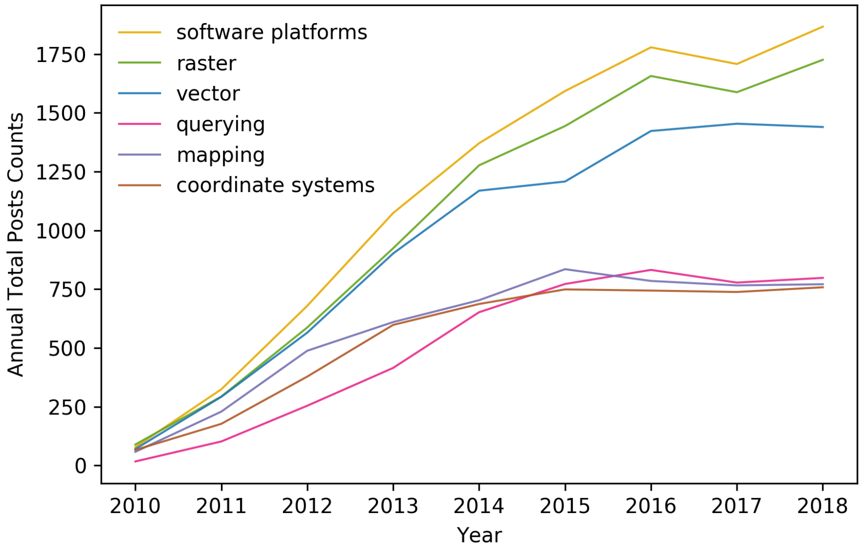
These topic clusters represent the aspects of GIS functionality talked about in posts, i.e. geospatial information representation (coordinate systems), geospatial information retrieval (e.g., geospatial querying), geospatial computing (e.g., raster and vector computations), and geospatial data visualization (e.g., mapping).

By assigning each of the 42,731 posts to the cluster with the highest probability, I was able to reveal a high level view of the temporal patterns of the topic clusters, as well as their popularity and difficulty. The definitions of popularity and difficulty are shown in Table 3.

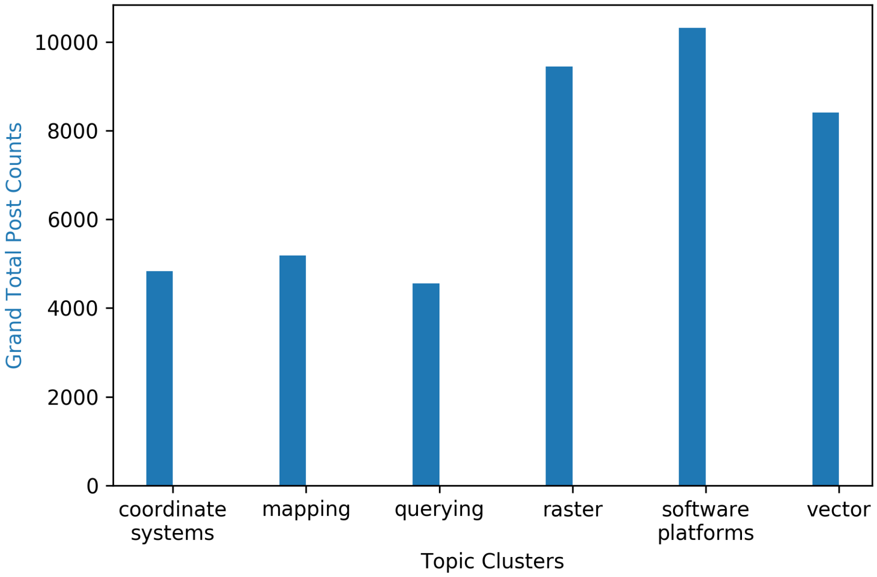
Table 3. Definitions of Cluster Popularity and Difficulty

	Metric	**Definition**
popularity	view counts	the median of the posts' view counts for a cluster
difficulty	unanswered rate	$\frac{\text{the number of posts that does not have accepted answers in a cluster}}{\text{the total number of posts in a cluster}}$
	answering time	the posts' median answering time in hours for a cluster

Concerning the temporal evolution of the topic clusters, Figure 4(a) shows the annual total number of posts per cluster from 2010 to 2018. *Raster* and *vector* topics have steadily grown over the years, along with *software platforms*. The number of posts related to *mapping*, *coordinate systems* and *querying*, however, reach a plateau after 2014. *Raster* and *vector* account for a large proportion of the total number of posts, which can also be seen in Figure 4(b) where the grand total number of posts in each cluster is shown.



(a)



(b)

Figure 4. (a) Annual total number of posts in each cluster between 2010 and 2018. (b) Grand total number of posts in each cluster between 2010-2018.

The median post view counts are shown in Figure 5. Those topic clusters with relatively fewer posts tend to have more views in general. *Coordinate systems* and *mapping* have much more views (in terms of median), even though the number of posts on them is fewer than others. My conjecture of it is that *coordinate systems* and *mapping* may have more commonly encountered and less specialized or software-specific questions, compared to questions about *raster*, *vector* and *software platforms.*

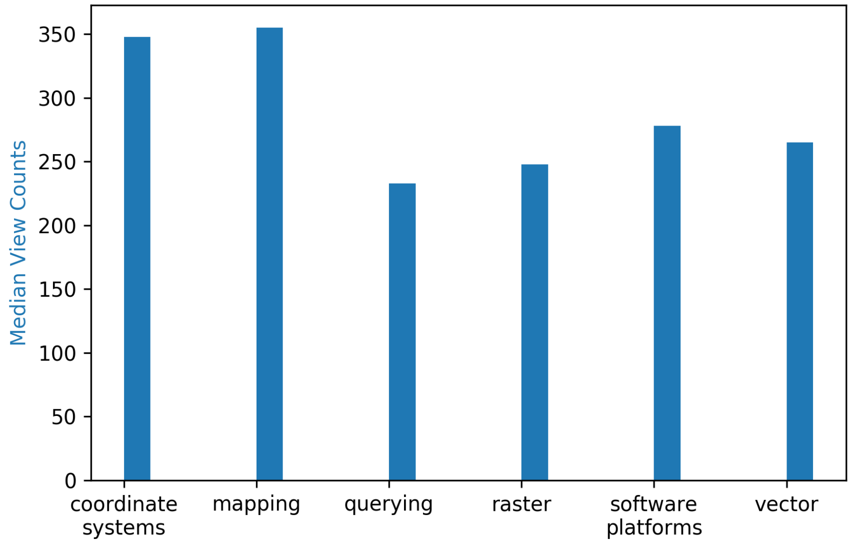


Figure 5. Median post view counts of each cluster.

To explore the difficulty of each topic cluster, the unanswered rates and median answer times of posts in each cluster are computed and shown in Figure 6. The unanswered rates (Figure 6(a)) are very even over topic clusters (except that *querying* has lower unanswered rate), indicating that no topic cluster in general is significantly more difficult than others. However, all the clusters have a very high rate of around 60% unanswered questions, suggesting that a large proportion of the posts is hard to answer. Nevertheless, the answer times (Figure 6(b)) in *querying* and *coordinate systems* are relatively shorter (1-3 hours) than those of other topics, and *mapping* has the longest answer times (over 5 hours), suggesting that for the questions that could be answered, *querying* and *coordinate systems* tend to be dealt with faster.

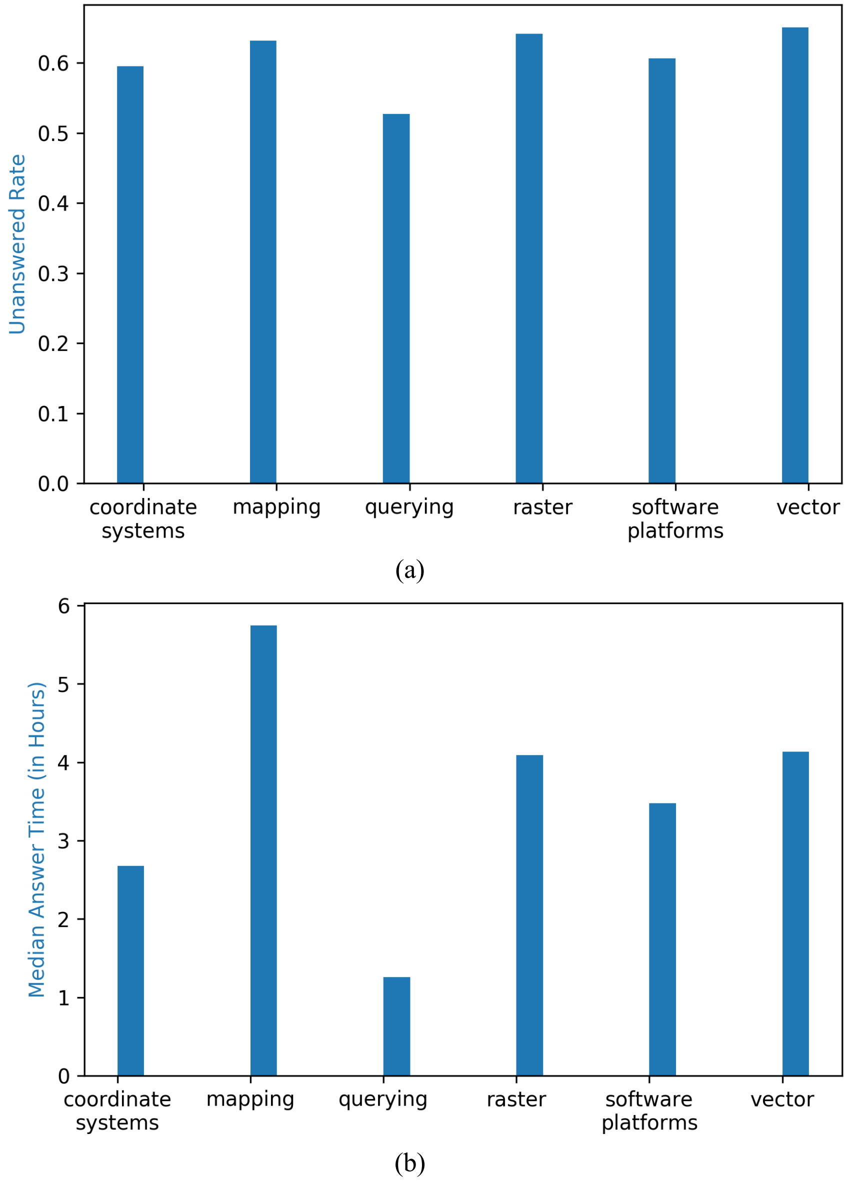


Figure 6. (a) Cluster unanswered rate. (b) Median post answer time of each topic cluster.

C. Tag Comparison

In the analysis so far, I found that coordinate systems and coordinates were the only two general geographic concepts among the 15 topics, with all others relating to data models and software procedures. To verify this observation of a very small conceptual basis for GIS on

Stack Exchange, I compared the 20 most frequently used tags in posts on GIS Stack Exchange with those on Cross Validated (i.e. Statistics) Stack Exchange because statistics has been broadly used across various fields thanks to its well-defined theories and methods (e.g. random variable, distribution, test, confidence interval). The results are shown in Table 4[5]. (Tags referring to theoretical concepts are in bold.)

Table 4. The 20 most frequent tags in GIS and Cross Validated (Statistics) Stack Exchange

	GIS Tags	Counts	Statistics Tags	Counts
1	qgis	23,607	r	19,615
2	arcgis-desktop	17,858	regression	17,613
3	arcpy	8,146	machine-learning	11,787
4	python	7,093	time-series	9,059
5	postgis	6,599	probability	7,040
6	raster	6,217	hypothesis-testing	6,101
7	coordinate-system	5,356	self-study	5,875
8	arcmap	4,862	distributions	5,835
9	geoserver	4,273	logistic	4,968
10	pygis	4,026	bayesian	4,638
11	shapefile	3,894	classification	4,610
12	gdal	3,798	correlation	4,187
13	arcgis-10.0	3,610	neural-networks	4,088
14	openlayers-2	3,318	statistical-significance	4,030
15	arcgis-10.1	3,125	mathematical-statistics	3,900
16	r	3,064	normal-distribution	3,651
17	arcgis-10.2	2,961	anova	3,621
18	postgresql	2,803	multiple-regression	3,370

[5] Data obtained on February 1st, 2019 from GIS Stack Exchange and Cross Validated Stack Exchange.

19	leaflet	2,681	mixed-model	3,176
20	polygon	2,663	clustering	2,941

One can observe that 16 of the top 20 tags for statistics are about theoretical concepts (e.g. probability, distributions, regression, hypothesis testing), while only one (coordinate system, again) in the top 20 tags for GIS is a theoretical concept, all other tags are software- or data model terms. This comparison of tags supports my findings from the LDA topic model. Reasons for this pronounced difference between the numbers of software-independent topics may include the variety of GIS interfaces and platforms including QGIS and ArcGIS as opposed to the dominant statistical platform R, and lack of theories that people can ask about. Still, this comparison suggests that the conceptual basis for GIS has not yet reached the level of clarity and consensus found in statistics, making it harder for GIS users to directly ask spatial analysis questions.

V. Conclusions and Future Work

The easy access to geospatial data in recent years has made spatial thinking and analysis applicable to scientific explanations across disciplines. However, how to use GIS for spatial analysis may not always be obvious. Many GIS users nowadays seek answers to the questions they face through online forums, such as Stack Exchange. Studying these questions helps uncover what users do not know or understand about GIS. In this study, over 40,000 posts on GIS Stack Exchange were investigated with a topic model LDA to show the topics that users ask about GIS. Fifteen topics were identified, including coordinate systems, geospatial querying, raster and vector computations, and (web) mapping. While the numbers of questions on raster and vector are increasing over time, they tend to have longer answer times and less view counts on Stack Exchange. The number of questions about coordinate systems is smaller, but they are typically viewed by more people and resolved in a shorter time. The reason might be that questions on raster and vector are more specialized and difficult and less likely to be resolved easily or quickly, while questions on coordinate systems are more common and easier to deal with. The comparison of the top 20 tags for posts on GIS Stack Exchange and Statistic Stack Exchange indicates a relatively immature conceptual basis of GIS compared to that of statistics.

These conclusions, however, need to be treated with caution. Several limitations affect this study. First, the results and conclusions are based on one single dataset — GIS Stack Exchange. The dataset can be biased without knowing exactly the user profiles. (GIS users on Stack Exchange could be more technical oriented.) Knowing user identity can offer us more information in interpreting the results and drawing conclusions. However, the bias is nearly inevitable because it is impossible to trace the large proportion of users who just view instead of creating or answering the posts. Inclusion of multiple forums such as GeoNet[6] and GIS Lounge[7], if available, may provide a more holistic view of the domain and mitigate the bias issues.

Second, LDA, as an unsupervised learning method, needs some human interventions and interpretation. For instance, the number of topics have to be specified in advance. In this paper, manual inspections on the results generated from different numbers of topics were carried out to determine the optimal number of topics.

[6] https://community.esri.com/

[7] https://www.gislounge.com/

Third, my definitions of popularity and difficulty of clusters are crude, only offering a rudimentary impression based on time to answer and frequency of viewing. Lastly, the comparison of tags on GIS and statistics is not entirely fair. GIS is software, while statistics is method. Still, one can observe the almost total lack of questions about relevant theories and methods in GIS posts, corresponding to those dominating the statistics forum. On the positive side, the methods used in this thesis are easily replicable and can be applied to other datasets where available.

This thesis is an initial exploration of user questions about GIS by mining crowdsourced (online forum) data in a quantitative and systematic way. The dataset and analyses reveals interesting findings that would not have been possible or made evident otherwise. The results depict a landscape of topics around which the current GIS users converse. As such, my findings provide guidance and directions to GIS educators and software developers in leveraging their knowledge to make GIS more accessible and easier to use across disciplines.

References

Albrecht, J. (1998). Universal analytical GIS operations: a task-oriented systematization of data structure-independent GIS functionality. In M. Craglia & H. Onsrud (Eds.), *Geographic Information Research: transatlantic perspectives*. (pp. 577–591). Taylor & Francis.

Allamanis, M., & Sutton, C. (2013, May). Why, when, and what: analyzing stack overflow questions by topic, type, and code. In *2013 10th Working Conference on Mining Software Repositories (MSR)* (pp. 53-56). IEEE.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research, 3*(Jan), 993-1022.

Cappelli, C. (2013). *The Language of Spatial Analysis*. Retrieved from https://www.esri.com/library/books/the-language-of-spatial-analysis.pdf

Chetty, R., Hendren, N., Kline, P., & Saez, E. (2014). Where is the land of opportunity? The geography of intergenerational mobility in the United States. *The Quarterly Journal of Economics, 129*(4), 1553-1623.

Christine, P. J., Auchincloss, A. H., Bertoni, A. G., Carnethon, M. R., Sánchez, B. N., Moore, K., ... & Roux, A. V. D. (2015). Longitudinal associations between neighborhood physical and social environments and incident type 2 diabetes mellitus: the Multi-Ethnic Study of Atherosclerosis (MESA). *JAMA internal medicine, 175*(8), 1311-1320.

Couclelis, H. (1992). People manipulate objects (but cultivate fields): beyond the raster-vector debate in GIS. In *Theories and methods of spatio-temporal reasoning in geographic space* (pp. 65-77). Springer, Berlin, Heidelberg.

Dangermond, J. (1983). A classification of software components commonly used in geographic information systems. In D. J. Peuquet & J. O'Callaghan (Eds.), *Design and Implementation of Computer-Based Geographic Information Systems* (pp. 70-91). Amherst, NY: IGU Commission on Geographical Data sensing and Processing.

Egenhofer, M. J. (1989, June). A formal definition of binary topological relationships. In *International conference on foundations of data organization and algorithms* (pp. 457-472). Springer, Berlin, Heidelberg.

Fisher, H. K. (1968). SYMAP: Synagraphic mapping system. *Cambridge, MA: Laboratory for Computer Graphics and Spatial Analysis, Harvard University.*

Frank, A. U. (1996). Qualitative spatial reasoning: Cardinal directions as an example. *International Journal of Geographical Information Science, 10*(3), 269-290.

Gao, S., & Goodchild, M. F. (2013, July). Asking spatial questions to identify GIS functionality. In *2013 Fourth International Conference on Computing for Geospatial Research and Application* (pp. 106-110). IEEE.

Goodchild, M. F. (1992). Geographical data modeling. *Computers & Geosciences, 18*(4), 401-408.

Goodchild, M. F., Egenhofer, M. J., Kemp, K. K., Mark, D. M., & Sheppard, E. (1999). Introduction to the Varenius project. *International Journal of Geographical Information Science, 13*(8), 731-745.

Goodchild, M. F., Yuan, M., & Cova, T. J. (2007). Towards a general theory of geographic representation in GIS. *International journal of geographical information science, 21*(3), 239-260.

Harris, Z. S. (1954). Distributional structure. *Word, 10*(2-3), 146-162.

Janelle, D. G., & Goodchild, M. F. (2011). Concepts, principles, tools, and challenges in spatially integrated social science. In T. Nyerges, H. Couclelis, & R. McMaster (Eds.), *The SAGE handbook of GIS and society* (pp. 27-45). London: SAGE Publications, Inc.

Jo, I., & Bednarz, S. W. (2009). Evaluating geography textbook questions from a spatial perspective: Using concepts of space, tools of representation, and cognitive processes to evaluate spatiality. *Journal of Geography, 108*(1), 4-13.

Krugman, P. (1991). Increasing returns and economic geography. *Journal of political economy, 99*(3), 483-499.

Kuhn, W. (2012). Core concepts of spatial information for transdisciplinary research. *International Journal of Geographical Information Science, 26*(12), 2267-2276.

Kuhn, W., & Ballatore, A. (2015). Designing a language for spatial computing. In *AGILE 2015* (pp. 309-326). Springer, Cham.

Marsh, M., Golledge, R., & Battersby, S. E. (2007). Geospatial concept understanding and recognition in G6–college students: A preliminary argument for minimal GIS. *Annals of the Association of American Geographers, 97*(4), 696-712.

Nystuen, J. (1963). Identification of some fundamental spatial concepts. *Papers of the Michigan Academy of Science, Arts and Letters, 48*, 373-384.

Peuquet, D. J. (1988). Representations of geographic space: toward a conceptual synthesis. *Annals of the Association of American Geographers, 78*(3), 375-394.

Ramage, D., Hall, D., Nallapati, R., & Manning, C. D. (2009, August). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1* (pp. 248-256). Association for Computational Linguistics.

Rhind, D. W., & Green, N. P. A. (1988). Design of a geographical information system for a heterogeneous scientific community. *International journal of geographical information system, 2*(2), 171-189.

Sinton, D. F., & Steinitz, C. F. (1969). GRID: a user's manual. *Cambridge, MA: Laboratory for Computer Graphics and Spatial Analysis, Harvard University.*

Sipe, N. G., & Dale, P. (2003). Challenges in using geographic information systems (GIS) to understand and control malaria in Indonesia. *Malaria journal, 2*(1), 36.

Tomlin, C. D. (1983). A Map Algebra. Harvard Computer Graphics Conference. *Cambridge, MA: Graduate School of Design, Harvard University.*

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association, 58*(301), 236-244.

Xia, X., Lo, D., Wang, X., & Zhou, B. (2013, May). Tag recommendation in software information sites. In *2013 10th Working Conference on Mining Software Repositories (MSR)* (pp. 287-296). IEEE.

Yang, X. L., Lo, D., Xia, X., Wan, Z. Y., & Sun, J. L. (2016). What security questions do developers ask? a large-scale study of stack overflow posts. *Journal of Computer Science and Technology, 31*(5), 910-924.

Appendix

The following are the 257 tags used in Section IV.A Topic Identification:

coordinate-system, polygon, field-calculator, dem, point, spatial-analyst, attribute-table, line, distance, convert, geometry, geoprocessing, coordinates, buffer, clip, qgis-processing, raster-calculator, merge, georeferencing, geocoding, intersection, attribute-joins, network-analyst, latitude-longitude, interpolation, attributes, spatial-join, import, elevation, extents, spatial-statistics, query, routing, area, classification, fields, topology, 3d-analyst, select-by-attribute, clustering, network, overlapping-features, contour, select, scale, ndvi, dissolve, feature-class, filter, splitting, overlay, zonal-statistics, statistics, linestring, select-by-location, heat-map, feature-layer, utm, polygon-creation, proximity, mosaic, point-in-polygon, qgis-modeler, grids-graticules, slope, time, snapping, rasterization, getfeatureinfo, analysis, expression, geolocation, raster-conversion, digitizing, linear-referencing, resolution, reclassify, digital-image-processing, route, kriging, address, viewshed, union, watershed, land-cover, polyline-creation, terrain, date, spherical-geometry, topography, vector-layer, events, cost-path, masking, geometric-network, vertices, accuracy, polygonize, tin, nearest-neighbor, measurements, centroids, spatial-query, length, animation, definition-query, hillshade, density, graph, extract, voronoi-thiessen, resampling, reverse-geocoding, relates, optimization, encoding, circle, direction, intersect, vectorization, xy, kernel-density, land-classification, distance-matrix, generalization, regression, geometry-conversion, modelling, xyz, gdal-merge, map-algebra, transportation, shortest-path, point-of-interest, query-layer, pixel, tracking, count, profile, sorting, extract-by-mask, smoothing, grouping, comparison, aggregation, angles, feature-extraction, identify, point-cloud, points-to-line, autocorrelation, land-use, model, delete, append, group-layer, navigation, volume, conditional, geostatistical-analyst, azimuth, combine, representation, spatial-analysis, change-detection, sampling, atmospheric-correction, affine-transformation, time-series, geotag, 3d-model, publishing, convex-hull, precision, flow, cloud-cover, compression, geodesy, stack, st-intersects, bearing, image-segmentation, erase, differences, flow-accumulation, reflectance, trace, solar-radiation, great-circle, orthorectification, where-clause, random-forest, correlation, cogo, multi-values, precipitation, radar, points, flow-map, dijkstra, catchment-area, validation, ellipsoid, altitude, field-properties, geographically-weighted-regression, nodes, migration, open-source-routing-machine, position, geodesic, dynamic-layer, create, indexing, vegetation-index, gdal-rasterize, variogram, trajectory, multipatch, multipoint, aspect, cross-section, pansharpening, geoid, dimensions, taudem, parallel-lines, concave-hull, loading, spline, focal-statistics, donut-polygons, machine-learning, diagram, spatial-etl, footprint, design, pan, unique-id, multimodal-network, normalize, geohash, shaded-relief, curvature, region, con, locator, stream-order, antimeridian, surface, spatial-adjustment, distribution, origin-destination, modis-reprojection-tool, slivers, service-area, trilateration