Geospatial Analysis and Modeling of Textual Descriptions of Pre-modern Geography

Von der Fakultät für Mathematik und Informatik
der Universität Leipzig
angenommene

DISSERTATION

zur Erlangung des akademischen Grades

DOCTOR RERUM NATURALIUM
(Dr. rer. nat.)

im Fachgebiet
Informatik
Vorgelegt

von M.Sc. Masoumeh Seydi
geboren am 22.09.1984 in Ray, Iran

Die Annahme der Dissertation wurde empfohlen von:
1. Prof. Dr. Gregory Crane, Universität Leipzig,
zusammen mit
Dr. Maxim Romanov (Universität Hamburg)
2. Prof. Dr. Øyvind Eide (Universität zu Köln)

Die Verleihung des akademischen Grades erfolgt mit Bestehen
der Verteidigung am 04.05.2022 mit dem Gesamtprädikat magna cum laude

Acknowledgments

This project would not have been possible without the support of many people. I must express my first gratitude towards my supervisor, Prof. Gregory Crane, for his continuous support during my research, patience, immense knowledge, and attention to details. I would also like to express deepest appreciation to Dr. Maxim Romanov for his valuable advice, invaluable contribution and suggestions. Your insightful feedback pushed me to sharpen my thinking and brought my work to a higher level. I should confess that I can not imagine a better and friendlier people to work with.

Besides, I would like to acknowledge Prof. Sarah Savant, Dr. Thomas Köntges, and Dr. Melinda Johnston for their valuable guidance and comments on my work. I want to thank you all for your patient support

In addition, I wish to thank my family, especially my parents, for their wise counsel and sympathetic ear and encouraging me throughout all my studies at university. You are always there for me.

Finally, I could not have completed this dissertation without the support, continued and unfailing love, and understanding of my beloved husband, Ali Rostami.

Transliteration

We use the following transliteration map to convert Arabic texts into Latin script. Quotes are written in the original transliterated phrases as in the corresponding source.

Arabic	Transliteration	Arabic	Transliteration	Arabic	Transliteration
ء	ᵓ	ذ	ḏ	ق	q
ا	ā	ر	r	ك	k
أ	a	ز	z	ل	l
إ	i	س	s	م	m
آ	ā	ش	š	ن	n
ب	b	ص	ṣ	و	w
ت	t	ض	ḍ	ؤ	ᵓ
ث	ṯ	ط	ṭ	ه	h
ج	j	ظ	ẓ	ى	y, ī
ح	ḥ	ع	ᶜ	ي	y, ī
خ	ḫ	غ	ġ	ئ	ᵓ
د	d	ف	f	ة	ŧ

Abstract

Textual descriptions of pre-modern geography offer a different view of classical geography. The descriptions have been produced when none of the modern geographical concepts and tools were available. In this dissertation, we study pre-modern geography by primarily finding the existing structures of the descriptions and different cases of geographical data. We first explain four major geographical cases in pre-modern Arabic sources: gazetteer, administrative hierarchy, routes, and toponyms associated with people. Focusing on hierarchical divisions and routes, we offer approaches for manual annotation of administrative hierarchy and route sections as well as a semi-automated toponyms annotation. The latter starts with a fuzzy search of toponyms from an authority list and applies two different extrapolation models to infer true or false values, based on the context, for disambiguating the automatically annotated toponyms.

Having the annotated data, we introduce mathematical models to shape and visualize regions based on the description of administrative hierarchy. Moreover, we offer models for comparing hierarchical divisions and route networks from different sources. We also suggest approaches to approximate geographical coordinates for places that do not have geographical coordinates—we call them *unknown place*s—which is a major issue in visualization of pre-modern places on map.

The final chapter of the dissertation introduces the new version of al-Ṯurayyā, a gazetteer and a spatial model of the classical Islamic world using georeferenced data of a pre-modern atlas with more than $2,000$ toponyms and routes. It offers search, path finding, and flood network functionalities as well as visualizations of regions using one of the models that we describe for regions. However the gazetteer is designed using the classical Islamic world data, the spatial model and features can be used for similarly prepared datasets.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> Iṣṭaḫr: with *kasraŧ* and *sukūn* after “ḫ” and a *nisbaŧ* from it is Iṣṭaḫrī and Iṣṭaḫrzī with the extra “z”. [It is] a city in Fārs from the third “climate.” Its length is seventy-nine degrees and its width is thirty-two degrees, and it is one of the most prominent fortresses, cities and regions of Fārs.... Al-Iṣṭaḫrī says: as for the city of Iṣṭaḫr, it is a city with the center of the capacity of a mile, which is one of the oldest and most famous cities of Fārs, and it was the residence of the king of Persia until Ardašīr turned to Gore...., and has a mosque known as the mosque of Sulaymān, peace be upon him. Some of the Persians claimed that the king who was before the Ḍaḥāk was Sulaymān the son of David, he said: it was in the old days on the city of Iṣṭaḫr fence and demolished, and built of mud, stones and plaster to the left of the ..., and Qantaraŧ Ḫurāsān out of the city on the door of the following Ḫurasan, and behind Qanṭaraŧ buildings and houses are not old.
>
> Yāqūt al-Ḫamawī, *Dictionary of Countries* (*Muᶜjam al-Buldān*)([1, p. 211])

The example in the pre-modern geographical source given above, offers geographical and cultural information about an extremely important historical location in the southwest of modern Iran, known as Iṣṭaḫr (Staḫr in Pahlavi). Yāqūt provides some regular categorical descriptions that allow us to compare all these settlements with each other and understand how they are related to each other. He also cites a major earlier geographer named al-Iṣṭaḫrī. Among the many things a medieval reader would have learned from his lexical efforts are onomastics, specifically, naming conventions. Yāqūt explains the pronunciation of the toponym by specifying the Arabic vowels or *ḥarakāt*. In addition, he explains the derived adjectives and modifiers (in Arabic, *nisbaŧ* equivalences): for instance, Iṣṭaḫrī or Iṣṭaḫrzī with the extra “z,” each of which designates someone or something associated with Iṣṭaḫr.

Beyond the extract cited above, we also learn administrative divisions of the Fārs province (a Persian speaking region very roughly equivalent to Modern Iran). At the widest level, there are climes (*iqlīm*, pl. *aqālīm*) that, as noted by [2], refers to regional entities of different size, shape, and provenance[1].

[1] The following is a complete list of *iqlīms* as per al-Iṣṭaḫrī’s division of the pre-modern Islamic world: Diyār al-ᶜArab (in modern Saudi Arabia), Baḥr al-Fārs (Persian Gulf), al-Maġrib (at Northwest Africa), Miṣr (Egypt), al-Šām (West Asia, east of the Mediterranean Sea), Baḥr al-Rūm, al-Jazīraŧ (the northern part of the area between the Euphrates and Tigris rivers, Upper Mesopotamia), al-ᶜIrāq (Iraq), Ḫūzistān (at the southwest of modern Iran,

Then, within each of these, there are regions, and within those there are settlements, such as Iṣṭaḫr. Al-Iṣṭaḫrī explains that Iṣṭaḫr is located in the third clime, which is called Fārs. Furthermore, he specifies this location down to the level of the village of Iṣṭaḫr, known as Darābağird, in the subregion of Iṣṭaḫr. The descriptions also include the type of divisions and settlements, such as province (*iqlīm*), district (*kūraŧ*), and village (*qaryaŧ*).

In addition, here are some lines that Yāqūt later provides to his readers with further information:

- First, about routes and distances between locations, as in this example: "... between Iṣṭaḫr and Šīrāz there are twelve *farsaḫ*s." *Farsaḫ* is a pre-modern Persian unit to measure distance based on time and was originally the distance that can be traveled by foot within one hour (students of Greek history may be familiar with *parsang*, the earlier form of this word). The term survives today and is still used in Iran, but it now designates a fixed length of approximately 6 km ([3]) rather than the inherently variable (and practical) measure that varies with the difficulty of the terrain: a *farsaḫ* up hill is much shorter than a *farsaḫ* on a flat and easy surface.

- Second is spatial information, by which we mean the physical dimensions of the city or a geographic entity itself. We learn that the length of Iṣṭaḫr is 79 degrees and its width is 32 degrees, or the city Iṣṭaḫr has "a center of the [capacity] of an [Arabian] mile."

- Finally, we learn the relative geographical locations of places. For instance, in the entry for Aṭrābulus, a city in al-Šām Province (in modern Syria), Yāqūt says ([1, p. 216]): "Aṭrābulus is located at the coast of Syrian see and between al-Lāḏiqiyyaŧ and ᶜAkkaŧ" and in another entry for a city with the same name in North Africa, he says ([1, p. 217]): "a city at the [western] end of Barqaŧ [Lybia] and beginning of [the region] Ifrīqīyyaŧ." This explanation specifies the location of a place and distinguishes the locations of the same name.

This specific entry within Yāqūt's geographical lexicon employs a technical vocabulary that is shaped around the same categories: administrative divisions, routes and distances, and relative spatial location. Other entries are patterned similarly, providing us information that makes it possible to do large scale comparative analysis. For example, when describing al-Šām Province ([1, p. 312]): "It (al-Šām) has five *jund*s[2]: *jund* of Qinnasrīn and *jund* of Dimašq and *jund* of al-Urdunn and *jund* of Filisṭīn and *jund* of Ḥimṣ."

It is the structural similarity in descriptions of all places of Yāqūt's and other geographers' categories that has made them such important resources. For example, one of their key applications in the non-digital field has been to provide the basis for modern encyclopedic efforts, such as the widely consulted Encyclopedia of Islam (EI). These encyclopedias are typically written formulaically, using categories derived from these geographies[3].

bordering Iraq and the Persian Gulf), Fārs (at the southwest of modern Iran), Kirmān (at the southeast of modern Iran), al-Sind (in modern Pakistan and the northwest of modern India), Arminīaŧ and al-Arrān and Azarbayjān (at Modern Azerbaijan, Armenia and northwest of modern Iran), al-Jibāl (at the west of modern Iran), al-Daylam (the mountainous regions of North Iran, at the southwest coast of the Caspian Sea), Baḥr al-Ḫazar (at the northwest coast of the Caspian Sea), Mafāzaŧ Ḫurāsān (the desert in the center of modern Iran), Sijistān (south of modern Afghanistan), Ḫurāsān (at the northeast of modern Iran and Afghanistan and southern part of Central Asia), and Mā-warāᵓ-l-nahr (at lower Central Asia roughly corresponding to modern-day Eastern Uzbekistan, Tajikistan, Southern Kazakhstan and Southern Kyrgyzstan).

[2] For the definition of *jund*, see [4]

[3] For example, as written in the entry of Iṣṭak̲h̲r in EI2 ([5]):

Historically, computer scientists have been using this data (including relationships between locations) to form datasets with structures to cover the geographical entities and relations. These datasets are the basis of the computational approaches we employ for modeling and visualizations of geographical concepts which illustrate the thinking at the heart of this dissertation.

Our goal, then, has been to develop a method of data extraction suitable for small- and medium-sized texts in Arabic (and other languages that use right-to-left scripts) that extracts data in ways that empowers humanists to pose and answer questions in relation to texts containing geographical information. More specifically, the developed method of data extraction and modeling enables what is sometimes called distant reading. According to [6, 7], distant reading is an approach that uses algorithms and tools provided by the computer science domain (i.e., to count, map, graph, and visualize) to analyze textual data instead of using traditional ways to read texts—so-called *close reading.* Close reading determines the central themes and analyzes their development to interpret a text passage and for humanists, the value of distant reading lies in helping make meaningful choices in terms of what they read closely [8].

To address this research goal, this dissertation introduces an iterative data annotation and extraction method for specific geographical cases that occur in similar patterns in textual descriptions. Once extracted, that data can then be manipulated and modeled, offering further insights in the distant reading of the textual data. We have also developed a gazetteer ([9]) and a spatial model based on similar data structures that researchers can extract using the method described in this dissertation. The methods, models, and visualizations provide a zoomed-out view of part of the information in textual sources and can be used in distant reading, as a complementary to close reading.

For example, and to further illustrate what follows, administrative divisions can be modeled as tree structures. Administrative divisions represent the way a land, which is the subject of a description, is divided into subordinate regions or encompasses other subregions and settlements in a top-down order. This means that the area in question (here the classic Islamic world) is divided into major provinces, each province is divided into subordinate regions, which in turn encompass either subregions or settlements. In other words, if we look at the way that division has occurred historically, we can see an area that includes smaller subregions, which in turn goes further, to include sub-units of settlements, and so on. This implies a hierarchical, and not a linear, relationship between the entities, which are super-regions, regions, subregions, and settlements.

Another example of such a structure could be an organizational structure (see Figure 1.1). The hierarchical relationship can be represented as a tree structure and the level of hierarchy depends on the granularity of the description. It has a root that is the highest-level entity and the only grandparent of all other entities. The root is the parent of a set of entities and each of those entities might be a parent of one or more entities. In other words, entities have either parent-child or sibling relationships; the root has no parents and siblings. Each entity is called a node. The

a town in Fārs. The real name was probably Stak̲h̲r, as it is written in Pahlavi; the Armenian form Stahr and the abbreviation ST on Sāsānian coins point in the same direction. The form with prosthetic vowel is modern Persian; it is usually pronounced Istak̲h̲ar or Iṣtaḥar, also with inserted vowel Sitak̲h̲ar, Siṭak̲h̲ar, Siṭark̲h̲; cf. Vullers, *Lex.Pers.- Lat.*, i, 94^{a}, 97^{a}, ii, 223, and Nöldeke in *Grundr. der Iran. Philol ., ii*, 192. The Syriac form is Iṣṭaḫr (rarely Iṣṭaḫr), in the Talmud probably Istahar (*Megilla* 13^{a}, middle). According to the statements of Persian authors, the town received its name from the lakes or swamps there. Perhaps, however, it is better to be derived with Spiegel (*Erânische Altertumskunde* , i, 94, note I) and Justi (*Grundr. der Iran. Philol.*, ii, 448) from the Avestan *stak̲h̲ra* “strong, firm”; for the latter word cf. Chr. Bartholomae, *Altiran. Wörterbuch* , p. 1591.).

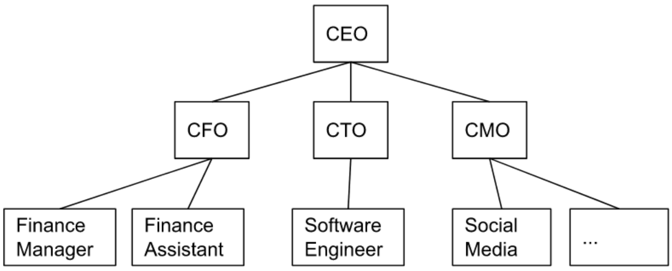


Figure 1.1: An imagined example of a tree structure to store an organizational hierarchy. The head of the organization is the CEO and all the other departments are positioned at a lower level. Each entity is a node and is connected to one or more entities by an edge that represents their relationship.

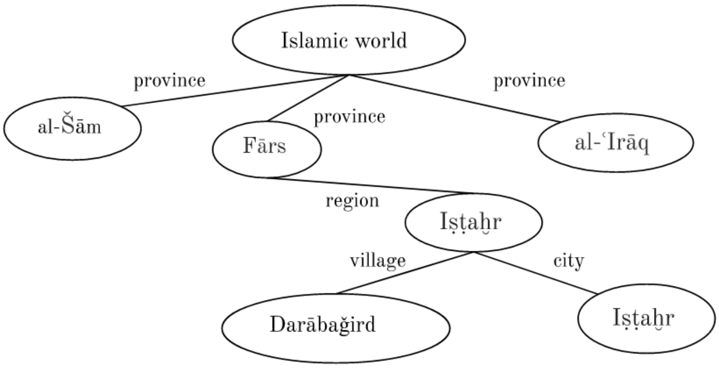


Figure 1.2: Part of the divisions according to al-Iṣṭaḫrī's description for the Fārs province. The area that he describes is divided into major provinces, such as Fārs, al-Šām, and al-ᶜIrāq. Iṣṭaḫr is then given as a region in Fārs, in which not only villages (such as village of Darābağird) are located but cities such as the city of Iṣṭaḫr are located. The hierarchical representation informs us in a formalized way that we have an ambiguity. It means, the same string, Iṣṭaḫr, can designate a region or a city. In annotating our sources, we need to distinguish (where possible and/or appropriate) between region and city. The type of divisions are shown as a property of the edges that connect the nodes. The higher level nodes are containers of the lower level nodes, as mentioned in the original text.

nodes are connected by edges that represent the relationship between each pair of connected nodes. Two connected nodes (parent-child) are connected by only one edge and no more.

In the quote with which this chapter began, both Yāqūt and al-Iṣṭaḫrī mention that Iṣṭaḫr is placed in a province (*iqlīm*) called Fārs. Since Fārs is one of the provinces according to al-Iṣṭaḫrī's divisions of the Islamic World, the tree will have the Islamic World as the root, Fārs as a child (internal tree node), and Iṣṭaḫr as a grandchild (tree leaf). Figure 1.2 illustrates this description.

Routes and distances connect the places (subjects of description). Each route section thus has a start and an end point and is of a specific length. These lengths can vary within sources, including time-based measures (e.g., a conventional day's travel) or more straightforward measures of length (e.g., *farsaḫ*). We may not be sure whether a unit defines time of travel or distance in space. By using descriptions such as those of al-Iṣṭaḫrī, we can access and define these lengths.

Beyond defining individual routes, a set of route sections, which is described as a variety of paths to connect a set of places, forms a network of routes. In other words, in a set of places, some pairs are related through routes. Therefore, there are two main sets of objects: places and routes. This description can be represented using an abstract data structure called *graph*, which models the relations among places in a more sophisticated way than just treating each route section individually. Similar to the tree structure described above, graphs represent places with

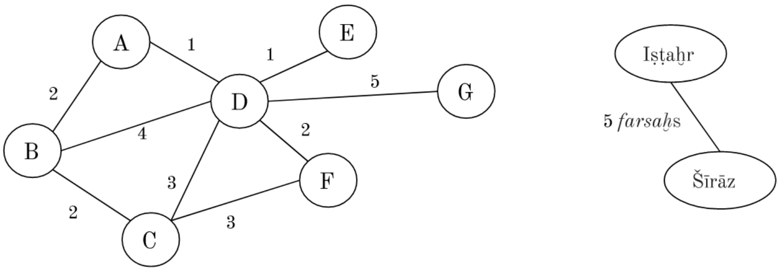


Figure 1.3: (Left): A sample graph showing the relation between a set of objects. In the context of a route network, each node (A to G) represents a place and each edge represents a path between the connecting nodes. Each edge also has a label that specifies the length of the edge or the distance of a route, which is called the weight. (Right): A (part of the) graph from the above pre-modern example representing the route (edge) between Iṣṭaḫr and Šīrāz (nodes) that is measured as five *farsaḫ*s

*node*s (also called *vertices*) and routes with *edge*s. Each place can be connected to zero, one, or any number of other places so that, unlike the tree structure, each pair of places is connected with multiple routes (edges). This means that multiple paths between two places that are explained in a text can be represented by a graph.

Additionally, a graph does not represent a hierarchical structure and, consequently, it has no root node. Accordingly, a graph properly models all the aspects of descriptions of routes in a connected way. Figure 1.3 (Left) illustrates an example of a graph in which each node represents a place and each edge is a route, characterized by a label that specifies the length of a route—distance between two places—called *weight.* Figure 1.3 (Right) is an individual route section taken from al-Iṣṭaḫrī that describes a route from Iṣṭaḫr to Šīrāz with the length of five *farsaḫ*s. As can be seen in this figure, the graph on the left shows the connectivity map of a sample area by bringing all the individual route sections together.

In this dissertation, we offer an approach for annotation, extraction, modeling, and visualization of data on the administrative divisions and route networks described in geographical sources. Our focus is on classical Arabic geographical sources, for which no significant advances have been achieved in the language-based tools to the best of our knowledge.

The current lack of tools is therefore addressed by this dissertation, which introduces specific models to fill the gaps in the datasets produced based on textual description, such as missing coordinates and representation of regions. Furthermore, this work enriches these new models by visualization and mapping, which in turn opens up exciting new possibilities for historians. Mapping not only visualizes narratives; it also becomes a narrative itself to connect various communities, regions, and locations that are difficult to see and to discover deeper ways to understand narratives. While these visualizations do not explicitly explain historical events and phenomena, they are a way of interpreting textual narratives and thus, the significance of the approaches we introduce in this dissertation is not just in relation to mapping and visualization; rather, it offers an approach that can expand the scale of the research that humanists undertake on textual sources. Geographical models (or new understanding of space) produced with these approaches can offer valuable geospatial context for the interpretation of historical sources.

Chapter 2 of this dissertation begins by taking a close look at the related work on gathering and using geographical concepts and information in humanities disciplines. This provides an overview of the state-of-the-art approaches that have been engaged in humanities studies to produce data and knowledge. The chapter starts with works that use Geographic Information System (GIS)[4] technologies as a common approach to managing geospatial data, mapping, and visualizations. GIS, according to [10], is the root for historical geographical information analysis. It has been used in humanities disciplines such as History for the last two decades and a great number of studies use GIS technologies for data storage and mapping.

The second section of this chapter reviews studies related to identifying and gathering geographical information from textual sources. This section covers annotation, Natural Language Processing (NLP), georeferencing, and geoparsing. A third section covers digital gazetteer development and explains a number of related projects to show how development has proceeded. In the final section, we review a number of studies on modeling geographical data and discuss the ways for which modeling approaches have been introduced to produce tools for information gathering, representation, and analysis.

Chapter 3 explains four major geographical use cases, to which we introduce an annotation scheme, data extraction, and modeling techniques. Each case is explained according to the structure and type of descriptions and the information they provide, together with examples. These cases are chosen from classical Arabic sources; however, similar cases might occur in other classical languages.

Chapter 4 discusses an exploratory approach for extracting geographical information from textual sources. Toponym detection approaches and standards are significantly different for modern sources since they tend to use more standardized place names. Toponym identification issues are similarly present in different languages; however, language-specific challenges still pose a significant number of problems. The aim of this chapter is therefore to develop a data annotation and extraction method that can be applied to cases that share similarities even though they are in different languages. Data extraction and analyses can facilitate distant reading of the text. The iterative nature of this method, which starts with annotation, is described in a few steps. Annotation facilitates data extraction and producing data structures from the specific information described in textual sources. We describe an annotation scheme to tag two major geographical cases, using OpenITI mARkdown, as an inline annotation approach. OpenITI mARkdown is designed to cover a wide range of information in a text, ranging from structural elements to morphological and semantic elements in small- and medium-sized texts. Using the mARkdown, we then describe the annotation of complex narrative patterns of geographical descriptions, such as administrative divisions and routes.

In addition to the manual annotation, in this chapter we will also explain an iterative semi-automatic annotation approach using machine learning techniques. This approach is an alternative method to annotated toponyms in longer texts to which manual annotation may not be efficiently applicable. Based on this annotation approach, we then propose a method for expanding the initial toponymic data used for toponym annotation. This means the data that is produced in this stage identifies new toponyms from the context, which can then be added to the initial toponymic data. Toponymic data can be used as a reference for pre-modern toponym identification and resolution [11] and can act like a reference gazetteer. Furthermore, this chapter explains the data extraction steps after annotation as well as the conventional data structures for the extracted geographical entities to produce datasets that can be used in digital tools and techniques.

[4] Geographical Information Systems is a framework to create, manage, analyze, and map all types of data. GIS is designed to connect geographical data to a map, integrating location data.For more information, see `https://www.esri.com/en-us/what-is-gis/overview`.

Chapter 5 offers applications of the data gathered from the geographical sources in a number of models and visualizations. This is the next step after data extraction in the iterative method, which is explained in Chapter 4. We introduce models related to administrative divisions, route networks, and so-called *unknown place*s[5], and aim at the following applications: 1) representing administrative divisions and route networks; 2) comparing descriptions of divisions and route networks in different sources; and 3) estimating geographical coordinates of *unknown place*s.

In the first section of this chapter, we introduce a number of mathematical models to represent the geographical position and extent of the pre-modern regions. For each model, we discuss whether it fits well into the textual description of the underlying data, the advantages and disadvantages, and ways to improve it. The second section explains comparison models for representing similarities and differences between different geographical descriptions of the same geographical entity. Two models are discussed in this subsection: 1) a comparison model for administrative divisions; 2) a simplification and comparison model for route networks. One can apply the first model to learn how two sources divide the same area at different levels of the hierarchy of geographical divisions. The second model facilitates creating a simplified version of a route network from the source description. Two simplified networks of an area, each based on an individual source description, can then be used in studying the differences and similarities of the network descriptions of that area. The third section introduces a model to estimate the geographical location of *unknown place*s. We describe the spatial calculations that are required to find the possible coordinates and implement the model using sample data.

Chapter 6 introduces al-Ṯurayyā, a gazetteer and a spatial model of the classical Islamic world, which is built on the earlier version of this work. This chapter shows another use case for the data extracted from the textual descriptions and how the data is applicable as a basis of models and implementations for explanatory use cases. The first part of this chapter explains the functions of gazetteers, which provide access to the places, route sections, and relevant technical information through the search function. The relevant available sources of toponymic information are accessible—following the idea of Linked Data.

The second part of Chapter 6 describes a spatial model that is designed to provide an analytical tool for historical events and phenomena by performing complex queries, such as pathfinding and network flood. Additionally, it provides visualization of pre-modern regions using one of the described models in Chapter 5. The current implementation adds interactivity to search and network query features, maps of the regions, and accessibility to a primary source and available external sources when showing the information of the places to the earlier version.

[5] Applying various methods of toponym identification might still leave toponyms that are not matched to any geographical latitude and longitude for various reasons, for example, if a place no longer exists or its name has changed over time. We call these toponyms *unknown place*s.

Chapter 2

Related Work

In this chapter we review related work in the Digital Humanities that deals with geographical information. The review covers the major topics that deal with gathering and preparing geographical data in textual sources and using this data in tools and models, as related to the work in this dissertation. We begin with the use of Geographic Information Systems (GIS) in the humanities because, as mentioned in the introduction, GIS is the root for historical geographical information analysis. There is a great potential to use GIS in various disciplines and all historians use GIS to some extent ([12]). We will then review studies relating to the production of geographical data through natural language processing, georeferencing, geoparsing, and annotation. This section also includes an explanation of toponym identification and resolution. The topics in this section focus on extracting geographical information and mapping places to geographical coordinates. We then review academic work on gazetteer development, which provides human and computer access to a dictionary of locations defined with place names and geographical coordinates. The final section focuses on the modeling of geographic data and reviews approaches that can be reproduced to help gain deeper insights into data.

2.1 GIS

GIS technology has been used in the applied sciences since its emergence in 1960, and over the last two decades has also been applied to digital research within disciplines such as History ([13]). A growing interest in the study of space, place, and landscape, and the refinement of the application of quantitative methods and computational approaches, as [13] states, has also increased the use of GIS technologies in literary studies ([14, 15, 7, 16, 17, 18, 19, 20]). This interest is largely rooted in GIS's availability ([21]). In this section, we will review a number of studies that use GIS as a common spatial technology for storing data and mapping to see how space and place are being analyzed, and how such technologies have been used. GIS technology is based on a simple model that gives each database object a location in space[1]. This model therefore offers opportunities to handle data by mapping it to an existing map or using it to populate a map ([22]). In fact, GIS

[1] According to [22, p. 11]: "Among the GIS community this term is used in a very similar way to the term location. Spatial data are data that refer to locations, and where a GIS book might say 'consider the role of space', a historian may well say 'consider the role of location'. There is, in fact, a slight difference in definition as space is a scientific way of defining location, usually through a coordinate system, thus 'a location in space' usually means a location that can be defined using one or more coordinates."

technology integrates different information related to a specific place to its geographical location on a map. Additionally, it facilitates visualization of locations (see [21]).

The application of digital mapping and GIS in the humanities has been influenced by recent developments in GIS technologies. This application has now converged on new methods called historical geographic information systems (HGIS) and also Spatial Humanities or Geohumanities ([23]). According to [24, p. 66]), "HGIS scholarship combines historical geography and spatial and digital history with databases that record both locations and time thus enabling maps (including animations) to illustrate changes over time" (see, for example, [21, 25, 26, 27]). [10, p. ix-xxii] mentions the use of geographical technologies when referring to Spatial Humanities, defining them as: "geographical technologies to develop new knowledge about the geographies of human cultures past and present." [10] also states that the origin of Spatial Humanities is HGIS.

The past decades, in particular, have witnessed the application of geographical technologies to research in humanities disciplines and projects, such as [28, 29]. The expression, the "spatial turn" (see [30, 31, 32, 33, 34, 25, 35, 36]), as [37] states, is often mentioned in relation to GIS technology and new geographical technologies, which make mapping easy and accessible. [38] discusses the importance of "spatial analysis" in humanities scholarship and states that it is "not just about mapping, but it is also about trying to broaden the reach of humanities research."

[35, 10, 21, 39, 40, 19] explain how spatial methodologies have been used in humanities in recent years. Some humanities disciplines, such as Archaeology, have used GIS technology since the early 1980s ([41]), while other disciplines, like History and Literary Studies, have only begun to use it more recently ([13]). Gathering information and creating datasets is an underlying step to data analysis and delivering applied contributions to knowledge. This is where a wide range of research topics cover the use of GIS, and, in general, geographic technologies, spatial analysis, and statistics in various disciplines, such as Archaeology, History, Literary Studies, and Linguistics.

Archaeology has a spatial nature and was the very first discipline to use geographical technologies. The spatial dimension of archaeological practice has led to the utilization of GIS (see [41, 42, 43]) and other spatial technologies, such as Remote Sensing and GPS (see [44, 45]). Spatial analysis has been widely used in studying the excavations and other trends in Archaeology. As a more recent phenomena in Archaeology, as [19] states, the study of spatial description in archaeological textual sources applies corpus and computational linguistics techniques to extract geographical and contextual information ([46]). Also, projects like [47] explore texts for archaeological place names through visualizations. [48] states that mapping has been recognized as a practice in archaeology. This study, considering the intersection of mapping and archaeology, explores the ways of generating archaeological data and contributing to methodological and theoretical problems in archaeology by mapping. According to [48], use of alternative cartographic methods in archaeology connects to the spatial humanities, spatial history, and digital archaeology ([49, 21, 50, 19, 39, 51]) and further engagement with mapping, visualization, and digital technology is being pointed out (see [52]) while [53] argues the lack of explicit engagement in historical archaeology despite the use of mapping methodologies.

Unlike Archaeology, history, as [19] states, has only more recently begun to use geographical methods. Similar to Archaeology, GIS technology has been used by historians to find methods that help find answers to particular research questions. Historical GIS (HGIS), as a dynamic and diverse subfield, has recently applied GIS to the qualitative analysis of historical sources (see [54, 55, 56]), although the initial focus was quantitative exploration of economic and political data such as [57, 58]. Furthermore, recent experimental research in HGIS combines spatial and corpus analysis ([59, 60]). [61] introduces a methodological experiment that combines geographical visualization, algorithmic thinking, and text mining to examine whether and how they advance our understanding of Chinese architecture in the literati's knowledge. Virtual 3D modeling and visualization of historical places is another example where GIS is integrated with other technologies ([19]); for example, see [62] and The Virtual St Paul's Cathedral Project[2].

In Literary Studies, as [19] states, the power of GIS and related technologies is changing the interpretation of the material, imaginative, and discursive geographies in both individual sources and large corpora. Digital literary atlas projects, such as the Cultural Atlas of Australia[3], [63], and [64], adopt the idea that literary critics use maps as "analytical tools that dissects the text in uncommon ways and bring to light relations that would otherwise remain hidden", as [14, p. 3] states. [19] notes that the value of maps in generating a kind of abstraction from textual sources underlies the methodological premise of such projects. According to [19], this abstraction helps to create new research questions and to guide critical examinations.

However, despite its popularity, the integration of Spatial Humanities approaches in literary studies faces some obstacles. One issue is the complexity of geography in literary works that makes it hard to model using GIS techniques ([65]). Also, in fictional literature, geography, as [66, p. 19] states, "can create its own space, without physical restrictions", thus mixing uncomfortably with non-fictional locations in projects such as the literary digital atlas. To overcome these issues, new practices and frameworks are required for interdisciplinary collaboration. Moreover, to show the contribution offered by the practices and frameworks employed in the analysis of literary texts, further scholarship is required. In this regard, Stanford Literary Lab models the use of crowdsourcing to generate an "emotional map" of the English metropolis to study literary geography by visualizing the significance of the toponyms mentioned in eighteenth- and nineteenth-century novels[4]. [67] at Lancaster University applies data mining approaches and GIS technologies to represent the English Lake District in literature. At the University of Edinburgh, the LitLong project[5] applies text mining approaches and georeferencing (see Section 2.2) to extracts of early modern and contemporary literary works and uses GIS technologies to generate a topography of Edinburgh as a literary city on a map. This map provides interactive explorations of the city via a web interface and mobile application.

2.2 Natural Language Processing (NLP), Georeferencing, Geoparsing, and Annotation

In this section we review the state of research on NLP, geoparsing, georeferencing, and annotation. Several projects have been dedicated to the production of digital material that facilitates search functionality and makes them available online (see the report in [68]). Many topics use digital material in location-based approaches that cover a long list of analyses related to places; for example, place of birth or death, stops on an itinerary described in a travelogue, locations of castles, and city borders. Production of digital material, advances in and availability of geographical technologies, and growing interest in and familiarity with the recent available tools (e.g., open-source

[2] https://vpcp.chass.ncsu.edu

[3] http://australian-cultural-atlas.info/CAA/index.php

[4] https://digitalhumanities.stanford.edu/projects

[5] https://litlong.org

GIS software[6]) highlights the potential of Spatial Humanities to advance humanities scholarship ([19]). As this study offers, advances in automated geoparsing ([69]), is an example that shows this potential. Applying geographical technologies to digital material allows one to apply another category of research topics: text analysis tools and approaches.

Geospatial analysis of texts, as a challenging task, entails techniques, such as geoparsing, to gather information as well as other digital methods to produce material for distant reading based on the gathered data. Such analysis, at a minimum, enables mapping, as noted by [19]. A considerable amount of literature has been published on exploration and recognition of geographical references in raw text, either manually or through customized Named Entity Recognition (NER) techniques in recent years, as [70] states. Previous attempts at gathering geographical information from textual sources were mostly focused on finding named entities, either by annotation or specialized NLP techniques to automatically summarize the geographical content or extract various patterns from unstructured texts. NER is a crucial constituent of NLP that enables identification of names of various domains, such as people and places, through the recognition of linguistic patterns, according to [71]. This approach has great potential for large-scale information extraction.

Geoparsing entails recognition of spatial terms, such as place names, in an unstructured text. Geoparsing is also known by other names, such as georecognition, toponym recognition, and geotagging ([72, 73]). It is useful since the results of geoparsing a text can be mapped to geographical coordinates—known as “geocoding” or “toponym resolution”—and can make use of geographical tools and technologies in analyses and visualizations. Geoparsing free text is widely used in information retrieval. Some of the geoparsing applications that are applied to modern material include: automated image tagging, web page annotation, and social media analytic ([74]). In fact, geoparsing and geocoding can be seen as complementary fields of recognition and geographic information recognition ([72, 75]). Geoparsing requires both toponym identification (e.g., Alexandria in a given passage refers to a place) and toponym disambiguation (e.g., Alexandria in that passage refers to the famous city in Egypt or a suburb of Washington, DC, or some other particular Alexandria). Toponym identification relies on two NLP approaches: NER (which identifies named entities such as place names and personal names) and Named Entity Matching (NEM). NER, according to [72] can be seen as a specific application of NLP and usually relies on linguistic properties such as part-of-speech tags[7] ([74]). As [73] states, NER is a crucial part of geoparsing ([77, 78]). Named Entity Matching (NEM) refers to toponym identification based on an authority list (similar to VIAF.ORG[8] for authors) such as a reference database, tag set, or gazetteer.

Toponym resolution (or geocoding) is a particular subset of the more general NEM task. Toponym resolution refers to the approaches that link a toponym (geographical entity) to a spatial interpretation in physical space for each geographic reference. A gazetteer is a special kind of authority list for geographic entities. In particular, it provides geographical coordinates expressed in latitude and longitude locating places and in some cases polygons bounding regions in space. Toponym resolution includes searching for a match among gazetteer entries and then disambiguation if there are multiple places with the same name (for instance, Washington designates in the US multiple cities as well as a state). We will describe the work on a gazetteer in Chapter 6.

[6] Examples are: QGIS, a free and open source geographic information system (`http://www.qgis.org`) and MapWindow (`http://www.mapwindow.org`).

[7] Georeferencing that is applied to free text also involves NER and toponym resolution approaches (see [76]).

[8] According to the description at `http://viaf.org`: The Virtual International Authority File combines multiple name authority files into a single name authority service. The goal of the service is to lower the cost and increase the utility of library authority files by matching and linking widely-used authority files and making that information available on the Web.

There are many gazetteers that serve different functions. Very large, contemporary gazetteers can contain both too much information and not enough. When we are scanning texts that were produced more than a thousand years ago, we do not want to be matching place names from North America (e.g., Alexandria, VA, or Troy, NY). At the same time, when we are working with historical sources, we may need to add information about places that do not appear in gazetteers for the contemporary world (e.g., Geonames[9]) or for other areas of the pre-modern period (such as Pleiades.org, which has focused primarily on the Greco-Roman world). Thus we have begun developing a gazetteer for the pre-modern Islamic world, which we will discuss in detail in Chapter 6. [72] describes the process of spatially encoded data—which includes geographic references—and surveys the techniques for extracting the information and geoparsing methods. It also provides an overview of the available tools that typical applications utilize for recognizing spatial language in textual sources.

Georeferencing ([69, 76]) assigns spatial footprints to geographic entities that appear on a map or in a text. Georeferencing on a map means using a modern map to locate points and polygons in a modern coordinate system. Georeferencing in a text is to link the place names to authority lists of place names, such as gazetteers, which have spatial info and allow us to map the place names. It has been a challenge in geospatial technologies ([33]), nevertheless, it is a common method to digitize the geographical information on maps, including historical maps that are not connected to any geographical infrastructure. It facilitates automatic spatial analyses of toponyms as geographical entities, such as placing the entities on a map or providing spatial search capability ([76]). Automatic georeferencing can be applied to the texts in natural languages in which the geographical information is in the form of place names.

In addition to NER, another class of NLP techniques allows automatic annotation within a text. In this regard, various work has been done on the identification of word classes, such as adjectives, nouns, verbs, and adverbs, in a text using part-of-speech (POS) annotation: POS categories help to find place names alongside the word form, such as proper names; adjectives help to catch the context related to landscape ([19]). Semantic annotation ([79]) adds another level of conceptual indicator to names or concepts in a text. Moreover, named entity disambiguation (NED, see [80]) connects to available data and knowledge bases to build a link between an identified toponym and a canonical reference. Similarly, toponym resolution provides geographical coordinates for the identified toponyms. A combination of NER and assigning coordinates is reflected in geoparsing attempts.

The increasing interest in using NER and NED technologies in digital humanities produced promising results ([81]); however, the application of NER technologies to classical corpora ([82, 83, 84, 85, 86, 87, 88]) poses numerous problems and few studies deal with historical issues, as [70] states. Additionally, NLP methods are not yet fully reliable in humanities scholarship and there are challenges to be met, such as losing contextual information or spelling variations in pre-modern corpora ([81]). A significant number of studies to date have mainly addressed issues of qualities of digital versions and OCR outputs, European or classical languages, or variant spelling ([89, 90, 91]). Problems of geoparsing[10] for toponym identification in text, mostly covering heuristics for toponym resolution, have also been discussed widely in prior studies ([92, 60, 93, 94, 95]). Many studies in this area have primarily drawn on rule-based or pre-existing statistical NER models to solve

[9] Examples are: GeoNames (`http://www.geonames.org/`) and GNIS (`http://nhd.usgs.gov/gnis.html`).

[10] Geoparsing, as a special toponym resolution approach, is an automatic process that resolves place names in natural language to toponyms in a geographic gazetteer with geographical coordinates.

toponym identification issues, mainly in the field of geographic information retrieval (GIR) and are well-known and widely discussed with regard to a variety of genres, including historical texts ([96, 94, 97]).

In historical corpora, as [70] states, combining GIS and NLP technologies has improved successful identification and analysis of toponymic information, often using relatively simple techniques that automatically identify toponyms and disambiguate them by matching them against reference gazetteer entries ([19, 60, 98, 20, 99, 100]). Historical collections require NER systems that can work with classical languages and gazetteers ([101, 102, 103, 19, 104, 105]). Most of these works use one NER system, whiles [70] introduces a combined system. In this regard, for advancing the computational analysis of vast historical corpora, [106] leverages the combined benefits of NLP, Corpus Linguistics, Machine Learning, and Spatial Analysis.

For less commonly studied languages, such as Arabic, the challenges are more salient. Despite the extensive literature on NER, few studies on the Arabic language could be found ([107, 108, 109, 110, 111, 112, 113, 114, 115]). Several works address the complexity of morphological and syntactic variations in modern Arabic NER. However, even fewer studies deal with classical Arabic, the main language of the written tradition of the Islamic world.

Beside NLP techniques—or computational linguistics, which is drawn from computer science—another category of techniques is rooted in corpus linguistics. This category includes the development of text analysis methods. Advances in the power and storage capacity of computers encouraged the development of methods that leverage new computational resources as well as new searching and analysis software to analyze large quantities of digitized texts.

According to [19], five major methods have contributed to the Corpus Linguistics discipline and enable researchers to discover linguistic patterns by offering semi-automated approaches that allow researchers to analyze far more textual data than they could with purely manual analysis. These methods are: (1) frequency lists, which show how frequent and, in an extended version, how distributed different words are in a text and enable researchers to concentrate on specific features of a corpus and help them to understand the importance of a word in a corpus; (2) concordances, which provide the context for each word in a corpus, and which in turn allow a researcher to discover patterns (and spatial patterns in a geographical context) that appear within that context; (3) n-grams, which are consecutive lists of a specific length (n) of words that appear within a text; (4) comparison of different frequency lists using keywords, which juxtaposes frequency lists for different texts (or corpora), which employs statistics to recognize words that occur more in one corpus versus another, and which can reveal useful topics (the list of words can also be used for further analysis); and (5) collocation, which detects which words regularly occur close to each other in a text and can be used by researchers, in a geographical context, to find topics related to the place names (or other geographical entities) in a corpus.

Corpora themselves are spaces and can be mapped. The combination of corpus linguistics methods and GIS, [59] explains a method to create, analyze, and visualize a database of place names and considers how the combination of collocation, keywords, and semantic analysis facilitates visualization of the meanings associated with particular place names. An example of the application of this method mentioned in this research is a map that shows the distribution of mentions of concepts, such as war or money, in a corpus.

Below are some examples of works that show how digital approaches aid extraction of geographical information in textual sources and how the information contributes to and informs the research in different fields of humanities.

- [104] focuses on geoparsing and outlines projects[11] that customize and use the Edinburgh Geoparser ([116]) for particular datasets and applications. The Edinburgh Geoparser is a system developed by the Edinburgh Language Technology Group (LTG) to automatically recognize toponyms in text and to disambiguate them. In the disambiguation step, several gazetteers, such as GeoNames[12] can be used to match the place names to the corresponding entries with latitude and longitude. Initially, it was configured for modern texts, but more recently the geoparser has been adapted for application to historical and ancient texts and modern-day newspaper text ([85, 117, 118, 119]).

- Using a georeferenced corpus consisting of texts from Royal Commissions in 1863 and 1893, [120] explains an approach to summarize the reports on fish stock and to evaluate the perception of changes in sea fishing stock. This study combines text mining approaches—computer-assisted qualitative data analysis (CAQDAS)—and GIS technology to compare the two texts and demonstrates the usefulness of this combination in historical analysis. To apply CAQDAS[13] to this corpus, this study starts by OCRing the corpus and generating the following material to use in analyses: meaningful units of evidence, word frequencies, selecting keywords and terms, tagging, locating and refining the co-occurrence of keywords in close proximity, and keywords in context. [120] discusses two inquiries based on the data extracted from the texts and how this approach can "launch a spatial history of practices, knowledge, and attitudes associated with sea fishing in Britain."

- [121] argues the importance of geographical tools and technologies in studying the literary representations of landscapes as a complement. Using LITSCAPE.PT ([122])—Atlas of Literary Landscapes of Mainland Portugal—project as an example, this study explains how integration of traditional reading practice with distant reading, GIS, relational databases, and collaborative works facilitates classification and analysis of excerpts of Portuguese literature (350 works dating from 1843 to 2014). The aim of LITSCAPE.PT, as [121, p. 59] states, is "to examine how literary representations of Portugal's landscapes have changed over time." The two outcomes that are presented as case studies, according to [121, p. 57], include: "documenting the evolving literary geography of Lisbon" and "exploring the representation of wolves in Portuguese literature."

- Geotagged sources and digitized historical texts offer new ways for exploring how places are described. [123] explains a project that explores notions of place in a range of digital sources, including modern user-generated content[14] as well as digitized historical documents such as Text+Berg ([124]). This combination, as [124] argues, is an opportunity that enables representation of different conceptualizations of the same place with theoretical underpinnings from, for example, Geography and Information Science. Ambiguity, vernacular names, ways of using concepts in different locations and by different groups, automatic generation of macro-maps from historical texts in space and time and, finally, techniques for characterizing and comparing regions based on the terms used to describe them are the issues that [124] mentions while exploring written representations of landscape. This study, through a set of

[11] The projects include: The Trading Consequences Project: Georeferencing Nineteenth Century Text, The GAP Project: Georeferencing Classical Texts, and The GAP Project: Georeferencing Classical Texts.

[12] `https://www.geonames.org`

[13] This study uses MAXQDA software (`https://www.maxqda.com/`).

[14] For example, Flickr (www.flickr.com) or Geograph (www.geograph.org.uk).

case studies, argues that engagement with written representations of space goes beyond the measurements of Euclidean space, which GIS facilitates, and yields a more exact conception of place ([19]). Using local information helps to gain insight into the ways in which a concept, such as a hill, has a different meaning in Great Britain than in Switzerland.

- Similar to [123], [125] states that a Euclidean conceptualization of space is not adequate for modeling networks of different natures and systems of knowledge exchange and that it is necessary to move beyond this. Using an example, [125] proposes the combination of features of GIS and network representations. This combination integrates intellectual and technological geographies—which show the spatial distribution of knowledge—and network representations of relevant actors and documents in the history of knowledge exchange in Early Modern Europe. [125] then argues that spatial humanities approaches are helpful in reconstruction of the spatial distribution of actors and documents, which might be more useful than trying to explain knowledge exchange by focusing on The Republic of Letters as one entity consisting of scholars. This study, inspired by "deep map" introduces a "deep network" that the combined method can help to construct.

2.3 Gazetteer

Recently, digital humanities scholars have been increasingly relying on digital historical gazetteers. There is a need to cover places and historical periods in gazetteers and provide human and automatic access to them [126] With improvements in digitization of historical texts and improvements in the algorithms of extracting place names, the potential of indexing the data on place names has become more and more achievable ([127, 128]). But, what is a gazetteer? They are defined as place name directories that provide geographic locations of place names as well as some description of the places themselves. Each entry includes at least one name, one feature type, and one location ([69]). According to [126], in many print atlases, place names, geographical coordinates, and a page location are listed in a gazetteer at the back of the volume. However, local gazetteers in China, for instance, have been a repository for many details about places, on a scale that ranges from provinces to towns to temples and rivers ([129]). Other gazetteers might have more information. For instance, Georgette Cornu's *Atlas* ([130]) is a collection of printed maps that shows the geographical position of places in the classical Islamic world as well as their type; for example, town or village. Additionally, maps of provinces specify which province the toponyms belong to. Accordingly, it also proposes divisions of the Islamic world, by categorizing toponyms in the provinces.

A number of attempts have simply developed digital versions of gazetteer texts and do not provide a map view, while other works have focused on online, map-based implementations of features to provide the immediate requirements of a gazetteer and/or expanding it with spatial models and functionalities. In what follows, we will provide an overview of a few projects of the latter type and will consider the following approaches as indicative of major examples that represent prominent methods. Moreover, we will briefly explain some related projects that serve similar purposes.

Pleiades ([131]), as a gazetteer, is a member projects of Pelagios[15] ([132]) that provides search-

[15] Pelagios is "a community-driven initiative that facilitates better linkage between online resources documenting the past" ([131, p. 1]).

able entries to toponyms of the classical world. This was initially limited to the Roman world, however, with Pelagios 3, it has begun to partially expand its scope, including to some parts of the classical Islamic world. The entries are enriched with a list of relevant information, such as geographical coordinates, URIs, references, various names, place types, and references, and are available in different formats, like JSON, KML, and XML. Pleiades has conventional approaches to link its contents consisting of text, images, media, 3D objects, online databases, and so on. One of these conventions regarding the name of a place is to reference it using the gazetteer URI; for example, referring to the possible location based on existing literature. Another important convention is that the resulting metadata of that place should be openly available online. Therefore, extensive information on a place can be extracted while it is visualized on the map.

Other similar projects employ the same functionality as Pleiades on a smaller scale, such as the PastPlace Historical Gazetteer[16] ([133]) and the Digital Atlas of the Roman Empire, which is extended to an atlas with searchable modern and ancient names. The Syriac Gazetteer[17] is yet another example of a geographical reference work for places of variant type or size relevant to Syriac studies. Data is provided in a database of all places with or without geographical coordinates and a series of maps is available to represent the places with geographical location.

The temporal gazetteer of Chinese history, CHGIS ([134]) is another approach to develop a gazetteer by integrating other types of information, such as administrative units and temporal concepts. It compiles the source texts for Chinese local gazetteers into "Dynastic Gazetteers." This project's objective is to offer a flexible tool, in the form of a database of places and administrative units. In doing so, it creates "a continuous time series of records that track changes in place name, administrative status, and geographic locations" (see `http://sites.fas.harvard.edu/~chgis/pages/intro`). Its data model requires administrative units in existence at a particular time, place names, administrative status (feature type), administrative hierarchy for a particular place instance, and changes to a particular administrative unit over time. Thus, when a new place is being added to the CHGIS database, it needs the following parts: change in place name; change in administrative status; change in location or boundary. Developed as a temporal gazetteer web service, it provides name, type, temporal span, spatial details (feature type and coordinates), and source and data source for each entry that one searches for. The underlying data enables searches for administrative units and capitals for any given time in Chinese history, building customized digital maps for particular times and places, or joining customized datasets for spatial analysis, thematic mapping, or other specialized statistical modeling of interest.

Combining other types of geographical data, such as trade routes and connections, enables network analysis and spatial models based on the underlying route network. The majority of digital gazetteers represent geographical information with a certain level of granularity for the entities represented—such as village, city, or province. Although this approach has proven successful in a large number of applications, this functionality, according to its primary aim, is inadequate when studying other major historical aspects, such as movement and travel, where geography and correlation of places play a fundamental role.

Integrating routes and places, ORBIS ([135]) introduces a historical geospatial network model of the Roman world with a primary focus on depicting different types of travel. The model reconstructs the costs of travel between different points both financially and time-wise to illustrate the possible paths to travel through the Roman Empire and offers primarily perspectives on the

[16] `http://www.pastplace.org`

[17] `https://syriaca.org/geo/index.html`

movements. Consisting of a digital archive of sites and routes using data from both primary sources and computational geography simulations about travel ([136]), ORBIS combines tools, archives, and publications to introduce a new genre and represent the nature of movement and the dynamic behind it (see [137, 138]).

The Heritage Gazetteer of Cyprus (HGC)[18] aims to gather the historic geography of Cyprus from different sources into a gazetteer. The gazetteer entries consist of two types: monuments, which are assigned geographical coordinates, and a group of geographical entities, such as villages, for which a polygon is assigned, since the precise location is unknown. Transliterations of place names are a challenge; resources are in various languages, which makes identification of place names challenging. This project is not limited to the classical world, as modern administrative regions and toponyms are named using the available transliteration references. It provides a rich description of place names and a specific place name is not preferred over another as a default name.

The PastPlace project provides entries for data on Wikidata as well as Vision of Britain Through Time[19] to prove the use of Wikidata as a gazetteer. It goes further than minimal gazetteer entries and provides place encyclopedias, describing localities and their relationships with other geographical entities.

In this dissertation, our approach to developing a gazetteer not only contains the gazetteer features (similar to the Pleiades), it also engages administrative units in the models and visualization, as implemented in CHGIS. Moreover, inspired by ORBIS, it combines a network of routes and implements a spatial model of the classical Islamic world in a set of functionalities.

2.4 Modeling

[15, p. 1-2] says that he aims to move “from texts to models, then; and models drawn from three disciplines with which literary studies have had little or no interaction: graphs from quantitative history, maps from geography, and trees from evolutionary theory.” Having in mind the common phrase “all models are wrong. Some are useful”, we now dig into the modeling in geographical data, specifically pre-modern geography, and explain the development of this concept in a number of works as a way to represent data for analysis.

[139, p. 255] defines modeling as “the heuristic process of constructing and manipulating models” and a model as “a representation of something for purposes of study,” or “a design for realizing something new.” This study explains that a model has two effects, which highlights the distinction between a model and an idea: first, it is computationally tractable to help us see the gap between what we know and what we can computationally obtain; and second, it is manipulable, which requires productions (e.g., diagrams and objects) that can be handled in a rather short time and underscores the process more than the product. A model, unlike a concept, attempts not to freeze a phenomena into a historical abstraction and instead, attempts to catch the dynamic and experiential aspects of it ([139]). Manipulability requires interactivity of a model and this means, a model has something, such as a diagram, that can be handled in a reasonably short time. To give a clearer idea of this time frame, it is useful to mention ([140]) where it, in the early significant developments of computer modeling in the 1950s, states that the lifetime of a model could be days or weeks, not years.

[18] http://www.cyprusgazetteer.org

[19] http://www.visionofbritain.org.uk

The definition of a model in [139, p. 257] represents modeling as “a temporary state in a process of coming to know rather than fixed structures of knowledge.” Therefore, the two phases of modeling, construction and manipulation, might be indistinct. This means, modeling intends to gain detailed knowledge and the knowledge gained from a model can be used in return to improve the model ([139]).

As [141] states, since history is closely connected to geography, spatial relationships are of great importance and historical explanations must recognize this connection. The potential roles of spatial models can be: 1) they help articulate spatial relationships; or 2) they provide missing data or knowledge by an inference capacity that hypothesized spatial relationships give.

[141] shows the application and potential of modeling techniques developed for contemporary geography in ancient geography. It gives examples of studies that have introduced models with little or no attention to spatial analysis of the information available in the maps, for instance [142]. A few studies, such as [143, 144], have used spatial analysis, but they are limited to drawing Voronoi polygons around centers and/or using statistical analysis to explore patterns or correlations between geographical entities such as settlements or archaeological sites. Spatial interaction models are occasionally used in a number of studies, but they often focus only on interaction and use primitive models, such as the gravity model ([145]). Introducing spatial interaction models, [141] states that there are broader categories of models for particular applications and notes how extending modeling concepts can be used for locational analysis and spatial interaction. Using the evolution of shopping center structures in Leeds as an example, [141] explains a model that describes this evolution. To show the general structure and potential application of the model in historical analysis, this study then extends the model to apply it to an ancient application: investigation of the emergence of the polis in Ancient Greece around the eighth century .

Site catchment analysis studies initially triggered the interest in modeling ancient movement in the 1970s ([146, 147]) and computer modeling of pathways and movement networks in Archaeology has been a subject of interest since the early 1990s, according to [148]. [149, 150] are examples of the first studies that introduce approaches to reconstruct ancient pathways using GIS. According to [148], a considerable amount of literature has focused on achieving reconstructions of ancient routes and movement patterns by incorporating the influential factors into the models. This study gives an overview of different approaches of pathways and movement networks in Archaeology and shows the importance and complexity of understanding ancient movement based on computer models. Additionally, this study summarizes the major issues of the approaches from a theoretical and methodological perspective. Below are samples of the studies in various areas related to modeling ancient paths and movement.

[151] works on territories of settlements and[152, 150] utilized cost surfaces for calculating and analyzing the movement around settlements. [149] adds calculation of least-cost paths (LCPs) to quantitative approaches of construction of ancient movement patterns and pathways. Other studies, such as [153] focused on movement capability. A number of studies focus on calculating other costs of movements, such as expenditure of energy for walking distances ([154, 155]), and movements through waterways ([148, 43, 156]). Many studies, as [148] states, focus on reconstruction of movement in ancient networks and model linear connections to predict movement using LCP toolboxes in GIS. Instead of just relying on the linear routes that a raster GIS environment provides, a number of studies have explored a more realistic approach for predicting movement patterns ([157, 158, 159, 51]). Another approach to model movement is agent-based modeling (ABM), which models goal-oriented and dispersal movements by simulation of the environment

to which the agents react, while also interacting with other agents ([160, 161, 162]).

A number of studies explore approaches to modeling potential movements to cover highly variable routes and incomplete datasets for modeling pathways. [163] develops a method to provide a movement proxy by calculating the average cost of moving from a grid cell (in the network) to other cells in an arbitrary radius. Similarly, [164, 165] calculate "potential path fields" and then sum accumulated cost surfaces for each grid cell. These two approaches can be used to analyze movement potential at various scales. As an alternative approach to produce movement potential maps, [166] offers a different weighting system based on a similar idea of calculating the costs from each node to other ones and assigning higher weight values to closer areas. [167, 168, 169, 158] experiment another approach by calculating multiple LCPs between settlements on the edges of each region and accumulating the values. [170, 171] calculate LCPs from starting points in a region to either the edges of the study region or to a number of settlements within a specified radius and [172] similarly calculates LCPs between random settlements that are positioned at variable distances. The idea of hydrological flow accumulation is being used by [173] to introduce an approach of accumulating cost surfaces, which [174] then experimented with to produce a focal mobility network.

Network analysis is another field that has been widely used in archaeological and historical studies (see [175, 176]). Studies in this field focus on topics like trade and communication networks (e.g., [177, 178, 179, 180, 181, 182]) and social network development (e.g., [183, 184, 185]). Reconstructing the network of connection, several studies have examined the potential of network analysis techniques for path modeling ([186, 173, 187, 188, 189, 190, 191, 192]).

Network (re)construction, as [148] states, requires different techniques to the path modeling used in GIS, as it limits simultaneous direct connections between multiple nodes. Common and efficient network construction techniques intend to limit the number of connections based on the number of the closest neighbors and/or their distance ([148, 193, 194, 195, 196, 197]). The gravity model is a common approach to limit the direct connections that can be used for network construction. [198, 199, 200, 201] have introduced methods to calculate the strength of interaction between the nodes based on the gravity model and limit the connections accordingly. It is very common that the data is not complete enough to construct the network. Studies such as [202, 203, 204, 191] have explored adding simulated nodes in the predicted proper locations in the network.

Network analysis techniques provide local and global metrics ([205, 206]). Local metrics, such as node degree, are primarily about the position of the individual nodes and can be used to study the hierarchical structure of the network. Global ones, such as clustering coefficients, help when studying the connectedness of the network. Analyses based on the network edges focus on the weight of the edges to understand the importance and strength of connection ([185]). Additionally, [186, 207] engage node size in assigning weight to the edges. Space syntax [208, 209] is an alternative approach, which uses the routes for analysis instead of the flows that go to and come from the nodes. [171, 188] combine space syntax and LCP modeling. Moreover, [184, 202, 190, 196, 203] combine node-based approaches and path models.

[210] introduces a methodology to model the travel costs of the Iberian Peninsula in the Roman transport network by assigning values to the network connectivity. The purpose is to reconstruct the transport conditions as a way of analyzing Roman infrastructure and transportation systems. Using Network Analysis and Social Network Analysis, the results of this study help with understanding the characteristics of geographical positions of centers and cities and explain some economic, political, and social dynamics.

Chapter 3

Classical Geographical Cases

Information related to places is ubiquitous in historical studies. Events, people, social and cultural phenomena, indeed almost any part of history, are related to one or more identifiable locations ([12]). It is then no surprise that in various historical studies and modes of analysis, geographical descriptions have been used in many ways by many scholars.

Geographical information in pre-modern sources is described based on the pre-modern understanding of the globe and functionality of the landscape. Pre-modern descriptions of geography not only in classical Arabic but in languages such as classical Chinese, Greek, and Latin, include various functional types of constructs that describe the landscape in terms of its use in pre-modern societies, such as for navigation and administration. Independent from the social and cultural context of the source, the linguistic and expressive basis of these constructs allows us to identify and specify them. The most important constructs across the classical sources can be classified as follows:

- Names of places (toponyms).
- Space segmentation: this can be geographic features (e.g., rivers, mountains) or on more arbitrary administrative or diplomatic decisions.
- Routes: these form a relation between places, most usually expressed with distance estimates. Other conditions, such as orientation and direction specifications, can often appear as additional context. This type of information can be part of a complete route path or itinerary, or part of a network (for this distinction, see 3.2, below).
- Geographical orientation: this can be contextualized indications of directions, movements, or placements in the landscape.

The constructs in various sources are dependent on the purpose and viewpoint of the sources. For instance, comprehensive geographies are more concerned with hierarchical data, functional to the definition of national boundaries, whereas route descriptions and distance estimates are much more relevant in travelogues. Two of the above categories are especially relevant, not only in terms of their role as individual geographical entities, but also for the purpose of modeling, which are hierarchical data describing administrative divisions as well as routes and distance.

In the following sections we will discuss prominent geographical cases as they appear in major pre-modern Arabic sources.

3.1 Gazetteer

Gazetteers stem from the pursuit of a form of encyclopedism, which wishes to provide descriptive geographical information for places. That is, the encyclopedic perspective of a gazetteer implies the intention to record a complete knowledge set. Accordingly, a gazetteer provides a typical dictionary-like list of entries of geographical names of any structure or composition, social statistics, and environmental features of an area or region, such as a continent, country, or town, together with the related information on those places. A gazetteer can also take the form of any structure of comprehensive descriptions of places that offer the above information. The information given might be enriched with a map or atlas and can involve the geographical location, direction, or position of a subject, measurements, and dimensions of geographical elements such as mountains and water bodies. A formal definition of a gazetteer is introduced in ISO 19112:2019[1] (p. 5), which defines it as:

> A directory of geographic identifiers describing location instances. It will contain additional information regarding the position of each location instance. It may include a coordinate reference, but it may also be purely descriptive. If it contains a coordinate reference, this will enable transformation from the spatial reference system using geographic identifiers to the coordinate reference system. If it contains a descriptive reference, this will be a spatial reference using a different spatial reference system with geographic identifiers, for example the postcode of a property.

According to this definition, the geographical location of a geographical place may be contained in a gazetteer entry. Place names may also be associated with one another through the ways in which the location and any further specification of a name is described.

Gazetteers of the pre-modern world obviously do not include information on modern concepts, and instead follow the classical understanding of geography in their descriptions. As an example, geographical locations and positions are described by cardinal directions in pre-modern societies.

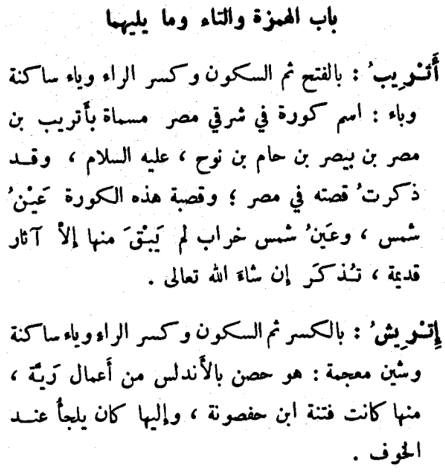
A prominent example of a gazetteer from the early Islamic word is *Dictionary of Countries (Mu^{c}jam al-buldān)* [1], written by Yāqūt b. ^{c}Abd Allāh al-Ḥamawī al-Rūmī (1179–1229 CE). This book sums up almost all medieval Islamic knowledge of the globe with more than $1,000,000$ words[2], which, as [211] notes, has made it an indispensable and regularly consulted book and an excellent source for reference even to the present day. For every place, arranged in alphabetical order, the author provides an entry in the form of a dictionary with philological information together with the exact spelling of the names. Additionally, it provides the geographical position and boundary of the listed entries as well as other available information such as mountains, deserts, seas, and islands. In addition to geography, it covers a vast range of information, drawn from fields including archaeology, ethnography, history, anthropology, and the natural sciences.

Figure 3.1 shows the first two entries in the *hamzahŧ and tāʾ* section. The first entry is a name of a *kūraŧ* (region) and he provides information about the name, geographical location (in eastern Miṣr/Egypt), and the capital city in this region. The second entry is a fortress in al-Andalus. Yāqūt also provides vocalizations for each entry.

[1] "Geographic information – Spatial referencing by geographic identifiers," International Organization for Standardization, Standard, 2003. For more information, see `https://www.iso.org/standard/26017.html`.

[2] The digitized edition is available at `https://raw.githubusercontent.com/OpenITI/0650AH/master/data/0626YaqutHamawi/0626YaqutHamawi.MucjamBuldan/0626YaqutHamawi.MucjamBuldan.Shamela0023735-ara1.mARkdown`.

Among other works of a similar nature, another example of the comprehensive collection of geographical knowledge and description is al-Muqaddasī's *The Best Division for the Knowledge of the Provinces (Aḥsan al-taqāsīm fī maᶜrifaŧ al-aqālīm)* [212, 213], which stands as a prominent representative of Arabic geography in the tenth century CE. Constructing his work with intentional thoroughness and following a different structure than Yāqūt, al-Muqaddasī provides a subjective and technical method of compiling his works and offering first-hand knowledge. Through a precise practice of the description, he organizes the knowledge by countries, instead of places, and represents the positions by descriptive language. Al-Muqaddasī describes the realm of Islam in fourteen regions; six Arab regions as well as eight regions of non-Arabs, including those in modern Iran, Afghanistan, and North Africa. The structure of his geographical description shows a multi-level hierarchy of administrative divisions, from the level of major provinces down to settlements. He starts at the top with the province, followed by their description, and then below comes the districts, with their major towns and other existing types of settlements/sites. Paying close attention to the districts and their categorization, he describes other types of micro-regions in the hierarchy of administration and provides all the local specific vocabulary.



باب الهمزة والتاء وما يليهما

أَتْرِيبُ : بالفتح ثم السكون وكسر الراء وياء ساكنة
وباء : اسم كورة في شرقي مصر مسماة بأتريب بن
مصر بن بيصر بن حام بن نوح ، عليه السلام ، وقد
ذُكرتُ قصته في مصر ؛ وقصبة هذه الكورة عَيْنُ
شمس ، وعَيْنُ شمس خراب لم يَبْقَ منها إلاّ آثار
قديمة ، تُذكَر إن شاءَ الله تعالى .

إِتْرِيشُ : بالكسر ثم السكون وكسر الراء وياء ساكنة
وشين معجمة : هو حصن بالأندلس من أعمال رَيَّة ،
منها كانت فتنة ابن حفصونة ، وإليها كان يلجأُ عند
الخوف .

Figure 3.1: The first two entries from Yāqūt's *Dictionary of Countries* ([1, p. 87]) at the beginning of the Section of *hamzaŧ* and *tāᵓ and what comes next.* The first entry reads: "Atrīb: with *fatḥ* then *al-sukūn* and *kasraŧ* for *al-rāᵓ* and *yāᵓ* with *sukūn* and *bāᵓ*: name of a *kūraŧ* in eastern Miṣr (Egypt). It is called Atrīb b. Miṣr b. Baysir b. Ḥam b. Nūḥ, peace be upon him, and his story was mentioned in Miṣr, and the capital of this area is ᶜAyn al-Šams, and ᶜAyn al-Šams is a ruin, of which only traces of ancient remain." The first part of the second entry reads: "Itrīš: with *kasraŧ* then *al-sukūn* and *kasraŧ* for *al-rāᵓ* and *yāᵓ* with *sukūn*, and *šīn muᶜjamaŧ*: it is a fortress in Andalusia ..."

Here, the encyclopedic approach to description tends to cover an exhaustive representation of all the above-mentioned components from various angles, such as geography, climate, productions, culture, weights and measures, trade and money, taxes, customs, juridical and theological schools Qurᵓānic readings, sacred places, power in place, and movements ([211]). Al-Muqaddasī describes the movements through connections and route sections that connect settlements and regions. For each province, he describes the network of routes, chiefly in a ramified shape, but also taking a linear form when describing itineraries. Detailing the routes with distances, he provides measurements in the regional and local units used in pre-modern society. By describing these connections, the route network of each region is thereby shaped as a graph of places and routes.

3.2 Routes and Travelogues

Connections and trade routes, which represent the movements in pre-modern societies, are a significant part of pre-modern spatial descriptions in the works of geographers, travelers, and historians. Generally, the introduction of connections and routes appears in the form of individual route sections through which at least two places are connected. The routes are mostly characterized by distances expressed in classical units. These can be miles in Roman society, as indicated in [214], or leagues, as an additional equivalent in the case of Britain and Gaul. They can also be *stadia*, such as the measurements in [214], or can be given as a stage (*marḥalaŧ*)[3], parasang (*farsaḫ*), day, or mile (*mīl*) in Muslim societies as the way al-Muqaddasī specifies distances. Additional details might be provided while describing the connections—such as indications of geographical orientation or conditions of travel—depending on the granularity and intention of the description. Despite this range, three parts are always essential to form an individual entity and define its structure: 1) the start or subject; 2) the end or object; and 3) the distance or predicate.

The description of connectivity by introducing route sections can conceptually be part of an itinerary (of any type) or a route network. These two entities cover major parts of such descriptions while their conceptual distinction might not be reflected explicitly in the text. In fact, the context of the description might be enough to imply its type (i.e., whether it is part of a route network or an itinerary). This distinction is functional to computational extraction and modeling workflows and we will discuss this in further detail in Section 4.1.1.2. Based on this distinction, a route network is a comprehensive description of the travel system, trade routes, and, in general, connectivity in terms of its spatial extent. Examples of a route network description could be a narrative describing several connecting routes in an area or one describing routes that have a central place in common.

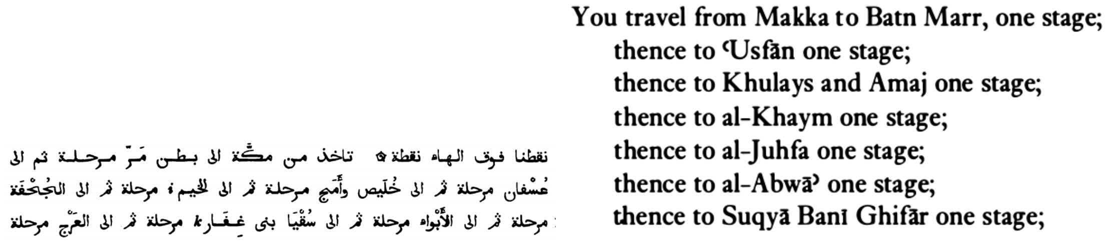
Al-Muqaddasī provides detailed and comprehensive descriptions of routes for every region that he introduces. As mentioned in Section 3.1, he structures his work by describing countries or provinces with various hierarchical administrative divisions. When it comes to the lower level divisions, he allocates significant room to describe routes that connect settlements and, in some cases, the neighboring regions. His descriptions mainly shape a network by bringing all the pieces together. He does this in various ways:

- He introduces a list of routes connecting a consecutive list of places to explain a path from one place to another, which, in fact, shapes an itinerary. The start and end places are often major cities through which many places are connected.
- He describes all the possible routes that start from a major city and connect to other places, either in the form of an itinerary or individual routes. The description follows a ramified pattern: a set of routes starting from one central point, which is often an important city.
- He describes individual routes that connect two places, which is not necessarily part of an itinerary or a network.

For the route sections, al-Muqaddasī often offers a distance and the measurement units given are provided according to the local area that he describes. However, he describes the routes in various forms and the descriptions offer, in general, a network of routes, which illustrate all the possible connections and the shape of connectivity in the corresponding area. Figure 3.2 is a short example (3.2a in Arabic and 3.2b in English) from this book, where the author describes the routes from Mecca (Makkaŧ) to Suqyā Banī Ġifār (and further to al-ᶜArǧ), with each section specified by a pre-modern unit, a stage of travel.

As an itinerary, a linear route is indicated in the shape of a path from a source to a destination along which the author lists places in a progressive way, often connected through the respective distances. According to the definitions, both network and itinerary route types have the same structure of description and building parts, but the method that an author uses to introduce and describe them indicates the distinction. That is, a linear form is used for an itinerary, whereas

[3] [212, p. 89]) defines a stage (*marḥalaŧ*) as: "We have reckoned the marhala (stage) at six *farsakh*s or seven."



نقطنا فوق الهاء نقطة ✿ تاخذ من مكّة الى بطن مَرّ مرحلة ثم الى
عُسْفان مرحلة ثم الى خُلَيص وأَمَج مرحلة ثم الى الخيم؛ مرحلة ثم الى الجُحْفَة
مرحلة ثم الى الأَبْواء مرحلة ثم الى سُقْيَا بنى غِفار؛ مرحلة ثم الى العَرْج مرحلة

(a) Arabic ([213, p. 106])

You travel from Makka to Batn Marr, one stage;
thence to ʿUsfān one stage;
thence to Khulays and Amaj one stage;
thence to al-Khaym one stage;
thence to al-Juhfa one stage;
thence to al-Abwāʾ one stage;
thence to Suqyā Banī Ghifār one stage;

(b) English, translated by Basil Collins [212, p.89]

Figure 3.2: A short passage from al-Muqaddasī's book describing route sections and distances in the Peninsula of the Arabs (Jazīraŧ al-ᶜArab)

a route network is ramified throughout multiple centers that are connected by multiple routes or paths.

Travel literature, also called a travelogue, is a literary genre that specifically describes itineraries. In this form, authors explain their own travels in general as one or multiple itineraries. Beside the route information, distances, travel conditions, and other valuable geographical details, a travelogue includes any related information of events, experiences, and places, as well as cultural, social, and geographical features and phenomena. The eleventh-century author Abū Muᶜīn Nāṣir b. Ḫusraw b. al-Ḥārith al-Qubāḏiyānī al-Balḫī's travelogue ([215]) is an example of the description of itineraries, here while he explains his travels through Iran, Syria, Palestine, and Egypt. Nāṣir-i Ḫusraw introduces the routes and places one after another along his journey. He often characterizes the routes by distances, which provides all the essential parts of this geographical entity. However, his description does not focus on the geographical aspect of the area that he visits; rather, his book provides important information on routes and connections in the form of an itinerary (See a part of his journeyfrom Naysābūr to Fusṭāṭ in Figure 6.8. The visualization is described in Section 6.3.2).

3.3 Administrative Hierarchy

Systematic division on an empirical territory (of any type, but particularly imperial) is fundamental to administrative and hierarchical descriptions in geographical sources. In fact, administration forms a concept that could, to some extent, be inspired by the geography of the territory. Administrative divisions are in fact non-atomic spatial entities, which can be defined as higher level classifications of places as atomic entities.

Premising various levels of hierarchy, we would argue that a general model of description can cover arbitrary levels, including those cases where one basic level of hierarchy is demanded by a specific requirement; for example, the area covered by the source as a "macro-region." Administrative hierarchical descriptions explain categorization of places as the lowest-level geographical entities within container entities. The higher levels of divisions contain lower-level divisions under the name of a division. In other words, a set of places are grouped in micro-regions and these regions form possible macro-regions. Each division can have its own type (geographical, conceptual, or political) as a property. The regions together shape a country or province according to the source. Such descriptions provide a comprehensive pattern of hierarchical descriptions.

One example of a comprehensive description is al-Muqaddasī's book ([212, 213]). It starts with provinces, as the highest-level regions, which include smaller regions as subregions, which in turn include more subordinate regions or settlements. This pattern forms a hierarchical tree from the provinces down to the settlements in various levels. For example, in the following passages al-Muqaddasī explains how The Peninsula of the Arabs, as a highest-level region/province, is divided into four major subregions of a specific type, and that the first subregion, al-Ḥiğāz, contains its capital as well as other types of settlements ([212, p. 64]):

> This is the form of the Peninsula of the Arabs. We have divided this region into four extensive provinces, and four large districts. The provinces are al-Ḥiğāz, al-Yaman, ʻUmān, Hajar. The capital of al-Ḥiğāz is Makkaŧ; among its towns are Yaṯrib, Yanbuʻ, Qurḥ, Ḫaybar, al-Marwaŧ, al-Ḥawrā’, Juddaŧ, al-Ṭāʾif , al-Jār, al-Suqyā (Yazīd), al-ʻAwnīd, al-Juḥfaŧ, and al-ʻUšayraŧ: these are the larger towns. Lesser towns are Badr, Ḫulayṣ, Amaj, al-Ḥijr, Badā Yaʻqūb, al-Suwariqiyyā, al-Furʻ, al-Sayraŧ, Jabalaŧ, Mahāyi^{c}, Ḥāḏaŧ.

In historical studies, administrative units are of particular interest, for which [127] gives a few reasons. One reason could be that the evolution of administrations/institutions and their relationships are the main focus when studying institutional history. The history of units and how their boundaries changed are interesting topics when studying politics and military history; governmental records are the focus of many archivists. Another reason is that many historians focus on the information mostly gathered by administrative units, and the boundary of the specific unit throughout the relevant data is required for longitudinal analysis of that information.

3.4 Geographical Aspects of Biographical Data: Toponyms Associated with People

Biographical data and collections, one of the most prominent genres of classical Islamic literature, contain a wide range of valuable information. This information covers individuals and social, occupational, and religious groups, historical events, as well as locations and time periods. The geographical aspect, expressed in toponymic data, is of great importance to further our understanding of the social geography of the Islamic world. Studying this aspect of biographical collections sheds light on the major social transformations that the Islamic community underwent during the early history.

As [216] states, various collections cover geographical and chronological scope in different sizes and organizations by mentioning toponyms related to biographies. In fact, toponymic data mentioned in biographies implies a connection between any type of individuals to places of birth/death, places they visited or traveled to, places they lived, or places to which they relate in any way. As an example, al-Ḏahabī's (d. 748/1348 CE) *History of Islam* (*Taʾrīḫ al-islām*), a major biographical collection, significantly covers the geography of corresponding biographies ([217]). Its geographical scope includes almost 340 toponyms with frequencies of five and higher ([216, p. 35]). It specifically covers the major populated regions and emphasizes Lower and Upper Mesopotamia (al-^{c}Irāq and al-Jazīraŧ), Northeastern Iran (Ḫurāsān), Syria (al-Šām), Egypt (Miṣr), Spain (al-Andalus), and Northwestern Africa (al-Maġrib). Other collections, such as *Šaḏarāt al-ḏahab fī aḫbār man ḏahab* of Ibn ʻImād al-Ḥanbalī (d. 1089/1678 CE), has similar geographical coverage in relation to the biographies.

Toponymic data enables researchers to track the importance of particular places and changes over time. Typically, biographies mention multiple places that represent many-to-many relations to be considered in mapping and geographically contextualizing the biographies. As noted earlier, mentioning certain places in a biography implies an affiliation between the protagonist of the biography and those places. Therefore, from a geographical point of view, locations of any type are the geographical entities, which are associated with people, time, and other locations. The locations can, efficiently, be reviewed by putting them on a series of maps to draw the big picture, yet studying the locations individually is neither logical nor does it necessarily provide a broad view of this kind of information in biographical collections.

Having established these basic cases, we will now turn to the process of annotating, extracting, and modeling the geographical data of various types. We will represent how each type can be annotated, formatted using proper data structures as well as a few modeling and visualization approaches for the types of the data we explained in this chapter.

Chapter 4

Geographical Textual Descriptions: Annotation and Extraction

Identification and extraction of geographical information from historical texts enables innovative spatial analysis. In this chapter, we will discuss an exploratory approach for extracting geographical information from textual sources. This approach, introduced by [217], combines quantitative and qualitative analyses, or, to use more recent terminology, distant and close reading (see Figure 4.1). Put simply, the approach takes advantage of meaningful linguistic patterns in geographical descriptions and is applied to the geographical types described in Chapter 3. The wider objective is thus to develop a method for data extraction that could be applied to similar use cases.

In this chapter, we will take the approach first suggested by [217] and refine it to gain results for two specific use cases: the descriptions of administrative hierarchy and the descriptions of routes. With this approach we will reduce text in a natural language to a machine-readable abstraction which is then analyzed. The resulting abstractions can be represented and visualized for the analysis of the relations between spatial entities, which can offer a new perspective on already studied texts.

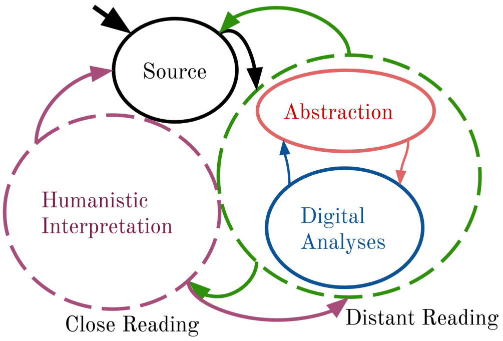


Figure 4.1: Iterative algorithmic analyses of textual resources ([218])

The approach that we use here is conceptualized in Figure 4.1 and, as can be seen, it is an iterative process, since it is impossible to resolve complexities of a natural language from the first attempt. The diagram makes clear that the steps are interconnected, reflecting the iterative nature of the entire process, through which the improvement of the process and results are ensured; that is, each iteration helps to improve the results. The process starts with working on a machine-readable version of a source (the back circle in Figure 4.1) with an initial humanistic understanding of the content and internal regularity from

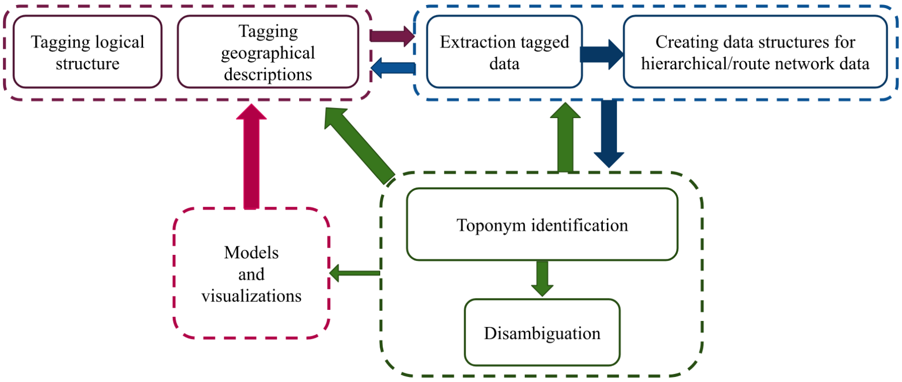

Tagging logical structure
Tagging geographical descriptions
Extraction tagged data
Creating data structures for hierarchical/route network data
Toponym identification
Disambiguation
Models and visualizations

Figure 4.2: Iterative process of analyzing geographical sources ([218])

which one can create an abstraction of the source. The creation of such an abstraction requires one to annotate the logical structure and the relevant information in the source, which will then facilitate automatic extraction of this information.

After annotation, one can proceed to extract the annotated data for transformation into a desired format, such as CSV (or TSV), and then represent the extracted data. In fact, the extracted data is an abstraction of the source to which digital analyses can be applied. The analyses can be used in humanistic interpretation of the source in close reading. Although the steps are explained in order, it does not necessarily force a strictly defined procedure following these ordered steps. As indicated in Figure 4.1, at each step, where enrichment of the earlier results is required, one can return to the previous step. In other words, this approach integrates close reading into the process of data extraction and analysis.

Figure 4.2 illustrates an example of this process applied to geographical sources that starts with tagging logical structures and geographical descriptions. The tagged data then can be extracted and converted into suitable data structures. The tagged data includes toponyms that need to be identified for further mapping and visualization. After the data is extracted from the textual source, we can employ computational approaches to match the toponyms against reference gazetteers. But, we always need to iteratively move back and forth between the steps to enrich the data by gathering more contextual information that may be required for the disambiguation of results (see [218]) for a detailed explanation of this process applied to two pre-modern geographical sources).

Table 4.1 shows how various types of geographical sources can be processed in a similar manner: starting with tagging of the required data, and then proceeding to extracting, representing the data with a proper data structure, and visualizing. For administrative hierarchies and route networks, we will explain a manual annotation approach to make use of meaningful linguistic patterns of geographical descriptions and access them through computational methods. For gazetteers and geographical data in biographical collections, We will discuss a semi-automatic process that relies on text-mining approaches. Regardless of the data extraction approach applied, we then take similar steps of data preparation and analysis, which are designed to be applicable to texts containing structurally similar descriptions.

	Administrative hierarchy	Routes	Toponyms associated with people	Maps	Gazetteer
Input	text	text	text	image	text (can be combined with geospatial data to generate maps)
Data extraction: OpenITI mARkdown	annotation	annotation	annotation	annotation	annotation
Data extraction: text mining approaches	–	–	text mining	–	text mining
Data representation	tree	network	table	tree, table, network	dictionary
Visualization	+	+	+	+	+

Table 4.1: Various geographical cases can be processed with a similar approach. This approach takes different types of inputs and includes similar steps of processing, while each step may use different techniques. Administrative hierarchy, routes, toponyms associated with people are the data types that live in texts, maps are visualization, and gazetteers are source texts that can be combined with geospatial data to generate maps. Gazetteers can include all three data types.

4.1 Annotation

Annotation is a method for providing meaningful indicators to textual information written in a natural language. The purpose of annotation is to facilitate automatic data extraction. Annotation can be "inline," where the data and annotation are stored together, or "external" (or, stand-off), which keeps the annotation separate from the text. In this section, we will focus on two major geographical cases and discuss an annotation approach that is specifically tailored so that right-to-left languages can tag geographical information in narrative sources.

4.1.1 Manual Annotation of Geographical Texts

Manual annotation is efficient for analytical tagging of semantic structures in relatively short texts. We apply manual annotation to the logical structure of a text as well as to hierarchical divisions and routes descriptions. As mentioned above, most efforts to annotate geographical data focus on tagging toponyms and references to geographical entities through manual or computational approaches, such as NER. An example of an annotation tool, which can be used to do manual and semi-automatic annotation, is Recogito[1], an online platform for annotating geographical texts and images. This approach, however effective, is still limited to toponyms as atomic geographical objects and only covers a limited subset of geographical concepts. Annotating complex structures requires substantial effort for constructing meaningful relations that can be manipulated by computational methods.

An appropriate vocabulary and standard is essential for annotating spatial narratives of pre-modern sources. However, so far, this has been missing from previous annotation attempts. Widely used standards like TEI XML or EpiDoc include basic schemes for named entities in general ([219]) and do not support annotation of semantic geographical relations and descriptions.

In addition to this gap, the complexity of working with left-to-right and right-to-left languages, paired symbols (such as angle brackets), and connected scripts in the same document, as [217] states, urges the use of a simpler and more efficient tagging scheme that simplifies the annotation process instead of adding more complications. This simplicity allows one to gain the best results as we aim at annotating almost all the corresponding material in the whole process.

[1] https://recogito.pelagios.org

A few attempts have been made toward semantic tagging. For example, [220] suggests a tagging method by engaging NLP and machine-learning approaches that take language-dependent characteristics into account. However, such approaches are limited to specific language-dependent patterns and are mainly introduced to be applied to modern sources.

In light of these limitations, the method proposed here offers a way to annotate geographical descriptions of administrative divisions and routes which, even though they provide similar information, may appear as individual structure and vocabulary. Here we do not focus on the existing computational approaches to annotate such texts, which represents another research perspective, rather, our way requires methods such as machine learning based on adequate amounts of training data on semantic patterns, where training models might be different for data from different textual sources. This approach has the distinct benefit that the outcome will not only provide a list of toponyms, which is the main focus in most annotation studies, rather it can offer more sophisticated geographical entities, such as a hierarchy of an empire or a trade network of an area.

OpenITI mARkdown ([221]) creates an easy-to-use and lightweight tagging scheme that consists of language-independent patterns built on regular expressions to assist the conversion of raw texts into machine-actionable format ([221]). It provides tagging options for a variety of logical, structural, and analytical patterns as well as recurrent and regular linguistic patterns, which is easily expandable. It also serves practical purposes through a bottom-up approach, where systematic patterns are tagged and classified without any intervention into the text. In addition, it can be converted to TEI XML for archival purposes. We utilize this approach to annotate descriptions of administrative hierarchies and route sections.

Annotation starts with the logical structure of the text; this is crucial in the whole process as it facilitates the ability to work with smaller units. The logical structure may also contain valuable information, and annotation makes them accessible for computational use, such as the major province names in the chapter headers in al-Muqaddasī's book. We will discuss this in more detail in the next section.

4.1.1.1 Administrative Hierarchy

Working with semi-structured textual descriptions that follow a fairly regular pattern is straightforward. Such descriptions are not always available for all historical contexts but they do exist in many languages. We illustrate how we can extract structured information from these natural language sources by capturing their semi-formal structure.

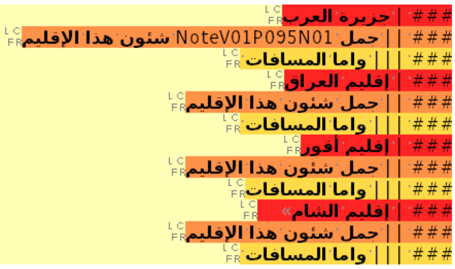
Here we take an example of comprehensive geography, al-Muqaddasī's book, that thoroughly describes the classical Islamic world, including each province and its geographical organization and route network. The annotation is designed to be used for other similar descriptions at any level of granularity. As mentioned in Section 3.3, al-Muqaddasī describes hierarchical divisions in detail and uses locally relevant terminology for major geographical units. He introduces the highest-level divisions in the logical structure of his book. Thus, tagging the logical structure gives an overview of the major provinces of the classical Islamic world as presented in this book.

Using OpenITI mARkdown, Figure 4.3 shows the headers of logical units of his text highlighted in distinct colors for headers of different levels (colors follow the rainbow spectrum). The current version is implemented in EditPad Pro[2], which supports custom highlighting and a navigation

[2] OpenITI mARkdown is not dependent on any particular editing environment. The current scheme implemented in EditPad Pro customizes the highlighting.

scheme. The red color corresponds to the chapters (tagged with ### | —level 1 heading), the orange color represents the sections included in a chapter (tagged with ### || —level 2 heading), and the yellow color points to the sub-subsection (tagged with ### ||| —level 3 heading). This method facilitates the arbitrary level of nested sections in the structure; for example, the subordinate section in the last section above (level 3) should be tagged with ### |||| (level 4) and so on.

The structure given in Figure 4.3 represents the list of the classical Islamic world provinces in the tenth century that al-Muqaddasī calls *iqlīm*, following the Greek *κλίμα*. We therefore tag the divisions as the following pattern in the scheme suggests:
WORLD: **PROVINCE** > **REGIONS** (of type) > **SETTLEMENTS** (of type).



```
### | جزيرة العرب
### || جمل NoteV01P095N01 شئون هذا الإقليم
### ||| واما المسافات
### | إقليم العراق
### || جمل شئون هذا الإقليم
### ||| واما المسافات
### | إقليم أقور
### || جمل شئون هذا الإقليم
### ||| واما المسافات
### | إقليم الشام«
### || جمل شئون هذا الإقليم
### ||| واما المسافات
```

Figure 4.3: Tagging logical units in al-Muqaddasī's book ([212, 213]). This figure is a folded view of a chapter, showing only a number of headings in the chapter that describes the Peninsula of the Arabs. The lines highlighted in red (tagged by ### |) are first-level section headings of different province, which are, in English, Jazīraṯ al-^{c}Arab (The Peninsula of the Arabs), iqlīm of al-^{c}Irāq (The region of Iraq), iqlīm of Aqūr (The region of Aqūr, and iqlīm of al-Šām (The region of al-Šām) in the order of appearance in the figure. The lines highlighted in orange (tagged by ### ||) are second-level (sub)section headings on general descriptions of each province, which are, in English, "A Summary Account of Conditions in this Region." The lines highlighted in yellow (tagged by ### |||) are third-level (sub-sub)section headings—descriptions of routes and distances within each province—which are, in English, "The Distances". Image source: [218].

In this pattern, the whole classical Islamic world is considered as WORLD, which is divided into PROVINCEs. Each PROVINCE consists of one or multiple subordinate regions of specific type(s). Each region may be divided into one or more subregions or may include a set of settlements of particular types. The order in the hierarchy of places and divisions is implied in a description of a region containing other regions or settlements. To clarify these implications, the scheme explicitly defines this order by means of predefined keywords. For example, REG1 (in the below patterns) indicates that the toponym is a region and is located at the first level. If this region is divided into subregions, those will then be tagged by REG2. TYPEs indicate the keywords that the author assigns to the toponyms while introducing them; these are preserved as in the source book. They also characterize the relations between the container and the included entities that can be used in data structures, such as trees, graphs, or more sophisticated structures like ontology, to preserve the type characteristics of toponyms. The patterns of this scheme for tagging divisions are listed below:

- #$#PROV **toponym** #$#TYPE ***type of region*** #$#REG1 (**toponym** #)+
- #$#REGX **toponym** #$#TYPE ***type of region*** #$#REGX (**toponym** #)+
- #$#REGX **toponym** #$#TYPE ***type of settlement*** #$#STTL (**toponym** #)+

For each piece of text that describes any type of divisions, we would choose an appropriate annotation pattern from the list above. Toponyms and types (of toponyms) in the above patterns are replaced with the names and terms from the source. In other words, terminology used in the source to categorize geographical entities is recorded as well (*toponym* and *type of region/settlement* in the above-mentioned patterns). Keywords, such as PROV, REG1, REG2, TYPE, STTL, act as indicators to produce information triples. Entities are connected through a TYPE, as OpenITI mARkdown offers, and all together they produce triples of SUBJECT > PREDICATE > OBJECT in the text. SUBJECT and OBJECT are the container and the contained toponyms, respectively, which are connected through a type (PREDICATE). As a result, we obtain a regulated system of indicating administrative hierarchies that can be manipulated computationally.

Figure 4.4 shows a fragment of al-Muqaddasī's book ([213, p. 68-69]) describing the divisions of the Peninsula of the Arabs tagged in the OpenITI mARkdown. The annotated lines are highlighted and each line holds both the original text, which appears immediately before this line, and between-the-lines annotation. For instance, the first highlighted line is the annotated version of the previous line where al-Muqaddasī talks about the major first-level divisions of the Peninsula of the Arabs, which are of two different types: *kūraŧ* (province) and *naḥiyaŧ* (district). Here is the corresponding passage from this book([212, p. 64]): "We have divided this region into four extensive provinces and four large districts. The provinces are al-Ḥijāz, al-Yaman, ᶜUmān, Hajar; the districts al-Aḥqāf, al-Ašḥār, al-Yamāmah, Qurḥ." The description give names of four *kūraŧ*s and four *naḥiyaŧ*s. The annotation of this information is as below:

> #$#PROV **Jazīraŧ al-ᶜArab** #$#TYPE ***kūraŧ*** #$#REG1 **al-Ḥijāz** # **al-Yaman** # **ᶜUmān** # **Hajar**

Next comes the districts (*naḥiyaŧ*s) that are first level regions and, similar to the provinces, we have tagged them using the REG1 keyword:

> #$#PROV **Jazīraŧ al-ᶜArab** #$#TYPE ***naḥiyaŧ*** #$#REG1 **al-Aḥqāf** # **al-Ašjār** # **al-Yamāmaŧ** # **Qurḥ**

In the next annotated lines, al-Muqaddasī mentions the capital (*qaṣabaŧ*) and the major cities of the al-Ḥijāz *kūraŧ*, respectively:

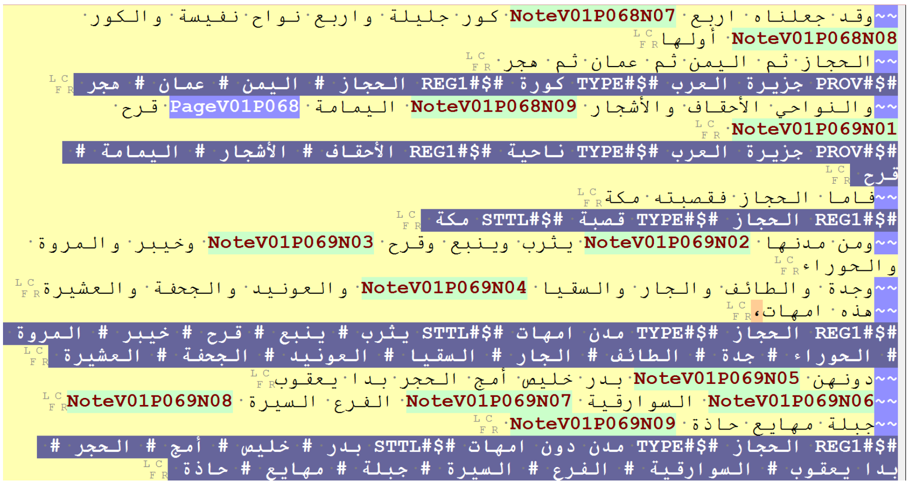
- #$#REG1 **al-Ḥijāz** #$#TYPE ***qaṣabaŧ*** #$#STTL **Makkaŧ**
- #$#REG1 **al-Ḥijāz** #$#TYPE ***mudun ummahāt*** #$#STTL **Yaṯrib** # **Yanbuᶜ** # **Qurḥ** # **Ḫaybar** # **al-Marwaŧ** # **al-Ḫawrā'** # **Juddaŧ** # **al-Ṭāᵓif** # **al-Jār** # **al-Suqyā (Yazīd)** # **al-ᶜAwnīd** # **al-Juḥfā** # **al-ᶜUšayraŧ**

Al-Muqaddasī then proceeds to list other minor cities and we have annotated the remainder in a similar manner. All these lines that we have added to the text are, in fact, individual data triples, and taken together they form the hierarchical division description.

4.1.1.2 Routes and Travelogues

Al-Muqaddasī gives a fairly regular and detailed description about how to get from one place to another. As with the hierarchical structures we can mine these routes as structured data. Similar to hierarchical data, the OpenITI mARkdown offers a pattern for annotating routes that connect settlements and have distance properties as shown in the annotation pattern below:

> #$#FROM **toponym** #$#TOWA ***toponym*** #$#DIST **distance as recorded**

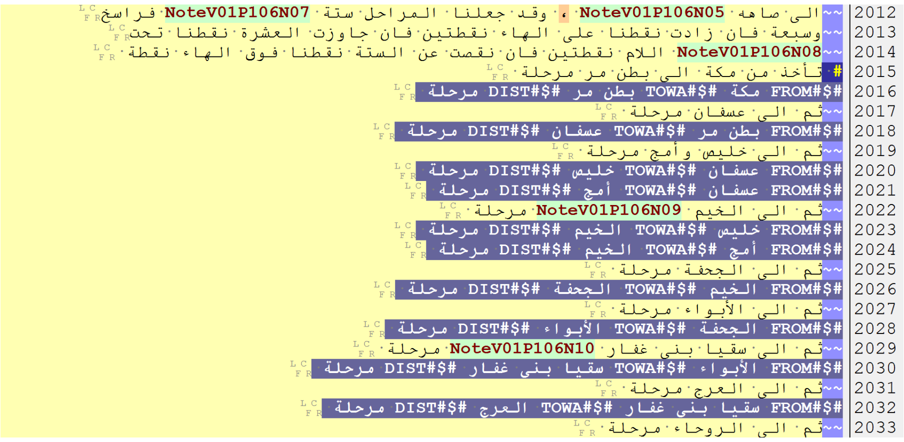


```
~~وقد جعلناه اربع NoteV01P068N07 كور جليلة واربع نواح نفيسة والكور
NoteV01P068N08 أولها
~~الحجاز ثم اليمن ثم عمان ثم هجر
#$#PROV جزيرة العرب #$#TYPE كورة #$#REG1 الحجاز # اليمن # عمان # هجر
~~والنواحي الأحقاف والأشجار NoteV01P068N09 اليمامة PageV01P068 قرح
NoteV01P069N01
#$#PROV جزيرة العرب #$#TYPE ناحية #$#REG1 الأحقاف # الأشجار # اليمامة #
قرح
~~فاما الحجاز فقصبته مكة
#$#REG1 الحجاز #$#TYPE قصبة #$#STTL مكة
~~ومن مدنها NoteV01P069N02 يثرب وينبع وقرح NoteV01P069N03 وخيبر والمروة
والحوراء
~~وجدة والطائف والجار والسقيا NoteV01P069N04 والعونيد والجحفة والعشيرة
~~هذه امهات،
#$#REG1 الحجاز #$#TYPE مدن امهات #$#STTL يثرب # ينبع # قرح # خيبر # المروة
# الحوراء # جدة # الطائف # الجار # السقيا # العونيد # الجحفة # العشيرة
~~دونهن NoteV01P069N05 بدر خليص أمج الحجر بدا يعقوب
~~NoteV01P069N06 السوارقية NoteV01P069N07 الفرع السيرة NoteV01P069N08
~~جبلة مهايع حاذة NoteV01P069N09
#$#REG1 الحجاز #$#TYPE مدن دون امهات #$#STTL بدر # خليص # أمج # الحجر #
بدا يعقوب # السوارقية # الفرع # السيرة # جبلة # مهايع # حاذة
```

Figure 4.4: Inline tagging of hierarchical data in al-Muqaddasī's book ([213, p. 68-69] and [212, p. 64]). For English translation of this passage see the cited paragraph in Section 3.3. Image source: [221].

This pattern also creates data triples, describing the start (FROM), the end (TOWA), and the length (DIST) of a route section, which records the value and units as mentioned in the source. The limited length of the keywords in the annotation patterns is intended to simplify the manipulation of the strings in the computations, since all the keywords will be treated as strings that need to be parsed. These pieces of information are the building blocks of a network or an itinerary. Figure 4.5 shows a few lines of al-Muqaddasī's book where he talks about the routes and distances in the Peninsula of the Arabs. Each highlighted line annotates an individual route section. For example, the first line introduces the route from Makkaŧ (Mecca) to Baṭn Marr and then, in the next line, from there to ᶜUsfān, with each taking one day of travel (*marḥalaŧ*). The annotation of these routes using the OpenITI mARkdown is given below:

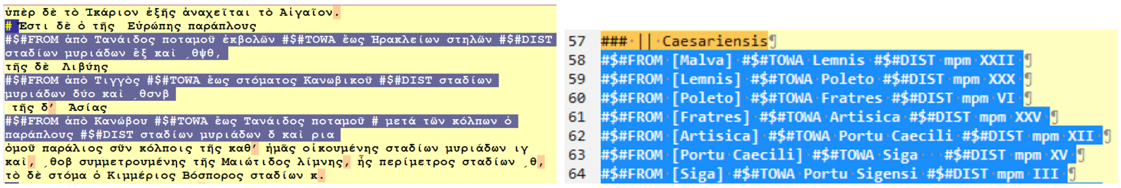
- #$#FROM **Makkaŧ** #$#TOWA **Baṭn Marr** #$#DIST *marḥalaŧ*
- #$#FROM **Baṭn Marr** #$#TOWA **ᶜUsfān** #$#DIST *marḥalaŧ*



```
~~الى صاهه NoteV01P106N05 ، وقد جعلنا المراحل ستة NoteV01P106N07 فراسخ
~~وسبعة فان زادت نقطنا على الهاء نقطتين فان جاوزت العشرة نقطنا تحت
~~NoteV01P106N08 اللام نقطتين فان نقصت عن الستة نقطنا فوق الهاء نقطة
# تأخذ من مكة الى بطن مر مرحلة
#$#FROM مكة #$#TOWA بطن مر #$#DIST مرحلة
~~ثم الى عسفان مرحلة
#$#FROM بطن مر #$#TOWA عسفان #$#DIST مرحلة
~~ثم الى خليص وأمج مرحلة
#$#FROM عسفان #$#TOWA خليص #$#DIST مرحلة
#$#FROM عسفان #$#TOWA أمج #$#DIST مرحلة
~~ثم الى الخيم NoteV01P106N09 مرحلة
#$#FROM خليص #$#TOWA الخيم #$#DIST مرحلة
#$#FROM أمج #$#TOWA الخيم #$#DIST مرحلة
~~ثم الى الجحفة مرحلة
#$#FROM الخيم #$#TOWA الجحفة #$#DIST مرحلة
~~ثم الى الأبواء مرحلة
#$#FROM الجحفة #$#TOWA الأبواء #$#DIST مرحلة
~~ثم الى سقيا بنى غفار NoteV01P106N10 مرحلة
#$#FROM الأبواء #$#TOWA سقيا بنى غفار #$#DIST مرحلة
~~ثم الى العرج مرحلة
#$#FROM سقيا بنى غفار #$#TOWA العرج #$#DIST مرحلة
~~ثم الى الروحاء مرحلة
```

Figure 4.5: Route sections with inline tagging in al-Muqaddasī's book using OpenITI mARkdown([213, p. 106] and [212, p. 89-90]). Image source: [221]. For English translation of the route sections, see Figure 3.2b.

In another example, Figure 4.6 shows the OpenITI mARkdown applied to two pre-modern



```
ὑπὲρ δὲ τὸ Ἰκάριον ἑξῆς ἀναχεῖται τὸ Αἰγαῖον.
# Ἔστι δὲ ὁ τῆς  Εὐρώπης παράπλους
#$#FROM ἀπὸ Τανάιδος ποταμοῦ ἐκβολῶν #$#TOWA ἕως Ἡρακλείων στηλῶν #$#DIST
σταδίων μυριάδων ἓξ καὶ ͵θψθ,
τῆς δὲ  Λιβύης
#$#FROM ἀπὸ Τιγγὸς #$#TOWA ἕως στόματος Κανωβικοῦ #$#DIST σταδίων
μυριάδων δύο καὶ ͵θσνβ
 τῆς δ’  Ἀσίας
#$#FROM ἀπὸ Κανώβου #$#TOWA ἕως Τανάιδος ποταμοῦ # μετά τῶν κόλπων ὁ
παράπλους #$#DIST σταδίων μυριάδων δ καὶ ρια
ὁμοῦ παράλιος σὺν κόλποις τῆς καθ’ ἡμᾶς οἰκουμένης σταδίων μυριάδων ιγ
καὶ, ͵θοβ συμμετρουμένης τῆς Μαιώτιδος λίμνης, ἧς περίμετρος σταδίων ͵θ,
τὸ δὲ στόμα ὁ Κιμμέριος Βόσπορος σταδίων κ.
```

(a) Greek source route sections ([222])

```
### || Caesariensis
#$#FROM [Malva] #$#TOWA Lemnis #$#DIST mpm XXII
#$#FROM [Lemnis] #$#TOWA Poleto #$#DIST mpm XXX
#$#FROM [Poleto] #$#TOWA Fratres #$#DIST mpm VI
#$#FROM [Fratres] #$#TOWA Artisica #$#DIST mpm XXV
#$#FROM [Artisica] #$#TOWA Portu Caecili #$#DIST mpm XII
#$#FROM [Portu Caecili] #$#TOWA Siga   #$#DIST mpm XV
#$#FROM [Siga] #$#TOWA Portu Sigensi #$#DIST mpm III
```

(b) Latin source itinerary ([214]). Image source: [218].

Figure 4.6: Using OpenITI mARkdown for annotating route sections in left-to-right languages. The mARkdown scheme is configured for left-to-right scripts in EditPad Pro.

Greek and Latin geographies. In these examples, the mARkdown scheme is configured for left-to-right scripts in EditPad Pro.The highlighted lines hold annotations, each representing individual sections. This sample represents an adaptation of the OpenITI mARkdown to other languages.

Another approach could be to keep the same keywords and add a fourth part to the triples that indicates the typology. For example, in the following patterns #$#ITIN and #$#NETW add type information to the route section:

- Network: #$#FROM **toponym** #$#TOWA **toponym** #$#DIST **distance** #$#ITIN
- Itinerary: #$#FROM **toponym** #$#TOWA **toponym** #$#DIST **distance** #$#NETW

The first tag set adds new information to the mARkdown without any intervention into the structure of annotations and triples, whereas the second adds more flexibility by adding various types without changing the existing keywords ([218]). The application of this typology might differ depending on the source and utilization of the existing information. For example, the geographical image of the Roman Empire in *Itinerarium Antonini* ([214]) is structured as a list of connections between places. In other words, the exhaustive representation of the geography happens through the description of the travel connections in a ramified pattern and, accordingly, we treat this itinerary as a network description. Furthermore, accurate information on the routes and distances is systematically given in alternative paths that encapsulate the travel routes and can be seen as micro-regions. These micro-regions are provided in macro-regions and provide a hierarchy of routes that suggest a hierarchy different from that which we derive from administrative descriptions ([223, 224, 218]). Therefore, tagging this information with hierarchical tag sets does not seem to be a suitable approach as it may construct a false picture. On the contrary, the above-mentioned patterns appear to be a suitable way to annotate this information, which might happen together with the route network data in geographical descriptions, to preserve the itineraries as macro-routes and also to better circumscribe an area. This circumscription, however when tagged as itineraries, can also be used as contextual information for toponym disambiguation ([218]).

4.1.2 Semi-Automatic Toponym Annotation

In this section, we will introduce a semi-automated process of toponym identification and disambiguation. The idea is to match all occurrences of toponyms in a text against a prepared list of toponyms in various forms (with all frequently occurring Arabic prefixes that combine different conjunctions and prepositions), and develop a process that offers manual and semi-automatic disambiguation as well as extrapolation models based on the disambiguated toponyms.

To do this we will employ gazetteer entries to identify and extract toponyms together with their contextual information; thus, in this process, the initial extracted data is a list of tokens that are matched against the gazetteer. We will then manually disambiguate matches to identify toponyms, classifying them manually as true or false (based on evaluation of the contextual information). Manual disambiguation provides data for implementing two extrapolation models of true and false toponyms. These extrapolation models use contextual information that occurs in disambiguated matches to extrapolate true or false matches on new data. Measurements in the extrapolation models rely on two features: frequencies of toponymic n-grams[3] and co-occurrences of neighboring tokens at specific positions relative to toponyms. True toponyms—either manually disambiguated or extrapolated—can then be tagged in the text, thus producing an annotated corpus.

Furthermore, we will use these models for defining context-driven features of a training set for a supervised machine learning (ML) model. In fact, we create training and test sets for ML models using the disambiguated data, which is produced either manually or by using the extrapolation models. The results of the test will show how reliable it could be for disambiguation purposes.

4.1.2.1 The Annotation Process

We will now explain the details of the above-mentioned annotation process. Given an initial list of gazetteer entries, we searched for all occurrences of toponyms, taking into account the Arabic morphological rules where nominals take several classes of clitics. From the list of Arabic clitics, the following can precede toponyms:

- Prepositions: *bi-* (by/with), *li-* (for, to)
- Conjunctions: *wa-* (and), *fa-* (and, then)

Combining these clitics and gazetteer toponyms, we generated a new list of tokens. This list involves the original form and possible variants of toponyms in combination with the clitics. For instance, for Baġdād, there are five entries: one entry for the original form, Baġdād, and four further entries for the combination of Baġdād and clitics as below:

- *bi* + Baġdād (ببغداد)
- *wa* + Baġdād (وبغداد)
- *wa* + *bi* + Baġdād (وببغداد)
- *li* + Baġdād (لبغداد)

The rationale is that when the clitics proceed a toponym, they form a single token and consequently they are recognized as a single token in the tokenization step. Since the gazetteer contains only the original form of toponyms, a matching process against the gazetteer entries will miss other frequent forms of toponyms (e.g., those in the Baġdād example above).

Adding the above-mentioned list to each gazetteer entry enlarges the initial list of toponyms to search for and enables one to identify most of the occurrences of a specific toponym. Moreover, it allows one to access the contextual information that occurs with all forms of toponyms, which is crucial as the context is the basis of premises in the extrapolation models. In other words, this approach is not based on string matching to identify the toponyms, which is widely covered within the existing literature, rather, it is based on the context that toponyms appear in and the words that often accompany the toponyms in a text. Hence, we extract all the toponyms in the

[3] [216, p. 76] gives examples of toponymic n-grams from four different sources.

above list together with the preceding and following trigrams. The positions of extracted tokens including toponyms are preserved to make them systematically accessible in the original text for tagging the true matches. The position of the matched toponym is calculated in the text and all the other surrounding tokens receive a relative position.

The identified toponyms are also automatically assigned a gazetteer URI, a name, and a link to the source text. This link in combination with the position of toponym in the text facilitates access to the toponym at the exact position where this match occurs. In addition, we define "status" and "score" fields for each match that specify the disambiguation status and the scores of the corresponding match that each extrapolation model calculates, respectively. Initially, we assign predefined values to these fields, which will be updated during the disambiguation process described later in this section. Below are the fields for a match in the dataset that we gather initially:

- Complete match: a phrase that holds the matched toponym together with the surrounding tokens. This means, we extract trigrams before and after the toponym and the toponym appears at the middle position of a phrase, punctuated by square brackets, as Yemen/al-Yaman in the below example.

  من مكة إلى [[اليمن]] فبلغ ذلك الحسن

  In English: "From Mecca to [[Yemen]], so reached that ...“

- Matched token: the matched toponym that is positioned in the middle of the complete match (punctuated with square brackets in the phrase above) and will be disambiguated.

- Surrounding tokens: each token in the four words before and after the matched token is preserved and specified with its relative position in the complete match.

- Matched position: the absolute position of the matched token (toponym) in the original text.

- Gazetteer URI and name: the URI and the name of the gazetteer entry matched to the token at the middle position.

- Link to the text to which the data belongs.

- Status and scores of extrapolation models: these values will be calculated by the extrapolation models. The status holds the disambiguation status—whether the match is disambiguated as a false or a true match—and the score shows a calculated numerical value that each mode yields.

After preparing the matches, we proceeded with the disambiguation process. The current implementation of the process brings the identified toponyms one by one together with the extracted context from the prepared dataset. It offers a set of predefined keywords for the status of the current match from which the user can choose and assign a true or false status to the match. The status is then saved and will be used in the extrapolation models.

It is also important to note that the disambiguation process is not fully manual as we have implemented a few semi-automatic elements. One idea is to automatically disambiguate similar cases. For instance, we use the "true always" status keyword to automatically set all the matches

من مكة إلى [[اليمن]] فبلغ ذلك الحسن

إلى [[اليمن]] فبلغ ذلك	مكة إلى [[اليمن]] فبلغ	من مكة إلى [[اليمن]]
إلى [[اليمن]] فبلغ ذلك الحسن	مكة إلى [[اليمن]] فبلغ ذلك	من مكة إلى [[اليمن]] فبلغ
[[اليمن]] فبلغ	مكة إلى [[اليمن]] فبلغ ذلك الحسن	من مكة إلى [[اليمن]] فبلغ ذلك
[[اليمن]] فبلغ ذلك	إلى [[اليمن]]	من مكة إلى [[اليمن]] فبلغ ذلك الحسن
[[اليمن]] فبلغ ذلك الحسن	إلى [[اليمن]] فبلغ	مكة إلى [[اليمن]]

Figure 4.7: Toponymic n-grams ($n = 2$ to $n = 7$) in the phrase at the top line (see the English translation in Section 4.1.2.1) that contain the toponym, which is the token in square brackets in the middle. For example, at the top right is 4-gram, which is 3-gram+toponym+0-gram; next down is 5-gram, which is 3-gram+toponym+1-gram; next down is 6-gram, which is 3-gram+toponym+2-gram. The toponyms will be replaced with a fixed string in the output list.

of a specific toponym that is always a toponym. In other words, all mentions of the tokens that are certainly toponyms in any context, such as Baġdād and its concatenated forms with clitics, will be automatically disambiguated and labeled as true. When the user disambiguates such a toponym for the first time and sets “true always” as the status value, this means that all occurrences of that token in the text are definitely toponyms. Therefore, the process will label all the matches of that toponym as “true” toponym and will exclude them from the disambiguation process.

Another way in which we have partially automated the disambiguation process is through the extrapolation models. That is, the models are developed to automatically analyze the contextual information of the matches and conclude the true or false status for each as the disambiguated data grows. In the following section, we will explain these to models.

4.1.2.2 Extrapolation Models

4.1.2.2.1 Frequency of Toponymic N-grams

Using the already disambiguated toponyms with their true or false status, the first extrapolation model is based upon the frequency of toponymic n-grams. It builds a dictionary of all toponymic n-grams from the disambiguated toponyms, which have true or false status value and thus can be called true and false toponyms, respectively. As an example, Figure 4.7 gives all the toponymic n-grams in the phrase at the top line of this table. This model creates all the possible n-grams in which the corresponding toponym is included. As in Algorithm 1, for each generated n-gram from the disambiguated matches (Line 4) it updates the frequency of times that the n-gram appears with true and false toponyms (Line 9) and calculates the precision value by dividing the frequency of cases that it occurs in true matches by its total frequency (Line 11). Each n-gram is thus assigned a precision value in the model, which offers a scoring system as a rubric for evaluating the ambiguous matches.

Having now scored each of the n-grams with a proportional value, we applied this model to the ambiguous matches—with initial status frequencies and score values of zero.The model receives the scored n-grams—generated in Algorithm 1—as an input. It then creates a dictionary of all toponymic n-grams from the ambiguous matches and checks them against the scored n-grams that are generated based on the disambiguated matches. If the n-gram is among the scored n-grams and has a precision value greater than 0.9, which means that in at least 90% of the cases this n-gram holds a toponym, then the corresponding toponym is evaluated as a true toponym. If the precision value is less than 0.1, which means that in less than 10% of the cases the n-gram appears with a toponym, the model extrapolates a false status for the toponym. For instance, if the phrase "he died in" (*tuwūffīya bi-*) is always followed by a toponym in the disambiguated cases, it means the next time that this phrase occurs in a sentence it will be followed by a toponym. In most of the cases, the toponyms are surrounded by particular verbs and prepositions, such as "lives in," "moved to," and "born in." The high frequency of specific n-grams increases the probability of occurrence of the same pattern, and vice versa. Those n-grams that have precision values greater than 0.1 and less than 0.9 will be tagged for further examinations.

Algorithm 1: This algorithm prepares the extrapolation model for disambiguation of toponyms based on the frequency of toponymic n-grams in the disambiguated data. For each n-gram, we count the number of times that it occurs with true and false toponyms. Then, dividing the frequency of the n-gram with true toponyms by its total frequency yields the precision value of each n-gram. The precision value will then be used to tag the undisambiguated toponyms as true or false matches.

```
function ngramExtrapolationModel ($disambiguatedData$)
Input  : The dictionary of disambiguated toponyms with corresponding data including
         the preceding and following trigrams.
Output: A dictionary of toponymic n-grams in the disambiguated matches, holding the
         frequency of their occurrences with true and false toponyms as well as the
         precision value.
$ngramsModelDict \leftarrow \{\}$
for $match$ in $disambiguatedData.items()$ do
    /* Generate the toponymic n-grams of the current disambigauted match.          */
    $ngramList \leftarrow generateAllNgrams(match)$
    for $n$ in $ngramList$ do
        if $n$ not in $ngramsModelDict$ then
            /* Initiate the n-gram key in the model dictionary.                      */
            $ngramsModelDict[n] \leftarrow ``true" : 0, ``false" : 0, ``total" : 0, ``prec" : 0$
        end
        /* Update the corresponding status (true or false) frequencies and precision values
           for the current n-gram in the model dictionary.                           */
        $ngramsModelDict[n][match[``status"]] += 1$
        $ngramsModelDict[n][``total"] += 1$
        $ngramsModelDict[n][``prec"] \leftarrow$
          $ngramsModelDict[``true"]/ngramsModelDict[``total"]$
    end
end
return $ngramsModelDict$                          // Return the model dictionary.
```

4.1.2.2.2 Co-occurrence Frequencies

The second extrapolation model is based on the co-occurrence frequencies of the same tokens at the same position relative to a toponym (see Algorithm 2). Following the idea of co-occurrence of similar surrounding tokens, this model focuses on the frequency of individual tokens that appear at a specific position in the neighborhood of toponyms. For instance, assume the token "in" happens x times at the position right before a toponym in a text, then the higher the value of x, the higher the probability of finding toponyms after this token.

Algorithm 2: This algorithm prepares the extrapolation model for disambiguation of toponyms based on the co-occurrence frequencies of tokens surrounding the toponyms. Each token at a specific position is assigned frequencies for the cases that occur with true or false toponyms and, accordingly, a ratio value by dividing the frequency of its occurrences with true toponyms by its total frequency. The ratio values are then used to extrapolate true or false status for undisambiguated matches.

```
function frequencyExtrapolationModel ($disambiguatedData$)
Input  : Dictionary of disambiguated matches with toponyms together with the
         preceding and following trigrams.
Output: A Dictionary of token-at-position that has frequency of occurrences in the true
         and false matches as well as the ratio of the keys that appear in true matches.
$freqModelDict \leftarrow \{\}$
for $match\ in\ disambiguatedData.items()$ do
    for $i\ in\ [0, 1, 2, 4, 5, 6]$ do
        /* match["string"] is an array of tokens of the complete match that contains the
           matched token together with the surrounding tokens.                          */
        $key \leftarrow match[``string"][i] + ``@" + i$
        if $key\ not\ in\ freqModelDict$ then
            /* Initiate the key of the current token@position in the model dictionary.  */
            $freqModelDict[key] \leftarrow ``true" : 0, ``false" : 0, ``total" : 0, ``ratio" : 0$
        end
        /* Update the corresponding status (true or false) frequencies and ratio values for
           the current token@position in the model dictionary.                          */
        $freqModelDict[key][match[``status"]] += 1$
        $freqModelDict[key][``total"] += 1$
        $freqModelDict[key][``ratio"] \leftarrow$
          $freqModelDict[key][``true"]/freqModelDict[key][``total"]$
    end
end
return $freqModelDict$                          // Return the model dictionary.
```

The model takes into account three positions before and after the matched token for each match and creates a dictionary of tokens at a specific position. This means that each match, such as the example in Figure 4.7, is an array of strings of length seven where the matched token is at the position 3. We then use the surrounding tokens at the positions $0, 1, 2, 4, 5$, and 6. Each token at these positions is treated separately. For example, if token "A" occurs immediately before a toponym in one match and immediately after a toponym in a different match, each will be an individual case, although the token is the same in both. In fact, the new position of the same token produces a different case and the occurrence of each token at a specific position is treated individually. Therefore, the combination of tokens and their position is unique. In the above example, we have taken into account token "A" at position 2 ("A@2") and token A at position 4 ("A@4").

The model creates a dictionary for each token at a specific position (Line 4) and, according to the frequency of occurrences of each key in the true and false matches, it then computes a ratio value for each (Line 7). For instance, if token "A") has appeared x times at position 2 in true matches and y times in false matches at the same position, the model will assign x and y values to "A@2"), as its "true" and "false" properties, respectively. The model then uses the true and false properties to computes the ratio of the true value x divided by the total value of true and false properties, $x + y$, and does the same calculation for all the token, such as token "A" at the position 3—"A@3"—or "B@1" and "C@2".

Having the dictionary of tokens for the disambiguated matches together with co-occurrence ratios, we next proceeded to create the same dictionary for the ambiguous matches. Tokens in ambiguous matches are treated according to their occurrence in the disambiguated matches. This means, if a key (i.e., token at a specific position) in the ambiguous matches exists in the disambiguated matches, it gets the existing ratio. The tokens in the ambiguous matches that do not exist in the disambiguated matches will have a ratio value of zero. Thus, each ambiguous match will have six ratio values. The average of the ratios, called the score, will then be used to disambiguate the ambiguous matches and extrapolate a true or false status for each. We assign false to the matches with the score in the first and second quartiles and true to the matches with the score in the forth quartile. The scores in the third quartile are marked for review.

4.1.2.2.3 A Supervised ML Approach

The heuristic approach that we explained in the previous subsection, produces a dataset for evaluation. As mentioned before, the dataset includes a list of matched toponyms, the surrounding tokens, and the disambiguation categories that the extrapolation models yield. The data that is shaped around each toponym can be seen as context-driven information for each match. The dataset is, in fact, a set of inputs mapped to outputs—$(input, output)$ pairs.

Furthermore, the $(input, output)$ pairs can act as a training set to learn a function that assigns new inputs to outputs ([225]), which are ambiguous matches and the true and false values, respectively. As [226] states, a supervised learning approach infers a function from a set of training examples in a labeled training set. It is best understood as approximating the inferred function f that maps input variables x to an output variable y by analyzing the training data, as shown in equation 4.1:

$$y = f(x). \tag{4.1}$$

Given a set of matches as examples, each labeled as a true or false match, we need a model that assigns true or false categories to unseen examples. This can be formulated as a classification problem using a Support Vector Machine (SVM) model ([227]). A classification problem seeks a model based on an existing set of examples. Each example is assigned a category. The model then places the new examples into one of the categories. The SVM model can be seen as a set of points in space so that their position forms clearly separated categories. Accordingly, the model will predict one of the categories for new points and correspondingly position them in space.

In this research example, there are two main categories of true and false matches and an SVM training algorithm builds a model to categorize the uncategorized matches. To apply SVM to our problem, a few steps need to be performed. Currently, there is a dataset, which is gathered through human experience (manual disambiguation) as well as measurements (the aforementioned extrapolation models). The dataset includes matches that are mapped to their disambiguation status, which the model will then use as inputs and outputs of a learned function, respectively. We therefore needed to determine the input feature representation of the learned function by transforming the input objects (matches) to feature vectors. Features involve representative information of the input objects. This representation is crucial since the model uses that to train the function. Consequently, it affects the accuracy of outputs that the function predicts.

We prepared the input features of the training set using the following information: part-of-speech (POS) of the matched token—at the middle position of a complete match and matched against the gazetteer entries—which has to be a noun; part-of-speech of the preceding token, which has to be a preposition; and the scores that the aforementioned extrapolation models provides. To determine the part-of-speech, we used the Stanford part-of-speech tagger[4] ([229]) and then represented the result tags of the token in question in binary values. Accordingly, if the token in the position of toponym is recognized as a noun, we assign 1 to this feature, otherwise, 0. Likewise, 1 and 0 specify the second feature, the part-of-speech of the token preceding the matched token. Value 1 shows that the preceding token is a preposition and 0 shows that the token is not a preposition. The third feature, as mentioned above, is the score measured by one of the extrapolation models. Similar to the part-of-speech values, we used binary values to represent the outputs and specify true or false categories of the matched toponyms by 1 and 0, respectively. The model then trains a function to distinguish between true and false matches according to the part-of-speech of the token in question and the preceding token as well as the measured score of the matches. In order to avoid an overfitting[5] problem, we hold out part of this prepared dataset from the training process and save it for testing. Therefore, we split the dataset into a training set for training the model and a test set for testing the trained model on the unseen samples. Without this process, the model would just repeat the labels of the samples that it has just seen and therefore obtain a perfect evaluation; however, attempts to predict correct categories on yet-unseen data might fail. The performance of predictions in the test then needs to be evaluated using various scoring metrics.

Having prepared the data (approximately $1,000,000$ records), in the first run we trained on 80% and tested on 20% of the data employing both linear and non-linear algorithms. Then, in the second run we decreased the training set size to 50% and ran a test on the other 50%. The results of each run are shown in Tables 4.2a and 4.2b, respectively. However, the results show more or

[4] In the time of implementing this idea, Stanford POS tagger was the best available tool to the best of our knowledge. [228] is another tool that was published shortly after this project.

[5] According to the Oxford dictionaries (`https://www.lexico.com/definition/overfitting`), in statistics, overfitting is defined as “The production of an analysis which corresponds too closely or exactly to a particular set of data, and may therefore fail to fit additional data or predict future observations reliably.”

less similar performance for both runs. This means the samples and the accuracy of the input features, even for a smaller portion of data, are enough to predict correct categories.

	linear	poly	RBF [230]	sigmoid
Accuracy	0.97	0.96	0.97	0.95
Precision(true category)	0.98	0.98	0.98	0.97
Precision(false category)	0.97	0.93	0.97	0.93
Recall(true category)	0.98	0.97	0.98	0.97
Recall(false category)	0.95	0.97	0.96	0.92
F-score(true category)	0.98	0.98	0.98	0.97
F-score(false category)	0.96	0.95	0.96	0.93

(a) Test size 0.2

	linear	poly	RBF	sigmoid
Accuracy	0.92	0.91	0.92	0.90
Precision(true category)	0.91	0.91	0.93	0.92
Precision(false category)	0.90	0.91	0.92	0.91
Recall(true category)	0.89	0.90	0.93	0.92
Recall(false category)	0.91	0.91	0.92	0.90
F-score(true category)	0.88	0.89	0.92	0.92
F-score(false category)	0.91	0.92	0.91	0.90

(b) Test size 0.5

Table 4.2: Evaluation of SVM model for toponym disambiguation using features of part-of-speech of the recognized token and its preceding token, and the score calculated to extrapolate true or false toponyms, as explained above.

The results in Tables 4.2, as mentioned above, use the part-of-speech of the matched token and that preceding it, and the score the second extrapolation method has measured based on the co-occurrence of the neighboring tokens. In order to test its effectiveness on the training process and the prediction performance of the function, we excluded the score from the training set. Similar to the previous test, we performed two runs: one trained on 80% of the data and tests on 20%; the other trained on 50% of the data and tests on 50%. According to the evaluation metrics of the runs in Tables 4.3, the performance, compared to the corresponding evaluations in Tables 4.2, declines. This shows that the score calculated to extrapolate true and false toponyms in the toponym disambiguation has a dramatic effect on training the model, while excluding the part-of-speech from the features list shows marginal changes in the performance, as can be seen in Tables 4.4.

The evaluation might still be overfitting on the test set since the evaluation parameters can be tweaked until the estimator performs optimally. Consequently, there is the risk of “leaking” knowledge about the test set into the model, in which case evaluation metrics would no longer report on generalization performance. One solution could be to hold out another part of the data for evaluation, the so-called “evaluation set,” so that the model is trained on the training set, evaluated on the evaluation set, and, if the evaluation shows successful results, the model will be tested on the test set. Partitioning data into three parts reduces the size of data on which the model can be trained. Therefore, we used cross-validation (CV) as an alternative validation strategy, in which there is no need to hold out the evaluation set and instead the training set is split into k partitions. Each time one partition is held out and a model is trained on the other $k-1$ partitions. The model is then evaluated on the hold-out partition and measurements give

	linear	poly	RBF	sigmoid
Accuracy	0.70	0.71	0.97	0.71
Precision(true category)	0.68	0.71	0.98	0.69
Precision(false category)	0.71	0.71	0.97	0.8
Recall(true category)	0.88	0.88	0.98	0.92
Recall(false category)	0.15	0.13	0.96	0.12
F-score(true category)	0.80	0.80	0.98	0.8
F-score(false category)	0.24	0.25	0.96	0.22

(a) Test size 0.2

	linear	poly	RBF	sigmoid
Accuracy	0.66	0.66	0.92	0.65
Precision(true category)	0.63	0.65	0.93	0.77
Precision(false category)	0.69	0.70	0.92	0.79
Recall(true category)	0.85	0.88	0.93	0.9
Recall(false category)	0.13	0.15	0.92	0.12
F-score(true category)	0.78	0.82	0.92	0.8
F-score(false category)	0.22	0.22	0.91	0.21

(b) Test size 0.5

Table 4.3: Evaluation of SVM model for toponym disambiguation using features of part-of-speech of the recognized token and its preceding token. This evaluation excludes the score calculated to extrapolate true or false toponyms, explained above, and test the effect of the score on the performance of the trained model.

	linear	poly	RBF	sigmoid
Accuracy	0.94	0.96	0.96	0.95
Precision(true category)	0.96	0.97	0.96	0.97
Precision(false category)	0.95	0.92	0.95	0.96
Recall(true category)	0.97	0.95	0.97	0.97
Recall(false category)	0.92	0.94	0.92	0.95
F-score(true category)	0.97	0.96	0.96	0.97
F-score(false category)	0.94	0.94	0.93	0.95

(a) Test size 0.2

	linear	poly	RBF	sigmoid
Accuracy	0.9	0.91	0.91	0.9
Precision (true category)	0.92	0.92	0.92	0.91
Precision (false category)	0.93	0.91	0.91	0.91
Recall (true category)	0.92	0.93	0.92	0.92
Recall (false category)	0.91	0.91	0.9	0.9
F-score (true category)	0.94	0.92	0.92	0.91
F-score (false category)	0.91	0.91	0.9	0.92

(b) Test size 0.5

Table 4.4: Evaluation of SVM model for toponym disambiguation using feature of the score calculated to extrapolate true or false toponyms. This evaluation excludes the POS of the recognized token and the preceding one to compare the effect of those features with the score.

Accuracy	0.97	0.97	0.97	0.97	0.97
Precision	0.96	0.97	0.97	0.97	0.97
Recall	0.96	0.97	0.97	0.96	0.96
F-score	0.96	0.97	0.97	0.97	0.96
Accuracy, using shuffle & split	0.97	0.97	0.97	0.97	0.97

Table 4.5: Cross-validation (CV): evaluating estimator performance of SVM model for toponym disambiguation using k-fold approach with 5 folds. Each column shows the score of the k_i fold as the test set.

Precision	Recall	F-score
0.96	0.97	0.96

Table 4.6: Performance metrics of a trained model of toponym recognition using statistical neural network

a list of k values that are the average of the values computed in the loop. Although there are various CV approaches, generally the same principle is followed in all.

Table 4.5 shows the evaluation measurements that we performed on the dataset. In this evaluation, we set $k = 5$, which splits the data into five partitions to generate possible (train, test) pairs and gives five average values for each pair in an evaluation round. Moreover, as an alternative to the basic partitioning approach, we evaluated the model using the shuffle and split approach[6], which iteratively produces random samples of the dataset for the training set and a test set of an arbitrary size. Using this approach gives more control over the number of iterations as well as the size of examples on each of the train/test split by randomly permuting the partitions of data. Therefore, the number of independent training/test sets can be manually defined and examples are first shuffled and then split into a (train, test) pair. The accuracy of this approach tested on the dataset is shown in the last row in Table 4.5.

As an alternative machine-learning algorithm to recognize toponyms, we performed a training process without predefined features, where instead, the model trains itself which toponyms are mentioned in a text and where. The model takes the texts in which the toponyms exist as well as the position of the toponyms and learns how to recognize the toponyms based on the context in which they are mentioned. To perform this algorithm, we employed a *SpaCy* library that uses statistical neural networks for various NLP applications, including NER. The samples that provide training and test sets are then prepared only by specifying the position of toponyms in each sample and there is no need to define features. As a result, we will have a list of sentences or passages together with the starting and ending positions of characters of toponyms. The difference to the previous prediction tasks that we discussed above is that here the problem is not formulated as a classification problem. Instead, we give the machine a list of disambiguated toponyms in passages to learn and predict the ambiguous toponyms in the rest of the dataset. Table 4.6 shows the performance of a language model trained on 80% and tested on 20% of data. This approach can be used as a way to recognize toponyms in a text that we have not been able to extract through gazetteer matching. In other words, the toponyms that we match against gazetteer entries and the disambiguation process produces a dataset by which we can train a model not only to extract more matches against the gazetteer entries, also to efficiently recognize the toponyms that are not in the gazetteer.

[6] An example implementation can be found at `https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.ShuffleSplit.html`

4.1.2.3 Summary

In this section we discussed approaches to identify toponyms in a text and to disambiguate the results. The approaches yield a list of disambiguated toponyms, which are identified through their position in the text and the URI of the corresponding text. This list together with the position of tokens in the text enables us to automatically access the toponyms, tag those occurrences in the text, and produce an annotated corpus. It also means that the first attempt of toponym matching and disambiguation produces a list of true toponyms that will be extended during the further iterations that use new texts. In other words, the process, in addition to the toponym identification and disambiguation, facilitates the tagging functionality of toponyms in the text.

4.2 Data Extraction and Structures

As mentioned before, annotating the information facilitates automatic data extraction and makes data usable in computational processes. After annotating the hierarchical descriptions and route network (Sections 4.1.1.1 and 4.1.1.2), we develop *python* scripts for extracting annotated information and creation of data structures. We use regular expressions to extract toponyms and complex entities in triples of data produced by annotating semantic patterns, such as route sections.

4.2.1 Administrative Hierarchy

Administrative hierarchy is characterized as a tree structure as explained before. The whole area that is described in a source can be the root of a tree, which includes all the regions and settlements. The area can be divided into multiple regions and each region can be divided into subregions. Multi-level divisions of this area constitute the tree's inner nodes, which can have other nodes and settlements as the tree leaves. Two consecutive levels of the tree represent a set of divisions as parent-child relationships. As described in Section 4.1.1.1, we annotate administrative divisions as triples of:

SUBJECT > PREDICATE > OBJECT.

Each triple represents a relationship between a subject and an object, and in the tree it will be represented as a parent-child relationship of two nodes at different levels. All the triples will then shape the whole area that is described according to the following pattern, as noted before:

WORLD: PROVINCE > TYPE > (REGION) > TYPE > SETTLEMENT

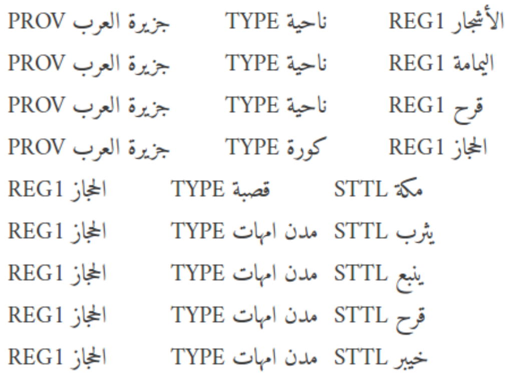
After extracting the annotated information, we construct the tree structure from a list of triples in a TSV (tab-separated values) file (see Figure 4.8), holding a multi-layer relation of the entities. Each triple describes individual relations between two regions or a region and a settlement: an upper level region of any type (of division) contains a lower level region or settlement. For instance, Figure 1.2 demonstrates a part of the tree structure that the few lines of data in Figure 4.8 represents. This tree shows a part of the province called Jazīraṯ al-ᶜArab and together with the rest of hierarchical information will shape the complete tree of this province.

This structure gives us the flexibility of expressing arbitrary levels of hierarchical descriptions. The tag set that we use in the annotation process (explained in Section 4.1.1.1, also shown in Figure 4.8) helps to follow the right order of toponyms when building the tree. We take the list of triple as input and initialize the tree by defining a root node, which is the whole area of description. To create the rest of the tree from the pieces of information in the triples, we gather all the highest-level nodes that exist in the dataset, which are provinces (tagged by PROV in Figure 4.8) in our dataset. Each province is then added to the tree as a child of the root. To add the descendants of each province, we extract the triples of data that form a path from this province to the constituent subregions and finally to the leaves, in a recursive procedure. Algorithm 3 shows this procedure that recursively builds the paths from a province to the lowest level settlements and writes each traversed path to the output file as a new line (Line 17). Therefore, each line in the output file will represent a path from a province to one of it is settlements. At the end, we will have a CSV file in which all the lines that start with a common root, constitute a macro-region (or province in our data) in the tree—as a sub-tree in the tree of the world. A sample line for the example in Figure 1.2 could be:

Fārs, Iṣṭaḫr, Darābağird

All the sub-trees together form the main tree. The output file can also be written into various formats, such as JSON, and then be converted to any proper format for visualizations. Algorithm 4 shows how we can generate a tree structure that includes all the provinces by using the recursive function in Algorithm 3.

Figure 4.9a illustrates the administrative divisions explained in al-Muqaddasī's book and how the descriptions nicely shape a tree. Figures 4.9b and 4.9c depict other visualizations of the first two levels of this structure. The former shows all provinces in the inner slices together with the corresponding subregions for each province in the outer slices. The latter is the zoomed out view of al-ᶜIrāq province with subregions in the inner slices and next level toponyms in the outer slices.



PROV جزيرة العرب	TYPE ناحية	REG1 الأشجار
PROV جزيرة العرب	TYPE ناحية	REG1 اليمامة
PROV جزيرة العرب	TYPE ناحية	REG1 قرح
PROV جزيرة العرب	TYPE كورة	REG1 الحجاز
REG1 الحجاز	TYPE قصبة	STTL مكة
REG1 الحجاز	TYPE مدن امهات	STTL يثرب
REG1 الحجاز	TYPE مدن امهات	STTL ينبع
REG1 الحجاز	TYPE مدن امهات	STTL قرح
REG1 الحجاز	TYPE مدن امهات	STTL خيبر

Figure 4.8: Examples of triples of data extracted from the annotated text in TSV format, showing the administration hierarchical data. The data is extracted from the annotated examples in Section 4.1.1.1. The first line in this figure read: PROV[INCE] Jazīraŧ al-ᶜArab (the peninsula of the Arabs) has a subregion of TYPE *nāḥiyaŧ* (district) REG[ION]1, which is called al-Ašjār. The transliterated names are available in the example given in Section 4.1.1.1 and in Figure 1.2.

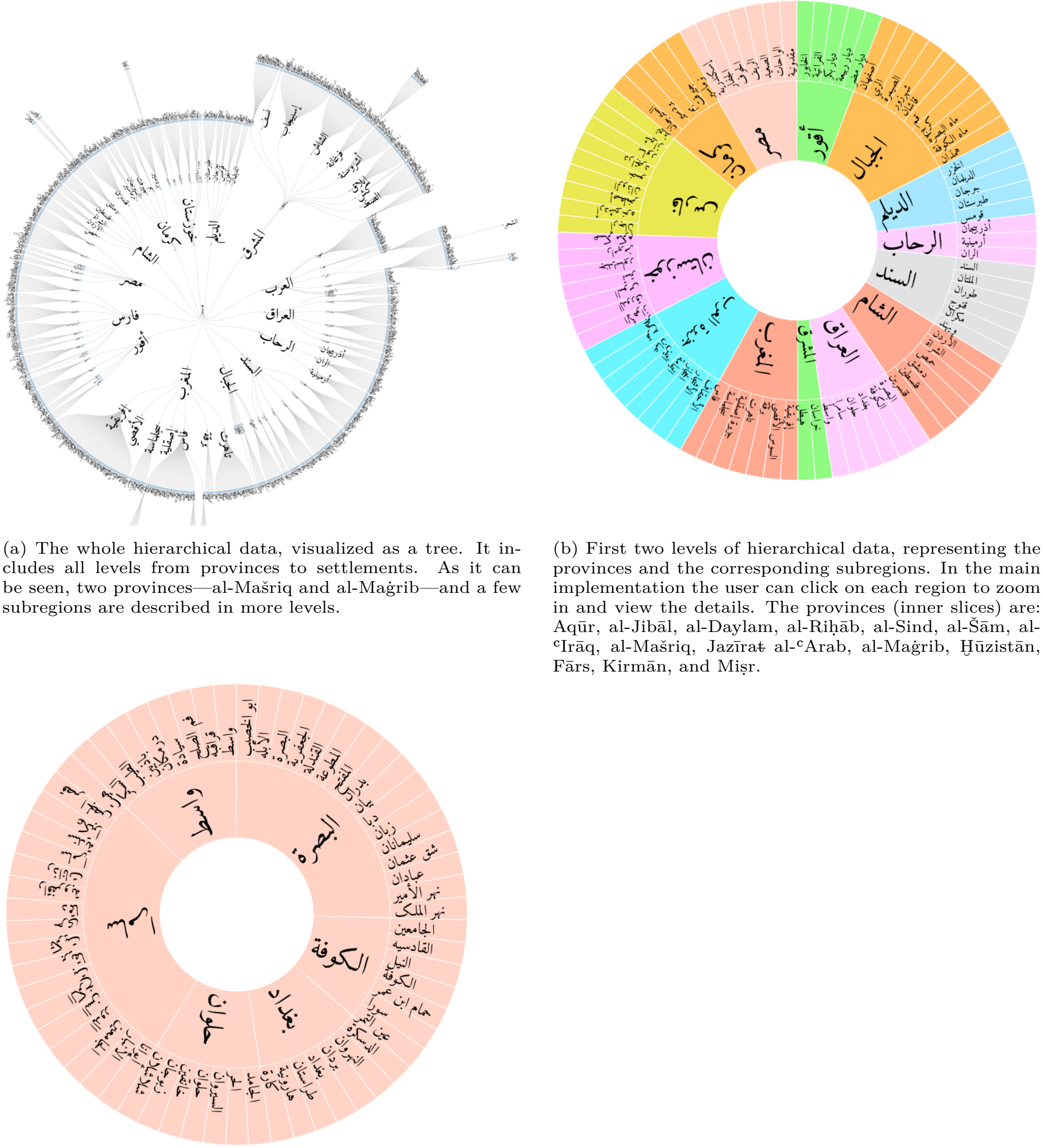
To represent the hierarchy of places and regions in common formats, such as GeoJSON ([231]), we consider two possible approaches as follows. One approach is to treat regions as individual objects separated from the contained places object, such as settlements. In other words, one can represent regions with structures such as polygons and multi-polygons that include sub-ordinate regions or settlements (see below for an example of a multipolygon *feature* in GeoJSON format, used to represent a region, as a container object):

```
{ "type": "Feature",
   "geometry": {
       "type": "MultiPolygon",
       "coordinates": [ [[[102.0, 2.0], [103.0, 2.0], [103.0, 3.0], [102.0, 3.0], [102.0, 2.0]]], [...], ...]
   },
   "properties": {
       "prop0": "value0", "prop1": {"this": "that"}
   }
}
```

Algorithm 3: This algorithm recursively generates a tree structure of a province. The province is the root of the tree and settlements are the leaves. The algorithm writes each traversed path in the tree to an output file so that each line shows a set of toponyms starting with a province and ending with a settlement. The available macro- and micro-regions are then placed between the province and the settlement. Types toponym are also preserved as edge labels in the tree, which then can be used in visualizations and computations.

```
function makeProvGraph ($graph, hierTriples, nodeLevel, nodeTraverse, pathsFile$)
Input  : The graph to which the nodes and edges will be added; list of hierarchical
         triples; current node level in the tree which is set to 2 in the first call of the
         function and increases by 1 in later calls; an array holding the current node and
         will be updated by adding its descendants; path to the file in which we save all
         the paths from the province to settlements.
Output: A network graph as a tree structure. The traversed paths in the hierarchical
         data will also be written to an output file.
$graph \leftarrow directedGraph()$
for $triple$ in $hierTriples$ do
    // Each triple has upper and lower level node, and lower level node label.
    $tri \leftarrow [upperLevelNode, lowerLevelNodeLabel, lowerLevelNode]$
    /* Search for triples of the current node                                  */
    if $tri[0]$ startswith $currentNodeLabel$ then
        /* $nodeLevel$ is a global variable to count the levels in the tree of all provinces.
           In the first call of this function for each province, this value is set to 2 (see
           Algorithm 4) and in each recursive call it increases by 1.            */
        $newNodeLevel \leftarrow nodeLevel$
        $graph.add\_node(newNodeLevel, label \leftarrow tri[2])$
        /* Generate an edge between the current (parent) node and new node.      */
        $graph.add\_edge(currentNodeID, newNodeLevel, label \leftarrow tri[1])$
        $nodeLevel \leftarrow nodeLevel + 1$
        /* If the current triple gives subregions, add it to the traversed path and
           recursively find the descendants for the current node. Then, remove it from the
           traversed path.                                                        */
        if $newNodeLable$ is not $STTL$ then
            $nodeTraverse.append$(the new node and its label)
            $makeProvGraph(graph, hierTriples, newNodeLevel, nodeTraverse, pathsFile)$

            $nodeTraverse.pop$(the new node and its label)
        end
        /* If current triple ends with a settlement, add the it to the list of descendants
           and add the current path from the province to the current settlement to the
           $outFile$.                                                             */
        else
            $nodeTraverse.append$(the settlement and its label)
            append $nodeTraverse$ to $outFile$
            $nodeTraverse.pop$(the settlement and its label)
        end
    end
end
return $graph$
```



(a) The whole hierarchical data, visualized as a tree. It includes all levels from provinces to settlements. As it can be seen, two provinces—al-Mašriq and al-Maġrib—and a few subregions are described in more levels.

(b) First two levels of hierarchical data, representing the provinces and the corresponding subregions. In the main implementation the user can click on each region to zoom in and view the details. The provinces (inner slices) are: Aqūr, al-Jibāl, al-Daylam, al-Riḥāb, al-Sind, al-Šām, al-ᶜIrāq, al-Mašriq, Jazīraŧ al-ᶜArab, al-Maġrib, Ḫūzistān, Fārs, Kirmān, and Miṣr.

(c) Zoomed view of al-ᶜIrāq province together with subordinate regions and settlements. The inner slices (subregions) are: al-Baṣraŧ, al-Kūfaŧ, Baġdād, Ḥulwān, Sāmarrāᵓ, and Wāsiṭ.

Figure 4.9: Hierarchical data visualization as described in al-Muqaddasī's book. In all visualizations, the inner slices show the higher levels of the hierarchy. The data is according to the parts that we have tagged and extracted.

Algorithm 4: This algorithm generates a tree of all the provinces (with a sub-graph for each province), using Algorithm 3

```
function buildHierarchicalTree ($hierTriples, pathsFile$)
Input  : Hierarchical triples and path to the file to which we write all the paths from
         provinces to settlements.
Output: Hierarchical tree of all provinces as a graph structure.
/* Generate the list of provinces.                                              */
$provinces \leftarrow getSetOfProvinces(hierTriples, "PROV")$
for $prov$ in $provinces$ do
    /* Initialize a directed graph for the current province.                    */
    $graph \leftarrow directedGraph()$
    /* Add the root node with id 1 and label $prov$                             */
    $graph.add\_node(1, label \leftarrow prov)$
    /* Global variable for node count starting at 2 as 1 is the root node.      */
    $nodeLevel \leftarrow 2$
    $graph.add\_node(nodeLevel, label \leftarrow prov)$
    /* Initialize the list of traversed path for the graph of this province by adding the
       $prov$. This list, which will be populated in $makeProvGraph$, holds all the paths from
       the root to leaves.                                                      */
    $provTraverse \leftarrow []$
    $provTraverse.append(prov)$
    $graph \leftarrow makeProvGraph(graph, nodeLevel, hierTriples, provTraverse, pathsFile)$
    write $graph$ to a file;
end
```

This approach, despite its applicability, adds complexities to the design, data management, modeling process, and the changes in the data. Moreover, it will make data unnecessarily voluminous and redundant. A simpler way is to represent all toponyms of various types (e.g. region, settlement) with the same object and interconnect them to represent the hierarchy. More clearly, we do not introduce new structures for container toponyms (i.e. regions) and use GeoJSON points as for settlements and use a property to connect settlements to the corresponding container toponyms. This is effective to introduce arbitrary levels of hierarchies without any need to define container objects for regions. Besides, it simplifies the process of applying changes in the hierarchical order and divisions. As an example, to add a new subregion at a specific level of the hierarchy, we just define a new object with a connection to the parent and set it as the parent of the containing regions and settlements. Accordingly, we can extract a specific level of hierarchy where multiple levels are available and construct regions and subregions in a simple process without pre-defining and storing them as separate data.

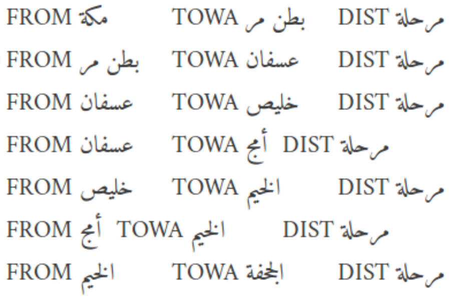
4.2.2 Routes and Distances

Route section descriptions have a similar triple structure of data: each explains a route of a specific length between two settlements. A network of routes can be represented by a graph data structure in which, unlike a tree structure, there is no ascending/descending order of the nodes. Hence, a list of route sections, as extracted from the annotated source, is adequate to build a network graph of routes. The network graph is, in fact, a set of nodes that are connected by edges, which can have properties, such as length, direction, name.

Figure 4.10 shows a few lines of the data file that is extracted from the tagged text of al-

Muqaddasī's book. Each line holds information on a single route section: source, destination, and distance, which provides all the required data to build a graph. In other words, each line represents an edge in the network graph and the places (sources and destinations) are the network nodes. Nodes (settlements) in a route network, unlike hierarchical data, are not connected only to their parents in a single connection, rather to multiple settlements to shape the neighborhoods and connectivity of the sites according to the original descriptions. The graph of routes is then created by adding each pair of nodes and the connecting edge. This algorithm can use any network creation package to build the graph by adding edges and nodes. The current implementation uses *NetworkX*[7] package that is developed in *python* for creating, manipulating, and analyzing the networks.

We use CSV for both input and output data formats while other common formats such as JSON, are easily adaptable. The same approach is applicable to form the itineraries since an itinerary is a linear path through a network graph. Hence, the principles of creating and manipulating are the same for both cases.



```
FROM مكة	TOWA بطن مر	DIST مرحلة
FROM بطن مر	TOWA عسفان	DIST مرحلة
FROM عسفان	TOWA خليص	DIST مرحلة
FROM عسفان	TOWA أمج	DIST مرحلة
FROM خليص	TOWA الخيم	DIST مرحلة
FROM أمج	TOWA الخيم	DIST مرحلة
FROM الخيم	TOWA الجحفة	DIST مرحلة
```

Figure 4.10: Examples of triples of tagged route sections extracted from al-Muqaddasī's book in TSV format. For example, the top line reads: FROM Makkat (Mecca) TOWA[RD] Baṭn Marr DIST[ANCE] is one-day-of-travel (*marḥalat*)

[7] https://networkx.github.io

Chapter 5

Modeling Geographical Data

In Chapters 3 and 4, we have established categories for geographical data and outlined approaches to data annotation and extraction and argued how one can apply those approaches to similar cases. The data extracted from narrative texts represent an abstract description of detailed explanations contained in the sources. These explanations are also represented in some pre-modern maps to illustrate regions, settlements, and other geographical entities. Figure 5.1 is an eleventh-century map from Maḥmūd al-Kašġarī in his book *Divânu Lügati't-Türk* in 1074 CE with North on the left and south on the right. He drew a circular map that has all the regions of the Turkish tribes from Europe to Chin as well as regions around the Persian Gulf. To better understand the pre-modern representations, we use the data that we extract from the pre-modern descriptions and redefine entities like regions according to our modern understanding of the region.

Having offered a detailed explanation of the process we used to extract this data, we will now introduce the computational approaches that provide the basis upon which to design and implement models for geospatial analysis and visualizations. For instance, (1) estimating the actual territory of an admin district from the point-level info about towns within it, and (2) estimating geographical coordinates for the lost places or the places that do not exist anymore.

In so doing we will build upon the overview of geographical types given in Chapter 3, where we argued that administrative hierarchies and route networks are higher level entities built upon atomic entities (settlements). The implementations and visualizations in this chapter use the test dataset derived from Cornu's *Atlas du monde arabo-islamique à l'époque classique: IXe–Xe siècles* ([130]) and, partly, al-Muqaddasī's book ([212]). Despite the specificity of this use case, the approaches offered here are not limited to any geographical space or language and can be applied to other datasets that supply similar types of data. Furthermore, the visualizations have been prepared in order to provide a visual insight into the relevant model using a test dataset and are enriched with a geographical map view.

In the first part of this chapter, we will propose models for representing, visualizing, and comparing complex geographical entities such as administrative districts and route networks. The first section introduces models for hierarchical data and evaluates them by considering advantages and disadvantages as well as discussing possible improvements and modifications. The second section offers approaches for comparing hierarchical data and route networks described in various sources. Additionally, in the final section, we will introduce a model to estimate geographical coordinates for pre-modern places whose geographical location is unknown for any reason (e.g., places that do not exist anymore).

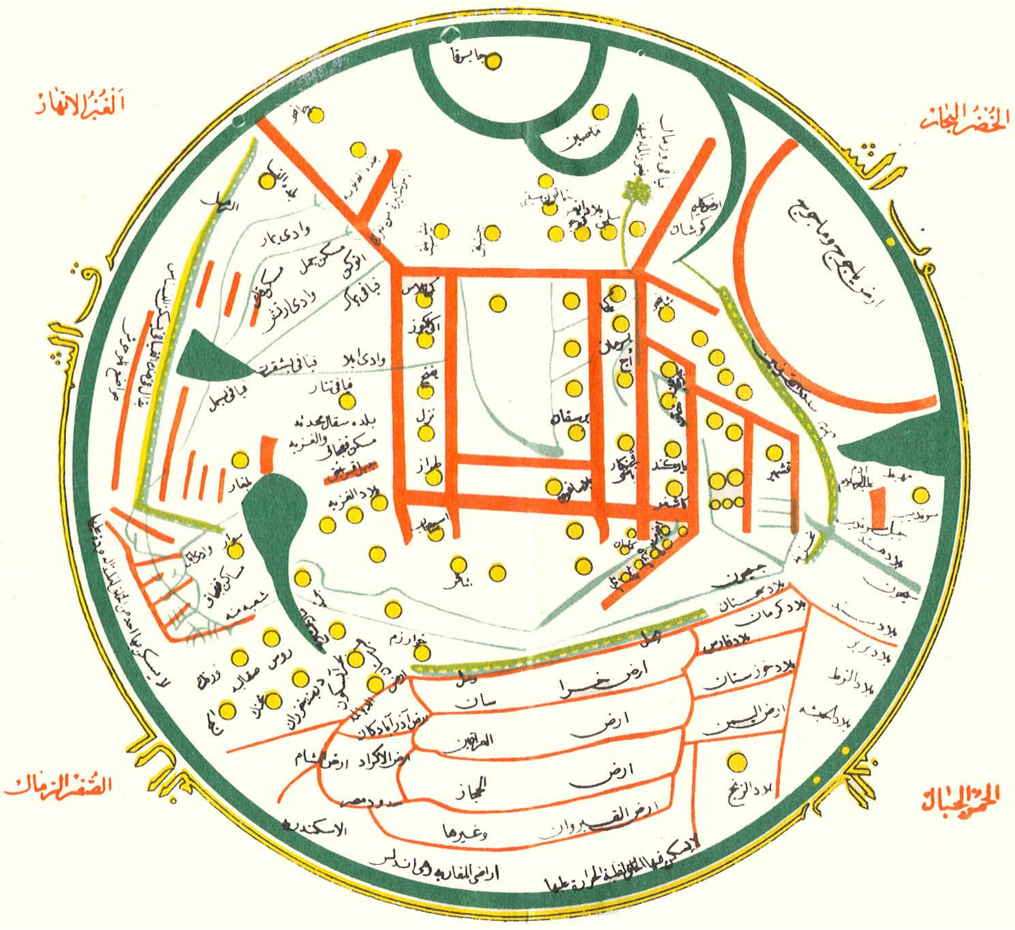


Figure 5.1: Maḥmūd al-Kašġarī's map of the world (*Divânu Lügati't-Türk*, translated into Turkish by Rifat Bilge, Istanbul 1917, 3 vols). Image source: [232].

5.1 Mathematical Models for Administrative Hierarchies

Administrative divisions and hierarchies have been variously described in different sources and different eras, because changes and developments in geography, space divisions, and administration occur in pre-modern societies over time. The models in this section allow us to investigate various aspects of regional, historical, and geographical topics, such as:

- The shape, geographical positions, and spatial extension of administrative hierarchies and divisions in the (Islamic) empire during different historical periods
- Geographical areas and locations that divisions cover
- Spatial and geographical changes over time
- Geographical interconnections between the regions

5.1.1 Sample Data

The sample dataset for the models described here can be in the following formats:

- GeoJSON ([231]): a format, based on JavaScript Object Notation (JSON), that encodes specific geometric types, such as point (e.g., locations), polygon (e.g., provinces and countries), line (e.g., boundaries, streets, and routes). Sets of geometric objects, such as MultiPoint, MultiPolygon, can also be represented as collections of objects. Objects can have non-spatial properties.

- TopoJSON[1]: an extension of GeoJSON, in JSON format, that reduces 80% or more of the size in GeoJSON.

- CSV: a format for storing comma separated (or any other separator) values. CSV has, in fact, a tabular structure.

In our case study, we have used the sample dataset derived from Cornu's *Atlas* in two data files in GeoJSON: settlements and routes. Settlements, as Point *feature*s in GeoJSON, are specified by geographical coordinates as well as several properties including various names (original, translated, and transliterated), the corresponding region. Below is a sample showing an arbitrary place as a Point:

```
{
  "type": "Feature", // feature for a point geometry in GeoJSON
  "geometry": {
    "coordinates": [44.35693, 33.35932], // geographical coordinate
    "type": "Point"
},
  "properties": {
    "althurayyaData": { // data based on the Cornu’s Atlas
        "URI": "BAGHDAD_443E333N_S", // unique identifier
        "coord_certainty": "certain",
        "language": ["ara","eng"],
        "names": {
            "ara": {
                "common": "   ",
                "common_other": "   ",
                "search": "   ", // used for search in Arabic
                "translit": "",
                "translit_other": ""
            },
            "eng": {
                "common": "",
                "common_other": "",
                "search": "Baghdad", // used for search in Latin script
                "translit": "Baġdād",
                "translit_other": "Baġdād"
            }
        },
        "region_URI": "Iraq_RE", // belongs to Iraq region
        "source": "cornuData",
        "top_type": "metropole" // type for major cities
        },
"references": {
    "primary": {
       // file identifier for the material from the primary source relevant to this place. The source book URI
       // is also included in the identifier.  The content of this file will be shown to the user.
"0900AbuCabdAllahHimyari.RawdMictar.Shamela0001043-ara1.000285-000.json": {
                "language_manifestation": "",
```

[1] https://github.com/topojson/topojson

```
                 "language_orig": "Arabic",
                 "match_rate": 100, // the match rate of the fuzzy
           search for the place in the current source. The value of 100
           shows that an exact match of this place has been found in the source.
                "match_status": "na",
                "title": "   " // title of the content from the source
            }
         },
    "secondary": { // available secondary sources related to this place
        "ei2_COM_0084.json": { // Unique identifier of the file in which we keep the search result.
        //The file name includes the unique id of the source, which in this example is Encyclopedia of Islam 2 (IE2).
                "language_manifestation": "",
                        "language_orig": "English",
                        "match_rate": "",
                        "match_status": ""
                  }
     }}}}
```

Route sections, as LineString in GeoJSON, are represented as a *feature* by a list of coordinates and a set of properties including source and destination. Source and destination are connected to the settlements data file through the place identifiers—*URI* in the above example. A sample route section in GeoJSON is given below, which connects Riḥā to Nābulus in al-Šām (Greater Syria) and is $48,742$ meters in length:

```
{
"type": "Feature",
"geometry": {
    "type": "LineString",
    "coordinates": [[...,...],[...,...], ...]]
},
"properties": {
    "sToponym": "RIHA_354E318N_S", // start toponym (point) of the line
    "eToponym": "NABULUS_352E322N_S", // end toponym (point) of the line
    "id": "CR0001_FROM354E318N_TO352E322N", // unique identifier
    "Meter": 48742 // length of the line in Meters
}
}
```

As noted above, Cornu's *Atlas* consists of provinces, which include settlements. Therefore, there are only two levels of hierarchy and there are no subregions. In the atlas the provinces are defined by separate maps—one map per province. Figure 5.2 is the map of the Peninsula of the Arabs province (Province D'Arabie) in this Atlas, which shows places and connections. The maps are georeferenced (by Dr. Maxim Romanov) and in the datafiles, each region consists of a set of settlements and route sections. This means that each place is associated with a region, as a property of the place, through a region id (*region_URI* in the above example). Nonetheless, the models are generalizable to multiple layers of hierarchical divisions where provinces are divided into subregions.

As explained before (see Section 4.2.1), a tree is an appropriate structure to represent hierarchical relations between entities. The root of a tree is the geographical area that covers our data; here we call it the *entire area*. The *entire area* is then divided into macro-regions and micro-regions (subregions) in the lower levels in which the settlements are placed as tree leaves. Implementing the tree structure, Figure 5.3a illustrates how the current sample dataset of provinces and subregions looks for the two levels of hierarchy that Cornu's *Atlas* provides. Figure 5.3b shows the subregions of the al-Šām (Greater Syria) province derived from the same dataset.

In what follows we will employ mathematical models to describe and visualize hierarchical

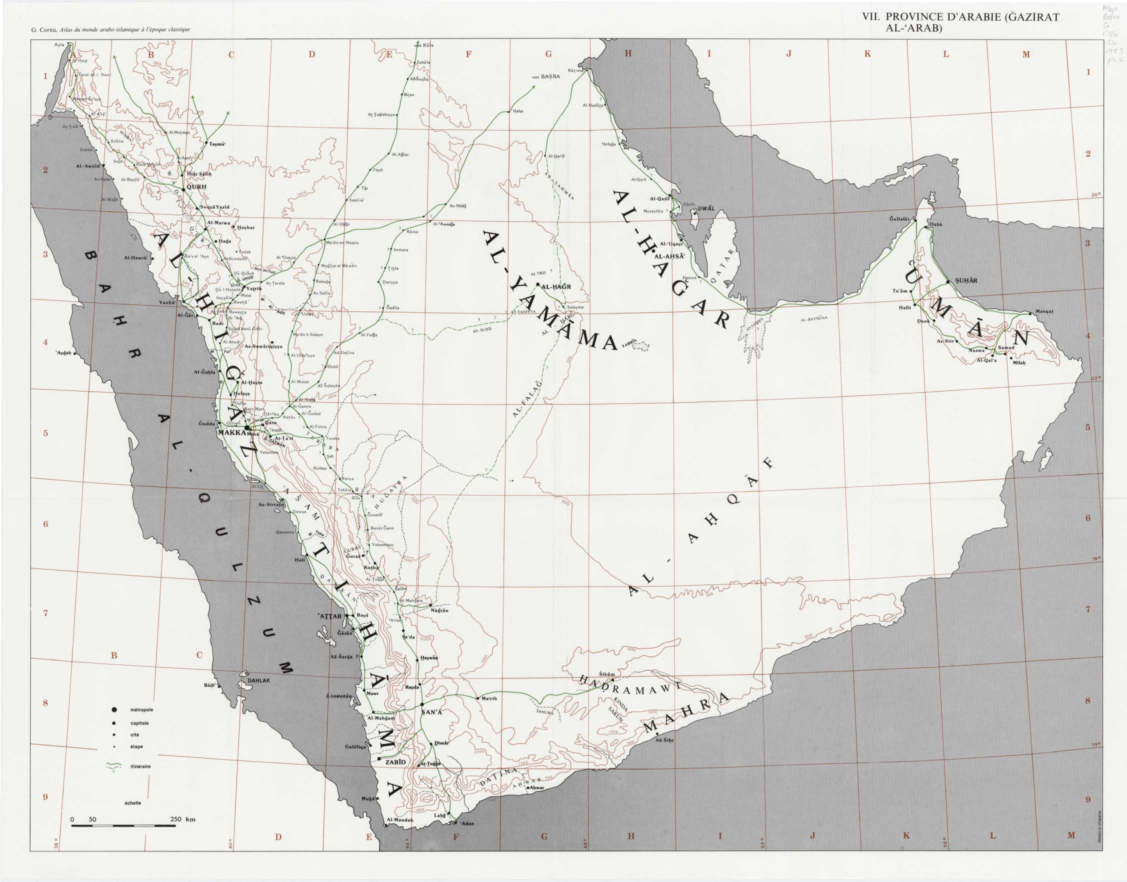
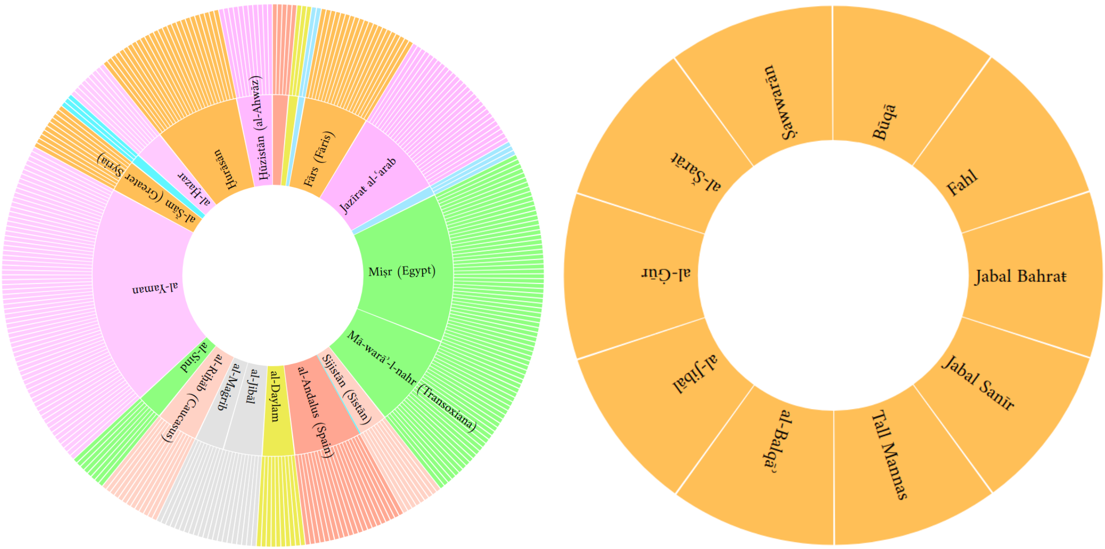


Figure 5.2: The map of the Peninsula of the Arabs (Jazīrat al-ᶜArab) from Cornu's *Atlas*, including places and route sections that belong to this region.



(a) Provinces and subregions. Note: Name of the smaller regions are not rendered at the current resolution level of the visualization.

(b) Subregions of al-Šām (Greater Syria) province

Figure 5.3: Hierarchical structure of data from Cornu's *Atlas*

divisions. For each model, we have implemented, using D3.js[2] library, a visualization using the sample dataset to illustrate and , if required, tweak the models or to discover a new one. Specifically, we offer the following mathematical models and examine them closely: quadtree clustering ([233]), Voronoi diagram ([234]), convex hull ([235]), and concave hull ([236]). After applying these models to our sample data, we introduce a new model that meets our requirements. In a few visualizations of the above models, we have partially reused implementations from Elijah Meeks's Blocks[3].

5.1.2 Quadtree

Sometimes gazetteers provide information about hierarchies: for instance, we can infer from such data that Leipzig is a city in Sachsen, which is a state in Germany, which is a country in Europe. But what happens when we do not have such information (i.e., all we have are the locations of settlements)? This is a common problem with archaeological data, especially from prehistoric periods. Quadtrees in spatial data help to form the possible regions according to the distances between the locations in the two dimensional space. This means, using quadtree we can infer regions for a set of locations.

Space divisions that are described in geographical texts produce at least two levels of hierarchy: regions and settlements. Quadtree that is based on a recursive subdivision of space, properly represents hierarchical data. It is a two-dimensional space segmentation method that recursively subdivides space into four equal-size regions with a square or rectangular (or arbitrary) shape at each level. The data in this space will then be placed within these regions. Each region has a certain maximum capacity of data to contain. The capacity can either be specified when the model is being implemented or the minimum capacity will be considered, which is one unit of data. Each region is divided when the capacity is reached and divides the containing space into four subdivisions, which are called NW, NE, SE, SW. Therefore, each node has either four or zero children. In other words, in the first division step, the root is divided into four child regions and then any generated region that contains more data than the capacity allows, will produce four subregions (subdivisions) by dividing the containing space.

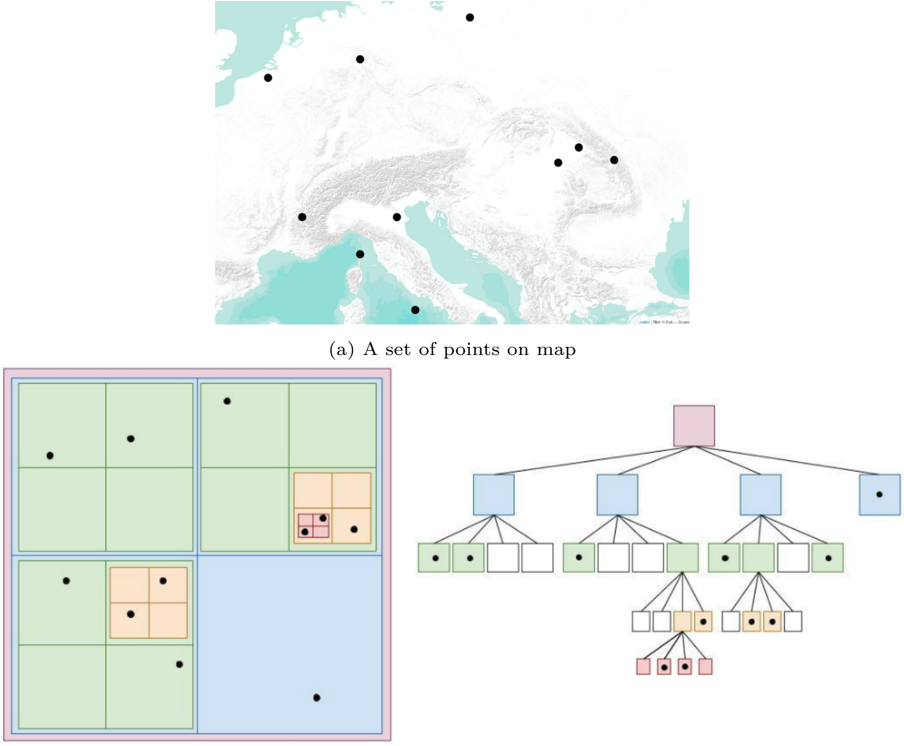
Figure 5.4 illustrates an example of space decomposition using quadtree together with the corresponding tree structure, in which only one particle is placed in each cell. Assuming each particle in this figure is a settlement in a two-dimensional space, quadtree partitioning will locate the settlements in the container regions at each level and divide the regions into four new subregions once they contain more than one settlement. At the end, each individual settlement will be placed within a region, which can be either an internal node or a leaf in the tree. As can be seen in this figure, only those (sub)regions that reach capacity are divided.

In the example in Figure 5.4, the quadtree space segmentation algorithm forms regions according to the spatial position of settlements and decides which settlement or set of settlements should be placed in the same region. The model might produce a multilevel structure: macro-regions appear at the highest-level, subregions and micro-regions at the internal levels, and settlements in the leaves. This generates various levels of regional and hierarchical divisions for a specific geographical area.

This model relies on a fully computational approach of space segmentation based on the po-

[2] `https://d3js.org`

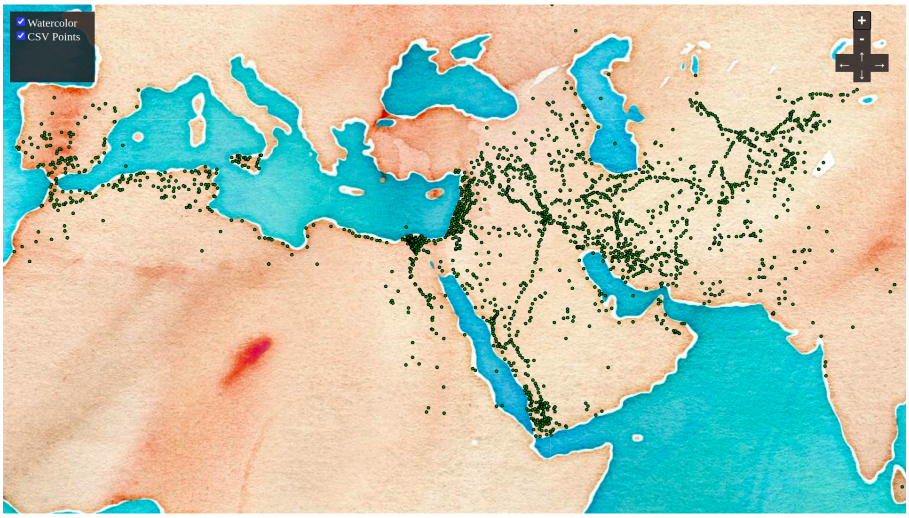
[3] `https://bl.ocks.org/emeeks`



(a) A set of points on map

(b) Quadtree spatial and logical space segmentation

Figure 5.4: Quadtree space segmentation example: (a) a set of locations (points) on map; (b) segmentation in spatial (left) and logical (right) arrangement using the above set of locations. This figure shows how the quadrants become as detailed as they need to be to place each location in a separate quadrant.



☑ Watercolor
☑ CSV Points

Figure 5.5: The settlements in Cornu's *Atlas* on which we implement the models in this chapter.

sitions and density of settlements in different parts of the space. In other words, it does not use hierarchical information that might be present in the underlying dataset. Applying this model to our sample dataset (initial points can be seen in Figure 5.5), Figures 5.6a and 5.6b illustrate how quadtree divides the *entire area* and groups the settlements into larger circular cells according to their positions in two zoom levels. Circles in the figures appear with different radiuses and shades of green color, according to the zoom level and the density of the underlying data at the corresponding zoom level. The smallest circles represent individual settlements and themselves also make up the larger circles in darker colors. Figure 5.6b demonstrates, for example, the way that the size, density, and shades of the circles represent Egypt (Miṣr) as a province. More densely settled areas contain larger circles in darker colors and less densely settled areas have smaller circles that show areas with fewer settlements.

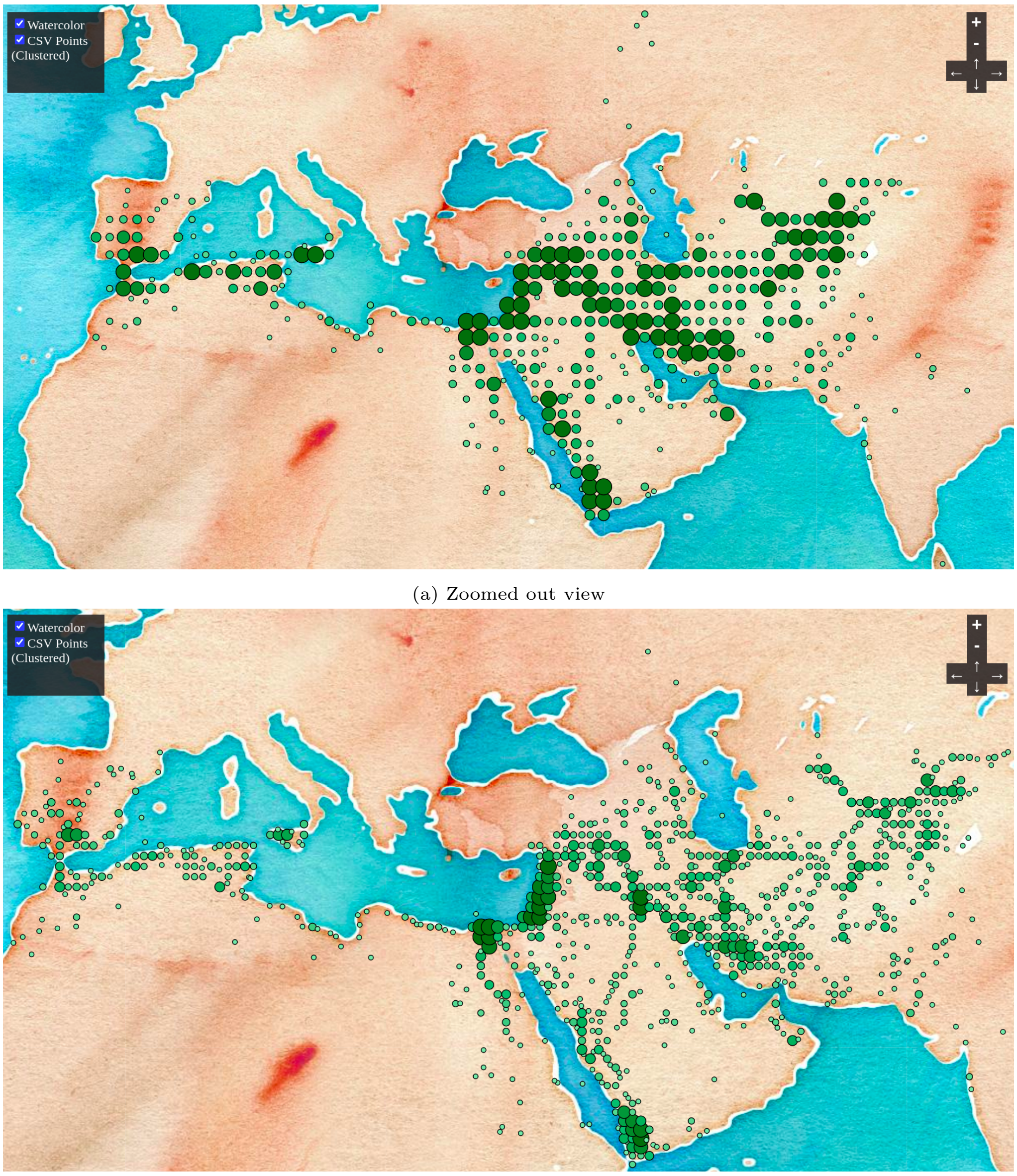
Quadtree clustering efficiently depicts the density of settlements in various parts of the space. It is well suited to research where one needs to contrast settled and unsettled areas. For instance, Northern Egypt, Yemen, and the Levant emerge as population centers while al-Andalus (Spain) has a range of settlements but less density. Furthermore, since this algorithm does not need hierarchical information and clustering is based on the spatial proximity of objects, the datasets that do not have hierarchical information (i.e. there is no mention of regions), can take advantage of the regions that this model offers. In the case of existing hierarchical information, the regions computed through this model can be compared with the existing regional divisions in the data.

However, while useful and appropriate for datasets without hierarchical information and for neighborhood operations[4] ([218]), we discuss other models in the following which engage hierarchical information to shape more precise regions.

5.1.3 Voronoi Diagram

When we discuss the extension of regions, an analytical view of the accessible area in a region can be of great value to model routes, divisions, and coverage. Voronoi diagram is another space

[4] Examples of neighborhood operations are: comparing—i.e., if neighboring cells share the same attribute; intersections—i.e., finding the intersection of cells; union—i.e., finding the combination of cells; and finding neighbors of a cell (or settlements).



(a) Zoomed out view

(b) Zoomed in view

Figure 5.6: Quadtree visualization of regions and settlements using the set of points in Figure 5.5. The size, density, and shades of the circles corresponds to the density of the initial places.

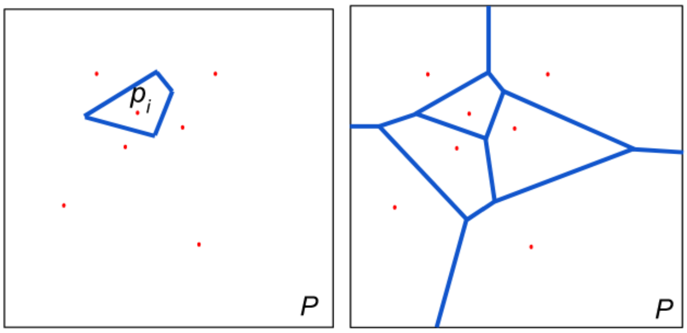


Figure 5.7: Voronoi partitioning and diagram in a two-dimensional space P. (Left): A Voronoi cell of the point p_{i}; (Right): Voronoi diagram of the points p_{i}.

partitioning algorithm in computational geometry, which is identified by a set of points. It has a wide range of uses, including geographical optimization, such as searching for the closest neighbor/feature to any given point and analyzing patterns of urban settlements[5].

A Voronoi diagram of a set of points (or settlements in the geographical context) is a collection of regions that divide the space in which the points are located. Voronoi diagrams shape regions according to their position in space. Each region contains exactly one settlement, called the seed. The regions divide the space so that all points in each region are closer to the seed in this region than to the seeds in other regions. In other words, each region is shaped around a specific settlement and contains all points whose distance to this settlement is shorter than to any other settlement. Below is the definition of a Voronoi diagram:

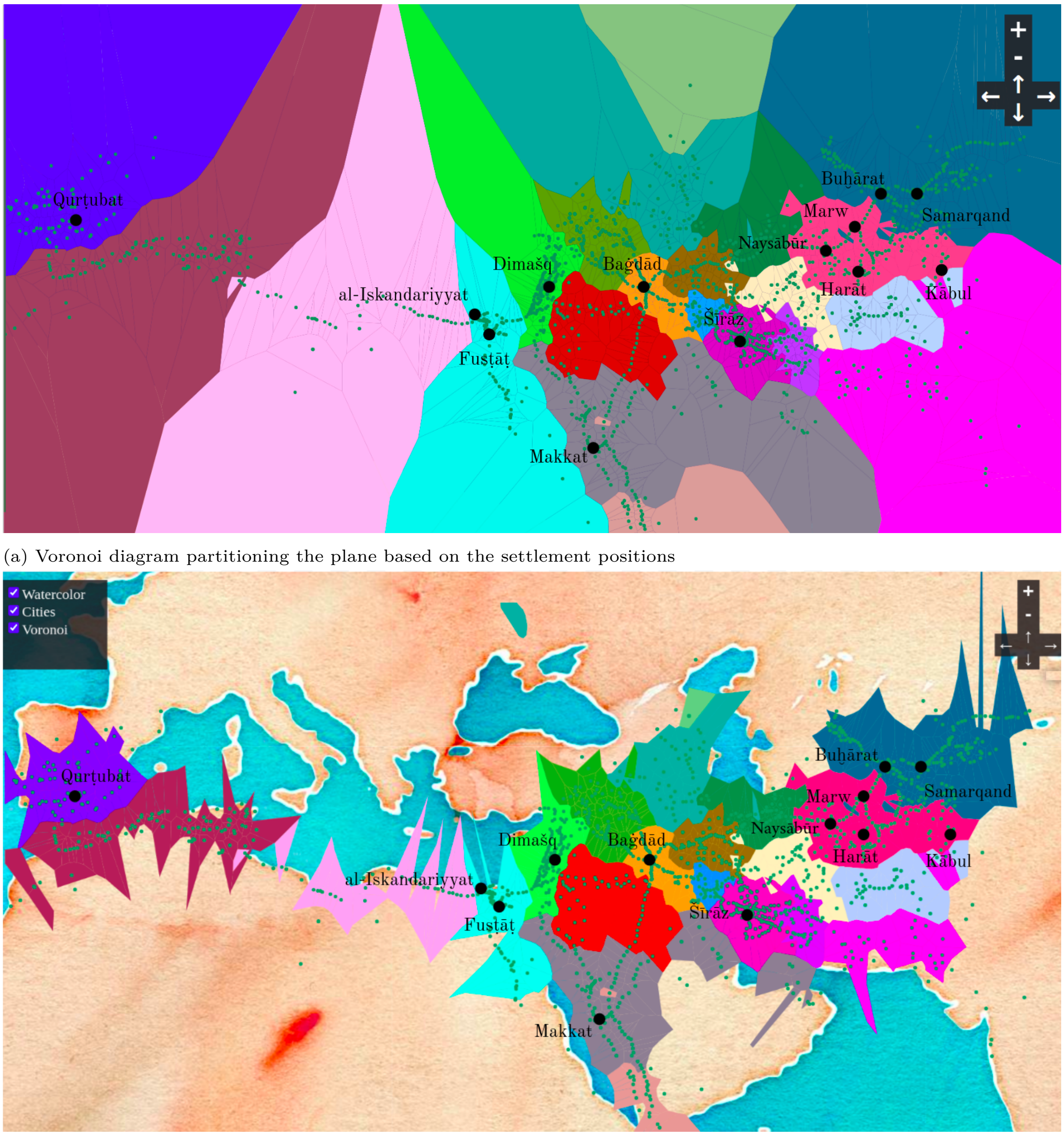
> Assume we have points $p_{\mathrm{i}}, 0 < i \leq n$, in the space P. A Voronoi cell is a set of all points from P that are closest to point p_{i} and a Voronoi diagram is the union of all Voronoi cells as shown in Figure 5.7 (Left) and Figure 5.7 (Right), respectively.

In geospatial studies, Voronoi diagrams are often used to show spheres of influence. Each Voronoi cell forms a region around a place so that the shortest spatial way to reach the different parts of the region is through this corresponding place. For instance, this could be a region that represents a district around its capital. The union of the calculated Voronoi cells of settlements that belong to the same region constitute a higher level region, like a country.

Applying this model to our sample dataset produced a Voronoi diagram of the settlements. In order to form regions for a group of settlements that belong to the same region, we identified all the polygons (Voronoi cells) of a region and applied the same color to them. The union of the Voronoi cells of settlements of a region thus shapes this region. The idea of merging cells or regions to shape higher level regions can be applied to different levels of hierarchy. Figure 5.8a depicts the Voronoi diagram implemented using the sample dataset. Cells of the same color represent specific regions.

The classic implementation of a Voronoi diagram corresponds to the whole plane on which the partitions are expanded. In the current implementation, the diagram covers the whole geographical space of Earth. However, our dataset is only limited to a relatively small area. As can be seen in Figure 5.8a, the diagram expands beyond the area that all the settlements in this dataset belong to and covers all of the Earth, with unexpectedly large partitions where the settlements are sparse.

[5] An example of finding the closest school in a neighborhood can be found at `https://github.com/shakasom/voronoi`. This implementation uses *python* libraries, such as Geopandas and Geovoronoi.



(a) Voronoi diagram partitioning the plane based on the settlement positions

(b) Voronoi diagram in which the cells larger than a specific value are filtered out in this visualization to hide very large Voronoi cells that completely cover large bodies of water. Filtering excludes those cells from the coloring process while the seeds for those regions are still shown on the map. Thus, the points that do not belong to any colored cells are in fact in the filtered cell. We selected the value for filtering by try and error to achieve the best view of the regions.

Figure 5.8: Implementation of Voronoi diagram for regions in the Cornu's *Atlas*

To solve this problem, we limited the Voronoi diagram to the *entire area* by clipping the Voronoi diagram using the bounding box of this area.

Limiting the calculated Voronoi diagram to the bounding box of the *entire area* solves this problem in specific cases where settlements are close to the borders of the bounding box. However, the size of partitions in some cases is still problematic. In Figure 5.8b, we have filtered out the cells whose area extends over a certain value, excluding unnecessarily large partitions from the visualization. We specify this value for visualization experiments considering the average area and the outliers of the polygons, which may vary for different datasets. This figure shows an example of filtering that removes some of the large cells from the sparse areas and water bodies.

For a set of places with specific classifications, such as capitals and metropoles, the corresponding Voronoi diagram can be used as an analytical tool to visualize travel routes. The classifications of places with Voronoi diagrams allows one to gain an insight into areas that potentially fall under the influence of a power located in a specific place.

Voronoi diagrams, unlike quadtree, utilizes the hierarchical data and the divisions it offers are based on the existing contextual data. Compared to quadtree, the shapes of regions are more meaningful. Regions are formed without any overlap. However, it leaves no unpartitioned space and, as a result, suggests more modern-type regional representation where borders divide the entire territory, leaving no “space” for “no man’s land”. Additionally, the model overextends regions into uninhabitable areas such as water bodies and deserts (although, for the former, a dataset with waterbody boundaries can be used to clip the Voronoi diagram or this can be done in GIS software using the clipping feature). To address these problems we suggest some modifications to this model in the following section.

5.1.4 Voronoi Clippings

Clipping a Voronoi diagram by a bounding box of the *entire area* produces a rectangular plane partitioned by Voronoi polygons. However, a box/rectangle is not a genuine demonstration of our *entire area*. A polygon that tightly encloses the area and encompasses all the sites would be a more reliable method to clip. Thus, we consider two methods to achieve this.

5.1.4.1 Convex Hull

A convex hull that encompasses all the settlements is a model that can be used to form a polygon that would “wrap” settlements belonging to a specific region. In computational geometry, a convex hull is a common structure and can be used to compute the envelope of a set of points on a plane. The convex hull of a set of points is defined as the smallest convex polygon that contains all the points (Figure 5.9). A particular property of a convex hull is that every straight line or edge that connects any two arbitrary points is inside the polygon (i.e., no internal angle is greater than 180°), as shown in Figure 5.10 (Left). According to the following definition, every set, or more specifically polygon, with this property is convex:

$$\text{A set } P \subseteq R \text{ is convex if } \overline{\text{pq}} \subseteq P, \text{ for any } \overline{\text{pq}} \in P, \tag{5.1}$$

where R is the space to which the set P belongs and pqis a straight line between two arbitrary points p and q. To visualize a convex hull, imagine a set of nails on the wall. Stretch a plastic rubber band around the nails (the black hull in Figure 5.10 (Right)) and let it go to snap around

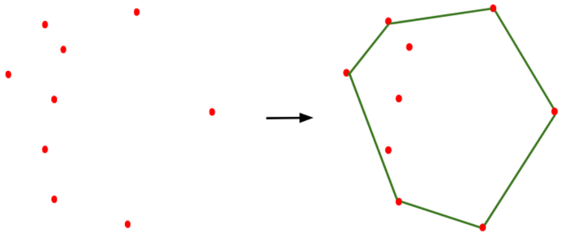


Figure 5.9: Example of convex hull: (Left) A set of points P; (Right) The convex hull of set P.

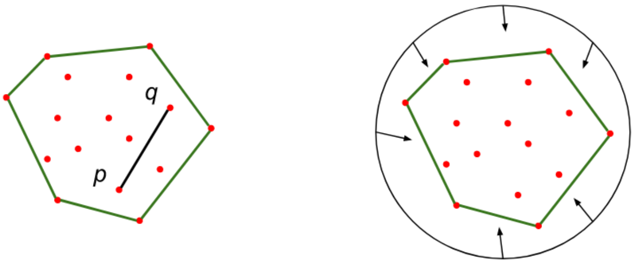


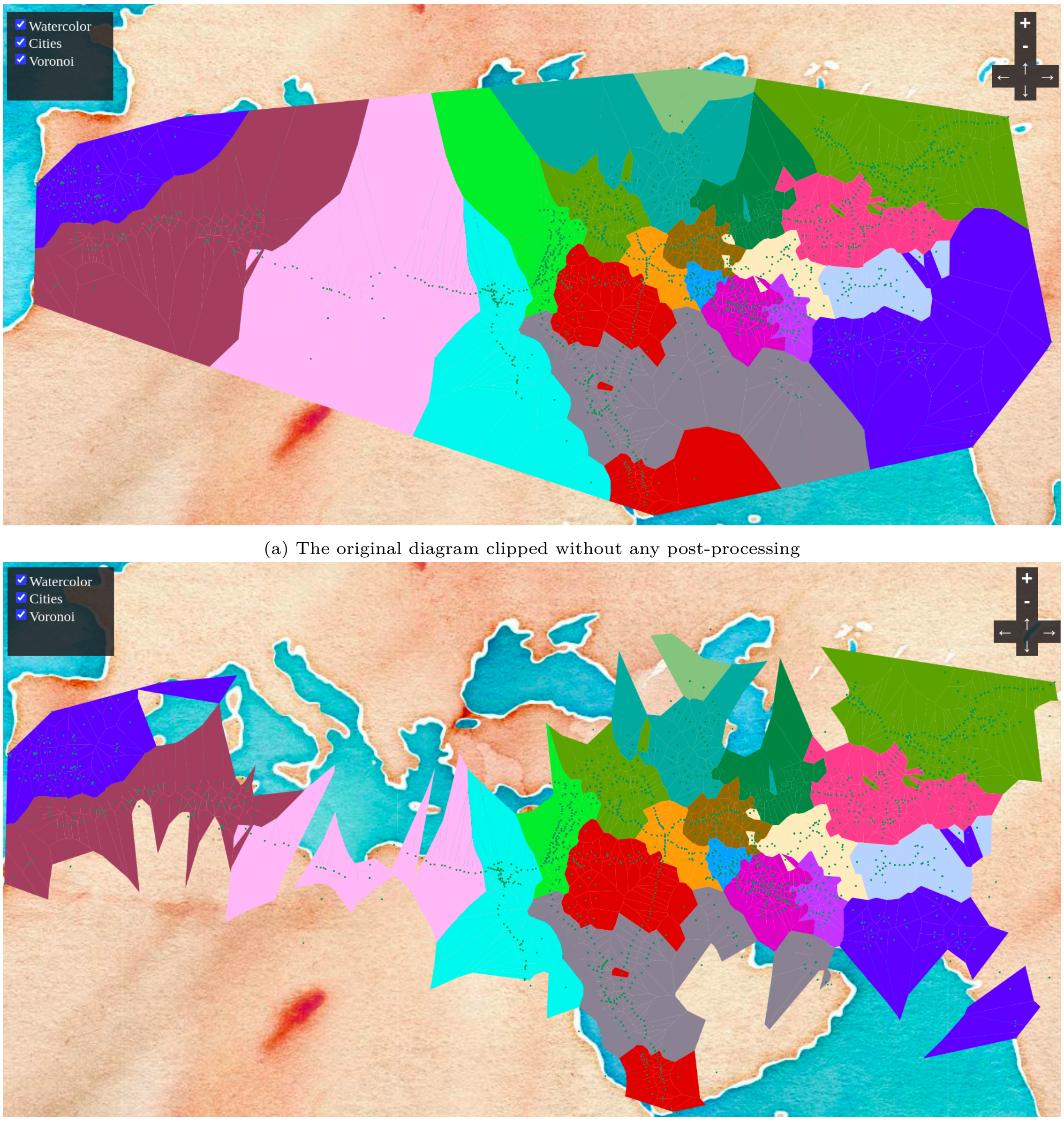
Figure 5.10: Convex hull: (Left) Every straight line between two arbitrary points is inside the convex polygon. In other words, there is no internal angle greater than 180°; (Right) the elastic rubber band analogy. Stretch a plastic rubber band around the nails (the black hull) and let it go to snap around the nails and enclose an area. The shape that the plastic band forms is the convex hull of the nail (point) set.

the nails and enclose an area. The shape that the plastic band forms is the convex hull of the nail (point) set.

We then used the convex hull of the settlements to clip the Voronoi diagram that is implemented in the previous section. As shown in Figure 5.11a, the diagram is drawn on a tighter area that is no longer rectangular, as it happens when we clip it by the bounding box of the *entire area*. The polygons are expanded and limited differently at the edges so that the shapes of the regions, in contrast to the previous implementation, are closer to what the underlying data expresses. Therefore, some regions that were hidden in the previous visualization due to the filtering of large polygons, are now present because they are pruned. However, some large empty areas are still formed as regions, as it can be seen in Figure 5.11b that shows a filtered view of the diagram. The reason for this issue, besides the large Voronoi cells, is the shape of the convex hull with which we clipped the Voronoi diagram. The convex hull of a set, depending on the position of the points, can leave some empty spaces around the points that it contains. In other words, some space in the enclosing area of a convex hull might be sparse or empty because the resulting hull has to be convex. For example, the convex hull in middle in Figure 5.12 produces an empty space inside, which is shown with the area covered by the inner polygon, while the actual coverage should be similar to the polygon on the right in this figure. The inner polygon represents the empty space that is the result of keeping the convex property of the result. In fact, the convex hull does not represent the boundaries of a given set well. Therefore, the clipping procedure might not be able to precisely cut out all the empty areas, which contain an insignificant amount of data (or no data at all) either at the edges or between the regions.

5.1.4.2 Concave Hull

The above-mentioned issues prompted us to investigate using concave hulls instead of convex hulls for clipping the Voronoi diagram. Concave hull, similar to convex hull, is a polygon that embraces



(a) The original diagram clipped without any post-processing

(b) Large cells are filtered by using transparent colors

Figure 5.11: Voronoi diagram clipping implemented for the Voronoi diagram of Cornu's *Atlas*, using a convex hull of the *entire area*. As it can be seen, in populated areas we can find fine-grained regions, such as Aqūr (al-Jazīraŧ), al-ᶜIrāq, Badiyyaŧ al-ᶜArab, al-Jibāl, and Ḫūzistān (in modern Iraq, Syria, and part of Iran) while in North Africa, which is more deserted compared to the eastern parts, region borders do not show the real shape of the regions, such as Miṣr (Egypt) and Barqaŧ (Lybia).

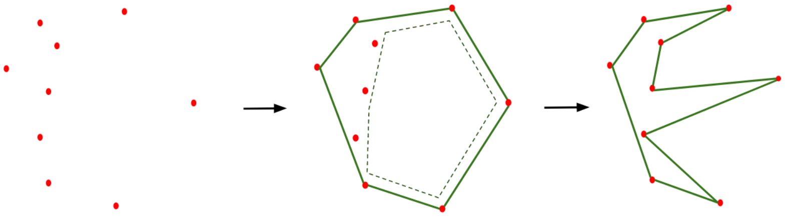


Figure 5.12: The area of a convex hull may not precisely show the area that the set of points covers. In this example, the convex hull (outer polygon in the middle) leaves some empty space, which is represented by the inner polygon drawn with dashed lines, while a polygon showing a tighter area of coverage should be similar to the left one.

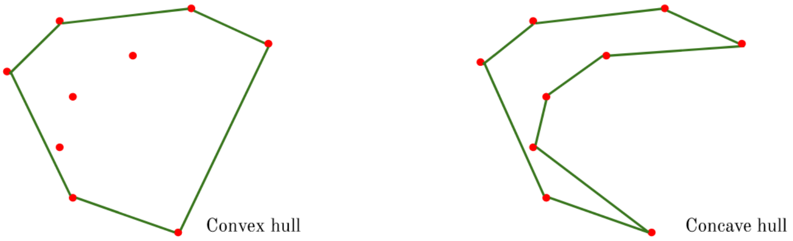


Figure 5.13: An example of convex and concave hulls of the same set of points. The convex hull encloses a much tighter area when compared with the convex hull.

an arbitrary set of points and computes a polygon with a smaller area, when compared to convex hull. Figure 5.13 demonstrates an example of both convex and concave hulls of a set of points. As this figure shows, the concave hull minimizes the enclosed area and shows the area occupied by the points more precisely.

Using concave hull, instead of convex hull, can improve the exclusion of larger areas and, to some extent, solve the problem of the outer boundaries. Compared to a convex hull, it does not enclose large empty spaces and, consequently, offers a much smaller polygon for the same points. However, it only trims the regions at the edges and the issue of inner boundaries still exists. We therefore need an approach that can be applied to each province instead of the *entire area*.

To achieve this, we employed the convex hulls of each province instead of the hull that contains the whole place. This model is a combination of two different models to improve the original idea of clipping: we form regions with the Voronoi diagram and hulls, and then utilize the hulls to clip the Voronoi diagram. In addition to excluding areas at the edges of the *entire area*, this approach removes areas that do not belong to any of the regions. Consequently, the whole diagram fits into the convex hulls of the regions. Figure 5.14 depicts how this model generates more compact regions, leaves out the areas that have no settlements, and, therefore, can be considered a more precise representation of our sample dataset. Nevertheless, the visualization shows that not all the empty parts are properly clipped and that there are overlaps of the hulls, which are formed due to the convex property of hulls.

5.1.5 Convex Hulls

Computing a polygon that accurately shows the area that a set of points occupy is a classic problem ([237]). Although convex hulls are not usable in combination with the Voronoi model, as demonstrated above, they can be used on their own as an effective model to represent divisions. We can therefore use convex hulls of a set of settlements that belong to a region as a model to represent that region. The implementation of this model on the sample dataset in Figure 5.15 uses settlements to represent the provinces in a two-level hierarchical dataset, as mentioned above. One can apply this model to any levels of hierarchy in the bottom-up approach. At each hierarchical level, points or subregions that belong to a higher-level container region, form a convex hull. As Figure 5.16 shows, there are subregions inside a province and each is represented by a convex hull. The implementation should consider building entities (settlements or regions) for each level to create the hulls.

Convex hull offers an easy-to-implement and practical model in visualizations; however, it may produce overlapping regions. Regions do not overlap in the real world and a settlement can only belong to one region at a time. Thus, overlaps are the major issue that needs to be resolved.

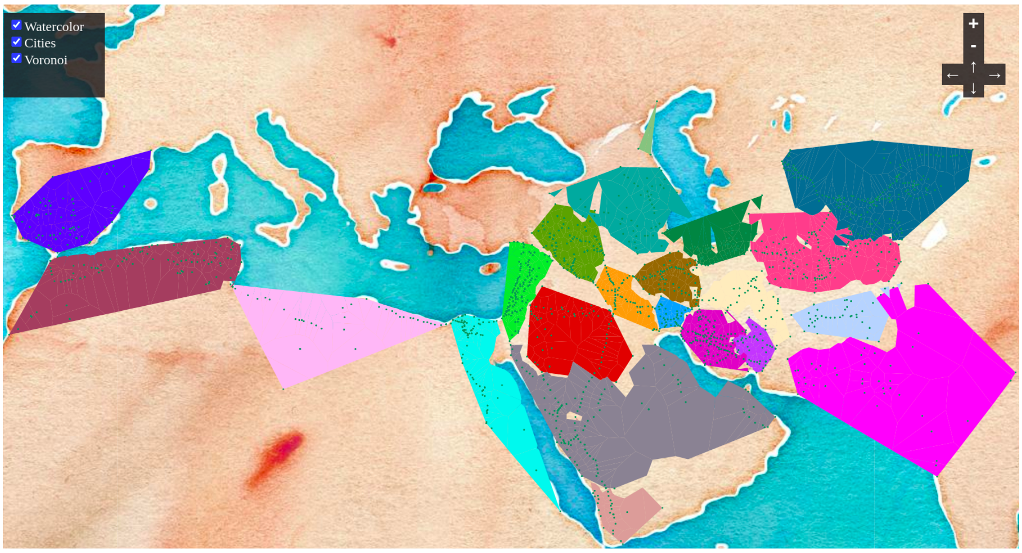


Figure 5.14: Voronoi diagram clipping by a convex hull of each region, implemented using Cornu's *Atlas* data. Each set of Voronoi cells that shape a region are clipped individually. Thus, the areas that contain no settlements are not shown as part of a region. Clipping removes irrelevant spaces, as, for example, in Miṣr (Egypt). Al-^{c}Irāq, Aqūr (Northern al-^{c}Irāq), al-Yaman (Yemen), Ḫūzistān, and Fārs (both at Southwest Iran) emerge with more clear borders.

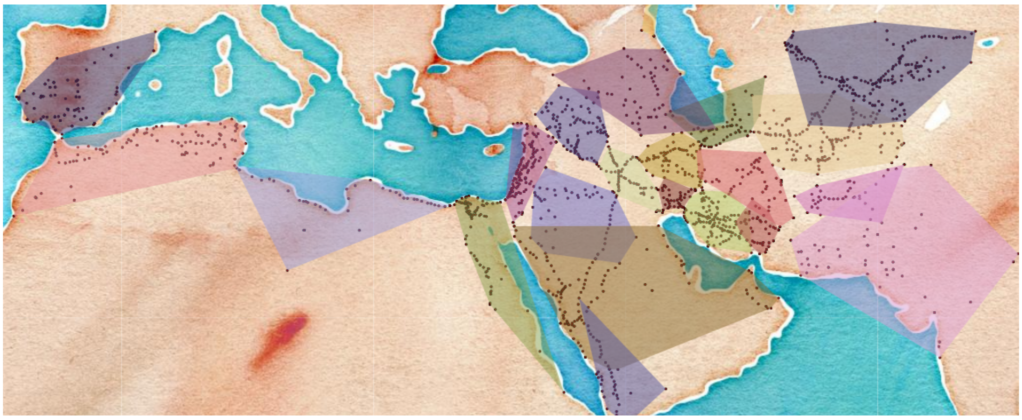


Figure 5.15: Convex hulls for visualizing regions of Cornu's *Atlas*

One major reason, as can be seen in Figure 5.15, is the property of the convex hull that leaves unnecessary empty spaces (see Figures 5.12 and 5.13) and empty areas that cause overlaps.

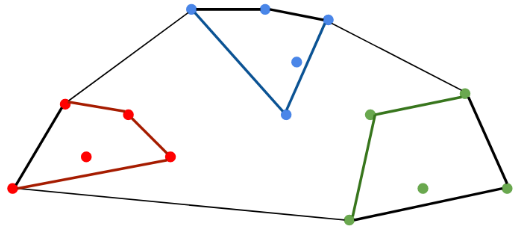


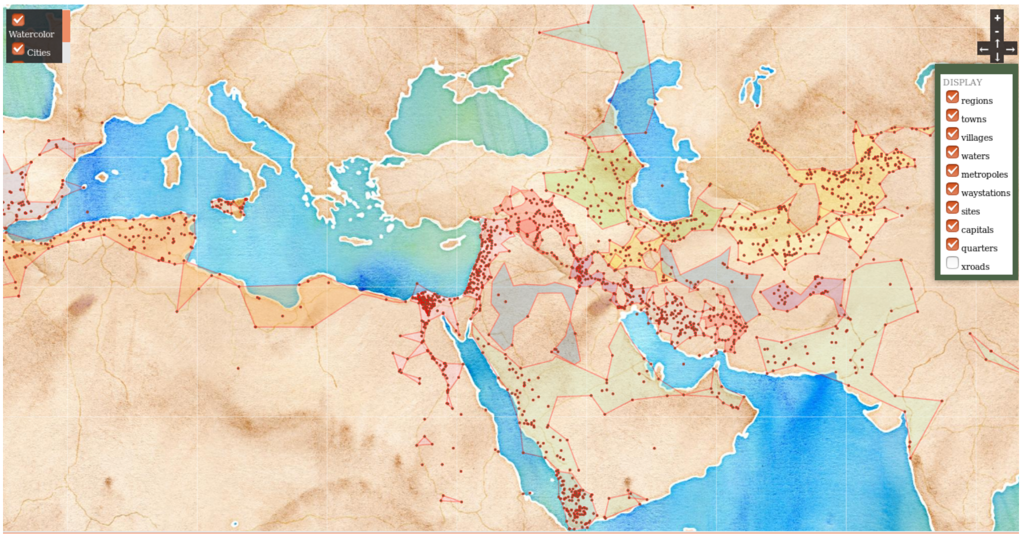
Figure 5.16: Nested convex hulls of subordinate regions

It should also be mentioned that the accuracy of data (i.e., all the coordinates are certain and accurate) that may cause overlap, is not our focus here. We assume that the data is clean and accurate and only evaluate the model according to the implemented visualization. However, this model could be a visual way to find those inaccuracies and problems in the data. A basic example of such problems is when a point of a region is misplaced in the middle of another region and causes an unnecessary overlap. This example could be a case to investigate any inaccuracies in the coordinates in the data.

5.1.6 Concave Hulls

We aim to tweak the process of computing a hull so that it effectively encompasses a set of points by replacing convex hulls with concave hulls. Concave hulls supposedly generate tighter polygons around the points. Therefore, the result, in comparison to the former models, will be regions with more nuanced shapes that do not overlap, as can be seen in Figure 5.17. Compared to the convex hull for the same set of points, the concave hull offers regions with minimum overlaps. Figure 5.18 provides a closer view of the two provinces, Si[ji]stān and al-Sind, represented with convex hulls (left) and concave hulls (right). As shown in this figure, the overlaps of the convex hull model are minimized using the concave hull model. Concave hulls include smaller empty areas, which contain no sites (see Figures 5.12 and 5.13). A perfect example of such areas are deserts in which no settlements are located: those areas are not computed as a part of a region in this model. The deserts in modern (Central) Iran and the Peninsula of the Arabs (shown with red ellipses in Figure 5.19) are left out of the concave hulls, while the convex hulls cover most of those areas. Thus, this model offers a more precise visualization of regions, effectively excluding “no man’s land”.

This model resolves almost all the major inefficiencies of the previous models. However, the regions still need trimming. Water bodies are partly covered by some regions while, based upon the underlying data, they are not part of the enclosing region. Moreover, closer study of the data reveals some discrepancies. In other words, some computed regions do not reflect the corresponding explanation of data. For example, the province of Miṣr (Egypt) is described along the delta and valley of the Nile and this model does not produce the precise representation of it (shown by a green ellipse in Figure 5.20). The calculation of the concave hull, which encompasses all the points in this province, attaches some extra parts to this province in the shape of triangles. It also leaves some settlements out of the relevant region. Therefore, the demonstration may go beyond the real territory or exclude some parts of the region. We observe this effect in other regions with complex geometries.



☑ Watercolor
☑ Cities

DISPLAY
☑ regions
☑ towns
☑ villages
☑ waters
☑ metropoles
☑ waystations
☑ sites
☑ capitals
☑ quarters
□ xroads

Figure 5.17: Concave hulls, visualizing the regions of the Cornu's *Atlas*

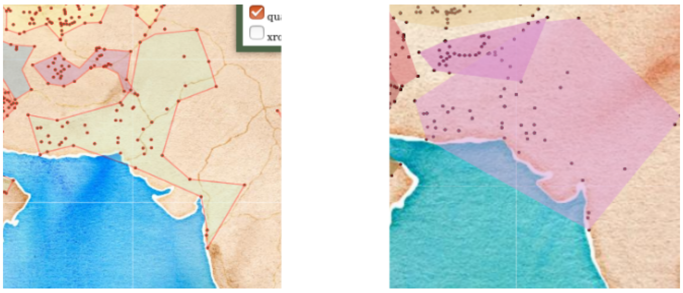


Figure 5.18: The Provinces of Si[ji]stān (upper, smaller) and al-Sind (lower, larger) visualized in two different ways: (Left) concave hulls; (Right) convex hulls. Concave hulls minimize the overlap.

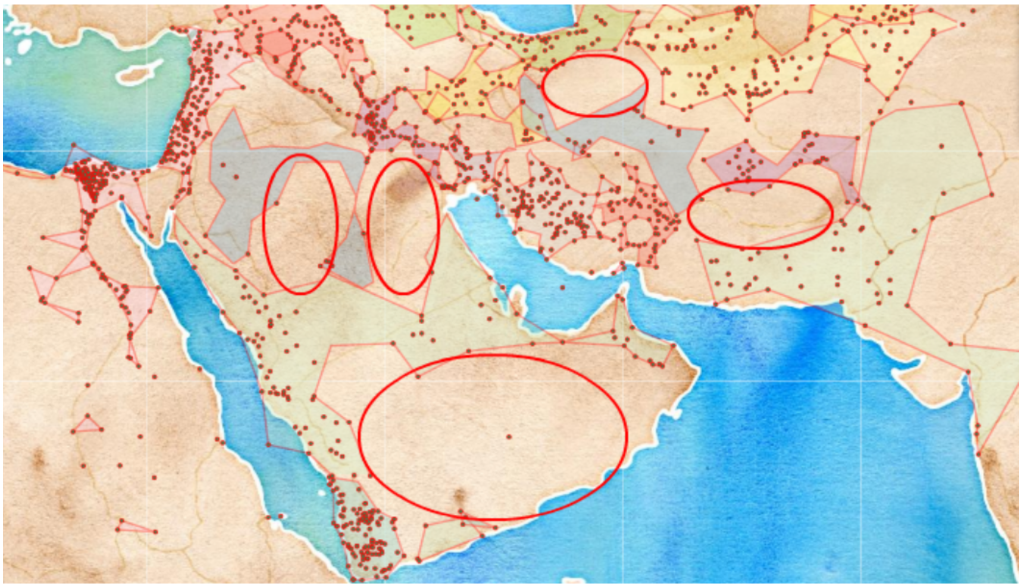


Figure 5.19: Examples of empty spaces (deserts in modern Iran and the peninsula of the Arabs with no settlements) that are excluded from regional representations computed with the concave hull model.

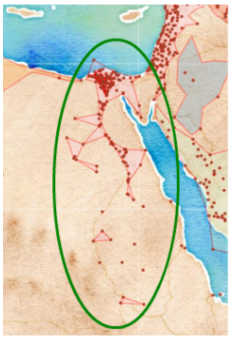


Figure 5.20: The province of Miṣr (Egypt), represented with the concave hull model.

5.1.7 Route Network

The models that we have discussed in the previous sections are based on the settlement data, coordinates, and their regional categorization. In other words, we have always used places and their position in space to model regions, as this is the most relevant information to define a region in the first place. Moreover, settlements are the building blocks of regions at various levels of the administrative hierarchy.

However, the models covered in the above sections also have disadvantages, as we discussed. In addition, they do not utilize the route network. A route network is an important entity that can be seen as another building block for regions in addition to settlements. In this section, we will therefore propose an approach that benefits from combining settlements and routes when modeling regions, by defining a region not only through settlements, but through routes.

To implement this model, we clustered entities that belong to the same region. We assigned a particular color to each set of nodes and edges that form a region. In a complete graph of nodes and edges of a route network, each set of nodes and edges of a region is assigned a distinct color. In fact, this approach models regions by highlighting objects that are described as belonging to it. It can thus be seen as the clustering of nodes and edges by specific properties.

Figure 5.21 illustrates the complete map of regions that this model produces. Each region is depicted by a set of nodes and edges in the same color. The route sections that connect to locations from two different regions are colored in gray (Figure 5.22). This approach inserts visual gaps between regions, without recourse to borderlines. The extent of each region is precise and clear without overlaps with other regions or water bodies. Additionally, areas that are clearly not parts of regions are completely pruned. Accordingly, this model reflects our data that does not concretely explain the expansion and shape of regions and offers precise visual and spatial representation, which is applicable to multiple levels of administrative hierarchy—for example, micro and macro regions.

5.1.8 Summary of Models for Administrative Hierarchy

In this section, we offered models for representing administrative hierarchy and provided implementations for each model. Through the development of adjustments and changes, the models evolve from analytical tools, which provide a deep insight into particular historico-geographical questions, toward a wider model that serves the general requirements of historical and geographical visualization. Thus, through the examples given above, we have shown that quadtree is an

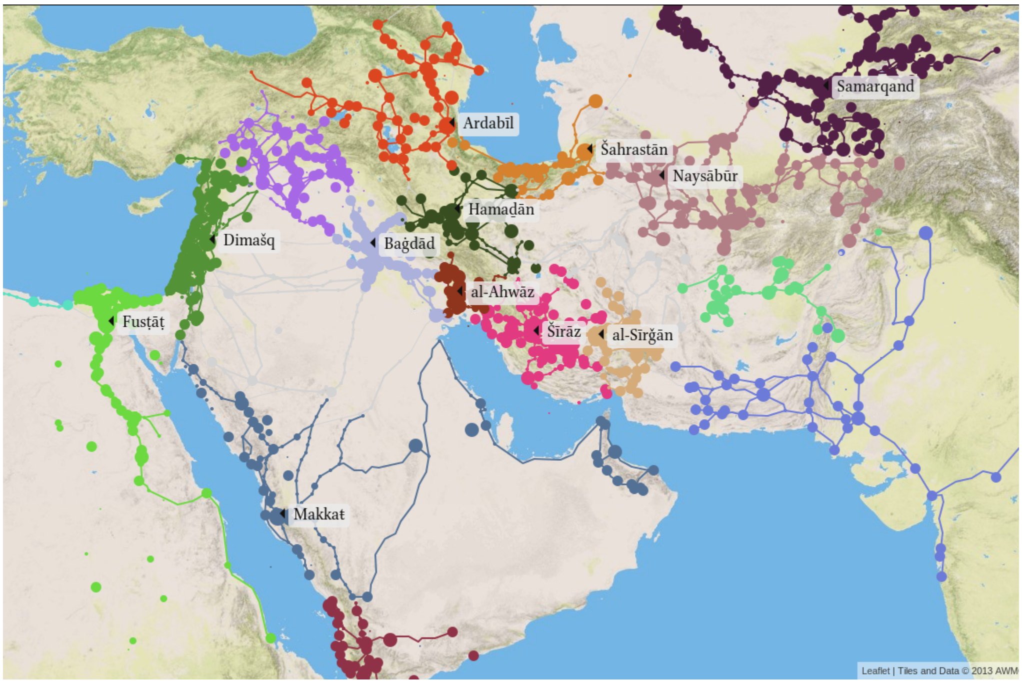


Figure 5.21: Modeling and visualizing the extent of provinces by assigning the same color to settlements and routes that belong to the same administrative unit (province).

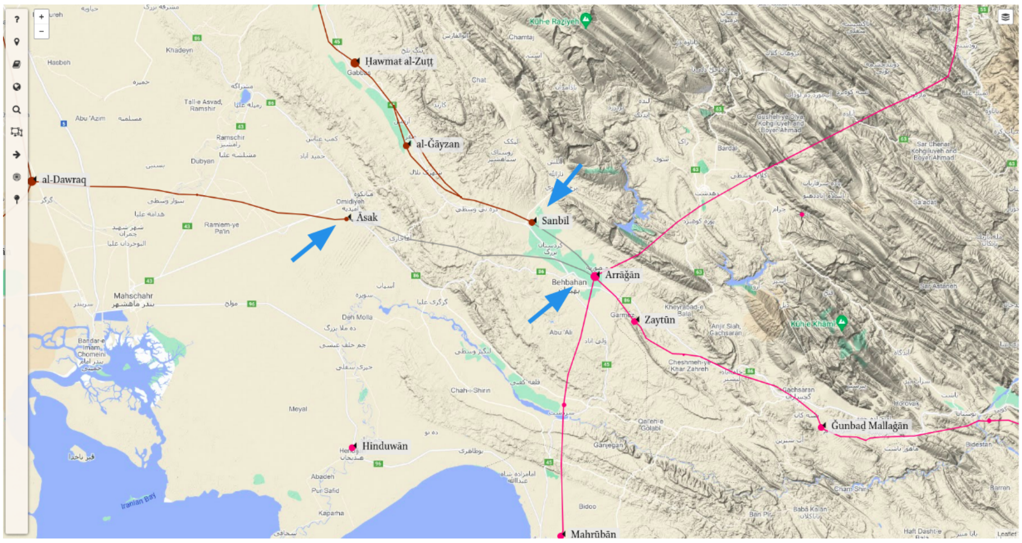


Figure 5.22: Route sections that connect two regions are gray to add a visual gap between regions and show a more precise representation of the area that belongs to the regions according to the data. This helps us to highlight the area of the regions with the corresponding settlements and routes in distinct colors. This example shows two routes that connect Sanbīl in Ḫūzistān province (in brown) and Āsak to Arrāǧān in Fārs province (in pink), are in gray. The above-mentioned settlements are shown with blue arrows.

effective representation of density of settlements in space and that Voronoi diagrams offer a powerful visual tool to analyze the changes of territories controlled by competing dynasties throughout the course of their rule. We have also shown that convex and concave hulls enclose the regions that encompass a set of settlements and, consequently, give effective results for adjusting Voronoi diagrams. Finally, by integrating route networks, we have shown that we can visualize the extent of regions, avoiding the imposition of the modern concept of borderlines.

5.2 Comparison Models

In this section, we will offer two models for the comparative study of geographical sources which describe geographically similar hierarchical data and routes. The first model converts hierarchical data into a matrix for a binary comparison. The second model proposes a method that simplifies route networks so that the visualization and comparison of different networks is more straightforward. The models allow one to identify and then interpret descriptive patterns and spatial understanding of various sources at the macro level. As in the first section, we will again use Cornu's *Atlas* and al-Muqaddasī's book for testing the entire process. What we offer here allows for not only the comparison of specific geographical sources, but also an approach to computationally study differences, resemblances, and corresponding relations between complex geographical descriptions from textual sources.

5.2.1 Hierarchical Data

We will now introduce a model to compare hierarchical data, which can be seen as an abstraction of the description of administrative hierarchies. This model structures the hierarchical data and includes lower-level objects in the containers as a set of binary vectors. The parent-child relation in a tree structure for hierarchical data is represented by the binary values of a matrix. The binary values are the minimum information required to similarly define the relations and make them computationally comparable. Therefore, the number of all toponyms and all regions in two (or more) sources will be m and n, respectively. This model proposes a two-dimensional binary matrix $A \in 0, 1^{\mathrm{m.n}}$ with toponyms as matrix rows and regions as columns to represent hierarchical descriptions. This means, for a source with five provinces, the corresponding matrix will have ten columns to hold the provinces and the total number of toponyms (in all provinces) will be the number of rows. Therefore, a cell $a_{\mathrm{i,j}}$ of the matrix A represents inclusion or exclusion of the toponym i in the corresponding region j with 1 and 0 values, respectively (see Table 5.1a).

One can generate a matrix for both sources, meaning the number of columns is the total number of regions in both sources and the number of rows is the total number of toponyms in both sources, as the example in Table 5.1b shows. In this table, region$_1$ and region$_2$, which occur in both sources, are represented in two columns: one for each source (*reg$_1$-s$_1$* and *reg$_1$-s$_2$* for region$_1$ and *reg$_2$-s$_1$* and *reg$_2$-s$_2$* for region$_2$). Also, those toponyms that are not common in both sources are indexed to represent the corresponding sources. For instance, *topo$_1$*, *topo$_2$*, and *topo$_4$* are common, while *topo$_i$-s$_2$* and *topo$_j$-s$_1$* are only present in source$_2$ and source$_1$, respectively. Instead of one matrix for both sources, we have generated a matrix for each source and have used both matrices in calculations. The union of toponyms as row labels is the key point to make the matrix columns—which are regions—comparable. We will explain the details later in this section. We use a matrix for both sources in the current implementation.

	reg$_1$	reg$_2$	reg$_3$	reg$_4$
topo$_1$	1	0	0	0
topo$_2$	0	1	0	0
topo$_3$	0	1	0	0
topo$_4$	0	0	0	0
topo$_5$	1	0	1	1

(a) A matrix representing toponyms in regions with binary values in one source.

	reg$_1$-s$_1$	reg$_2$-s$_1$	reg$_1$-s$_2$	reg$_2$-s$_2$
topo$_1$	1	0	1	0
topo$_i$-s$_2$	0	0	0	1
topo$_j$-s$_1$	1	0	0	0
topo$_3$	0	1	0	1
topo$_4$	1	0	0	1

(b) A matrix representing regions and toponyms in two sources. Rows and columns are all the toponyms and regions in both sources, respectively.

Table 5.1: Examples of matrices for representing hierarchical data in a single source (a) and in two sources (b). Value 1 in a cell means the toponym in the corresponding row is in the region in the corresponding column and 0 means the opposite.

	reg$_1$	reg$_2$	subregion$_1$	subregion$_2$	subregion$_3$
toponym$_1$	1	0	1	0	0
toponym$_2$	0	1	0	1	0
subregion$_1$	1	0	0	0	1
subregion$_2$	0	1	0	0	0
subregion$_3$	1	0	0	0	0

Table 5.2: A matrix that shows a hierarchical division. In addition to the toponyms that are in regions (or subregions), it shows which subregions belong to higher level regions.

A matrix that thoroughly represents an administrative hierarchical tree will include all toponyms from various levels of the tree. Rows are all settlements and subregions that belong to a higher level region. Columns are all "container toponyms" including regions and subregions (all container toponyms). Each container toponym will have a column to represent included settlements and other subregions as well as a row to represent higher level regions in which they are included. As in Table 5.2 *subregion$_1$* and *subregion$_3$* are in *reg$_1$*, *subregion$_2$* is in *reg$_2$*, and *toponym$_1$* is in *subregion$_1$*. To use this structure, one should select two consecutive hierarchical levels in comparison, which we will describe later in this section.

Given two hierarchical datasets, we implement the above-mentioned comparison model by building a binary matrix of both datasets. Then, we mathematically compare matrix columns as pairs of vectors, where each vector is an abstract mathematical notation that we create from textual descriptions. As mentioned above, we only need the sub-matrix, which includes only toponyms from the same levels of corresponding hierarchical trees to compare descriptions of the same regions in different datasets. We can take any two consecutive levels from both hierarchical trees that correspond to each other. For example, in the hierarchical tree in Figure 5.23 we can compare either *provinces* and *subordinate regions* or *subordinate regions* and *settlements* of these trees. In other words, the model takes two container objects from two data sources, which are regions from any level in a hierarchical tree, and compares their divisions.

We populate row and column labels with toponyms and regions from both sources and initialize a matrix with 0 and 1 values. The union of toponyms and regions of both sources are computed for the row and column labels, respectively. This will produce extra rows and columns for the toponyms and regions of a source that are not mentioned in the other one. As an example, toponyms in *Region$_i$* in both sources A and B in Table 5.3a form vectors as in Table 5.3b. As in this table, the string values of toponymic names are then replaced by 0 and 1 values and the

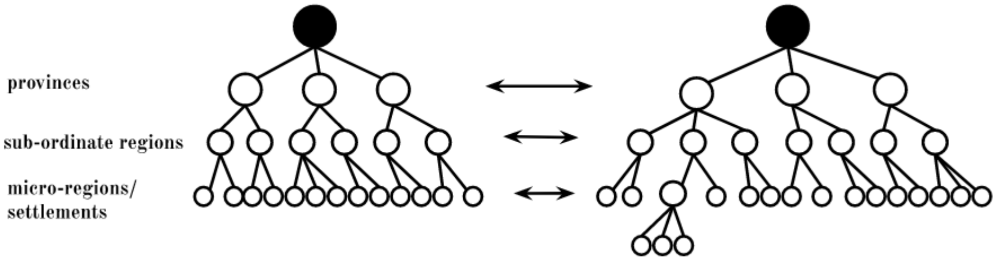


Figure 5.23: Comparing the hierarchical trees occurs at two consecutive levels to show how similarly or differently two higher-level regions (tree nodes) are described in two different sources.

Region$_i$ A	Region$_i$ B
toponym$_1$	toponym$_3$
toponym$_2$	toponym$_4$
toponym$_3$	

(a) Toponyms of a region in two different sources A and B

	Region$_i$ A	Region$_i$ B
toponym$_1$	1	0
toponym$_2$	1	0
toponym$_3$	1	1
toponym$_4$	0	1

(b) Vector structure of each region, holding all toponyms of both sources

Table 5.3: An example showing the creation of vectors for each region in a matrix structure for two sources. The comparable vertices have the same length.

comparison process will deal with binary values instead of words and strings, which considerably simplifies the comparison process.

The higher level nodes—the container regions/objects—in each comparison case (e.g., the *sub-ordinate regions* level in Figure 5.23 in comparing each region's subregions or settlements) should be an identical list for both datasets, as mentioned above, to make the comparison meaningful. This means that each comparison case shows how a certain level of divisions is described. If the lists of regions are not identical in the sources, a preprocessing step is required to prepare a list of provinces for the matrix. Differences in the list of regions might occur for various reasons: two sources might label the same geographical and administrative region differently; a region might not be in the hierarchical description of one source while described in another; or sources might describe the divisions of the same geographical area differently (for instance, an area is described as one region in one source while another source describes the same area with multiple regions). One can prepare such a list by mapping regions of both sources that map each region in the first source to zero, one, or multiple region(s) in the second source (or vice versa).

5.2.1.1 Test Data

As mentioned earlier, we have used this model to compare Cornu's *Atlas* with two levels of hierarchy and al-Muqaddasī's book with multiple levels of hierarchy. In order to make these two datasets comparable, we simplified the multi-level structure of the al-Muqaddasī's data by removing subregions (inner nodes) of the hierarchical tree and then connecting the higher level regions directly to the settlements. Figure 5.24 illustrates the tree nodes that we have eliminated. In this implementation, this means that we converted the input file, which has the whole tree data structure, into a CSV file with three columns: "province," "type," and "settlement". Then, we created a dictionary of containers (here, provinces) in which lower level toponyms (here, settlements) are located. This modification gives a tree structure of depth three for both datasets.

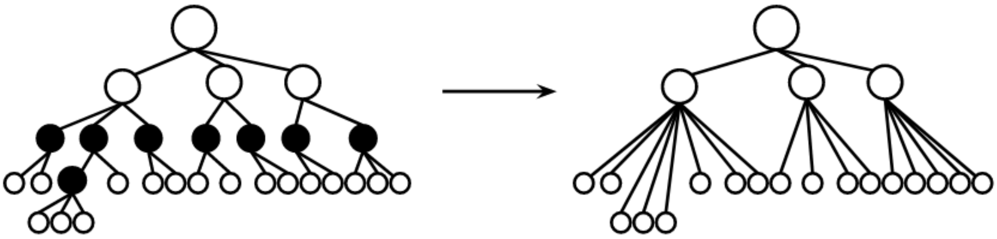


Figure 5.24: Removing the inner nodes (subregions) to simplify the hierarchical data and create corresponding levels for comparison.

While preparing hierarchical data, we need to make sure that geographical entities are consistently identified in both datasets. The current implementation assumes that all the toponyms are normalized and uniformly identified, which is done in the pre-processing step. This step is different for various datasets, depending on the language used in datasets and required normalization. For example, various spellings of the same toponyms and multiple names for the same place need to be normalized and/or unified[6]. The map of regions will be the column labels and cell values are then populated using both the regions and toponyms list.

We stored the data from Cornu as a JSON file from which we prepared a dictionary of regions containing toponyms. Similarly, we prepared the al-Muqaddasī data using the type column in the above-mentioned CSV file. Then, we create a dictionary for each region containing all the toponyms as dictionary keys assign 1 and 0 values to the keys according to their inclusion/exclusion in the corresponding region. The result can be a matrix for both sources or separate matrices for each source, depending on the structure of the input data. Figure 5.25 is an example of a column vector in a matrix that we created for the provinces of Miṣr (Egypt) and al-ᶜIrāq from the al-Muqaddasī data: settlements with value 1 belong to the corresponding province and those with value 0, do not. As mentioned earlier, in a matrix holding both sources, for each common region in either of the sources a separate column is allocated. For example, Miṣr (Egypt) is a province in both sources in our dataset and two columns point to this province: one to Miṣr (Egypt) in Cornu's *Atlas* and the other to Miṣr (Egypt) in al-Muqaddasī's book.

Using the matrix of top-level provinces, researchers will now be able to undertake one-to-one or one-to-many mapping of the corresponding provinces. Figure 5.26 depicts the comparison between the provincial divisions in both sources, which we have built from the comparison matrix created above. The inner circle shows provinces in Cornu and the outer one corresponds to al-Muqaddasī. This comparison depicts how two groups of provinces, al-Mašrīq and al-Maġrib, are represented in al-Muqaddasī as macro-regions. Al-Muqaddasī considers Ḫurāsān, Sijistān, and Mā warāᵓ al-nahr (Transoxiana) of Cornu's *Atlas* as al-Mašriq.

This type of comparison and visualization (Figure 5.26) is meaningful for provincial level description, where the number of toponyms are limited and data preparation is scalable. However, it is not suitable for comparing hierarchical data at the toponym level. Therefore, we have offered another application for comparison of sources at the lower levels of hierarchical trees (e.g., set-

[6] In our experience, the preparation step, even when done manually, might not be perfect. For example, there might be some cases that we miss or we cannot recognize the multiple names of a place. This comparison model can be used as a normalization/verification tool to identify what is missing, make corrections, and then repeat the whole comparison process. This underscores the iterative nature of the algorithm where the model, which is based on the data, helps to study the context and improve data consistency. Particularly, development of the theoretical approaches helps data preparation and, conversely, the models evolve as the data improves.

	al-^{c}Irāq	Miṣr
تاغليسية	0	0
بورة	0	0
محلة كرمين	0	1
بغداد	1	0
فغرسين	0	0
ده نوجيكت	0	0
أيوانة	1	0
سدر	0	1
الواردة	1	0
سامراء	1	1

Figure 5.25: A two-dimensional matrix structure for hierarchical data for Miṣr (Egypt) and al-^{c}Irāq provinces in al-Muqaddasī's book. The data partially shows two columns in the matrix and the corresponding rows dedicated to the regions. Settlements are the row labels and the provinces are the column labels and 1 and 0 values specify if a settlements is in the corresponding province or not, respectively.

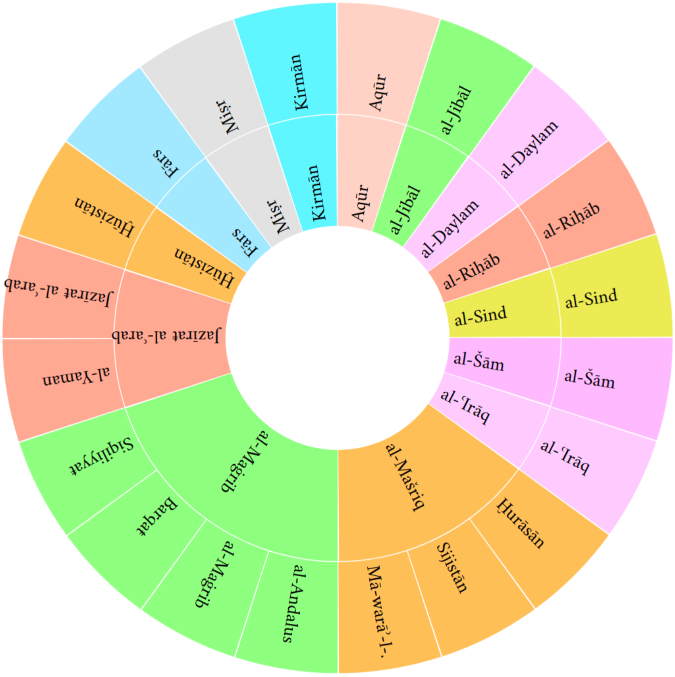


Figure 5.26: Provincial divisions in Cornu's *Atlas* (outer circle) and al-Muqaddasī's book (inner circle). As it can be seen, both sources often describe the same region, such as al-^{c}Irāq, al-Jibāl, and al-Šām. Cornu (outer slices) sometimes represents a single region in al-Muqaddasī's book as multiple regions and never has a region that overlaps two regions in al-Muqaddasī's book.

tlements) in Figure 5.27a. The underlying matrix of this implementation explains the provinces composed of settlements and based upon that we develop the one-to-one comparison of the regions. This application depicts how similar or different are Cornu's *Atlas* and al-Muqaddasī's book in describing settlements in each region by showing absolute numbers of common and different toponyms for each source. Each group of three bars in this figure corresponds to a region or a set of regions. The red and blue bars indicate the number of toponyms mentioned for the corresponding region in each source. The green bars show the number of toponyms that both sources describe in the corresponding region, stating the number of common toponyms. It should be noted that in this example, the number of toponyms from Cornu's *Atlas* is $2,175$ while we have only $1,090$ toponyms from the annotated parts of al-Muqaddasī's book[7], thus, makes it hard to draw any concrete comparative conclusion. However, this example shows the applicability of the model. Figure 5.27b illustrates a normalized view of this data, as a complementary visualization.

This model, as discussed above, offers a new perspective on the geographical source case studies and an effective approach to develop comparative workflows of hierarchical divisions data. It can also work as a general model for comparison of multiple sources. In this case, assume we have $n > 2$ number of sources and each has multiple regions R and each region has a set of toponyms T. The process of generating the comparison matrix of regions and toponyms starts with preparing the list of toponyms of all regions and mapping the regions. The mapping step creates a map of regions of the same administrative divisions from different sources in order to make a comparable list of regions. As mentioned before, a region in a source might map to zero, one, or multiple regions in another source, as described below:

- If a specific geographical area is not described in one of the sources, then we can either exclude it from the comparison process or include it in the process, but all the row values for that column in the comparison matrix will get the value of 0. It then will show zero overlap of those regions in the comparison, which is a way of representing the difference in describing an area in both sources.

- For the regions in one source that map to exactly one region in the other source, we yield a one-to-one mapping of the corresponding regions. This case is straightforward, as discussed above. The only complication in this case could be differences in the names of corresponding regions, which should be removed by normalizing and mapping the names.

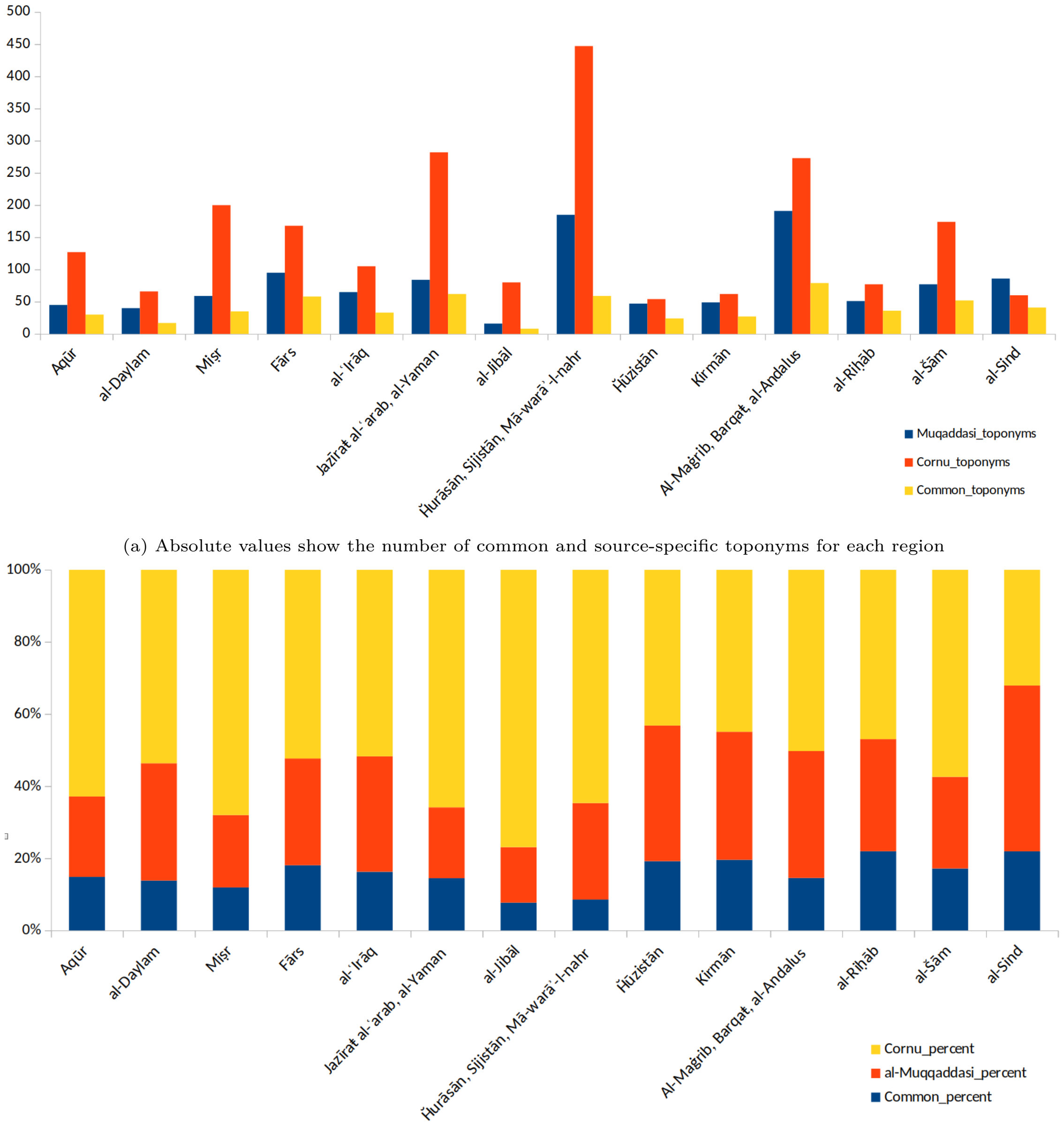
- The other case is that one (or multiple) region(s) in one of the sources correspond to multiple (or a different number of) regions in the other sources at the same level of the hierarchy. This can be resolved by mapping the region(s) in one source to the region(s) in the other source so that each mapped case describes the similar geographical and administrative area. This will create a one-to-many or many-to-many mapping of the regions.

The preparation process, as described above, can then be followed by creating the matrix structure, as described, and applying the comparative computations and visualizations.

5.2.2 Route Networks

In this section we will introduce a model to assist with the comparison of route network descriptions. This model converts curved paths in a network to more straightened paths which will allow

[7] The annotated file is available at: `https://raw.githubusercontent.com/OpenITI/0400AH/master/data/0390Muqaddasi/0390Muqaddasi.AhsanTaqasim/0390Muqaddasi.AhsanTaqasim.MSG20191024-ara1`.



(a) Absolute values show the number of common and source-specific toponyms for each region

(b) Normalized view, shows percentages of common and source-specific toponyms for each source

Figure 5.27: Application of the model for comparing hierarchical data. This chart represents the absolute number of common and different toponyms that Cornu's *Atlas* ([130]) and annotated parts of al-Muqaddasī's book ([212, 213]) describe for each region, with 2, 157 and 1, 090 toponyms from each source, respectively.

for the production of less complicated network visualizations. The benefit of this is that simplified networks make it much easier to compare different networks with each other.

This model is based on the assumption that descriptions of routes between major cities in both sources are similar. For instance, both Cornu and al-Muqaddasī describe the paths that connect the capital cities in Islamic provinces. Although the detailed descriptions of routes between destination points are not the same, there are still certain similarities in descriptions, which our model can use to simplify both networks and align them with each other. As just mentioned, simplified versions of networks can be used for visual comparison of networks in order to identify similarities and differences between them, which otherwise are not easy to spot.

Simplification in the context of this model focuses on the visual aspect. The idea, as described earlier, is to reshape the network so that its edges (route sections) appear in less curved shapes, while the routes still represent the same connectivity as the original network. This model, in fact, uses route sections and their lengths together with the positions of nodes in a network to redraw curved paths into straighter paths by repositioning the nodes along the paths. Despite the changes that occur in the spatial positions of the nodes while reshaping the network, the model is designed to preserve the proportional distances as a fundamental property of the route network and to produce a scaled version of the network. Unlike line simplification algorithms ([238, 239, 240]) this model does not reduce or eliminate the nodes, but rather employs a different method of polygonal approximation of the simplified paths in a network; however, a general analogy between the core ideas reveals similar objectives shared by this model and the line simplification algorithms.

In graph theory, the degree of a vertex (node) of graph $G = (V, E)$, where V is the set of vertices and E is the set of (undirected) edges, is the number of incident edges to that vertex. In the example in Figure 5.30, vertices u and v have the degree of 3 and 2, respectively. Given graph G, the incident edge is defined as below:

> Let $u, v \in V$ be vertices of G. Let $e = u, v \in E$ be an edge of G. Then, $e = u, v$ is incident to u and v, or joins u and v. Similarly, u and v are incident to e, as Figure 5.28 illustrates.

The implementation of this model starts with reconstructing the paths between the high-degree nodes[8] in a network. The rationale is that such nodes are often major centers or settlements and are usually mentioned in various geographical descriptions as are the paths connecting them to other centers or settlements. For example, Baġdād, Dimašq, and Fusṭāṭ are all high-degree nodes in both Cornu's and al-Muqaddasī's route networks. In addition, high-degree nodes, because of their centrality, are often identified by the geographical coordinates on which we base our calculations.

The idea is to redraw the paths between the centers on a straight line and repeat this process until we cover as many network nodes as possible. In other words, we reconstructed the paths by putting the locations along the path on a straight line from the source to destination (see Figure 5.29). We did this by placing the geographical coordinates and bearing of the nodes together with a path that connects them. We have assumed that the distance of the route sections are given. Since these paths partially or completely—depending on the network data—shape the network, a simplified form of the paths will shape the simplified network.

[8] To implement the algorithm, we started with nodes that have the highest degree and then, according to the number of the nodes that the algorithm can generate with those initial nodes, we chose lower degrees. This threshold may vary based on the characteristics of the network, minimum and maximum degree of the network graph, and number of vertices with minimum or maximum degree. One can try the algorithm with various threshold values and see which value gives the best result by analyzing the resulting graph. The aim is to traverse the graph so that the simplified graph holds all the vertices and edges of the original.

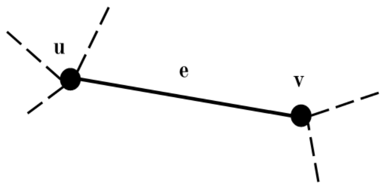


Figure 5.28: Incident edge in a graph. Edge $e = u, v$ is incident to u and v, or joins u and v. Similarly, u and v are incident to e.

The process starts with the high-degree nodes in a network to find all the existing paths between them. Using the exact coordinates of the high-degree nodes (i.e., source and destination), we calculated the shortest geographical distance[9] between the source and the destination, which is the length of the straight line connecting them. Thus, we generated a hypothetical line between the source and the destination on which we can then pin places along the path. To achieve this, we need to know how far we should move from the current location toward the destination to reach the new location of the next place along the path. To calculate this, we assumed that there is a line connecting n_1 to n_2 with length D_d, as in Figure 5.29 (Left). The proportional distance of point n_1 from P_1, d'_1, on the straight line is calculated in 5.2:

$$d'_1 = \frac{d_1 \times D_d}{D_o}, \tag{5.2}$$

where d_1 is the original distance of the route section from P_1 to d_1 and D_o if the length of the original path from P_1 to P_2, which is sum of the length of all route sections along this path. Then, we calculate the geographical bearing (navigation)[10] (Figure 5.30) of P_2 (destination) from P_1 (source) to know towards which direction we should move to find a new coordinate of n_1 at distance d'_1.

Now, we have derived all the required values to calculate the new location of (n_1) along the path on the above-mentioned hypothetical line: coordinates of the current point P_1, the distance and bearing from P_1. This process repeats every time that new coordinates are attained by replacing the previous location, P_1 in the above example, with the newly calculated coordinates (n_1) to find the next location along the path (n_2) until it reaches the destination (P_2). The details of the process are given below:

1. Choose the source and destination.
2. Find a path from the source to the destination.
3. Find the source and destination coordinates.

[9] Geographical distance is measured along the surface of the earth and can be based on flat, spherical, and ellipsoidal abstractions of that surface. The flat surface measurement may be useful over small distances and the accuracy of the measurement decreases with greater separation of the points and in areas close to the geographic poles. Spherical and particularly ellipsoidal approximations with less error rate. However, calculating the exact distance is unattainable if one attempts to account for every irregularity on the surface of the earth. Here, we have used The Great Circle distance ([241, p. 322-326]) in the calculations with almost 0.5% error rate (not more than 0.5% for latitude and 0.2% for longitude ([242, p. 10])). There are various methods and formulations to calculate this distance ([243]), from which we have derived the following formulation:
Let ϕ_1, λ_1 and ϕ_2, λ_2 be the geographical latitude and longitude in *radians* of two points 1 and 2, $\Delta\phi$, $\Delta\lambda$ be their absolute differences, and R the radius of Earth; then the distance can be computed as

$$2R \cdot \arcsin\sqrt{\sin^2(\frac{\delta\phi}{2}) + \cos\phi_1 \cdot \cos\phi_2 \cdot \sin^2(\frac{\delta\lambda}{2})}.$$

[10] Bearing is the measurement of direction, with reference to North, towards which we move to reach a geographical location form the current location. It is expressed in miles or degree that specifies the horizontal angle clockwise to North between two locations or objects. The direction in calculating the bearing is crucial. It means the bearing of location A to location B is completely different from location B to A. Thus, we have to set the source and destination carefully in calculating the bearing (see Figure 5.30). Otherwise, we may locate the places in wrong geographical direction.

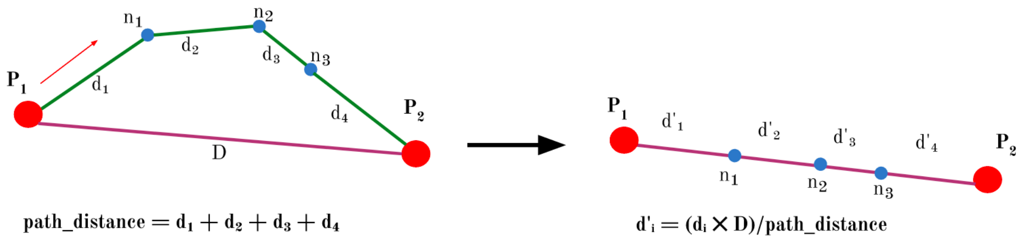

n1
n2
d2
d3
n3
P1
d1
d4
P2
D
path_distance = d1 + d2 + d3 + d4
P1
d'1
d'2
d'3
d'4
P2
n1
n2
n3
d'i = (di × D)/path_distance

Figure 5.29: Calculating new distances for lower-degree nodes based on the direct distance between two higher-degree nodes (D) and original distances mentioned in the source (d_i). (Left): An arbitrary path in a network; (Right): Simplified version of the path on a straight line. The original distances (d_i) on the original path are converted to new distances ($d\ _i$) on the straight line with the same proportions.

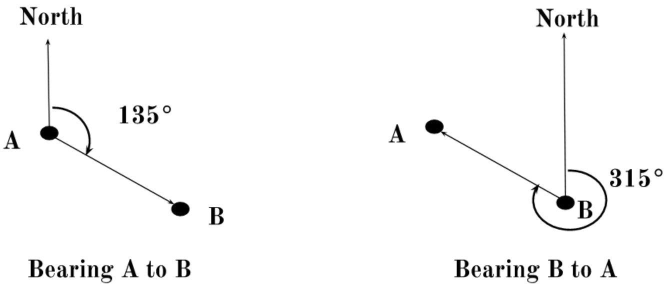


Figure 5.30: Bearing shows the direction from one location to another, as an angle measured clockwise from the North direction. (Left): Bearing of B from A; (Right): Bearing of A from B.

4. Calculate the direct distance, which is the length of the connecting direct line (Euclidean distance).

5. Calculate the geographical bearing/navigation between the source and destination.

6. Calculate the length of the path by accumulating the length of the individual edges from the source to destination in the distance in the network.

7. If the next (neighbor) node on the path is not yet processed—source and destination nodes are recognized as processed nodes—perform the following steps:

   (a) Calculate the distance to the next node in the path: if the previous node(s) has/have been skipped because they had been processed in previous iterations, the calculation must consider the skipped distances. Otherwise, the distance needs to be calculated between the previous node to the current node.

   (b) Calculate the proportional distance to the next node on the direct line from source to destination according to the original distance between the node and the next (neighbor) node on the path and length of the direct line.

   (c) Calculate the possible coordinate according to the computed proportional distance and the bearing.

   (d) Add the node to the list of processed nodes.

8. If the node is already processed:

   (a) Update the distance variable by adding the distance of the current node.

   (b) Move to the next node in the path and repeat 7 until the destination is reached.

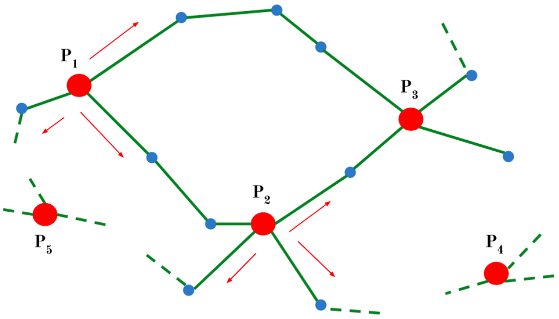


Figure 5.31: Traversing paths between higher-degree nodes (in red) to meet lower-degree nodes (in blue).

9. Start the next iteration from 2 until all possible paths are traversed.

10. Iterate over the list of the initial node list in 1 to process all the high-degree nodes.

Figure 5.29 illustrates the calculation of proportional distances and the new coordinates for lower-degree nodes based on the graph traversal (Figure 5.31). The traversal of all the paths between higher degree nodes meets lower degree nodes. One can set a minimum value for the node degree.

5.2.2.1 Post-processing

The above process traverses all the possible paths between the high-degree nodes to position the nodes in the first iteration. However, it might not cover the whole network since, depending on the network, it might leave the paths connected to the low-degree nodes non-traversed. Thus, it requires a post-processing step to identify the nodes that are not yet processed. This step is meant to populate the network with as many missing nodes as possible. Keeping in mind that a node without exact coordinates in the network might alter the process, the underlying premise of the model assumes that all (source/destination) nodes have coordinates. However, we only need source and destination coordinates in each path. Then, we can discuss modifications to the model that should be considered if a network includes nodes without coordinates. Below are the steps to perform:

1. Gather the list of unprocessed nodes.

2. Order the list based on the ascending node degrees; thus the lowest-degree nodes, namely of degree 1, will be processed first. This is done to traverse the longer paths, which process more nodes along the way. Therefore, the path traversal iterations and thus the whole process will be shorter. Besides, traversing a longer path once puts the nodes on straighter lines than traversing the same path in smaller subordinate paths, which constitute the longer path, in multiple iterations.

3. Find all the paths from high-degree nodes to these nodes.

4. Follow steps 3 to 10 in the process described above.

5. Update the source and destination and repeat the iteration from step 3.

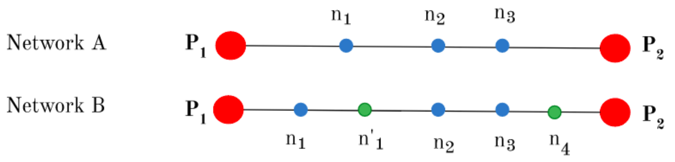


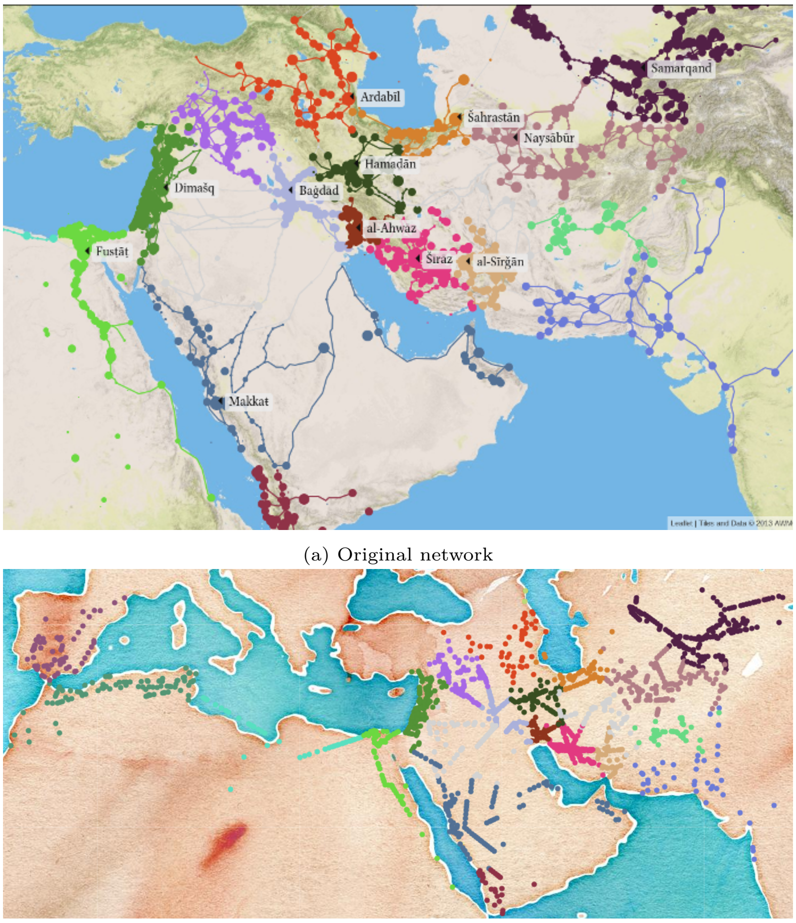
Figure 5.32: An example illustration of applying simplified networks in network comparison.

As mentioned above, this procedure might not thoroughly re-create the network if there are nodes without geographical coordinates at the source or destination positions in pathfinding, since the shortest distance and geographical bearing calculations require coordinates. This leads to missing the nodes along the paths that do not have coordinates for source and/or destination points. To minimize this outcome, we processed those paths by searching for the nearest node to the source/destination that has coordinates and by traversing the shortened path that ends at this node instead of the destination. As soon as the path is traversed up to the latest node with coordinates, the rest of the nodes along the path can be located based on the forthcoming distances as the path is being traversed at the same geographical bearing. This method, consequently, covers the nodes in a path along which at least two nodes have geographical coordinates. Despite the fact that the new coordinates might not be as accurate as the paths with source and destination coordinates, this re-creates a simplified version of a network that is intended to be used in comparisons rather than for locating the exact coordinates of places for which we do not have coordinates—here called *unknown place*s. In other words, the model is originally based on the approximate location of the places and proportional distances. For this reason, the above-mentioned issue does not affect the objective inasmuch as the model is able to create simplified versions of networks suitable for comparison.

5.2.2.2 Applications

This model, as mentioned, can be applied in comparative visualizations of various networks to represent similarities and differences. A simplified version of a network offers a more schematic view, but without the loss of major information. Figure 5.32 illustrates an example of visualization to compare part of the data from two networks. In this figure, points n_1, n_2, and n_3, occur on the path $\overline{P_1P_2}$ in Network A and the same points occur on the same path in Network B, but n_1 is in a different distance to P_1 and n_2. In addition, Network B has two more points on the same path: n'_1 and n_4. This example shows how the model allows one to eliminate all the unnecessary visual complexities and details to represent the similarities and differences. Figure 5.33 is a visualization of Cornu's *Atlas* route network in its original (Figure 5.33a) and in a simplified shape (Figure 5.33b).

Another application of this model is estimating the geographical location of *unknown place*s, but whose place in the network is specified. Since the simplification model relocates the places along the paths, all of these places will be assigned a coordinate, regardless of whether they have coordinates or not. Therefore, places that already have coordinates will be given a new coordinate and *unknown place*s will be assigned geographical coordinates. The coordinates, as mentioned above, are proportionally computed by using the distances between the *unknown place*s and their neighbors.



(a) Original network

(b) Simplified version

Figure 5.33: The route network of Cornu's *Atlas*

5.3 Unknown Places

The process of identifying the classical toponyms—after the extraction, matching, and disambiguation steps—often leaves a list of *unknown place*s, as first mentioned in Section 5.2.2.1. *Unknown place*s are those without any exact geographical locations, as there is no possible geographical reference to locate them on a map by means of straightforward automatic or manual approaches. Places that no longer exist and for which we cannot find any reference in historical gazetteers are examples of such places. *Unknown place*s are significant from a research perspective because they leave gaps in data structures, models, and visualizations and may interrupt computational processes. However, some contextual information in the textual description of these places could be used to identify them. In this section, we will introduce another model[11] to approximate geographical locations for *unknown place*s based on the underlying route network.

Depending on the contextual information and the purpose of the study, the approach toward identifying *unknown place*s may vary. For example, in archaeological studies, where the result of mapping will be used in excavations, the approach should aim to obtain precise coordinates. ([244]) introduces an approach to find the exact location of sites mentioned on a tablet to be used in historical and archaeological investigations. This approach takes into account the mere mention of two toponyms on the same tablet as the contextual information for *unknown place*s. It then defines a relation between them where there is no itinerary. Tobler assumes that places that are frequently mentioned together are geographically closer than places that are not frequently mentioned together. Hence, he fills the gap of geographical distance by calculating the social interaction between the places using the gravity model ([245]).

Distance is a basic geographical relation. The model we have introduced here adopts the distances and existing coordinates in a network to estimate the geographical coordinates. In other words, it utilizes the existing quantitative information, which has been already extracted and is applicable to similar datasets. It is applicable to the places with a measured connection to one or more places in the underlying network or itinerary data.

There are various cases of connections between the unknown and coordinated places in a network. The first case is that the toponym in question is connected to only one coordinated toponym. Figure 5.34 (Left) illustrates this case, where u_1, an *unknown place*, is only identified by a connection of length d_1 to p_1, an exact location. In this case, u_1 could be any point on the circle $C(p_1, d_1)$, as Figure 5.34 (Right) depicts.

The second case is when an *unknown place* is connected to at least two other coordinated places. Suppose we have two points, p_1 and p_2, at d_1 and d_1 distance from a third point, an unknown point, respectively. Accordingly, the possible points at which the unknown point could be located are the geometric intersection points of two circles, one with center p_1 and radius d_1, $C(p_1, d_1)$, and the other with center p_2 and radius d_2, $C(p_1, d_1)$. Figure 5.35 (Left) shows a case in which the unknown point u_1 is connected to two identified places p_1 and p_2. In Figure 5.35 (Right) the intersection of the two circles $C(p_1, d_1)$ and $C(p_2, d_2)$ represents the possible positions of u_1.

We test this approach on a sample dataset from al-Muqaddasī's route networks, implementing Algorithm 5. The algorithm creates a graph of nodes (Line 2) and then searches for the nodes without coordinates that have more than one neighbor with exact locations to start the compu-

[11] The first model is discussed in Section 5.2.2.2, where we introduced the application of network simplification model.

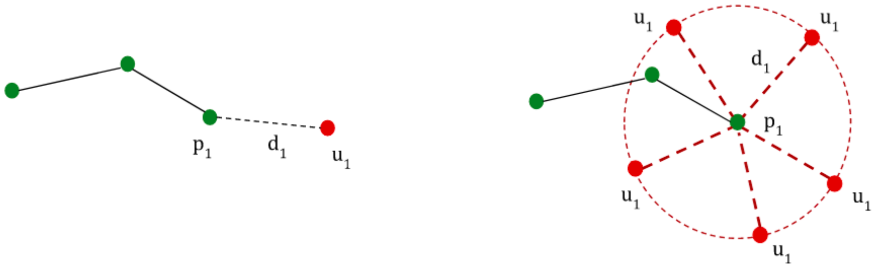


Figure 5.34: (Left): *Unknown place* u_1 (in red) is defined by a single connection to the identified location p_1 at the certain distance d_1; (Right): All the points on the circumference on the circle $C(p_1, d_1)$ represent potential locations of u_1.

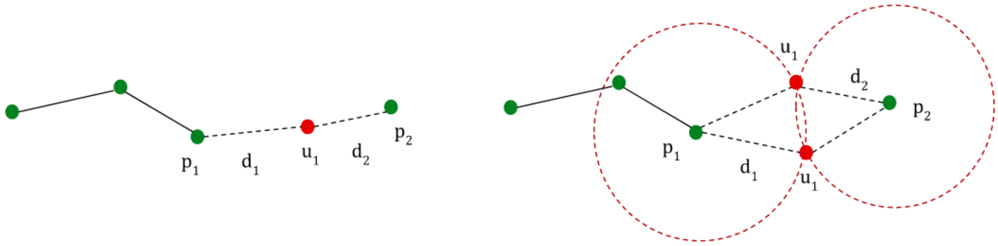


Figure 5.35: (Left): *Unknown place* u_1 (in red) is defined by two connections to the identified locations p_1 and p_2 at the distances d_1 and d_2, respectively; (Right): Two intersection points of the circles $C(p_1, d_1)$ and $C(p_2, d_2)$ are the possible locations of u_1.

tation (Lines 6 to 9). These nodes have "null" values for the coordinate property in the graph g. Having the coordinates of their first two closest neighbors, the algorithm projects them using the predefined projection coordinate system[12] (see Figure 5.36). We use *epsg:26915* projection system in the example (Line 3). Then, it draws a circle with a radius of the distance from the neighbors and intersects the circles (Lines 22 and 23). If the circles intersect, it re-projects the intersection points back to the *WGS* system,[13], which yields the coordinates on earth (Lines 24 to 26). This obtains a pair of coordinates for each place of null coordinates. Then, a quick calculation for finding the coordinates on the middle of the line that connects the two new coordinates is done to assign the new coordinates to the unknown node in the graph as a property (Line 27).

The sample dataset has 1,146 nodes, according to the annotated data that we have generated, of which 447 nodes are unknown. The number of nodes that are connected to zero, one, two, three, and four coordinated neighbors are 110, 212, 112, 10, and 3, respectively. We ran the algorithm on the nodes that have two neighbors, which yields new coordinates for more than 43% of the *unknown place*s (49 new coordinates out of 112 *unknown place*s). Figure 5.37 visualizes the result of the implementation in which known places are represented in green and new coordinates for the *unknown place*s are represented by red circles. Figure 5.38 represents a closer view on part of

[12] A Projected Coordinate System (PCS) is based on a Geographic Coordinate System (GCS) and is defined on a flat, two-dimensional plane. It is used to translate the physical objects on the three-dimensional spherical and curved surface of earth to a two-dimensional surface (see Figure 5.36). The translation is commonly called "projection" and produces constant lengths, angles, and areas across the two dimensions. Ellipsoidal or spherical surfaces work best to represent and approximate the surface of earth. These surfaces are then formed into cylinders, cones, or flat maps. From the mathematical point of view, it is the transformation of geodesic positions into rectangular grid (Cartesian) coordinates (x, y) values, called the x- and y-coordinate, respectively. The former represents the horizontal position and the latter is the vertical position of an entity relative to the center of a grid. The center of the grid is at the $(0, 0)$ position.

[13] *WGS* is a Geographic Coordinate System (GCS). A GCS defines the locations on earth using a spherical three-dimensional surface that is used to specify the real world points. WGS belongs to this category.

Algorithm 5: This algorithm uses the triangulation approach to approximate the location of the *unknown places* that are connected to two places with exact coordinates in a route network. It yields two coordinates from which one can either choose one of them or use a custom function to specify a point between them as the new coordinates and assign it as a property to the node in the underlying network.

```
function findLostCoordinates ($routesFile$)
Input  : CSV file of route sections with source and destination coordinates of known
         places, and distances.
Output: A graph of routes including newly found coordinates for the unknown places.
$g \leftarrow createGraph(routesFile)$
$proj \leftarrow Proj(epsg : 26915)$
$found \leftarrow numberOfNullCoordinates$
$prevFound \leftarrow 0$
while $found \neq prevFound$ do
    $prevFound \leftarrow found$
    for $node\ in\ g$ do
        if $node\ has\ null\ coordinates\ and\ len(neighbors(node)) > 1$ then
            $neighbors[node] \leftarrow []$
            for $neighbor\ of\ node$ do
                if $neighbor\ has\ coordinates$ then
                    $neighbors[node].append(n)$
                end
            end
            $nei1 \leftarrow closestNeighbor[0]$
            $nei2 \leftarrow closestNeighbor[1]$
            $x1, y1 \leftarrow proj(nei1[coordinate])$
            $x2, y2 \leftarrow proj(nei2[coordinate])$
            $r1 \leftarrow distance(node, nei1)$
            $r2 \leftarrow distance(node, nei2)$
            $circle1 \leftarrow Point(x1, y1).buffer(Decimal(r1)).boundary$
            $circle2 \leftarrow Point(x2, y2).buffer(Decimal(r2)).boundary$
            if $circle1.intersects(circle2)$ then
                $newCoord1 \leftarrow proj(intersectionPoint1, inverse = True)$
                $newCoord2 \leftarrow proj(intersectionPoint2, inverse = True)$
                // Find the middle point between the new coordinates and set it as the new
                   coordinate of the node.
                $node[coordinates] \leftarrow findMiddlePoint(newCoord1, newCoord2)$
                $found \leftarrow found - 1$
            end
        end
    end
end
return $g$
```

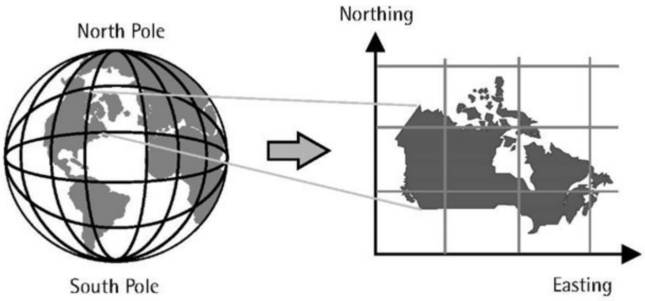


Figure 5.36: Map projection is a transformation of the entities on the spherical surface of earth to a two-dimensional plane (Image source: `http://what-when-how.com/gps/datums-coordinate-systems-and-map-projections-gps-part-2`).

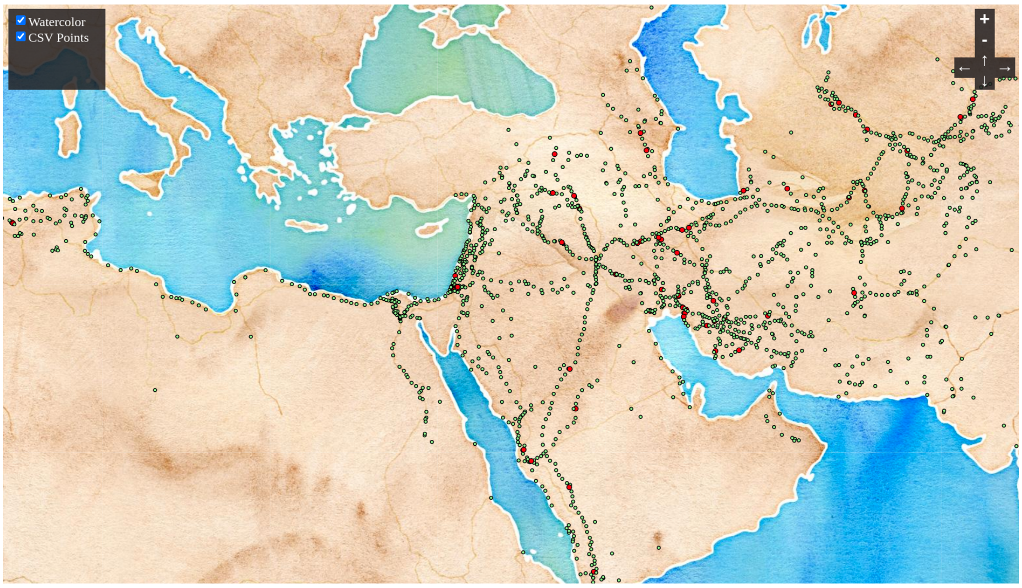


Figure 5.37: Result of the triangulation approach in Algorithm 5 tested on a sample dataset from al-Muqaddasī's book. The red points represent the new estimated locations for the *unknown place*s that are connected to two coordinated places in the network.

the data that is visualized on different map tiles with modern cities names available.

The third case is when an unknown point is connected to more than two coordinated locations; for example, to three points (Figure 5.39). The new information that the third connection adds is the lost piece of the puzzle by which we can find the exact coordinates of the *unknown place*. The premise in this approach is that the distances and all the calculations—such as converting classical distance values to modern values and all the geometric computations and projections—are precise enough to compute the coordinates. Nevertheless, if the circles do not intersect for any reason, such as the imprecise distances or projection of the coordinates, this approach may not be able to estimate the location.

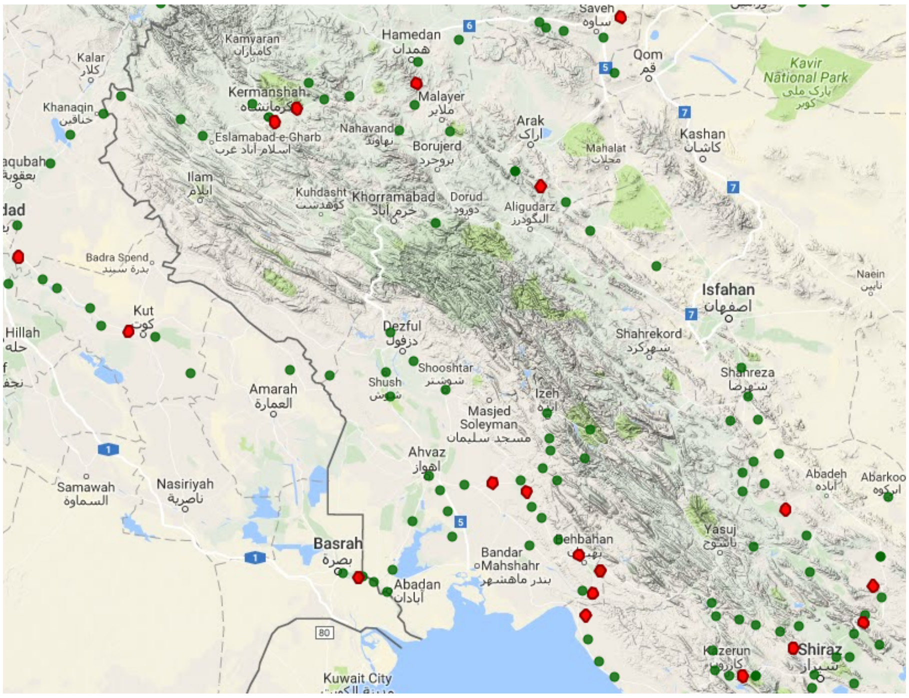


Figure 5.38: A closer view of a part of the map in Figure 5.37, visualized on different map tiles.

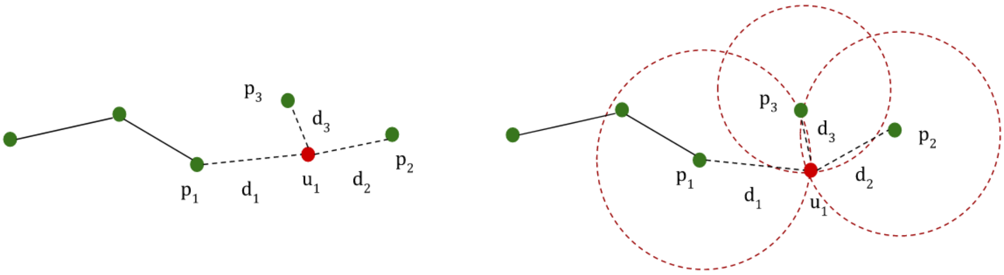


Figure 5.39: (Left): *Unknown place* u_1 (in red) is connected to three exact locations p_1, p_2, and p_3 at the distances d_1, d_2, and d_3, respectively; (Right): The intersection point of the circles $C(p_1, d_1)$, $C(p_2, d_2)$, and $C(p_3, d_3)$ is the possible location of u_1.

Chapter 6

Al-Ṯurayyā, a Gazetteer and a Spatial Model of the Classical Islamic World

In historical research, gazetteers play an important role which has not been largely acknowledged. Many historical places no longer exist and place names and spellings might have significantly changed through time. Digitization of historical texts and improvements in the algorithms used for extracting place names makes the enormous potential of indexing that data more feasible. Gazetteers provide an authoritative reference to geographical entities (not just "place names") stored in the form of stable URIs. They map spatial entities to real available geographical places and may also provide a classification of place types. Using maps, gazetteers also allow researchers to place people and events in spatial context. Modern geographical studies and investigations are often based on maps which make the gazetteers a complement to maps ([127]). An increasing interest in digital gazetteers in recent years has been building bridges between the geospatial web and the semantic web. This is the first reason and motivation of the growing interest in developing and using digital gazetteers, as ([127]) explains. More and more digital humanities research projects have been extensively using digital gazetteers as a backbone and an established resource.

In this chapter, we introduce a gazetteer, al-Ṯurayyā[1] ([9]), which also features a spatial model of the classical Islamic world, including over $2,000$ toponyms and route sections georeferenced from Cornu's *Atlas*. According to ([127]), since the information on the web is not necessarily geographical but semantic, the idea of linked data ([246]) proposes connecting related information. Al-Ṯurayyā, therefore, is also designed to facilitate collecting and engaging related geographical information by providing links to external sources.

[1] The name al-Ṯurayyā, "Pleiades" in Arabic, is a tribute to Pleiades Gazetteer (`https://pleiades.stoa.org/`), which was the main source of inspiration at the early stage of development of this project.

6.1 Introducing al-Ṯurayyā

Al-Ṯurayyā is the new version of the project[2], using Cornu's *Atlas* to represent geographical information of the classical Islamic world in different provinces from Spain and North Africa to India. The project's original aim was to develop an extensive gazetteer of the Islamic Empire where a primary attribute of collection objects is their location, represented by their geographical coordinates. Coordinates can be attached to any type of available information about a geographic location—any georeferenced piece of information. Providing the above-mentioned essential elements of a gazetteer entry, al-Ṯurayyā supports several functions to answer the "where is" question by providing contextual information and key factors as attestations, like administrative hierarchy of locations. The current implementation expands the gazetteer features and serves as a spatial model through which one can visualize geographical entities; that is, places, routes, itineraries, paths, and networks. The programmatic shell of the gazetteer can be reused with different data: that is, researchers can use their own data sources prepared according to al-Ṯurayyā's data model and benefit from the full functionality for that specific dataset. Data preparation, which has been done in the early version, has taken the contextual factors into account and represented them in the underlying models and algorithms.

Visualization of travel and movement allows one to explore the Islamic Empire from different perspectives. Additionally, one can perform pathfinding inquiries, such as searching for the shortest—whether the shortest path is always the desired one depends on the travel condition—and/or the shortest safe path to travel from one location to another. More abstract questions and aspects could be covered according to this model, such as visualization of reachability from a center in the entire Islamic Empire and territories of control. Al-Ṯurayyā combines simulation models for navigation with distance information from Cornu's *Atlas*. Its data model takes into account connectivity and spatial relations of entities, various types of settlements and divisions in classical geography, and additional evidence and references provided by the context. It is designed to serve as a starting point for visual representation and analysis of spatial data in the written sources, and as a tool for answering meaningful, complex research questions about the pre-modern world, its geographical shaping, and spatial mentality. Al-Ṯurayyā is also designed to be connected to available secondary sources related to the geographical entities as well as to primary sources.

The list of settlements and routes as well as the distances are produced during the digitization and georeferencing process of Cornu's *Atlas* using QGIS software[3]. Settlements are assigned a category that specifies a geographical/administrative type, such as capital, metropole, city, town, or village, and the routes are specified with the connecting points and length (lengths of path sections were calculated in QGIS). The data model also enables representation of multiple levels of hierarchical divisions, such as provinces, regions, settlements. In the current version of al-Ṯurayyā, the spatial model is limited to a static representation of pathfinding and network reachability.

6.2 Gazetteer

Creating a map with relevant geospatial details for the classical Islamic world requires an effort to design a proper data model and representation of the data, while for modern transportation

[2] The previous version, developed at Tufts University is [247]

[3] A free and open source Geographic Information System (`https://www.qgis.org/en/site`).

networks it is relatively straightforward by using modern Application Programming Interfaces (APIs) and Geographical Information Systems (GIS). Precise coordinates of places for primary use in al-Ṯurayyā are obtained by georeferencing from Cornu's *Atlas*. Each location is stored together with technical and historical information from the source atlas with links to contextual information in primary and secondary sources. Al-Ṯurayyā provides the required components of a gazetteer entry: a geographic name; a geographic location represented by coordinates; and a type designation. With these attributes, it can function as a tool for indirect spatial location identification through names and types. Below is a detailed description of the gazetteer's features:

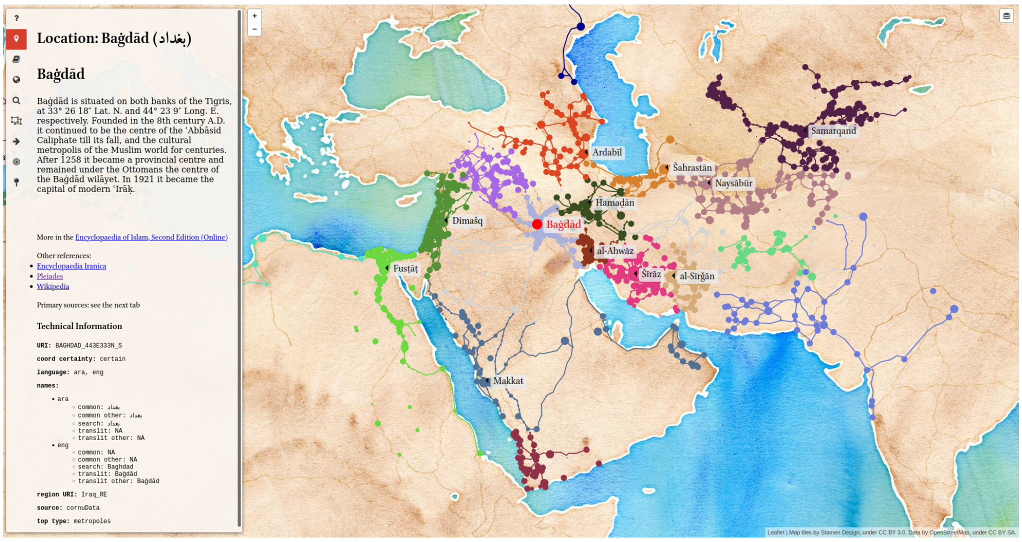
- Information of the places as in Figure 6.1:
  - Coordinates of the underlying geometry
  - URI of the current gazetteer entry
  - Coordinate certainty
  - The region URI to which the toponym in question belongs
  - The source, from which the technical data is provided
  - Settlement type according to the source, such as metropole, capital, town, and village
  - Names in English and Arabic, and their transliterated forms with the possibility to add other languages
  - Any additional available information and description in the data (e.g., Figure 6.1 shows a short passage about Baghdād together with the technical data.)
- Descriptions of the toponym in question from primary sources and links to the available secondary sources: al-Ṯurayyā facilitates inclusion of references to preprocessed entries from primary and secondary sources. The entries are stored and retrieved in separate data files. The reference information is then displayed together with other technical information on a toponym in question (Figure 6.2). In the current dataset[4], matching records from the primary sources are automatically matched using the *fuzzywuzzy*[5] *python* library. The certainty of the matching record is shown as a percentage value next to each item (see Figure 6.2b).
- Search a toponym in Arabic or transliterated form (see Figure 6.3)

6.3 Spatial Model

Building a spatial model requires more information than that which gazetteers usually provide. Routes that connect sites are the key elements for generating a route network, which can be used for simulating and visualizing travels and movements. Accordingly, al-Ṯurayyā offers a set of functionalities implemented in two main modules of the spatial model. The model then facilitates studying the space segmentation, administrative divisions, and various forms of pathfinding. One module represents paths and itineraries of various complexities and the other one models the flood network of one or multiple center(s). This section positions this spatial model as a tool

[4] The current data provides the description only from al-Ḥimyarī's *Rawḍ al-Miʿṭār*. However, any number of other sources can be added to the data model.

[5] `https://pypi.org/project/fuzzywuzzy`



Location: Baġdād (بغداد)

Baġdād

Baġdād is situated on both banks of the Tigris, at 33° 26 18″ Lat. N. and 44° 23 9″ Long. E. respectively. Founded in the 8th century A.D. it continued to be the centre of the ʿAbbāsid Caliphate till its fall, and the cultural metropolis of the Muslim world for centuries. After 1258 it became a provincial centre and remained under the Ottomans the centre of the Baġdād wilāyet. In 1921 it became the capital of modern ʿIrāḳ.

More in the Encyclopaedia of Islam, Second Edition (Online)

Other references:

- Encyclopaedia Iranica
- Pleiades
- Wikipedia

Primary sources: see the next tab

Technical Information

URI: BAGHDAD_443E333N_S

coord certainty: certain

language: ara, eng

names:

- ara
  - common: بغداد
  - common other: بغداد
  - search: بغداد
  - translit: NA
  - translit other: NA
- eng
  - common: NA
  - common other: NA
  - search: Baghdad
  - translit: Baġdād
  - translit other: Baġdād

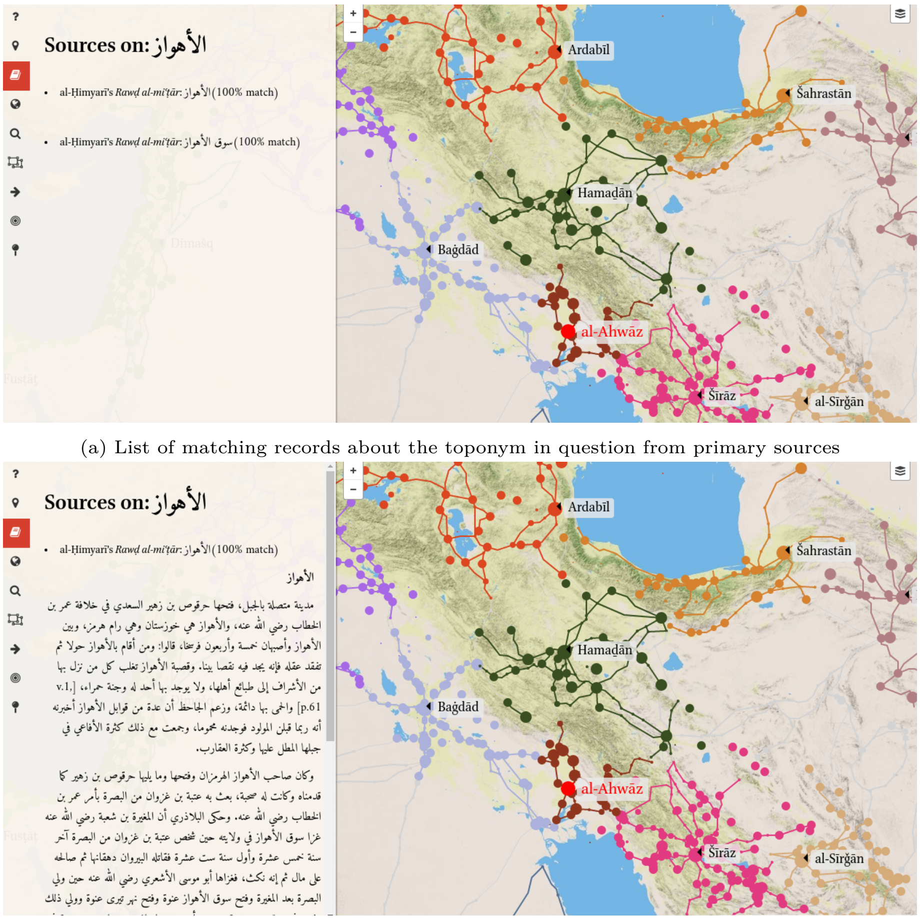
region URI: Iraq_RE

source: cornuData

top type: metropoles

Ardabil
Šahrastān
Samarqand
Naysābūr
Hamaḏān
Dimašq
Baġdād
al-Ahwāz
Fusṭāṭ
Šīrāz
al-Sīrǧān
Makkat
Leaflet | Map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under CC BY SA.

Figure 6.1: A place (Baġdād/Baghdad) is selected in al-Ṯurayyā and the available information of the place is shown on the left panel, including a short description of the place (at the top), links to the existing secondary sources, an the technical information (at the bottom).



Sources on: الأهواز

- al-Ḥimyarī's *Rawḍ al-miʿṭār*: الأهواز (100% match)
- al-Ḥimyarī's *Rawḍ al-miʿṭār*: سوق الأهواز (100% match)

(a) List of matching records about the toponym in question from primary sources

Sources on: الأهواز

- al-Ḥimyarī's *Rawḍ al-miʿṭār*: الأهواز (100% match)

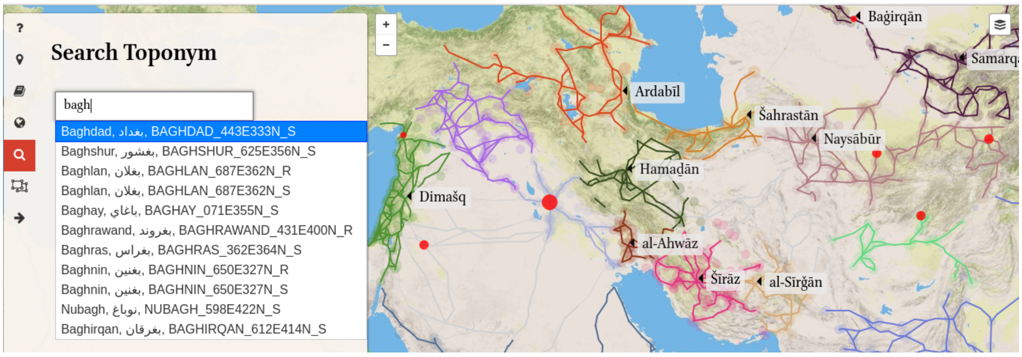
الأهواز

مدينة متصلة بالجبل، فتحها حرقوص بن زهير السعدي في خلافة عمر بن الخطاب رضي الله عنه، والأهواز هي خوزستان وهي رام هرمز، وبين الأهواز وأصبهان خمسة وأربعون فرسخا، قالوا: ومن أقام بالأهواز حولا ثم تفقد عقله فإنه يجد فيه نقصا بينا. وقصبة الأهواز تغلب كل من نزل بها من الأشراف إلى طبائع أهلها، ولا يوجد بها أحد له وجنة حمراء، [v.1, p.61] والحمى بها دائمة، وزعم الجاحظ أن عدة من قوابل الأهواز أخبرنه أنه ربما قبلن المولود فوجدنه محموما، وجمعت مع ذلك كثرة الأفاعي في جبلها المطل عليها وكثرة العقارب.

وكان صاحب الأهواز الهرمزان وفتحها وما يليها حرقوص بن زهير كما قدمناه وكانت له صحبة، بعث به عتبة بن غزوان من البصرة بأمر عمر بن الخطاب رضي الله عنه. وحكى البلاذري أن المغيرة بن شعبة رضي الله عنه غزا سوق الأهواز في ولايته حين شخص عتبة بن غزوان من البصرة آخر سنة خمس عشرة وأول سنة ست عشرة فقاتله البيروان دهقانها ثم صالحه على مال ثم إنه نكث، فغزاها أبو موسى الأشعري رضي الله عنه حين ولي البصرة بعد المغيرة وفتح سوق الأهواز عنوة وفتح نهر تيرى عنوة وولي ذلك

(b) Description from a source in expanded view

Figure 6.2: Entries from the primary sources matching a selected toponym in al-Ṯurayyā. (a) List of primary sources from which relevant information to the place is being shown; (b) Expanded view of one of the sources, showing the relevant content from that source.


Search Toponym

bagh

Baghdad, بغداد, BAGHDAD_443E333N_S
Baghshur, بغشور, BAGHSHUR_625E356N_S
Baghlan, بغلان, BAGHLAN_687E362N_R
Baghlan, بغلان, BAGHLAN_687E362N_S
Baghay, باغاي, BAGHAY_071E355N_S
Baghrawand, بغروند, BAGHRAWAND_431E400N_R
Baghras, بغراس, BAGHRAS_362E364N_S
Baghnin, بغنين, BAGHNIN_650E327N_R
Baghnin, بغنين, BAGHNIN_650E327N_S
Nubagh, نوباغ, NUBAGH_598E422N_S
Baghirqan, بغرقان, BAGHIRQAN_612E414N_S

Figure 6.3: Search panel and possible matches from the list of toponyms in al-Ṯurayyā

for answering meaningful and complex research questions about the classical Islamic world and a starting point for visual analysis of similar data.

We will explain this model through a general overview of features. The underlying data is prepared with regard to the historical evidence as described in ([248]). It contains sites of various types organized in different administrative divisions and routes, each specified with a length. Route networks and hierarchical geographical entities are complex objects modeled from atomic entities—sites and route sections.

6.3.1 Provinces and Administrative Divisions

As discussed in Section 4.2.1, administrative divisions build the relations between atomic entities (i.e., settlements), without introducing complex objects, such as polygons, in the dataset. In other words, we leveraged this relation to model divisions and depict them as individual higher-level geographical entities. Thus, places that belong to the same category of divisions form a region. Additionally, the route sections also represent another type of relation between places in the same region.

Regions in al-Ṯurayyā are represented in a map of regions/provinces using the model described in Section 5.1.7, where each region is visualized with a specific color. Thus, the places and routes in the same color illustrate the geographical position and administrative extent of a region (Figure 6.4). A list of regions is also populated on the left panel, from which the user can select a specific region to highlight and display an isolated view of a region (see Figure 6.5).

This feature is not only a simple visualization of the data, but also a model for administrative divisions. The current data specifies only one level of division, which is considered as the main provinces for a specific era. However, the data model can be used for more granular levels of hierarchical divisions with arbitrary types of macro- and micro-regions. This model suggests a method for representing regions which perfectly fits the existing data without causing any abnormalities that interfere with modeling and analyzing the real world phenomena and historical evidence (see Section 5.1.7).

6.3.2 Pathfinding and Itineraries

Route sections data underlies the route network upon which we implemented the pathfinding features to visually describe travels and movements. The features include finding the shortest and

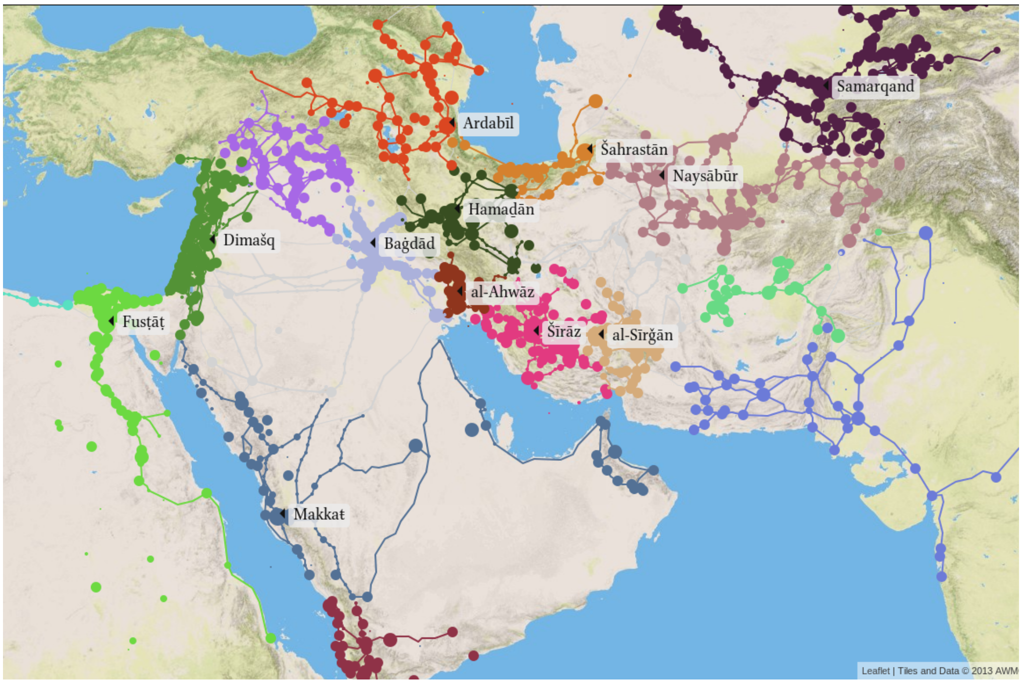
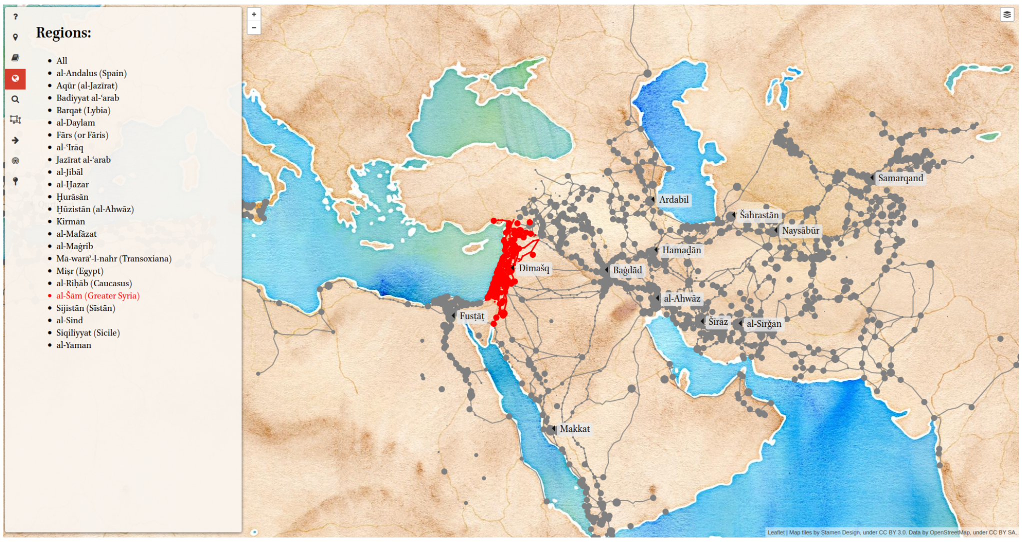


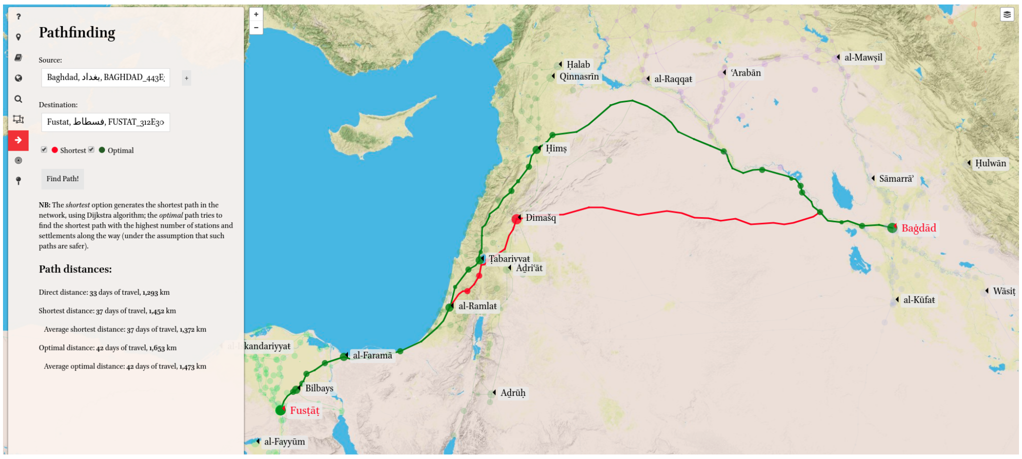
Figure 6.4: The map of regions in al-Ṯurayyā where each region is represented by a specific color.


Regions:

- All
- al-Andalus (Spain)
- Aqūr (al-Jazīrat)
- Badiyyat al-ʿarab
- Barqat (Lybia)
- al-Daylam
- Fārs (or Fāris)
- al-ʿIrāq
- Jazīrat al-ʿarab
- al-Jibāl
- al-Ḫazar
- Ḫurāsān
- Ḫūzistān (al-Ahwāz)
- Kirmān
- al-Mafāzat
- al-Maġrib
- Mā-warāʾ-l-nahr (Transoxiana)
- Miṣr (Egypt)
- al-Riḥāb (Caucasus)
- al-Šām (Greater Syria)
- Sijistān (Sīstān)
- al-Sind
- Siqiliyyat (Sicile)
- al-Yaman

Samarqand
Ardabīl
Šahrastān
Naysābūr
Hamaḏān
Dimašq
Baġdād
al-Ahwāz
Šīrāz
al-Sīrǧān
Fusṭāṭ
Makkat
Leaflet | Map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under CC BY SA.

Figure 6.5: The map of regions, now focused on the province al-Šām (Greater Syria) in al-Ṯurayyā, showing a user's selection (shown at the left and highlighted in red).



Pathfinding

Source:

Baghdad, بغداد, BAGHDAD_443E;

Destination:

Fustat, فسطاط, FUSTAT_312E30

☑ ● Shortest ☑ ● Optimal

Find Path!

NB: The *shortest* option generates the shortest path in the network, using Dijkstra algorithm; the *optimal* path tries to find the shortest path with the highest number of stations and settlements along the way (under the assumption that such paths are safer).

Path distances:

Direct distance: 33 days of travel, 1,293 km

Shortest distance: 37 days of travel, 1,452 km

Average shortest distance: 37 days of travel, 1,372 km

Optimal distance: 42 days of travel, 1,653 km

Average optimal distance: 42 days of travel, 1,473 km

Figure 6.6: Pathfinding feature in al-Ṯurayyā computing both the shortest (red) and optimal (green) paths from Baġdād (Baghdad) to Fusṭāṭ (Cairo).

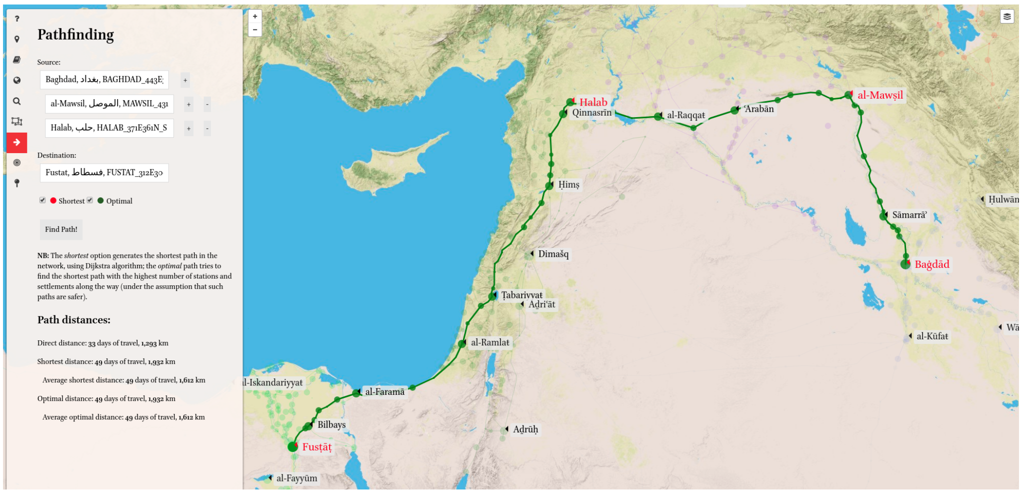
optimal paths between a source and a destination in the underlying route network as well as forming an itinerary with specified stops along the way. The shortest path is calculated and visualized by implementing Dijkstra's single source shortest path model ([249]). The implementation is also adjusted to contain a number of arbitrary stops on the way from the source to the destination that the user chooses to shape an itinerary.

The optimal path is a modified version of the shortest path where the constituent routes are not longer than a threshold value. It enforces the length of a day of travel as the threshold in the computations. This limitation is to avoid the routes that do not take the travelers to a way station within a day to overnight and survive the brutal conditions of travel and, thus, are considered "dangerous." Accordingly, the calculated optimal path has the highest number of stops along the way. This approach obtains a more realistic view of travel in the pre-modern period, where different paths of travel had to be considered and the facilities, infrastructures, and conditions limited the choices available.

The shortest-path algorithm constructs the path step by step by selecting the shortest possible route section to take at each step. The optimal path tweaks this algorithm and biases the selection of routes by engaging the threshold value. This value is in fact within-a-day travel distance, which has been calculated by [248] based on the descriptions of the routes and distances in relevant primary sources.

Figure 6.6 is an example of pathfinding in al-Ṯurayyā where the algorithm computes the shortest and optimal paths from Baġdād to Fusṭāṭ, colored in red and green respectively. In this example, the former takes travelers through the Syrian desert without considering any safety measures while the latter computes a "safer" shortest path. The optimal path avoids the risk of taking the longer routes through the desert with a smaller number of settlements along the way. It is also the shortest safe path, thus called *optimal*.

In addition to the pathfinding options, al-Ṯurayyā models more complex paths with arbitrary stops. This model finds and visualizes itineraries which pass through the stops (maximum ten) that users can specify. The implementation customizes the pathfinding computation model in a similar manner to the optimal path calculation with a number of stops. It chooses the shortest route section to take at each step so that the path goes through the stops. Figure 6.7 shows the same example of pathfinding where the user seeks a path from Baġdād to Fusṭāṭ (Cairo) along



Pathfinding

Source:

Baghdad, بغداد, BAGHDAD_443E; +

al-Mawsil, الموصل, MAWSIL_431 + -

Halab, حلب, HALAB_371E361N_S + -

Destination:

Fustat, فسطاط, FUSTAT_312E30

☑ ● Shortest ☑ ● Optimal

Find Path!

NB: The *shortest* option generates the shortest path in the network, using Dijkstra algorithm; the *optimal* path tries to find the shortest path with the highest number of stations and settlements along the way (under the assumption that such paths are safer).

Path distances:

Direct distance: 33 days of travel, 1,293 km

Shortest distance: 49 days of travel, 1,932 km

Average shortest distance: 49 days of travel, 1,612 km

Optimal distance: 49 days of travel, 1,932 km

Average optimal distance: 49 days of travel, 1,612 km

Figure 6.7: Plotting an itinerary from Baġdād to Fusṭāṭ (Cairo) via al-Mawṣil (Mosul) and Ḥalab (Aleppo) in al-Ṯurayyā.

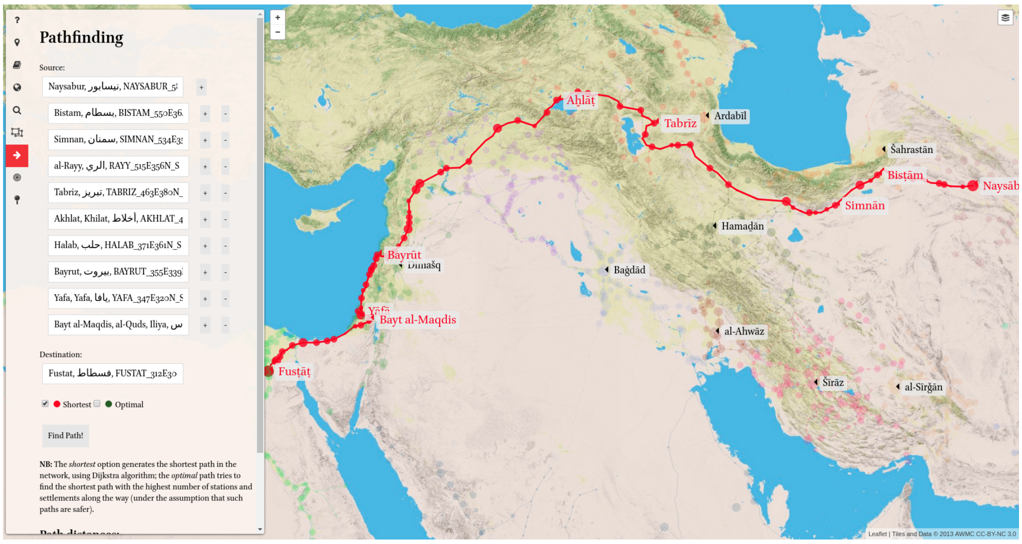
which one can stop at al-Mawṣil (Mosul) and Ḥalab (Aleppo). In another use case, Figure 6.8 plots the itinerary that Nāṣir-i Ḫusraw took from Naysābūr (Nishapur) to Fusṭāṭ (Cairo), as he describes in his *Book of Travels (Safar-nāmah)* ([215]). The itinerary runs through the places that he mentions (the ten places to stop, which is given as input on the left tab). Furthermore, it also suggests locations that he has not mentioned, but that he may have visited on his way.

The pathfinding feature also calculates the distance information of both shortest and optimal paths and average distance of each as well as the direct distance between the source and the destination, as can be seen in Figures 6.6 and 6.7.

6.3.3 Flood Network

Al-Ṯurayyā introduces a method to model the reachability in the underlying network of routes. The flood network forms a network of reachable settlements from a specified settlement. It shows how reachable the other places are from a center, and also shows the unreachable places. It underscores both reachability and limitations of reach from a selected center to find answers to abstract questions related to: the spread of power, unity, and reach of an empire in the historical context, such as seizing the territory of control after conquering a metropolitan center by a ruler.

Representing a flood network, compared to pathfinding, requires more complex graph analysis and visualization. The flood network of a given center (or a set of centers) shapes layers or categories of locations based on their distances from selected locations(s). Each layer represents the same distance or level of accessibility from the center to the locations in the layer. More clearly, each layer is a category of places that can be reached within the same number of days [of travel]. The calculated flood network represents how travel was possible from a central place to other parts of the network. For example, Figure 6.9 demonstrates the flood network of Baġdād that is reachable within five days' travel. Each layer in the figure is displayed in a particular color to highlight the corresponding area in which all places lie the same number of days of travel away from Baġdād. The geographical extent of each layer is determined by the value of the number of days to travel in each layer that the user specifies. Therefore, the number of days to reach the places in each layer from the center is calculated by this value multiplied by the number of layers



Pathfinding

Source:

Naysabur, نيسابور, NAYSABUR_5[+

Bistam, بسطام, BISTAM_550E36. + -

Simnan, سمنان, SIMNAN_534E3[+ -

al-Rayy, الري, RAYY_515E356N_S + -

Tabriz, تبريز, TABRIZ_463E380N_ + -

Akhlat, Khilat, أخلاط, AKHLAT_4 + -

Halab, حلب, HALAB_371E361N_S + -

Bayrut, بيروت, BAYRUT_355E339[+ -

Yafa, Yafa, يافا, YAFA_347E320N_S + -

Bayt al-Maqdis, al-Quds, Iliya, س + -

Destination:

Fustat, فسطاط, FUSTAT_312E30

☑ ● Shortest □ ● Optimal

Find Path!

NB: The *shortest* option generates the shortest path in the network, using Dijkstra algorithm; the *optimal* path tries to find the shortest path with the highest number of stations and settlements along the way (under the assumption that such paths are safer).

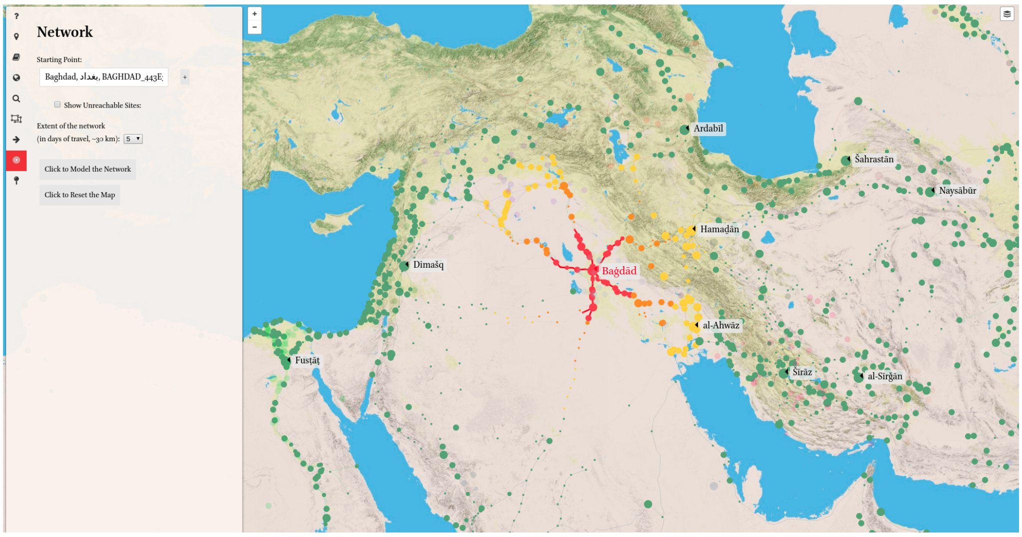
Figure 6.8: Plotting the itinerary from Naysābūr (Nishapur) to Fusṭāṭ (Cairo) from Nāṣir Ḫusraw's *Book of Travels (Safar-nāmah)* in al-Ṯurayyā.

closer to the center. For instance, if this value is set to two days, it means that places in the closest layer to the center are reachable within five days, and the next farthest layers are reachable within four, six, and eight days respectively. As Figure 6.9 depicts, locations in red, orange, yellow, and green are reachable within five, ten, and twenty days, respectively. The pale-colored points are unreachable locations; that is, they are out of reach locations according to the search criteria in the underlying network of routes. The implementation engages the value of a day of travel in meters in Dijkstra's single source shortest path algorithm. Then, the classification is defined based on the calculated distances from the center.

This model also facilitates computation and visualization of the flood network of multiple centers, as shown in Figure 6.10. The calculated network describes the same concept of reach from multiple centers instead of one. Accordingly, the example in this figure shows layers which are reachable within certain days of travel from either of the centers: Naysābūr or Marw. The implementation incorporates the network of multiple sources to calculate the integrated flood network. This method can also be combined with other models, such as the Voronoi diagram (Section 5.1.3), to simulate the accessibility of an area.

6.3.4 Path Alignment Tool

Using the spatial data that al-Ṯurayyā provides based on the Cornu's *Atlas*, we implement a tool for matching the location of the *unknown place*s to a toponym or approximating its coordinate as an alternative way of identifying *unknown place*s. The objective is to align a path from a source to the same/a similar path in Cornu and to facilitate the coordinate selection for an *unknown place*. Given aligned paths between a geographical source and the Cornu's *Atlas* which is visualized on the gazetteer map, we can see the differences or similarities of the path descriptions and neighbors of an *unknown place* in both. We take advantage of al-Ṯurayyā's spatial model to integrate this tool and represent a source path—a path from a source text—as a list of consecutive places and the similar or equivalent reference path—the path that al-Ṯurayyā's pathfinding algorithm finds on the map.



Network

Starting Point:

Baghdad, بغداد, BAGHDAD_443E;

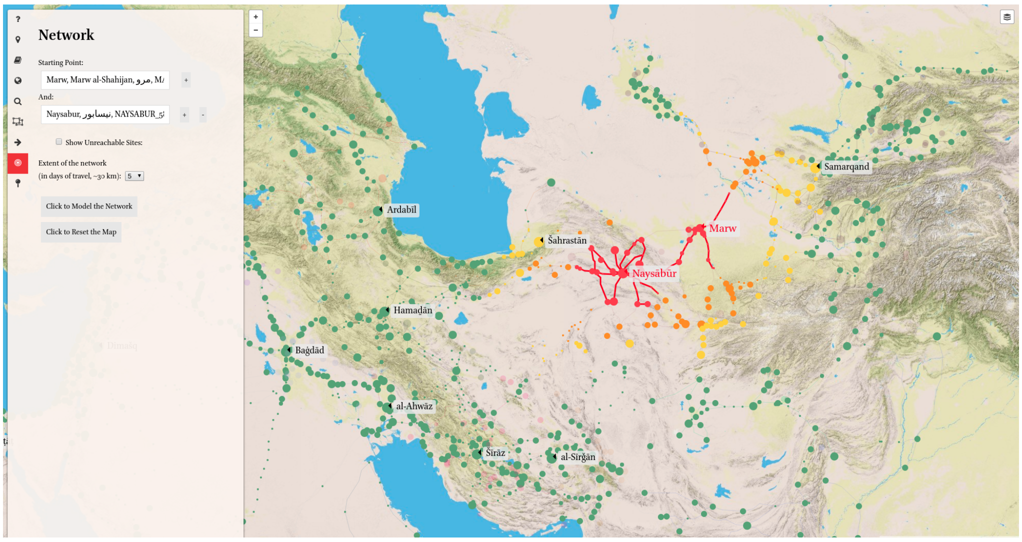
Show Unreachable Sites: □

Extent of the network
(in days of travel, ~30 km): 5

Click to Model the Network

Click to Reset the Map

Figure 6.9: Flood network of Baġdād within five days of travel: the red points show the places that can be reached in five days, orange points are those that can be reached in ten days, the yellow points are in fifteen days of travel distance, green ones can be reached in twenty days, and the pale-colored points are places that can not be reached from Baġdād in this network of routes.



Network

Starting Point:

Marw, Marw al-Shahijan, مرو, M/

And:

Naysabur, نیسابور, NAYSABUR_5

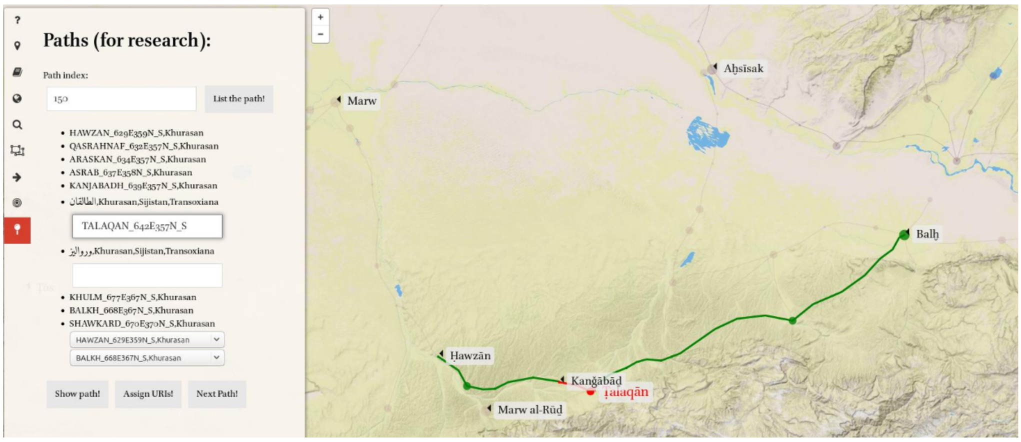
Show Unreachable Sites: □

Extent of the network
(in days of travel, ~30 km): 5

Click to Model the Network

Click to Reset the Map

Figure 6.10: Flood network of multiple arbitrary centers, Naysābūr and Marw (Mary): the red points show the places that can be reached in five days, orange points are those that can be reached within ten days, the yellow points are in fifteen days of travel distance, green ones can be reached in twenty days, and the pale-colored points are places that can not be reached from these centers in this network of routes.



Paths (for research):

Path index: 150

- HAWZAN_629E359N_S,Khurasan
- QASRAHNAF_632E357N_S,Khurasan
- ARASKAN_634E357N_S,Khurasan
- ASRAB_637E358N_S,Khurasan
- KANJABADH_639E357N_S,Khurasan
- الطالقان,Khurasan,Sijistan,Transoxiana

TALAQAN_642E357N_S

- ورواليز,Khurasan,Sijistan,Transoxiana
- KHULM_677E367N_S,Khurasan
- BALKH_668E367N_S,Khurasan
- SHAWKARD_670E370N_S,Khurasan

HAWZAN_629E359N_S,Khurasan

BALKH_668E367N_S,Khurasan

Figure 6.11: Alignment of a path, including *unknown places* (shown together with an input box), from al-Muqaddasī on the left and the same path from Cornu's *Atlas* on the map in al-Ṯurayyā. The places with a gazetteer identifier and geographical coordinates in the Atlas are shown with the corresponding id while *unknown places* (i.e., without a gazetteer identifier and geographical coordinates) are shown together with an input box to insert the newly found id and coordinate and save it to a file.

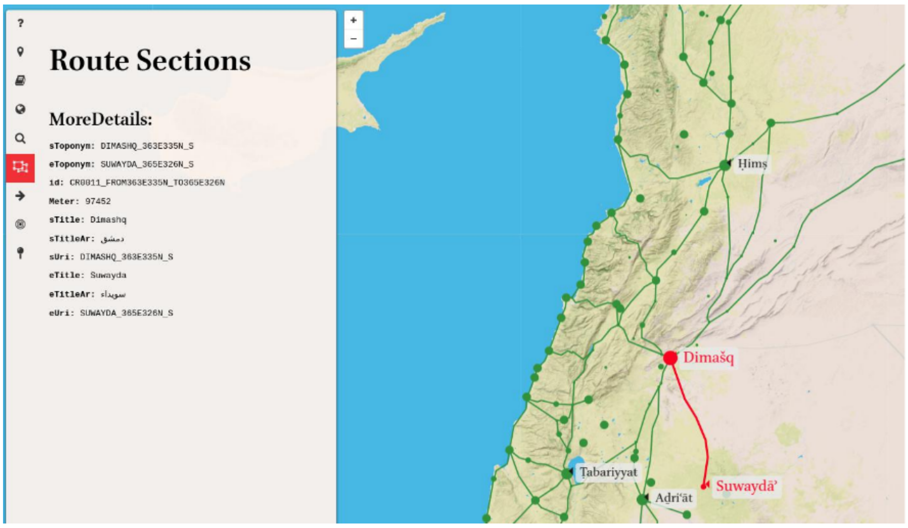
Assume places A, B, and C are consecutively mentioned in a source path with definite distances while on the map there is only one path from A to C and, therefore, B is considered as an *unknown place*. Hence, one can estimate the location of B according to the distance to A and C on the route between A and B. The example in Figure 6.11 shows a path from Ḥawzān to Šawkard from al-Muqaddasī's book on the left while the same path from Cornu's *Atlas* is visualized on the map. The map visualization represents the geographical extent and neighboring places around which an *unknown place* might be located. At the same time, the consecutive list of places in the left tab represents the *unknown places* along the path (from al-Muqaddasī's book), common places, and their neighbors in order.

Investigating the position of an *unknown place*, it can be seen in this figure that the first *unknown place* along the path is mentioned right after Kanǧābāḏ and the same happens in the path on the map. As a part of the contextual data, the corresponding region of the places is also mentioned together with the toponyms which disambiguates the similar place names in various regions. Accordingly, Ṭalaqān that comes after Kanǧābāḏ on the map is the exact match to the *unknown place* on the left path.

Having various paths, along which an *unknown place* happens, enables investigating an *unknown place* description with more geographical contextual data that each path partially provides, such as neighbors. Furthermore, visualizing the reference paths on the map in some cases helps to approximate the locations more accurately. For instance, if a path connects two places on different sides of a water body and one is seeking a place along this path, then the approximate location cannot be over the water, unless it is an island, that might happen in fully computational approaches. However, this tool is limited to be used for small datasets.

6.3.5 Data Structure

As stated in Section 6.2, data is stored in GeoJSON to benefit its spatial specification as a common format for geospatial purposes and to enable expansion and flexibility of the data for each geospatial feature. The data structure is expandable for revisions and future modifications. It



Route Sections

MoreDetails:

```
sToponym: DIMASHQ_363E335N_S
eToponym: SUWAYDA_365E326N_S
id: CR0011_FROM363E335N_TO365E326N
Meter: 97452
sTitle: Dimashq
sTitleAr: دمشق
sUri: DIMASHQ_363E335N_S
eTitle: Suwayda
eTitleAr: سويداء
eUri: SUWAYDA_365E326N_S
```

Figure 6.12: Detailed information of a route section connecting Dimašq to Suwaydāʾ in al-Ṯurayyā.

allows one to augment the contextual information of locations and route sections as references and geographical features, such as other names in different languages, if available.

6.3.5.1 Places

On the map, places are represented by GeoJSON features of type Point. Relevant data is part of the properties, including names, identifiers, type, the region to which the places belong, and links to references to primary and secondary sources. References are stored in external data files in JSON format to avoid unnecessary complexity of data files and keep the size of data as small as possible, which also makes the process of data preparation for both GeoJSON and reference data files easier and cleaner. Moreover, for future developments, we can keep the same structure and links to the references to store the data in a document-oriented database (e.g., MongoDB[6]).

Besides the references and geographical specification, places are specified by various names of a toponym in the available languages, an identifier (URI), the type of the toponym, and the identifier of the region it belongs to. The region information in fact shows the relations between geospatial entities. The examples of data for places in Section 5.1.1 represents the structure of a single location of type Point in GeoJSON format. We add to this structure an archive record that enables adding data from different sources in future.

6.3.5.2 Routes and Distances

Similar to the places, route sections are stored as GeoJSON features of type LineString (see the example in Section 5.1.1). Beside the geospatial specification of the geometry, the identifier and length of the section in meters, as well as the locations that are connected through the section are included in the data. By clicking on a route section on the map, one can access the corresponding information of the route on the left panel (Figure 6.12).

[6] https://www.mongodb.com

Chapter 7

Conclusions and Further Work

Pre-modern geographical descriptions provide various categories of information as textual descriptions. In this dissertation we explain ideas, based on the prior research in the field of Geospatial Humanities, that take us from the descriptions to the maps, modern tools, and visualizations, which helps us to get meaningful insights into findings. We discuss four major cases of geographical data, which include gazetteer, administrative hierarchy, routes and distances, and toponyms associated with people. We closely explain administrative hierarchy and routes based on the examples from pre-modern sources. To start the data gathering process, we use OpenITI mARkdown to annotate sophisticated patterns of descriptions of hierarchical administrations as well as route sections. The mARkdown is specifically useful for right-to-left languages, where the existing standards, such as TEI XML, make the annotation process overlay complicated. The mARkdown is however convertible to the conventional standards, such as TEI XML. Additionally, we discussed an ML-based approach, SVM approach, to identify the toponyms and represent the effectiveness of this approach. As a future work, one can examine other machine learning approaches to the pre-modern corpus that we use here as well as more recent POS tagger tools, such as CAMeL Tools, and compare the results to the SVM approach.

Annotation makes the data ready to extract. After the extraction, we represent the annotated data with proper data structures, which then can be used in the state-of-the-art tools and approaches. We completely implement the annotation and extraction process on a classical Arabic source and partially on a Greek and Latin source to show the generality of this approach. To expand the mARkdown, a future work is to fully apply the process to the sources in other languages to cover new patterns of descriptions as well as the language-specific patterns.

Having ready the geographical data, we then introduce models for representing administrative hierarchy. Regions, territories, and administrative divisions in pre-modern sources do not fit into the modern concept of political maps. The models that we explain here help us visualize the pre-modern description of regions and divisions on modern maps, based on our modern knowledge. In doing so, we discuss the pros and cons of each model after visualizing them on a map to revise the ideas to achieve improved models. The models include the quad-tree, Voronoi diagram, convex hulls, concave hulls, and a model that uses both settlements and route networks. Since the models, except the last one, cover some water bodies, one can use the waterbody boundaries to cut those areas. Another work is to implement the models for multiple levels of hierarchical data.

The modeling chapter also offers comparison models for administrative divisions and route networks. They could be used to show the similarities and differences of various descriptions of the same area. We use a sample dataset to implement the models and show the practicality and generality of the model. For administrative divisions, the model converts textual descriptions to a 2D matrix with toponyms as rows and regions as columns. The zero and one values in the matrix show which toponym belongs to which region. This means, each region is a matrix column that is a two-dimensional vector. We then show how we can compare the region by mathematically comparing the vectors and achieve a normalized view of the corresponding sources. Vectors offer various ways to compare the sources on various levels of hierarchy and can produce different levels of statistical data. For instance, one can compare the number of the divisions at each level of hierarchical data and find the overlapping ones. Similarly, this can happen at the settlement level, as the model can give an overview of the (non)overlapping settlements in a region in different sours. An example of this phenomenon is changes of borders of a country through time. The data structure of this model provides enough information to develop visualization of the overlapping borders on the map as a future application. The model can be applied to multiple sources. Matrix structure is capable of providing a high-level view of multiple sources in comparison, which can be implemented in future work. With multiple schemes compared, we can cluster the schema and potentially build a consensus tree of sorts that would tell us which geographers are closer to each other based on how they describe regions.

We propose a model for comparing the route networks as an approach for simplifying the networks and preparing them for comparison as well as a method to estimate a geographical location for unknown (lost) places. The model rebuilds the network by connecting the high-degree nodes with straight lines and putting the lower-degree nodes on the lines at proportionate distances based on the original distances, thus, converts the paths to straight paths where applicable. The shape of the simplified network then depends on the number of high-degree nodes and the position of the nodes. Therefore, the model would work best, in our experiment, on small parts of the network and helps to use the relatively straight paths in visualizations. The shape of the network differs by setting different values for high-degree nodes to start the process. Depending on the network, this can be done by trying different values. Another variable that affects the simplification process is the number of *unknown place*s, if there are any in the network, specifically among the high-degree nodes as this model relies on the existing certain coordinates to build the network. A future work is to implement the comparison process for this model. Also, interactivity in the implementation of this model could be another work to do as selection of the nodes to start the simplification process is critical. This model also offers an approach to locate the *unknown place*s with approximate coordinates that are computed, during the simplification process, based on the proportionate distances of the nodes.

In the last section of Chapter 5, we introduce an alternative approach for estimating geographical coordinates for *unknown place*s using triangulation. A sample implementation shows the output of the model for the places that are geographically connected, at a specified distance, to two known places (i.e., places that have geographical coordinates). As a future work, one can test this method on places where all the places have coordinates, but withhold a subset of coordinates (e.g., by randomly removing the location data for a part of the data and then approximate the locations) to evaluate the accuracy of the new coordinates when the answer is available. Also, one can implement and evaluate the last part of the model that covers the *unknown place*s connected to more than two places.

The last chapter introduces al-Ṯurayyā, a gazetteer and a spatial model for the classical Islamic word. The gazetteer includes more than 2,000 toponyms as well as route sections and distances. It provides search function, region visualization, path finding, and flood network. As a future work, time dimension can be added to al-Ṯurayyā in order to cover the data from various eras so that users can select a time period and accordingly the data will be loaded. This requires the time specification in the dataset as well as new data from various periods to make the implementation of this feature possible. Also, as a gazetteer of the pre-modern Islamic world, al-Ṯurayyā can cover more corresponding areas, such as the Ottoman Empire and relevant places from the Pleiades gazetteer. At the moment, al-Ṯurayyā works with the dataset from Cornu's *Atlas*. However, the structure is ready to digest data from different sources and can be expanded by adding more data from different sources. Another approach to increase the area of coverage of the gazetteer as well as the functionalities that we offer is to add import functionality. This approach will allow the users to load their data into the gazetteer and use the existing models and features.

Bibliography

[1] Y. ibn ^{c}Abd Allāh al Ḥamawī, *Mu^{c}jam al-buldān.* Bayrūt: Dār Ṣādir, 1977.

[2] Z. Antrim, *Routes and Realms: The Power of Place in the Early Islamic World.* OUP USA, 2012.

[3] W. Hinz, *Encyclopaedia of Islam, Second Edition.* Brill, 2012.

[4] D. Sourdel, *Encyclopaedia of Islam, Second Edition.* Brill, 2012.

[5] M. Streck and G. Miles, *Encyclopaedia of Islam, Second Edition.* Brill, 2012.

[6] F. Moretti, "Conjectures on World Literature," *New Left Review*, vol. 1, pp. 54–68, 01 2014.

[7] ——, *Distant Reading.* London: Verso, 2013.

[8] S. Jänicke, G. Franzini, M. F. Cheema, and G. Scheuermann, "On Close and Distant Reading in Digital Humanities: a Survey and Future Challenges," in *Eurographics Conference on Visualization (EuroVis) - STARs*, R. Borgo, F. Ganovelli, and I. Viola, Eds. The Eurographics Association, 2015.

[9] M. Seydi and M. Romanov, "Al-Ṯurayyā, the Gazetteer and the Geospatial Model of the Early Islamic World," *Digital Humanities 2019 Conference Papers*, 2019.

[10] I. Gregory and G. Alistair, *Introduction: From Historical GIS to Spatial Humanities: Deepening Scholarship and Broadening Technology.* Indiana University Press, 2014, pp. ix–xxii.

[11] J. L. Leidner, "Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding," *SIGIR Forum*, vol. 41, no. 2, p. 124–126, December 2007.

[12] I. Gregory, K. Kemp, and R. Mostern, "Geographical Information and Historical Research: Current Progress and Future Directions," *History and Computing*, vol. 13, pp. 7–23, 03 2001.

[13] P. Murrieta-Flores, C. Donaldson, and I. Gregory, "GIS and Literary History: Advancing Digital Humanities Research through the Spatial Analysis of Historical Travel Writing and Topographical Literature," *Digital Humanities Quarterly*, 2017.

[14] F. Moretti, *Atlas of the European Novel 1800-1900.* London, New York: Verso, 1997.

[15] ——, *Graphs, Maps, Trees: Abstract Models for Literary History.* London, New York: Verso, 2007.

[16] M. Jockers, *Macroanalysis: Digital Methods and Literary History.* Urbana: University of Illinois Press, 2013.

[17] R. Tally Jr., *Spatiality.* New York: Routledge, 2013.

[18] C. Donaldson, I. Gregory, and P. Murrieta-Flores, "Mapping 'Wordsworthshire': a GIS Study of Literary Tourism in Victorian Lakeland," *Journal of Victorian Culture*, vol. 20, pp. 287–307, 07 2015.

[19] I. Gregory, C. Donaldson, P. Murrieta-Flores, and P. Rayson, "Geoparsing, GIS, and Textual Analysis: Current Developments in Spatial Humanities Research," *International Journal of Humanities and Arts Computing*, vol. 9, pp. 1–14, 03 2015.

[20] D. Cooper, C. Donaldson, and P. Murrieta-Flores, Eds., *Literary Mapping in the Digital Age*, 1st ed. London: Routledge, 2016.

[21] D. Bodenhamer, J. Corrigan, and T. Harris, *The Spatial Humanities: GIS and the Future of Humanities Scholarship*, ser. The Spatial Humanities. Indiana University Press, 2010.

[22] I. Gregory, *A Place in History: a Guide to Using GIS in Historical Research*, 2nd ed. Oxbow Books, 2005.

[23] A. Muri, "Chapter 14-Beyond GIS: On Mapping Early Modern Narratives and the Chronotope." *Digital Studies/Le Champ Numérique*, vol. 6, 2016.

[24] D. G. Cole, "Placing History: How Maps, Spatial Data, and GIS Are Changing Historical Scholarship," *Cartographic Perspectives*, no. 63, pp. 66–69, Jun. 2009.

[25] A. K. Knowles, "Introduction," *Social Science History*, vol. 24, no. 3, p. 451–470, 2000.

[26] P. Bol, "Toward Spatial Humanities: Historical GIS and Spatial History." *Journal of Interdisciplinary History*, vol. 46, pp. 266–267, 08 2015.

[27] A. Lünen and C. Travis, *History and GIS: Epistemologies, Considerations and Reflections.* Springer, 2013.

[28] C. Gordon, *Mapping Decline: St. Louis and the Fate of the American City*, ser. Politics and Culture in Modern America. University of Pennsylvania Press, Incorporated, 2008.

[29] R. White, *Railroaded: The Transcontinentals and the Making of Modern America.* W. W. Norton, 2011.

[30] A. R. H. Baker, *Geography and History: Bridging the Divide*, ser. Cambridge Studies in Historical Geography. Cambridge University Press, 2003.

[31] P. Bol, *Creating a GIS for the History of China.* Redlands, CA: ESRI Press, 2008, pp. 27–60.

[32] M. Dear, J. Ketchum, S. Luria, and D. Richardson, Eds., *GeoHumanities: Art, History, Text at the Edge of Place*, 1st ed. London: Routledge, 2011.

[33] I. N. Gregory and P. S. Ell, *Historical GIS: Technologies, Methodologies, and Scholarship*, ser. Cambridge Studies in Historical Geography. Cambridge University Press, 2007.

[34] A. Knowles, Ed., *Past Time, Past Place: GIS for History.* Redlands, CA: ESRI Press, 2002.

[35] A. Knowles and A. Hillier, Eds., *Placing History: How Maps, Spatial Data, and GIS Are Changing Historical Scholarship.* Redlands, CA: ESRI Press, 2008.

[36] B. Warf and S. Arias, Eds., *The Spatial Turn: Interdisciplinary Perspectives.* London: Routledge, 2008.

[37] J. Guldi, "The Spatial Turn in History," https://spatial.scholarslab.org/spatial-turn/the-spatial-turn-in-history/, 2011.

[38] E. Cordulack, "What is the 'Spatial Turn'? A Beginner's Look," http://at.blogs.wm.edu/what-is-the-spatial-turn-a-beginners-look, 2011.

[39] I. Gregory and A. Geddes, Eds., *Toward Spatial Humanities: Historical GIS and Spatial History.* Indiana University Press, 2014.

[40] D. R. Montello, K. Grossner, and D. G. Janelle, Eds., *Space in Mind: Concepts for Spatial Learning and Education.* Cambridge, MA: The MIT Press, 2014.

[41] M. Aldenderfer and H. Maschner, Eds., *Anthropology, Space, and Geographic Information Systems.* New York: OUP, 1996.

[42] A. Fotheringham, C. Brunsdon, and M. Charlton, *Quantitative Geography: Perspectives on Spatial Data Analysis.* SAGE Publications Ltd, 01 2000.

[43] D. Wheatley and M. Gillings, *Spatial Technology and Archaeology: the Archaeological Applications of GIS.* CRC Press, 01 2002.

[44] E. K. A. Robert Nash Parker, *GIS and Spatial Analysis for the Social Sciences: Coding, Mapping, and Modeling.* New York: Routledge, 2008.

[45] M. Goodchild and D. Janelle, "Toward Critical Spatial Thinking in the Social Sciences and Humanities," *GeoJournal*, vol. 75, pp. 3–13, 02 2010.

[46] J. Richards, S. Jeffrey, S. Waller, F. Ciravegna, S. Chapman, and Z. Zhang, *CHAPTER 1: the Archaeology Data Service and the Archaeotools Project: Faceted Classification and Natural Language Processing.* eScholarship University of California, 2011, p. 31–56.

[47] E. Barker, S. Bouzarovski, C. Pelling, and L. Isaksen, "Mapping an Ancient Historian in a Digital Age: the Herodotus Encoded Space-text-image Archive (HESTIA)," *Leeds International Classical Studies*, vol. 9, 01 2010.

[48] A. Warner-Smith, "Mapping the GIS Landscape: Introducing "Beyond (within, though) the Grid"," *International Journal of Historical Archaeology*, vol. 24, 12 2020.

[49] D. J. Bodenhamer, *Beyond GIS: Geospatial Technologies and the Future of History.* Dordrecht: Springer Netherlands, 2013, pp. 1–13.

[50] T. Evans and P. Daly, *Digital Archaeology: Bridging Method and Theory.* London: Routledge, 2006.

[51] M. C. Howey and M. Brouwer Burg, “Assessing the State of Archaeological GIS Research: Unbinding Analyses of Past Landscapes,” *Journal of Archaeological Science*, vol. 84, pp. 1–9, 2017.

[52] J. T. Herrmann, “Special Issue on Digital Domains: Introduction,” *Advances in Archaeological Practice*, vol. 2, no. 3, p. 145–146, 2014.

[53] E. González-Tennant, “Recent Directions and Future Developments in Geographic Information Systems for Historical Archaeology,” *Historical Archaeology*, vol. 50, pp. 24–49, 2016.

[54] B. Donahue, *Mapping Husbandry in Colonial Concord: GIS as a Tool for Environmental History.* Redlands, CA: ESRI Press, 2008, pp. 151–178.

[55] A. K. Knowles, W. Roush, C. Abshere, L. Farrell, A. Feinberg, T. Humber, G. Kuzzy, and C. Wirene, *What Could Lee See at Gettysburg?* Redlands, CA: ESRI Press, 2008, pp. 235–266.

[56] R. Schwartz, I. Gregory, and J. Marti-Henneberg, *History and GIS: Railways, Population Change, and Agricultural Development in Late Nineteenth Century Wales.* Routledge, 2011, pp. 251–266.

[57] J. Galloway, *Reconstructing London’s Distributive Trade in the Later middle Ages: the Role of Computer-assisted Mapping and Analyses.* Glasgow: Association for History and Computing, 2000, pp. 1–24.

[58] I. Gregory, “Longitudinal Analysis of Age- and Gender-specific Migration Patterns in England and Wales: a GIS-based Approach,” *Social Science History*, vol. 24, no. 3, p. 471–503, 2000.

[59] I. N. Gregory and A. Hardie, “Visual GISting: Bringing Together Corpus Linguistics and Geographical Information Systems,” *Literary and Linguistic Computing*, vol. 26, no. 3, pp. 297–314, 05 2011.

[60] P. Murrieta-Flores, A. Baron, I. Gregory, A. Hardie, and P. Rayson, “Automatically Analyzing Large Texts in a GIS Environment: The Registrar General’s Reports and Cholera in the 19$^{\text{th}}$ Century,” *Transactions in GIS*, vol. 19, no. 2, pp. 296–320, 2015.

[61] C.-X. Shu, *Digital Humanities and GIS for Chinese Architecture: a Methodological Experiment.* Leuven University Press, 2019, pp. 301–346.

[62] T. M. Harris, S. Bergeron, and L. J. Rouse, *Humanities GIS: Place, Spatial Storytelling, and Immersive Visualization in the Humanities.* Routledge, 2011, pp. 226–240.

[63] B. Piatti, H. R. Bär, A.-K. Reuschel, and L. Hurni, “Die Geographie der Fiktion — Das Projekt “Ein literarischer Atlas Europas”,” *Journal of Cartography and Geographic Information*, vol. 58, p. 287–294, 2008.

[64] C. Travis and R. Breen, “Digital Literary Atlas of Ireland,” http://cehresearch.org/DLAI, 2017.

[65] S. Bushell, "The Slipperiness of Literary Maps: Critical Cartography and Literary Cartography," *Cartographica: The International Journal for Geographic Information and Geovisualization*, vol. 47, pp. 149–160, 09 2012.

[66] B. Piatti and L. Hurni, "Cartographies of Fictional Worlds," *The Cartographic Journal*, vol. 48, pp. 218–223, 11 2011.

[67] D. Cooper and I. N. Gregory, "Mapping the English Lake District: a literary GIS," *Transactions of the Institute of British Geographers*, vol. 36, no. 1, pp. 89–108, 2011.

[68] J. Gerhard and W. van den Heuvel, "Survey Report on Digitisation in European Cultural Heritage Institutions 2015," *EUMERATE Thematic Network*, 2015.

[69] L. Hill, *Georeferencing: The Geographic Associations of Information.* Cambridge, MA: The MIT Press, 2006.

[70] M. Won, P. Murrieta-Flores, and B. Martins, "Ensemble Named Entity Recognition (NER): Evaluating NER Tools in the Identification of Place Names in Historical Corpora," *Frontiers in Digital Humanities*, vol. 5, 03 2018.

[71] D. Nadeau and S. Sekine, "A Survey of Named Entity Recognition and Classification," *Lingvisticae Investigationes*, vol. 30, 08 2007.

[72] J. L. Leidner and M. D. Lieberman, "Detecting Geographical References in the Form of Place Names and Associated Spatial Natural Language," *SIGSPATIAL Special*, vol. 3, no. 2, p. 5–11, jul 2011.

[73] M. Gritta, M. T. Pilehvar, and N. Collier, "A Pragmatic Guide to Geoparsing Evaluation: Toponyms, Named Entity Recognition and Pragmatics," *Language resources and evaluation*, vol. 54, no. 3, p. 683—712, 2020.

[74] S. Middleton, G. Kordopatis-Zilos, S. Papadopoulos, and Y. Kompatsiaris, "Location Extraction from Social Media: Geoparsing, Location Disambiguation and Geotagging," *ACM Transactions on Information Systems*, vol. 36, no. 4, pp. 1–27, June 2018.

[75] L. Moncla, W. Renteria-Agualimpia, J. Nogueras-Iso, and M. Gaio, "Geocoding for Texts with Fine-grain Toponyms: an Experiment on a Geoparsed Hiking Descriptions Corpus," in *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - SIGSPATIAL '14.* Dallas, Texas: ACM Press, 2014, pp. 183–192.

[76] O. Parisot and T. Tamisier, "A Corpus of Narratives Related to Luxembourg for the Period 1945-1975," *14th International Workshop on Technologies for Information Retrieval (TIR 2017)*, 2017.

[77] G. DeLozier, J. Baldridge, and L. London, "Gazetteer-independent Toponym Resolution Using Geographic Word Profiles," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, no. 1, Feb. 2015.

[78] M. Gritta, M. T. Pilevar, N. Limsopatham, and N. Collier, "What's Missing in Geographical Parsing?" *Language Resources and Evaluation*, vol. 52, 06 2018.

[79] P. Rayson, D. Archer, S. Piao, and T. Mcenery, “The UCREL Semantic Analysis System,” in *Beyond Named Entity Recognition Semantic Labeling for NLP Tasks in LREC’04*, 01 2004.

[80] S. Cucerzan, “Large-scale Named Entity Disambiguation Based on Wikipedia Data,” in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL).* Prague, Czech Republic: Association for Computational Linguistics, Jun. 2007, pp. 708–716.

[81] P. Rayson, A. Reinhold, J. Butler, C. Donaldson, I. Gregory, and J. Taylor, “A Deeply Annotated Testbed for Geographical Text Analysis: the Corpus of Lake District Writing,” in *Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities*, ser. GeoHumanities’17. New York, NY, USA: ACM, 2017, pp. 9–15.

[82] L. Borin, D. Kokkinakis, and L.-J. Olsson, “Naming the Past: Named Entity and Animacy Recognition in 19th Century Swedish Literature,” in *Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007).* Prague, Czech Republic: Association for Computational Linguistics, Jun. 2007, pp. 1–8.

[83] S. Mac Kim and S. Cassidy, “Finding names in Trove: named entity recognition for Australian historical newspapers,” in *Australasian Language Technology Association Workshop 2015*, B. Hachey and K. Webster, Eds., vol. 13. Australasian Language Technology Association, 2015, pp. 57–65.

[84] K. Byrne, “Nested Named Entity Recognition in Historical Archive Text,” in *International Conference on Semantic Computing (ICSC 2007)*, 2007, pp. 589–596.

[85] C. Grover, S. Givon, R. Tobin, and J. Ball, “Named Entity Recognition for Digitised Historical Texts,” in *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008*, 2008, pp. 1343–1346.

[86] J. Brooke, A. Hammond, and G. Hirst, “GutenTag: an NLP-driven Tool for Digital Humanities Research in the Project Gutenberg Corpus,” in *Proceedings of the Fourth Workshop on Computational Linguistics for Literature.* Denver, Colorado, USA: Association for Computational Linguistics, Jun. 2015, pp. 42–47.

[87] R. Sprugnoli, G. Moretti, S. Tonelli, and S. Menini, “Fifty Years of European History through the Lens of Computational Linguistics: the De Gasperi Project,” *Italian Journal of Computational Linguistics*, vol. 2, pp. 89–99, 12 2016.

[88] S. van Hooland, M. De Wilde, R. Verborgh, T. Steiner, and R. Van de Walle, “Exploring Entity Recognition and Disambiguation for Cultural Heritage Collections,” *Digital Scholarship in the Humanities*, vol. 30, no. 2, pp. 262–279, 11 2013.

[89] K. Kettunen, E. Mäkelä, T. Ruokolainen, J. Kuokkala, and L. Löfberg, “Old Content and Modern Tools - Searching Named Entities in a Finnish OCRed Historical Newspaper Collection 1771-1910,” *Digital Humanities Quarterly*, vol. 11, 11 2017.

[90] C. Neudecker, L. Wilms, W. J. Faber, and T. van Veen, “Large-scale Refinement of Digital Historic Newspapers with Named Entity Recognition,” *Proceedings of the IFLA Newspapers/GENLOC Pre-conference Satellite Meeting*, 2014.

[91] B. Batjargal, G. Khaltarkhuu, F. Kimura, and A. Maeda, "An Approach to Named Entity Extraction from Historical Documents in Traditional Mongolian Script," in *2014 IEEE/ACM Joint Conference on Digital Libraries (JCDL)*. Los Alamitos, CA, USA: IEEE Computer Society, sep 2014, pp. 489–490.

[92] J. Clifford, B. Alex, C. M. Coates, E. Klein, and A. Watson, "Geoparsing History: Locating Commodities in Ten Million Pages of Nineteenth-century Sources," *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, vol. 49, no. 3, pp. 115–131, 2016.

[93] B. Wing, "Text-based Ddocument Geolocation and Its Application to the Digital Humanities," PhD dissertation, The Faculty of the Graduate School of The University of Texas at Austin, 2015.

[94] J. Santos, I. Anastácio, and B. Martins, "Using Machine Learning Methods for Disambiguating Place References in Textual Documents," *GeoJournal*, vol. 80, p. 375–392, 06 2015.

[95] T. Brown, J. Baldridge, M. Esteva, and W. Xu, "The substantial Words Are in the Ground and Sea: Computationally Linking Text and Geography," *Texas Studies in Literature and Language*, vol. 53, p. 324–339, 2017.

[96] F. Melo and B. Martins, "Automated Geocoding of Textual Documents: A Survey of Current Approaches," *Transactions in GIS*, vol. 21, no. 1, pp. 3–38, 2017.

[97] R. Purves and C. B. Jones, "Geographic Information Retrieval," *ACM SIGSPATIAL Special*, vol. 3, pp. 2–4, 2011.

[98] C. Porter, P. Atkinson, and I. Gregory, "Geographical Text Analysis: a New Approach to Understanding Nineteenth-century Mortality," *Health and Place*, vol. 36, pp. 25–34, Nov. 2015.

[99] P. Murrieta-Flores and I. Gregory, "Further Frontiers in GIS: Extending Spatial Analysis to Textual Sources in Archaeology," *Open Archaeology*, vol. 1, no. 1, 2015.

[100] L. E. da Silveira, "Geographic Information Systems and Historical Research: an Appraisal," *International Journal of Humanities and Arts Computing*, vol. 8, no. 1, pp. 28–45, 2014.

[101] M. F. Goodchild and L. L. Hill, "Introduction to Digital Gazetteer Research," *International Journal of Geographical Information Science*, vol. 22, no. 10, pp. 1039–1044, 2008.

[102] H. Manguinhas, B. Martins, and J. Borbinha, "A Geo-temporal Web Gazetteer Integrating Data from Multiple Sources," in *2008 Third International Conference on Digital Information Management*, Nov 2008, pp. 146–153.

[103] M. Berman, R. Mostern, and H. Southall, Eds., *Placing Names: Enriching and Integrating Gazetteers*, ser. Spatial Humanities. United States: Indiana University Press, 8 2016.

[104] B. Alex, K. Byrne, C. Grover, and R. Tobin, "Adapting the Edinburgh Geoparser for Historical Georeferencing," *International Journal of Humanities and Arts Computing*, vol. 9, pp. 15–35, 03 2015.

[105] P. de Soto, R. Simon, E. Barker, and L. Isaksen, “Linking Early Geospatial Documents, One Place at a Time: Annotation of Geographic Documents with Recogito,” *e-Perimetron*, vol. 10, pp. 49–59, 01 2015.

[106] D. Jiménez Badillo, P. Murrieta-Flores, B. Martins, I. Gregory, M. Favila-Vázquez, and R. Liceras-Garrido, “Developing geographically oriented NLP approaches to sixteenth–century historical documents: digging into early colonial Mexico,” *Digital Humanities Quarterly*, vol. 14, no. 4, Dec. 2020.

[107] S. Darren, “Automatic Arabic Named Entity Extraction anf Classification for Information Retrieval,” *International Journal on Natural Language Computing*, vol. 9, pp. 1–22, 12 2020.

[108] C. Helwe, G. Dib, M. Shamas, and S. Elbassuoni, “A Semi-supervised BERT Approach for Arabic Named Entity Recognition,” in *The 5th Arabic Natural Language Processing Workshop*, 10 2020.

[109] O. Asbayou, “Arabic Location Name Annotations and Applications,” *9th International Conference on Natural Language Processing (NLP 2020)*, pp. 51–65, 2020.

[110] W. Etaiwi, A. Awajan, and D. Suleiman, “Statistical Arabic Name Entity Recognition Approaches: a Survey,” *Procedia Computer Science*, vol. 113, pp. 57–64, 09 2017.

[111] M. Oudah and K. Shaalan, “NERA 2.0: Improving Coverage and Performance of Rule-based Named Entity Recognition for Arabic,” *Natural Language Engineering*, vol. -1, pp. 1–32, 05 2016.

[112] A. Dandashi, J. Al Jaam, and S. Foufou, “Arabic Named Entity Recognition—a Survey and Analysis,” in *Intelligent Interactive Multimedia Systems and Services 2016*, ser. Smart Innovation, Systems and Technologies, L. Jain, L. Jain, G. De Pietro, L. Gallo, R. Howlett, and L. Jain, Eds. Germany: Springer Science and Business Media Deutschland GmbH, 2016, pp. 83–96.

[113] M. Althobaiti, U. Kruschwitz, and M. Poesio, “Combining Minimally-supervised Methods for Arabic Named Entity Recognition,” *Transactions of the Association for Computational Linguistics*, vol. 3, pp. 243–255, 2015.

[114] H. Elsayed and T. Elghazaly, “A Named Entities Recognition System for Modern Standard Arabic using Rule-based Approach,” in *2015 First International Conference on Arabic Computational Linguistics (ACLing)*, 2015, pp. 51–54.

[115] K. Shaalan, “A Survey of Arabic Named Entity Recognition and Classification,” *Computational Linguistics*, vol. 40, no. 2, pp. 469–510, 06 2014.

[116] B. Alex, K. Byrne, C. Grover, and R. Tobin, “A Web-based Geo-resolution Annotation and Evaluation Tool,” in *Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop.* Dublin, Ireland: Association for Computational Linguistics and Dublin City University, Aug. 2014, pp. 59–63.

[117] C. Grover, R. Tobin, K. Byrne, M. Woollard, J. Reid, S. Dunn, and J. Ball, “Use of the Edinburgh Geoparser for Georeferencing Digitized Historical Collections,” *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, vol. 368, pp. 3875–89, 08 2010.

[118] L. Isaksen, E. Barker, E. Kansa, and K. Byrne, “GAP: a NeoGeo Approach to Classical Resources,” *Leonardo*, vol. 45, pp. 82–83, 02 2012.

[119] B. Alex and C. Grover, “Labelling and Spatio-temporal Grounding of News Events,” in *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*. Los Angeles, California, USA: Association for Computational Linguistics, Jun. 2010, pp. 27–28.

[120] R. Schwartz, “Digital Partnership: Combining Text Mining and GIS in a Spatial History of Sea Fishing in the United Kingdom, 1860 to 1900,” *International Journal of Humanities and Arts Computing*, vol. 9, pp. 36–56, 03 2015.

[121] D. Alves and A. Queiroz, “Exploring Literary Landscapes: From Texts to Spatiotemporal Analysis through Collaborative Work and GIS,” *International Journal of Humanities and Arts Computing*, vol. 9, pp. 57–73, 03 2015.

[122] M. Almas, A. Rodrigues, N. Correia, A. Queiroz, and D. Alves, “LITESCAPE.PT - a Portuguese Literary Atlas,” in *Congresso de Humanidades Digitais em Portugal*, oct 2015.

[123] R. Purves and C. Derungs, “From Space to Place: Place-based Explorations of Text,” *International Journal of Humanities and Arts Computing*, vol. 9, pp. 74–94, 03 2015.

[124] M. Volk, “How Many Mountains Are There in Switzerland? Explorations of the Swiss Toponame List,” in *Searching Answers: a Festschrift for Michael Hess on the Occasion of his 60th Birthday*, S. Clematide, M. Klenner, and M. Volk, Eds. Münster: Monsenstein und Vannerdat, 2009, pp. 127–140.

[125] C. Heuvel, “Mapping Knowledge Exchange in Early Modern Europe: Intellectual and Technological Geographies and Network Representations,” *International Journal of Humanities and Arts Computing*, vol. 9, pp. 95–114, 03 2015.

[126] K. Grossner, K. Janowicz, and C. Kessler, *Place, Period, and Setting for Linked Data Gazetteers*, ser. The Spatial Humanities. United States: Indiana University Press, 2016, pp. 80–96.

[127] H. Southall, R. Mostern, and M. L. Berman, “On Historical Gazetteers,” *International Journal of Humanities and Arts Computing*, vol. 5, no. 2, pp. 127–145, 2011.

[128] P. K. Bol, *What Do Humanists Want? What Do Humanists Need? What Might Humanists Get?* Routledge, 2011, pp. 296–308.

[129] R. Mostern, “Historical Gazetteers: an Experiential Perspective, with Examples from Chinese History,” *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, vol. 41, no. 1, pp. 39–46, 2008.

[130] G. Cornu, *Atlas du monde arabo-islamique a l’epoque classique IXe-Xe siecles: repertoires des toponymes*. Leiden: E.J. Brill, 1985, no. Bd. 2.

[131] E. Barker, R. Simon, L. Isaksen, and P. de Soto Cañamares, “The Pleiades Gazetteer and the Pelagios Project,” in *Placing Names: Enriching and Integrating Gazetteers*, M. L. Berman, R. Mostern, and H. Southall, Eds. Bloomington: Indiana University Press, August 2016, pp. 97–109.

[132] R. Simon, E. Barker, and L. Isaksen, “Exploring Pelagios: a Visual Browser for Geo-tagged Datasets,” in *International Workshop on Supporting Users’ Exploration of Digital Libraries*, 2012.

[133] H. Southall, M. Stoner, and P. Aucott, “3.08 - PastPlace Historical Gazetteer,” in *Comprehensive Geographic Information Systems*, B. Huang, Ed. Oxford: Elsevier, 2018, pp. 110–118.

[134] J. Åhlfeldt, M. Berman, and M. Wick, *Historical Gazetteer System Integration: CHGIS, Regnum Francorum Online, and GeoNames*, ser. The Spatial Humanities. United States: Indiana University Press, 2016, pp. 110–125.

[135] W. Scheidel, “Orbis: the Stanford Geospatial Network Model of the Roman World,” https://orbis.stanford.edu/orbis2012/ORBIS_v1paper_20120501.pdf, 05 2015.

[136] E. Meeks, “The Design and Implementation of ORBIS: The Stanford Geospatial Network Model of the Roman World,” *Bulletin of the Association for Information Science and Technology*, vol. 41, no. 2, pp. 17–21, 2015.

[137] S. Dunn, “Review of Orbis,” http://journalofdigitalhumanities.org/1-3/review-of-orbis-project-by-stuart-dunn/, 2012.

[138] E. Meeks and K. Grossner, “Orbis: an Interactive Scholarly Work on the Roman World,” http://journalofdigitalhumanities.org/1-3/orbis-an-interactive-scholarly-work-on-the-roman-world-by-elijah-meeks-and-karl-grossner, 2012.

[139] W. McCarty, *Modeling: a Study in Words and Meanings.* John Wiley & Sons, Ltd, 2004, ch. 19, pp. 254–270.

[140] M. Minsky, “Conscious Machines,” in *“Machinery of Consciousness”, Proceedings, National Research Council of Canada, 75th Anniversary Symposium on Science in Society*, 1991.

[141] T. Rihll and A. Wilson, “Spatial Interaction and Structural Models in Historical Analysis: Some Possibilities and an Example,” *Histoire & Mesure*, vol. 2, no. 1, pp. 5–32, 1987.

[142] B. K. Roberts and R. E. Glasscock, Eds., *Villages, Fields and Frontiers: Studies in European Rural Settlement in the Medieval and Early Modern Periods.* Oxford, England: B.A.R., 1983.

[143] B. J.L., *Iron Age Europe in the Context of Social Evolution from the Bronze Age through to Historic Times.* University of Bradford, 1984, pp. 157–226.

[144] D. Knight, *Late Bronze Age and Iron Age Settlement in the Nene and Great Ouse Basins.* BAR, 1984.

[145] I. Hodder and C. Orton, *Spatial Analysis in Archaeology*, ser. New Studies in Archaeology. Cambridge University Press, 1979.

[146] C. Vita-Finzi, E. S. Higgs, D. Sturdy, J. Harriss, A. J. Legge, and H. Tippett, “Prehistoric Economy in the Mount Carmel Area of Palestine: Site Catchment Analysis,” *Proceedings of the Prehistoric Society*, vol. 36, p. 1–37, 1970.

[147] E. S. Higgs and C. Vita-Finzi, "Prehistoric Economies: a Territorial Approach," in *Papers in Economic Prehistory*, 1972, p. 27–36.

[148] P. Verhagen, L. Nuninger, and M. R. Groenhuijzen, *Modelling of Pathways and Movement Networks in Archaeology: an Overview of Current Approaches.* Cham: Springer International Publishing, 2019, pp. 217–249.

[149] P. v. Leusen, "Cartographic Modelling in a Cell-based GIS," *Computing the Past. Computer Applications and Quantitative Methods in Archaeology. CAA92*, vol. 45, pp. 105–124, 1993.

[150] P. Verhagen and J. McGlade, *Some Criteria for Modelling Socio-economic Activities in the Bronze Age of Southeast Spain.* London: Taylor and Francis, 1995, pp. 187–209.

[151] J. E. Ericson and R. Goldstein, *Work Space: a New Approach to the Analysis of Energy Expenditure.* Department of Anthropology, University of California, 1980, pp. 21–30.

[152] V. Gaffney and Z. Stančič, *GIS Approaches to Regional Analysis: a Case Study of the Island of Hvar.* David Brown Book Company, 1991.

[153] Y. Kondo and Y. Seino, "GPS-aided Walking Experiments and Data-driven Travel Cost Modeling on the Historical Road of Nakasendō-Kisoji (Central Highland Japan)," in *Making History Interactive. Computer Applications and Quantitative Methods in Archaeology (CAA). Proceedings of the 37th International Conference*, B. Frischer, J. Webb Crawford, and D. Koller, Eds., 2020, pp. 158–165.

[154] R. G. Soule and R. F. Goldman, "Terrain Coefficients for Energy Cost Prediction." *Journal of Applied Physiology*, vol. 32, no. 5, pp. 706–708, 1972.

[155] Pandolf, K. B. and Haisman, M. F. and Goldman, R. F., "Metabolic Energy Expenditure and Terrain Coefficients for Walking on Snow," *Ergonomics*, vol. 19, no. 6, pp. 683–690, 1976, pMID: 1009916.

[156] T. Whitley, I. Moore, G. Goel, and D. Jackson, "Beyond the Marsh: Settlement Choice, Perception, and Spatial Decision-making on the Georgia Coastal Plain," in *Making History Interactive. Computer Applications and Quantitative Methods in Archaeology (CAA). Proceedings of the 37th International Conference*, 2010, pp. 380–390.

[157] M. C. Howey, "Multiple pathways Across Past Landscapes: Circuit Theory as a Complementary Geospatial Method to Least Cost Path for Modeling Past Movement," *Journal of Archaeological Science*, vol. 38, no. 10, pp. 2523–2535, 2011.

[158] P. Murrieta-Flores, "Understanding Human Movement through Spatial Technologies. The Role of Natural Areas of Transit in the Late Prehistory of Southwestern Iberia," *Trabajos de Prehistoria*, vol. 69, pp. 103–122, 06 2012.

[159] R. J. van Lanen, M. C. Kosian, B. J. Groenewoudt, and E. Jansma, "Finding a Way: Modeling Landscape Prerequisites for Roman and Early-Medieval Routes in the Netherlands," *Geoarchaeology*, vol. 30, no. 3, pp. 200–222, 2015.

[160] I. Herzog, "Dispersal versus Optimal Path Calculation," in *Keep the Revolution Going. Proceedings of the 43rd Annual Conference on Computer Applications and Quantitative Methods in Archaeology*, 2016, pp. 567–578.

[161] M. Lake, "The Use of Pedestrian Modelling in Archaeology, with an Example from the Study of Cultural Learning," *Environment and Planning B: Planning and Design*, vol. 28, no. 3, pp. 385–403, 2001.

[162] C. D. Wren, J. Z. Xue, A. Costopoulos, and A. Burke, "The Role of Spatial Foresight in Models of Hominin Dispersal," *Journal of Human Evolution*, vol. 69, pp. 70–78, 2014.

[163] M. Llobera, *Understanding Movement: a Pilot Model Towards the Sociology of Movement.* IOS Press, 2000, pp. 65–84.

[164] D. Mlekuž, "Time, Geography, GIS and Archaeology," in *Fusion of Cultures, Proceedings of the 38th Conference on Computer Applications and Quantitative Methods in Archaeology*, 2013, pp. 359–366.

[165] ——, *Exploring the Topography of Movement.* De Gruyter, 2014, pp. 5–22.

[166] I. Herzog, "Least-Cost Kernel Density Estimation and Interpolation-based Density Analysis Applied to Survey Data," in *Proceedings of the 38th Annual Conference on Computer Applications and Quantitative Methods in Archaeology, CAA2010*, 2016, pp. 367–374.

[167] T. Whitley and L. M. Hicks, "A Geographic Information Systems Approach to Understanding Potential Prehistoric and Historic Travel Corridors," *Southeastern Archaeology*, vol. 22, pp. 76–90, 2014.

[168] K. Zakšek, E. Fovet, L. Nuninger, and T. Podobnikar, "Path Modelling and Settlement Pattern," in *Layers of Perception. Proceedings of the 35th International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA)*, 2008, pp. 309–315.

[169] K. Z. Élise Fovet, "Path Network Modelling and Network of Aggregated Settlements: a Case Study in Languedoc (Southeastern France)," *Computational approaches to the study of movement in archaeology*, vol. 23, pp. 43–72, 2014.

[170] D. A. White and S. B. Barber, "Geospatial Modeling of Pedestrian Transportation Networks: a Case Study from Precolumbian Oaxaca, Mexico," *Journal of Archaeological Science*, vol. 39, no. 8, pp. 2684–2696, 2012.

[171] J. Verhagen, "On the Road to Nowhere? Least Cost Paths, Accessibility and the Predictive Modelling Perspective," in *Fusion of Cultures. Proceedings of the 38th Annual Conference on Computer Applications and Quantitative Methods in Archaeology, Granada, Spain, April 2010 (BAR International Series 2494)*, F. Contreras, M. Farjas, and F. Melero, Eds. Archaeopress, 2013, pp. 383–389.

[172] M. Llobera, *Working the digital: some thoughts from landscape archaeology.* Material Evidence: Learning from Archaeological Practice, 01 2015, pp. 173–88.

[173] M. Llobera, P. Fábrega-Álvarez, and C. Parcero-Oubiña, "Order in Movement: a GIS Approach to Accessibility," *Journal of Archaeological Science*, vol. 38, no. 4, pp. 843–851, 2011.

[174] R. Opitz, "Prospecting for Archaeological Features with Ikonos Satellite Images. A Case Study Around Falerii Novi (VT)," *Archaeological Computing Newsletter*, vol. 64, pp. 2–6, 2006.

[175] C. Knappett, Ed., *Network Analysis in Archaeology: New Approaches to Regional Interaction.* Oxford: Oxford University Press, 2013.

[176] T. Brughmans, A. Collar, and F. Coward, Eds., *The Connected Past. Challenges to Network Studies in Archaeology and History.* Oxford: Oxford University Press, 2016.

[177] G. Earl and S. Keay, "Urban Connectivity of Iberian and Roman Towns in Southern Spain: a Network Analysis Approach," in *Digital Discovery. Exploring New Frontiers in Human Heritage. CAA2006. Computer Applications and Quantitative Methods in Archaeology. Proceedings of the 34^{th} Conference*, 2007, pp. 77–86.

[178] S. M. Sindbæk, "The Small World of the Vikings: Networks in Early Medieval Communication and Exchange," *Norwegian Archaeological Review*, vol. 40, no. 1, pp. 59–74, 2007.

[179] S. Sindbæk, "Northern Emporia and Maritime Networks. Modelling Past Communication Using Archaeological Network Analysis," *Harbours and maritime networks as complex adaptive systems*, vol. 2, p. 105–118, 2015.

[180] C. Knappett, T. Evans, and R. Rivers, "Modelling Maritime Interaction in the Aegean Bronze Age," *Antiquity*, vol. 82, no. 318, p. 1009–1024, 2008.

[181] C. Knappett, R. Rivers, and T. Evans, "The Theran Eruption and Minoan Palatial Collapse: New Interpretations Gained from Modelling the Maritime Network," *Antiquity*, vol. 85, no. 329, p. 1008–1023, 2011.

[182] C. Carreras and P. D. Soto, "The Roman Transport Network: a Precedent for the Integration of the European Mobility," *Historical Methods: a Journal of Quantitative and Interdisciplinary History*, vol. 46, no. 3, pp. 117–133, 2013.

[183] B. J. Mills, J. J. Clark, M. A. Peeples, W. R. Haas, J. M. Roberts, J. B. Hill, D. L. Huntley, L. Borck, R. L. Breiger, A. Clauset, and M. S. Shackley, "Transformation of Social Networks in the Late Pre-Hispanic US Southwest," *Proceedings of the National Academy of Sciences*, vol. 110, no. 15, pp. 5785–5790, 2013.

[184] F. Coward, "Grounding the Net: Social Networks, Material Culture, and Geography in the Epipaleolithic and Early Neolithic of the Near East (21–6,000 cal BCE)," *Network Analysis in Archaeology: New Approaches to Regional Interaction*, pp. 247–280, 01 2013.

[185] M. Golitko and G. M. Feinman, "Procurement and Distribution of Pre-Hispanic Mesoamerican Obsidian 900 BC–AD 1520: a Social Network Analysis," *Journal of Archaeological Method and Theory*, vol. 22, pp. 206–247, 02 2015.

[186] L. Kaddouri, "Structures spatiales et mises en réseaux de villes pour la régionalisation des territoires," PhD dissertation, Université Paul Valéry, 2004.

[187] I. Herzog, *Least-cost Networks.* Amsterdam University Press, 2013, pp. 237–248.

[188] P. Verhagen, T. Brughmans, L. Nuninger, and F. Bertoncello, *The Long and Winding Road: Combining Least Cost Paths and Network Analysis Techniques for Settlement Location Analysis and Predictive Modelling.* Amsterdam University Press, 2013, pp. 357–366.

[189] J. Verhagen, S. Polla, and I. Frommer, “Finding Byzantine Junctions with Steiner Trees,” in *Computational Approaches to Movement in Archaeology. Theory, practice and interpretation of factors and effects of long term landscape formation and transformation*, ser. Volume 23, S. Polla and P. Verhagen, Eds. De Gruyter, 2014, pp. 73–98.

[190] M. Groenhuijzen and J. Verhagen, “Exploring the Dynamics of Transport in the Dutch Limes,” *eTopoi. Journal for Ancient Studies*, no. Special Volume 4, pp. 25–47, 2015.

[191] ——, “Testing the Robustness of Local Network Metrics in Research on Archeological Local Transport Networks,” *Frontiers in Digital Humanities*, vol. 3, no. 6, 2016.

[192] H. A. Orengo and A. Livarda, “The Seeds of Commerce: a Network Analysis-based Approach to the Romano-British Transport System,” *Journal of Archaeological Science*, vol. 66, pp. 21–35, 2016.

[193] L. J. Gorenflo and T. L. Bell, *Network Analysis and the Study of Past Regional Organisation.* Cambridge University Press, 1991, pp. 80–98.

[194] D. J. Badillo, “A Method for Interactive Recognition of Three-dimensional Adjacency patterns in Point Sets, Based on Relative Neighbourhood Graphs: an Archaeological application.” PhD dissertation, University of London, 2004.

[195] R. Rivers, C. Knappett, and T. Evans, *What Makes a Site Important? Centrality, Gateways, and Gravity.* OUP, 2013.

[196] M. R. Groenhuijzen and P. Verhagen, “Comparing Network Construction Techniques in the Context of local Transport Networks in the Dutch Part of the Roman Limes,” *Journal of Archaeological Science: Reports*, vol. 15, pp. 235–251, 2017.

[197] F. Fulminante, L. Prignano, I. Morer, and S. Lozano, “Coordinated Decisions and Unbalanced Power. How Latin Cities Shaped Their Terrestrial Transportation Network,” *Frontiers Digital Humanities*, vol. 4, p. 4, 2017.

[198] S. Plog, *Measurement of Prehistoric Interaction Between Communities.* New York: Academic Press, 1976, p. 255–272.

[199] C. L. Crumley, “Three Locational Models: an Epistemological Assessment for Anthropology and Archaeology,” *Advances in Archaeological Method and Theory*, vol. 2, pp. 141–173, 1979.

[200] I. Hodder, “Some Marketing Models for Romano-British Coarse Pottery,” *Britannia*, vol. 5, pp. 340–359, 1974.

[201] M. Jochim, Ed., *Hunter-gatherer Subsistence and Settlement: a Predictive Model.* Academic Press, 1976.

[202] A. Bevan and A. Wilson, “Models of Settlement Hierarchy Based on Partial Evidence,” *Journal of Archaeological Science*, vol. 40, no. 5, pp. 2415–2427, 2013.

[203] E. Paliou and A. Bevan, “Evolving Settlement Patterns, Spatial Interaction and the Socio-political Organisation of late Prepalatial South-central Crete,” *Journal of Anthropological Archaeology*, vol. 42, pp. 184–197, 2016.

[204] T. Davies, H. Fry, A. Wilson, A. Palmisano, M. Altaweel, and K. Radner, "Application of an Entropy Maximizing and Dynamics Model for Understanding Settlement Structure: the Khabur Triangle in the Middle Bronze and Iron Ages," *Journal of Archaeological Science*, vol. 43, pp. 141–154, 2014.

[205] S. Wasserman, K. Faust, C. U. Press, M. Granovetter, U. of Cambridge, and D. Iacobucci, *Social Network Analysis: Methods and Applications*, ser. Structural Analysis in the Social Sciences. Cambridge University Press, 1994.

[206] A. Collar, F. Coward, T. Brughmans, and B. J. Mills, "Networks in Archaeology : Phenomena, Abstraction, Representation," *Journal of Archaeological Method and Theory*, vol. 22, no. 1, pp. 1–32, 2015.

[207] L. Nuninger, L. Sanders, F. Favory, P. Garmy, C. Raynaud, C. Rozenblat, L. Kaddour, H. Mathian, and L. Schneider, "La modélisation des réseaux d'habitat : trois expériences," *Mappemonde*, vol. 3, p. 1–28, 10 2006.

[208] B. Hillier and J. Hanson, *The Social Logic of Space.* Cambridge University Press, 1984.

[209] S. Bafna, "Space Syntax: a Brief Introduction to Its Logic and Analytical Techniques," *Environment and Behavior*, vol. 35, no. 1, pp. 17–29, 2003.

[210] P. de Soto, *Network Analysis to Model and Analyse Roman Transport and Mobility.* Cham: Springer International Publishing, 2019, pp. 271–289.

[211] A. Miquel, *Encyclopaedia of the History of Arabic Science.* London: Routledge, 1996, ch. Geography, pp. 796–812.

[212] al-Muḥammad ibn Aḥmad al Muqaddasī, *The Best Divisions for Knowledge of the Regions: a Translation of Aḥsan al-taqāsīm fī ma^{c}rifaŧ al-aqālīm.* Reading, UK: Centre for Muslim Contribution to Civilization: Garnet Pub., 1994.

[213] ——, *Kitāb Aḥsan al-taqāsīm fī ma^{c}rifaŧ al-aqālīm*, ser. Bibliotheca Geographorum Arabicorum. Leiden: Brill, 1906, vol. 1-3.

[214] O. Cuntz and G. Wirth, Eds., *Itineraria Romana: Itineraria Antonini Augusti et Burdigalense.* Berlin, Boston: De Gruyter, 1990.

[215] E. Yarshater, Ed., *Naser-e Khosraw's Book of Travels: (Safarnama).* Albany, NY: Bibliotheca Persica, 1986.

[216] M. Romanov, "Computational Reading of Arabic Biographical Collections with Special Reference to Preaching in the Sunnī World (661–1300 ce)," PhD dissertation, University of Michigan, Horace H. Rackham School of Graduate Studies, 2013.

[217] ——, "Algorithmic Analysis of Medieval Arabic Biographical Collections," *Speculum*, vol. 92, no. S1, pp. S226–S246, 2017.

[218] M. Seydi, M. Romanov, and C. Palladino, "Premodern Geographical Description: Data Retrieval and Identification," in *Proceedings of the 11th Workshop on Geographic Information Retrieval*, ser. GIR'17. New York, NY, USA: ACM, 2017, pp. 4:1–4:10.

[219] T. Consortium, "Names, Dates, People, and Places." *TEI P5: Guidelines for Electronic Text Encoding and Interchange (version 4.2.2)*, pp. 456–509, 2021.

[220] L. Moncla, M. Gaio, J. Nogueras-Iso, and S. Mustiere, "Reconstruction of Itineraries from Annotated Text with an Informed Spanning Tree Algorithm," *International Journal of Geographical Information Science*, vol. 30, pp. 1–24, 01 2016.

[221] M. Romanov, "OpenITI mARkdown," https://alraqmiyyat.github.io/mARkdown/, 2017.

[222] A. Diller, "Agathemerus, Sketch of Geography," *Greek Roman and Byzantine Studies*, vol. 16, pp. 59–76, 1975.

[223] M. Calzolari, *Introduzione allo studio della rete stradale dell'Italia romana: l'itinerarium Antonini.* Roma: Accademia Nazionale dei Lincei, 1996.

[224] N. Reed, "Pattern and Purpose in the Antonine Itinerary," *The American Journal of Philology*, vol. 99, no. 2, pp. 228–254, 1978.

[225] S. Russell and P. Norvig, *Artificial Intelligence: a Modern Approach*, 3rd ed. Upper Saddle River, NJ: Prentice Hall Press, 2009.

[226] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*, 2nd ed. Cambridge, MA: MIT Press, 2018.

[227] C. Cortes and V. Vapnik, "Support-vector Networks," in *Machine Learning*, vol. 20, 1995, p. 273–297.

[228] O. Obeid, N. Zalmout, S. Khalifa, D. Taji, M. Oudah, B. Alhafni, G. Inoue, F. Eryani, A. Erdmann, and N. Habash, "CAMeL Tools: an Open Source Python Toolkit for Arabic Natural Language Processing," in *Proceedings of the 12th Language Resources and Evaluation Conference.* Marseille, France: European Language Resources Association, May 2020, pp. 7022–7032.

[229] S. Green and C. Manning, "Better Arabic Parsing: Baselines, Evaluations, and Analysis." in *Coling 2010 - 23rd International Conference on Computational Linguistics, Proceedings of the Conference*, vol. 2, 01 2010, pp. 394–402.

[230] Y.-W. Chang, C.-J. Hsieh, K.-W. Chang, M. Ringgaard, and C.-J. Lin, "Training and Testing Low-Degree Polynomial Data Mappings via Linear SVM," *Journal of Machine Learning Research*, vol. 11, pp. 1471–1490, 2010.

[231] H. Butler, M. Daly, A. Doyle, S. Gillies, and T. Schaub, "The GeoJSON Format," RFC 7946, August 2016.

[232] M. Mahdipoor, "Map from Mahmud al-Kashgari's Diwan (11th Century)," https://commons.wikimedia.org/wiki/File:Mahmud_al-Kashgari_map.jpg, 2013.

[233] H. Samet, "The Quadtree and Related Hierarchical Data Structures," *ACM Computing Surveys*, vol. 16, no. 2, pp. 187–260, June 1984.

[234] F. Aurenhammer, "Voronoi Diagrams—Survey of a Fundamental Geometric Data Structure," *ACM Computing Surveys*, vol. 23, no. 3, pp. 345–405, Sep. 1991.

[235] M. de Berg, O. Cheong, M. van Kreveld, and M. Overmars, *Computational Geometry: Algorithms and Applications.* Springer Berlin Heidelberg, 2008.

[236] H. Edelsbrunner, D. Kirkpatrick, and R. Seidel, "On the Shape of a Set of Points in the Plane," *IEEE Transactions on Information Theory*, vol. 29, no. 4, pp. 551–559, July 1983.

[237] A. Moreira and M. Santos, "Concave Hull: a K-nearest Neighbours Approach for the Computation of the Region Occupied by a Set of Points." in *GRAPP 2007, Proceedings of the Second International Conference on Computer Graphics Theory and Applications*, 01 2007, pp. 61–68.

[238] D. H. Douglas and T. Peucker, "Algorithms for the Reduction of the Number of Points Required to Represent a Digitized Line or Its Caricature," *Cartographica: The International Journal for Geographic Information and Geovisualization*, vol. 10, pp. 112–122, 1973.

[239] M. Visvalingam and J. Whyatt, *Line Generalisation by Repeated Elimination of the Smallest Area.* Cartographic Information Systems Research Group, University of Hull, 1992.

[240] U. Ramer, "An Iterative Procedure for the Polygonal Approximation of Plane Curves," *Computer Graphics and Image Processing*, vol. 1, no. 3, pp. 244–256, 1972.

[241] L. M. Kells, W. F. Kern, and J. R. Bland, *Plane and Spherical Trigonometry.* McGraw Hill Book Company, Inc., 1940.

[242] G. B. M. of Defence (Navy), *Admiralty Manual of Navigation*, ser. B.R Series. HMSO, 1992.

[243] C. Carter, "Great Circle Distances: Computing the Distance Between Two Points on the Surface of the Earth," https://www.inventeksys.com/wp-content/uploads/2011/11/GPS_Facts_Great_Circle_Distances.pdf, 2002, accessed: 2019-01-24.

[244] W. Tobler and S. Wineburg, "A Cappadocian Speculation," *Nature*, vol. 231, 1971.

[245] W. Isard, *Methods of Regional Analysis.* John Wiley and Sons, Inc., 1960.

[246] C. Bizer, T. Heath, and T. Berners-Lee, "Linked Data - The Story so far," *International Journal on Semantic Web and Information Systems*, vol. 5, no. 3, pp. 1–22, 2009.

[247] A. T. C. Jackson, M. Romanov, "Al-Ṯurayyā Gazetteer Ver. 0.2," https://althurayya.github.io/previous/althurayya_02/.

[248] C. Jackson, "An Interactive Map of the Classical Islamic World (v.1)," 2014.

[249] E. Dijkstra, "A Note on Two Problems in Connexion with Graphs." *Numerische Mathematik*, vol. 1, pp. 269–271, 1959.